Towards Trusted AI: A Blueprint for Ethics Assessment in Practice

Christoph Tobias Wirth[1]
Smart Data & Knowledge Services, German Research Center for Artificial Intelligence (DFKI GmbH), Kaiserslautern, Germany

Mihai Maftei
Ethics Team, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany

Rosa Esther Martín-Peña
Educational Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH), Berlin, Germany

Iris Merget
Agents and Simulated Reality, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany

Abstract

The development of AI technologies leaves place for unforeseen ethical challenges. Issues such as bias, lack of transparency and data privacy must be addressed during the design, development, and the deployment stages throughout the lifecycle of AI systems to mitigate their impact on users. Consequently, ensuring that such systems are responsibly built has become a priority for researchers and developers from both public and private sector. As a proposed solution, this paper presents a blueprint for AI ethics assessment. The blueprint provides for AI use cases an adaptable approach which is agnostic to ethics guidelines, regulatory environments, business models, and industry sectors. The blueprint offers an outcomes library of key performance indicators (KPIs) which are guided by a mapping of ethics framework measures to processes and phases defined by the blueprint. The main objectives of the blueprint are to provide an operationalizable process for the responsible development of ethical AI systems, and to enhance public trust needed for broad adoption of trusted AI solutions. In an initial pilot the blueprinted for AI ethics assessment is applied to a use case of generative AI in education.

2012 ACM Subject Classification Computing methodologies → Artificial intelligence; Social and professional topics → Codes of ethics; Human-centered computing → Collaborative and social computing; Applied computing → Arts and humanities

Keywords and phrases Trusted AI, Trustworthy AI, AI Ethics Assessment Framework, AI Quality, AI Ethics, AI Ethics Assessment, AI Lifecycle, Responsible AI, Ethics-By-Design, AI Risk Management, Ethics Impact Assessment, AI Ethics KPIs, Human-Centric AI, Applied Ethics

Digital Object Identifier 10.4230/OASIcs.SAIA.2024.7

Category Academic Track

Funding The Federal Ministry of Education and Research (BMBF) is funding the project *Artificial Intelligence for Arts Education (AI4ArtsEd)* within the funding measure Cultural Education in Social Transformations as part of the federal research program for Empirical Educational Research.

Acknowledgements We thank Samantha Morgaine Prange and Lisa-Marie Goltz from the DFKI Ethics Team for their valuable contribution throughout the entire paper writing process.

[1] corresponding author

Symposium on Scaling AI Assessments (SAIA 2024).
Editors: Rebekka Görge, Elena Haedecke, Maximilian Poretschkin, and Anna Schmitz; Article No. 7; pp. 7:1–7:19
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Artificial intelligence (AI) holds the promise of transforming our world. However, the development of AI technologies leaves also place for unforeseen ethical challenges. Unethical use of AI can lead to various negative outcomes, such as biases and discrimination, privacy and human rights violations, and unintentional harm.

Furthermore, AI practitioners often possess an abstract and somewhat limited understanding of ethical principles and how to translate them into practice effectively. Although their primary motivation is implementing ethical guidelines or principles within practical designs that meet legal requirements, this does not necessarily ensure that AI products are ethically or socially acceptable. Legal compliance alone does not guarantee that AI technologies align with broader societal values or adequately address ethical concerns.

One argument explaining this phenomenon is that new laws often have an extended lead time and cannot keep up with rapidly changing social norms or values. They are not designed to address or adapt to swift shifts in societal expectations. This gap highlights the need for practical ethics to guide practitioners in *operating in the grey areas* [12]. The concept of the grey area refers to ethical dilemmas that emerge when society repeatedly suffers from poor decisions not addressed by existing legislation. These dilemmas often pressure the legal system to adapt and consider new social realities outside existing legal frameworks.

Examples of unethical AI use include Amazon's recruiting algorithm, which displayed a gender bias favoring male applicants over female ones [25]. Another study revealed that AI-based gender classification technology tends to be less accurate for skin types of darker color [5]. Incidents like these can rapidly undermine public trust in AI models' safety, security, reliability, and ethical standards. Without trust, people may fear that AI systems will produce incorrect, inconsistent, or harmful outcomes.

The concept of *Trusted AI* can be explored from multiple distinct perspectives. From the multiplicity of definitions, we understand the term "Trusted AI" as the evaluation of artificial intelligence concerning its reliability and effectiveness in individual applications from the user's perspective, also considering the specific cultural context and values of the community in which the AI system is embedded.

To enhance user trust in AI applications we need to ensure that AI systems are conformant to ethics quality metrics. For this purpose, the German Research Center for Artificial Intelligence (DFKI) Ethics Board has developed a Blueprint for AI Ethics Assessment. In this paper, we present our Ethics-By-Design-based approach aimed at proactively and reactively mitigating the ethical challenges an AI system may encounter during design, development, and deployment.

2 Current global state of AI Ethics implementation

Countries around the world define national AI strategies to leverage the rapid advancement of AI technology. Executing an AI strategy needs governance that includes oversight mechanisms to address risks such as bias, privacy infringement and misuse, but also to build and maintain trust in AI, while at the same time enables AI innovation and research. On international level, the United Nations laid out foundations for the first global architecture for AI governance based on international cooperation [36]. An effective AI governance framework provides a structured approach based on the pillars of regulation, sound AI policy, supporting standards for compliance, and innovation measures. Figure 1 illustrates the building blocks of an AI governance framework. This structure highlights how every element depends on a strong ethical foundation. The AI Strategy represents, in the context of a state, a government's approach to the development, deployment, and regulation of AI technologies and from a corporate perspective, represents the enterprise AI roadmap. Below, the concept of AI Governance defines the structural support required to operationalize the pillars (regulation, policies, standards, and innovation), aligning them under a unified framework. The four pillars are grounded on a structural basis represented by the foundational aspects of rule of law, human rights, and democracy. At the very bottom, ethics serves as a fundamental grounding, upon which every component and the entire structure as a whole is developed and sustained. This section provides an overview of the current global landscape of AI ethics, examining how different Digital Empires are responding to the challenges posed by AI. Different regulatory approaches, ethical guidelines, and policy initiatives that have been implemented to ensure that AI technologies are developed and deployed responsibly will be explored. The Digital Empires create a pull effect on other countries in adapting their regulatory approach commonly denoted as Brussels, Beijing, and California effect. The following overview only presents the current point-in-time snapshot of the operationalization potential for AI ethics by selected global digital powers. The choice of geographies is not meant to be biased and presented in alphabetic order.

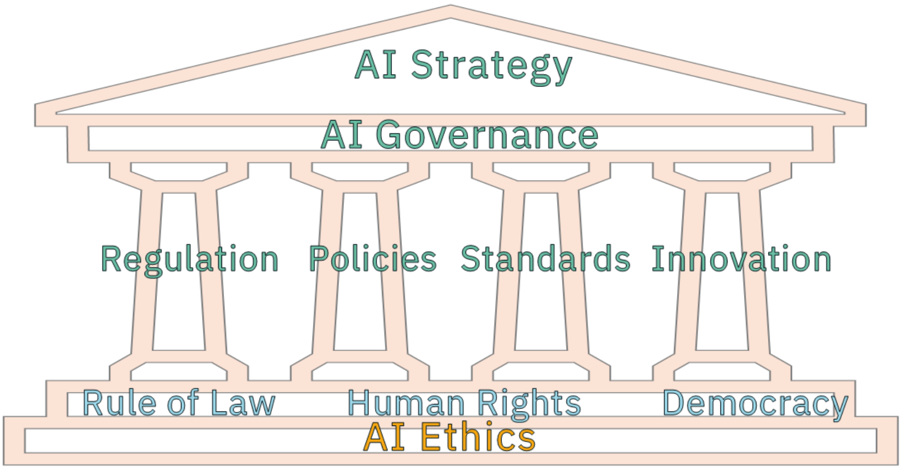


Figure 1 Building blocks of a national AI strategy comprise of its governance structure and the functional pillars of regulations, policies, standards and innovation supported by the foundational layer of ethics providing the fundament for rule of law human rights, and democratic values.

Africa

The African Union's (AU) "Continental AI Strategy" prioritizes "economic growth, social progress, and cultural renaissance" [1]. with the help of AI systems. The principles focus on local first and people-centeredness as well as ethics and transparency, inclusion and diversity, human rights and dignity, peace and prosperity, cooperation and integration, and skills development, public awareness and education. This strategy puts forward an Africa-centric and development-oriented and inclusive approach around five focus areas notably: harnessing AI's benefits, building AI capabilities, minimizing risks, stimulating investment and fostering cooperation. It is part of the AU Agenda 2063 which aims to further peace, prosperity, self-governance, and international cooperation. The strategy is divided into 5 areas of actions which should be implemented between 2025 and 2030, they are the following: Maximizing AI Benefits, Building Capabilities for AI, Minimizing AI Risks, African Public and Private Sector Investment in AI, and Regional and International Cooperation and Partnerships. Additionally, South Africa has published the "National Artificial Intelligence Policy Framework" [24] and Nigeria its corresponding "National Artificial Intelligence Strategy" [15], both in August 2024.

Canada

In June 2022 the Canadian Government submitted the "Artificial Intelligence and Data Act (AIDA)" [14] under the "Digital Charter Implementation Act" [13], following the "Pan-Canadian AI Strategy" [6] launched in 2017. AIDA adheres to the OECD regulations, the EU AI-Act and the NIST [18] Risk Management Framework reflecting the influence of the Brussels Effect in the Canadian AI strategy, but also the interest in aligning with international standards and ethics requirements to strengthen international/economic relations. AIDA is an addition to existing laws like consumer protection and human rights and will probably come into force in 2025 with administration and enforcement responsibilities lying with the Minister of Innovation, Science, and Industry. In the incipient stages of implementation, the emphasis will be on education, setting up guidelines, and assisting businesses in voluntarily adhering to the new regulation. The government plans to provide sufficient time for the ecosystem to adapt to the new framework before initiating any enforcement action.

China

The National Governance Committee for the New Generation Artificial Intelligence published the "Ethical Norms for the New Generation Artificial Intelligence" [23] in September 2021. The norms for the AI life cycle include fairness, justice, harmony, and security, preventing bias, discrimination, and privacy/information leaks. China has launched the Global AI Governance Initiative (GAIGI) [8] as part of its Belt and Road Initiative, promoting international cooperation in AI governance. Unlike the EU AI Act, China has been regulating specific AI applications individually, such as internet recommendation algorithms, deep synthesis technology, and generative AI. This approach allows China to address specific issues with correspondent rules, building new policy tools and regulatory expertise with each regulation. After the release of ChatGPT the Cyber Space Administration of China (CAC) reacted within 6 months with Draft Measures for Generative AI [37]. China's AI regulations are designed to be iterative, allowing for quick updates in response to rapid AI developments. The "Interim Administrative Measures for Generative AI Services" [22] exemplify this iterative approach, with the expectation that AI regulation remains highly adaptive.

Europe

In August 2024 the world's first regulation on AI, the EU AI Act, went into force. This Regulation shall support the EU objective of being a "global leader in the development of secure, trustworthy and ethical AI" [11] and it shall "ensure the protection of ethical principles" [11]. Recognition on the international level of the European legislation reflects the global interest and adaptiveness to the EU regulatory framework, generating the Brussels effect [3]. The AI Act's binding rules are built on a risk-based approach. However, the implementation of ethics principles for providers and deployers of AI is left on a voluntary basis. The AI Act suggests that for voluntary ethics codes of conduct to be effective, they should be based on clear objectives and key performance indicators to measure the achievement of those objectives. The AI Act does not explicitly mention that an ethics assessment framework for trustworthy AI must be applied. The AI Act encourages to implement ethics processes in AI system development. In this regard, the EU issued both independently and in collaboration with international bodies multiple ethics principles, guidelines, and assessment frameworks, such as: (i) The High-Level Expert Group on Artificial Intelligence (HLEG) Ethics Guidelines for Trustworthy AI [9], (ii) the Assessment List for Trustworthy Artificial Intelligence (ALTAI) [10],(iii) UNESCO Ethical impact assessment [33].

India

The Indian Government released in 2018 the National Strategy on AI [19]. India focus lies on: healthcare, education, agriculture, smart cities and mobility. Those needs are based on the seven ethics principles: safety and reliability, equality, inclusivity and non-discrimination, privacy and security, transparency, accountability, and protection and reinforcement of positive human values. These frameworks are not binding, but, for example, the copyright law has been adjusted for AI-generated content. One of the lawsuits against deepfakes was issued after the incident of the Bollywood Actor, Anil Kapoor. His persona had been faked to use for merchandise to earn money. The court agreed with Kapoor since this was a violation of his rights [27]. Furthermore, developments in legislation have been made. The Digital Personal Data Protection Act (DPDPA) was issued in 2023 to ensure the safe usage of personal data to train AI systems [16].

Singapore

Though Singapore does currently not have any binding regulation on AI, the Singaporean government has developed variety of sector-specific and voluntary frameworks to guide the responsible use of AI and to safeguard public interest in AI ethics and governance. In the following two frameworks are introduced, one for financial institutions and the other one for the deployment of generative AI. In 2022 the Monetary Authority of Singapore published assessment methodologies for the fairness, ethics, accountability and transparency (FEAT) principles, to guide the responsible use of AI by financial institutions [17]. The fairness assessment methodology ensures that the AI-assisted decision-making process does not systematically disadvantage individuals or groups of individuals, without appropriate justification. The fairness principle is checked throughout the lifecycle of the AI system's development process based on the key concepts such as selection of personal attributes, types of bias and their mitigation methods, and fairness objectives and their metrics. In 2024 Singapore released the "Model AI Governance Framework for Generative AI" [21] which addresses risks related to Generative AI and provides guidance on practices for safety evaluation of Generative AI models. The framework is based on the core principles of accountability, transparency, fairness, robustness and security and it extends the previous version from 2019 developed for Traditional AI.

U.S.A.

In October 2023 the White House released the Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence. The Biden Administration focuses on eight principles, such as: Safety/Security, Robustness, Reliability, and Repeatability. AI must be standardized and testable before its use to diminish risks. Furthermore, constant monitoring is necessary to ensure ethical development, resilience against misuse, and compliance with Federal laws [31]. The next step is the Blueprint for an AI Bill of Rights, with the principles: safe and effective systems, algorithmic discrimination protection, data privacy, notice and explanation, and human alternatives, consideration and fallback [28]. Although this is a voluntary framework Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and OpenAI have offered their commitment [29]. Additionally, 28 healthcare providers and payers have committed to the responsible use of AI in healthcare [30]. The different states can also make their own laws to regulate AI [2]. The Artificial Intelligence Risk Management Framework was published in January 2023 by the National Institute of Standards and Technology (NIST). NIST uses a modified version of the AI lifecycle from the OECD Framework for the Classification of AI systems. After the release of ChatGPT NIST has published the Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile in July 2024.

2.1 Implications for AI ethics assessment – The need for a process blueprint

As evidenced by the information presented above, all countries except for the EU AI Act have voluntary regulations or soft-laws when it comes to AI systems. The guidelines often focus on the same principles with security being at the forefront. Privacy and protection are always among the principles, but their understanding differs between countries. As an effect, different court outcomes might appear. In the Indian case mentioned above the court decision was favoring the actor, but in a similar incident in the U.S., when Scarlett Johannson wrote to OpenAI about illegally using her voice, the company stopped the use of her persona, but on a legal level no measures have been taken [26]. This shows that AI governance and ethical frameworks vary across the globe in regard to regional, legal and cultural values, and even more when it comes to strategic interests in shaping digital power.

There are three competing regulatory models, each reflecting a different approach for the digital economy. The United States adopts a market-driven model, focusing on flexible frameworks, China follows a state-driven approach, emphasizing control, security, and social stability in AI development, and the European Union takes a rights-driven stance, prioritizing ethical standards [4]. These three distinct models – market, state, and rights-driven – illustrate that the global landscape of AI ethics is not only a mere reaction of technological advancements but also a manifestation of the underlying political, economic, and cultural dynamics that concretize each region's approach to AI governance.

In summary, a global ethical framework, with the objective to guide the deployment of trusted AI and to promote the responsible use of AI, implies the need of a process blueprint. The blueprint for an AI ethics assessment must fulfill two acceptance criteria. The first criterion refers to its high level of independence, which implies it is agnostic to the underlying regulatory model, to the deployed AI algorithm, to the technology in which the AI model is embedded in, and it is agnostic to the needs of the industry sector or to the business model or scale of business. The second criterion of the blueprint allows for adaptivity to varying comprehension of ethical principles and values. As already been pointed out, the interpretation or choice of ethical principles depends not only on the cultural perspective, but it is also tailored to specific industry needs and it also aims to maximize the space for AI innovation for which most national AI strategies of countries define a leading position. Lastly, the AI ethics process blueprint that fosters a trusted AI ecosystem cannot be static. The blueprint itself requires a review and update process that adapts to advancements in AI.

3 The Blueprint for AI Ethics Assessment in Practice

3.1 Motivation: Blueprint for the entire AI lifecycle

While most AI assessment solutions comprise high-level ethics principles and evaluation tools [7], [20], they miss the practical aspects needed for operationalization in the cycle from idea-to-AIOps deployment. Therefore, our aim is to build a generic AI Ethics Assessment Blueprint for the evaluation of the entire lifecycle of an AI system, from design and development to deployment.

The Blueprint's adaptable framework integrates ethical principles and their associated assessment tools as inputs, leading to a materiality analysis of the AI system. To achieve our goal, we utilized the UNESCO Ethics Principles [32] and the UNESCO Ethical Impact Assessment Tool [33]. We chose the UNESCO ethics framework for two reasons, first, it is congruent with the EU definition of trustworthy AI and, second, it is a global reference standard, adopted by all 193 UNESCO member states in November 2021. An overview of the UNESCO Ethics Principles is provided in appendix A.

The Blueprint for AI Ethics Assessment serves as a facilitator, ensuring that the development process and lifecycle of an AI system are supported rather than constrained. It is designed to enhance and ease the ethical evaluation process, but also to support the ethical and responsible design, development, and deployment of AI systems, providing a structured approach that does not hinder the AI system different lifecycle phases.

3.2 Key Requirements: Successful Implementation of AI Ethics Assessment

In accomplishing operationalization, an AI ethics assessment framework must contain at least the following three components: (i) high-level ethics principles, (ii) an ethics assessment tool corresponding to the ethics principles, and (iii) a set of evaluation measures relating to key performance indicators (KPIs).

Ethics metrics or their defined thresholds provide an important instrument in the decision-making process, for example in selecting mitigation strategies as part of the results of an ethics assessment. Without outcome-driven ethics metrics along the AI lifecycle pathway the operationalization of an ethics assessment framework remains a challenging milestone. To solve the challenge, we propose to develop a phased approach which is described in the next section.

3.3 Structure: The need for a phased approach aligned to the AI lifecycle

The decision to implement a five-phase process in the Blueprint for AI Ethics Assessment is rooted in the need to establish a structured approach to addressing ethical challenges throughout the entire lifecycle of an AI system. This phased approach was chosen to ensure that ethics are not treated as an afterthought or a box-ticking exercise but are an integrated part of AI development and deployment. Because AI technologies present complex and multifaceted ethical dilemmas that require ongoing, context-sensitive assessment, a single-stage process would be insufficient to capture the nuanced and evolving nature of the ethics issues.

The five phases are based on two motivational drivers: first, to reflect the ethics principles and second, to incorporate the technicalities of the software-engineering needs. Furthermore, focusing on concrete phases such as system design, data management, model development, system monitoring, and human-centric evaluation, the Blueprint can ensure that ethical issues like fairness, transparency, and accountability are assessed in relation to the actual system development process.

Each of the five phases corresponds to a distinct aspect of the AI system's development, from design to human-centricity evaluation, allowing for a step-by-step integration of ethics in an iterative manner, with each phase building upon the previous one. The rationale behind dividing the process into five phases is to break down the complexities of AI ethics into manageable components, each targeting specific risks and challenges that might emerge at different stages of the AI lifecycle. This phased approach enables continuous feedback loops, ensuring that ethical compliance is not static but evolves alongside the system itself, thus creating a more dynamic and responsive framework. A detailed description of each phase will be presented below.

3.4 Specification: Detailing out the five phases of the Blueprint

Our framework addresses the three main first level stages of an AI system lifecycle - the system design stage, the development stage, and the deployment stage. On the second level, we define five phases where each phase entails two processes on the third level and one result as outcome. Every process includes methodologies and practices aimed at addressing the specific needs and challenges of its corresponding stage within the AI system lifecycle. A schematic of the Blueprint for AI Ethics Assessment in Practice is shown in figure 2.

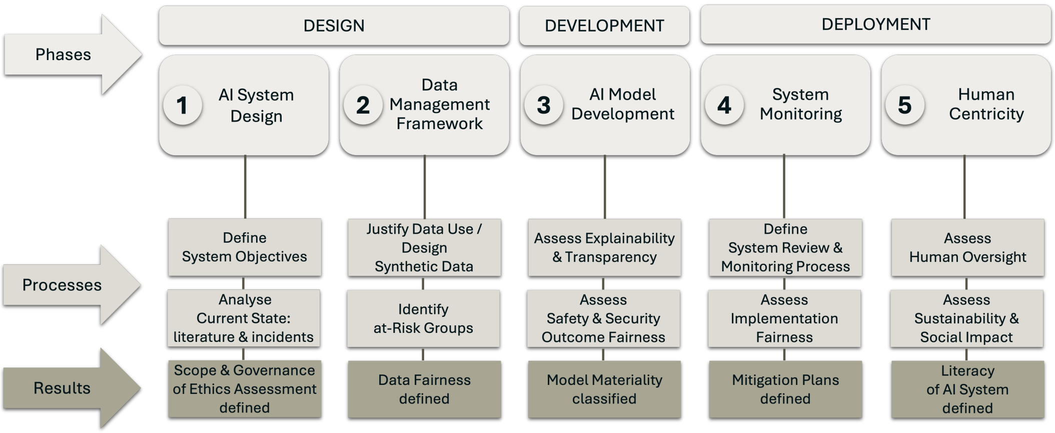

Phases
DESIGN
DEVELOPMENT
DEPLOYMENT
1 AI System Design
2 Data Management Framework
3 AI Model Development
4 System Monitoring
5 Human Centricity
Processes
Define System Objectives
Justify Data Use / Design Synthetic Data
Assess Explainability & Transparency
Define System Review & Monitoring Process
Assess Human Oversight
Analyse Current State: literature & incidents
Identify at-Risk Groups
Assess Safety & Security Outcome Fairness
Assess Implementation Fairness
Assess Sustainability & Social Impact
Results
Scope & Governance of Ethics Assessment defined
Data Fairness defined
Model Materiality classified
Mitigation Plans defined
Literacy of AI System defined

Figure 2 Schematic of the AI Ethics Assessment Framework for the responsible design and achievement of Trusted AI.

Below, the detailing out the five phases of the Blueprint are presented:

The 1st phase. AI System Design starts with the first process, Define System Objectives, during which the stakeholders define together with the ethics board the objectives and purpose of the AI system. In the second process, Analyze Current State: literature and incidents, an examination of existing literature and relevant incidents is performed to identify potential ethical challenges and best practices to address associated risks. As a result, the scope and governance of the ethics assessment are defined.

The 2nd phase. Data Management Framework focuses in the first process, Justify Data Use / Design Synthetic Data, on the procedural assessment of data use and (when applicable) on the design of synthetic data with the aim to enhance privacy protection and to facilitate controlled experimentation without compromising sensitive information. The second process, Identify At-Risk Groups, analyzes the prospective data-related ethical issues and it ensures that social justice and equity is promoted. This includes addressing the needs of diverse age groups, cultural and linguistic communities, persons with disabilities, gender diversity, and disadvantaged, marginalized, and vulnerable individuals.

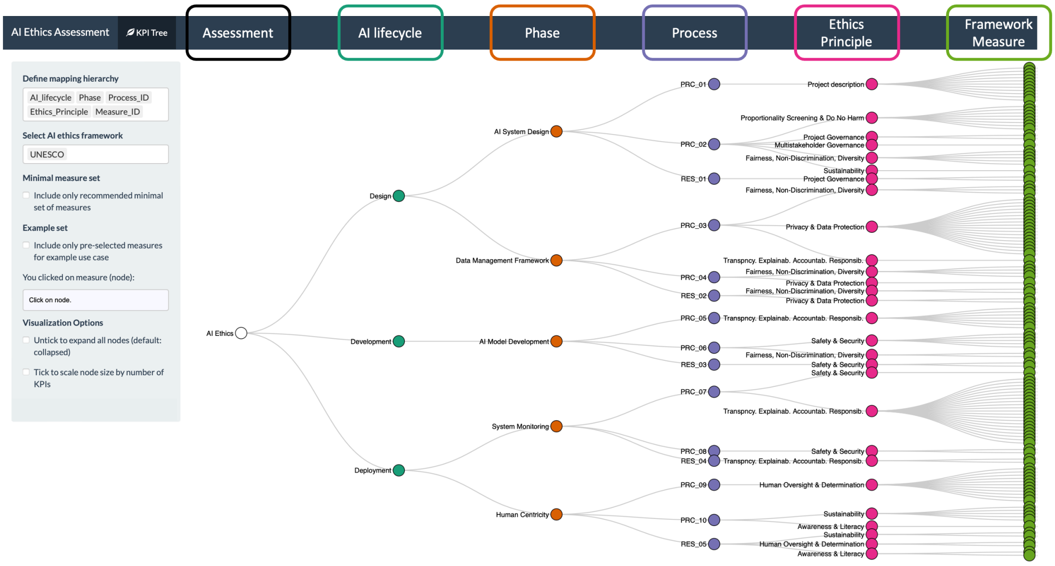
The 3rd phase. AI Model Development starts with the first process, Explainability and Transparency Assessment, which is designed to monitor system outputs and AI-supported decisions to ensure that outputs are explainable, transparent, and aligned with ethics guidelines and stakeholder expectations. This process shall identify checkpoints for feedback collection and continuous monitoring to align the system with human needs and ethical standards. The second process, Assess Safety and Security, selects measures to ensure safety and security of the AI system within a set of defined categories, such as data safety, system robustness, functionality, and detection of potential vulnerabilities of (cyber)security.

The 4th phase. System Monitoring focuses on the first process, Define System Review and Monitoring Process, on the development of a protocol for system evaluation. System monitoring is designed as a continuous process and its goal is to ensure that the AI system operates ethically throughout the whole AI system lifecycle. The second process, Assessment of Implementation Fairness, will evaluate if social justice, fairness and non-discrimination are safeguarded in all AI system structures and layers, for example data intake, algorithmic processing and decision-making.

The 5th phase. Human Centricity starts with the first process, Assessment of Human Oversight, to ensure that the AI system includes different dimensions of oversight which include developer oversight, public oversight, user oversight, and reviewer oversight. During this process, a documented procedure for collecting and analyzing user feedback shall be developed to detect and address ethical challenges in all AI system lifecycle stages. The second process, Assessment of the Sustainability and Social Impact, ensures the continuous assessment of human, social, cultural, and environmental impacts of the AI system. This process shall identify sustainable practices which in turn can address adverse effects on societal and environmental levels.

3.5 Applicability of the AI Ethics Assessment Blueprint

The blueprint is designed to enable the operationalization of AI ethics assessment. The applicability of the blueprint is manifold. It addresses AI systems working alongside human subjects, for example in robotic-assistance or in AI-supported decision-making. It assesses impact along the AI supply chain where downstream AI-driven products or solutions are built around a (generative) AI model offered by an upstream provider. To allow for diverse applications to be assessable, we established a harmonized terminology by mapping assessment questions of a chosen ethics framework to our phases and processes of the AI ethics assessment blueprint. Our approach focused initially on measures of the UNESCO ethical impact assessment with about 160 questions or measures, but it can easily be extended to other frameworks, for example to the European Commission's Assessment List for Trustworthy AI (ALTAI). The mapping of measures is visualized in the dendrogram in figure 3. Answers to the ethics framework assessment questions will then contribute to the outcomes of the blueprint in practice. An effective and timely assessment will require a screening and application or use case specific selection of framework questions. It is not required to answer all questions to summarize in a meaningful outcomes report. To guide the screening process of questions for relevant outcomes a definition of the outcomes measures of the blueprint is presented in figure 3.



AI Ethics Assessment | KPI Tree

Define mapping hierarchy: AI_lifecycle, Phase, Process_ID, Ethics_Principle, Measure_ID

Select AI ethics framework: UNESCO

Minimal measure set

□ Include only recommended minimal set of measures

Example set

□ Include only pre-selected measures for example use case

You clicked on measure (node): Click on node.

Visualization Options

□ Untick to expand all nodes (default: collapsed)

□ Tick to scale node size by number of KPIs

Assessment
AI lifecycle
Phase
Process
Ethics Principle
Framework Measure
AI Ethics
Design
Development
Deployment
AI System Design
Data Management Framework
AI Model Development
System Monitoring
Human Centricity
PRC_01
PRC_02
RES_01
PRC_03
PRC_04
RES_02
PRC_05
PRC_06
RES_03
PRC_07
PRC_08
RES_04
PRC_09
PRC_10
RES_05
Project description
Proportionality Screening & Do No Harm
Project Governance
Multistakeholder Governance
Fairness, Non-Discrimination, Diversity
Sustainability
Project Governance
Fairness, Non-Discrimination, Diversity
Privacy & Data Protection
Transpncy. Explainab. Accountab. Responsib.
Fairness, Non-Discrimination, Diversity
Privacy & Data Protection
Fairness, Non-Discrimination, Diversity
Privacy & Data Protection
Transpncy. Explainab. Accountab. Responsib.
Safety & Security
Fairness, Non-Discrimination, Diversity
Safety & Security
Safety & Security
Transpncy. Explainab. Accountab. Responsib.
Safety & Security
Transpncy. Explainab. Accountab. Responsib.
Human Oversight & Determination
Sustainability
Awareness & Literacy
Sustainability
Human Oversight & Determination
Awareness & Literacy

Figure 3 Selection tool for AI ethics assessment blueprint with hierarchical mapping of ethics framework measures to ethics principles, processes, phases and stages of the AI lifecycle.

3.6 Outcomes catalogue for the AI Ethics Assessment Blueprint

Given the diversity of AI use cases the blueprint can address we cannot define application specific output measures and instead we propose an outcomes category for each of the 5 phases. For each outcomes category we provide a set of selectable outcome measures which we denote as AI Ethics key performance indicators (KPIs). The AI Ethics KPIs will then ensure the responsible design, development, and deployment of AI systems. The following outcomes catalogue is exemplary and has no intention of being complete.

3.7 AI Ethics KPIs for outcomes category of phase 1: “Scope and Governance of Ethics Assessment defined”

- **Define governance:** Assemble an Ethics Board, with roles, responsibilities, system objectives, and accountability structures defined within the first few months of project initiation.
- **Identify potential incidents:** Analyze current literature to identify potential incidents, and related proposed mitigation measures for the specific use case.
- **Define review process:** Establish regular review mechanism for ethics clearance by process of multi-stakeholder collaboration including ethics advisors.
- **Define scope:** Select AI system features that must undergo ethical screening by the Ethics Board.

3.8 AI Ethics KPIs for outcomes category of phase 2: "Data Fairness defined"

- **Identify at-risk-groups:** Identify at-risk groups which may be systematically disadvantaged by the AI system.
- **Define fairness:** Select fairness objectives and associated fairness metrics to measure consequences of biased or unfair data on model outputs with respect to harms and benefits which (at-risk) individuals may receive by use of the AI system.
- **Implement bias detection:** Implement bias detection processes throughout the AI lifecycle at defined bias checkpoints based on selected fairness metrics. Definition of processes for mitigation of bias present in data or outputs that impact fairness of the AI system.
- **Implement fairness audit:** Define regular screening and data audits to ensure compliance with fairness data guidelines and ethics principles.

3.9 AI Ethics KPIs for outcomes category of phase 3: "Model Materiality classified"

- **Assess transparency:** Document performance and uncertainty of the AI model with respect to fairness objectives. Justify personal attributes in the data that are used for fairness assessment.
- **Assess explainability:** Conduct explainability assessment of the AI system to ensure that the system's potential decision-influencing processes are clear and understandable to stakeholders of the system and to users and addresses of the system's outcome. Explainability assessments are of paramount importance for decision-assist systems. Here the focus is on detecting decision boundaries and deriving concrete recommendations for actions in gray area situations or for high-stakes decisions.
- **Assess safety and security:** Conduct a safety and security evaluation of the AI system to ensure all identified risks to fairness and operational safety are documented and classified by materiality prior to deployment.
- **Identify AI materiality:** Document all risks associated with the AI model's materiality and categorize all impacts with respect to severity and likelihood. Identify mitigation strategies for each risk.
- **Update materiality classification:** Define process for post-hoc assessment or audit of the AI model's materiality classification. Flag any newly ob-served risks and resolve in continuous system improvement initiatives.

3.10 AI Ethics KPIs for outcomes category of phase 4: "Mitigation Plans defined"

- **Define system monitoring:** Define system monitoring and review process to detect abnormal operation of the AI system. Address all identified ethical, operational, and security risks so that potential system impacts are aligned with fairness objectives.
- **Measure incident metrics:** Track incident response metrics so that monitoring enables fast time to detect (TTD) and fast time to resolve (TTR).
- **Define mitigation plan:** Define fallback or mitigation plan in case of trigger events from system monitoring or review.

- **Update mitigation strategies:** Evaluate effectiveness of mitigation plans by measuring incidents after a post-implementation phase. Align updates of mitigation measures with new risks or evolving model behaviors.

3.11 AI Ethics KPIs for outcomes category of phase 5: "Literacy of AI System defined"

- **Guide safe and responsible use:** Develop operationalization guidelines for the AI system. Implement AI literacy and ethics awareness program to en-sure that all stakeholders understand how to use and interact with the AI system considering ethics, and system limitations. Ensure that users of a collaborative AI system understand the embodied ethics under normal operation and ethical boundaries. Assess regularly (by user feedback or questionnaires) stakeholders' ability to exercise human oversight over the AI system. Train developers and system users in recognizing and mitigating ethical risks.
- **Evaluate human centeredness:** Analyze by regular post-deployment reviews that human-AI interaction and human oversight remain effective. These reviews will track how effectively the system supports AI-assisted decision-making.
- **Establish AI training for professional development:** Ensure regular updates of the AI literacy and sustainability training. Adjust training programs based on explaining observed versus expected outcomes, on system improvements, user and stakeholder feedback, and on advancement in state-of-art and energy-efficient technologies. Ensure that employees acquire sufficient knowledge in developing, improving, deploying or using the AI system throughout the entire life cycle.
- **Measure impact on social goals:** Identify Social Development Goals (SDGs) also known as Global Goals adopted by the United Nations [35], societal benefits/social goals, or sustainability goals where the AI system can create impact on. Ensure regular screening so that the system's social impact aligns with ethical standards and long-term social benefits, and that the environmental issues are mitigated through sustainable practices during system operations.
- **Measure energy consumption:** Measure energy consumption and related costs during training and inference stages. Identify options to minimize the system's carbon footprint, for example by choosing a smaller (foundational) AI model or by effective finetuning. Compare effects of model hosting on premise, on cloud and on edge (device).

4 Use Case: *StableArtists* - Generative Art in Education

4.1 Objectives of the AI system *StableArtists*

We propose an AI system, generative AI for Arts Education with the acronym StableArtists. The technical realization of the system is based on a custom-trained text-to-image AI model that generates images based on a given prompt or textual description. The custom-trained model is obtained through a fine-tuning process where a pre-trained base model is trained further on curated data of digitalized artwork which was previously created by students. The fine-tuned model adjusts the weights of the base model so that it can now produce images in the artistic style of the peer group of students who contributed with their artwork. The workflow for image generation by the *StableArtists* app is presented in figure 4. The main goal of this system is to build AI literacy by helping students to acquire the knowledge necessary to understand AI from a technical, ethical and user or business needs perspective as described in UNESCO's AI competency framework for students [34].

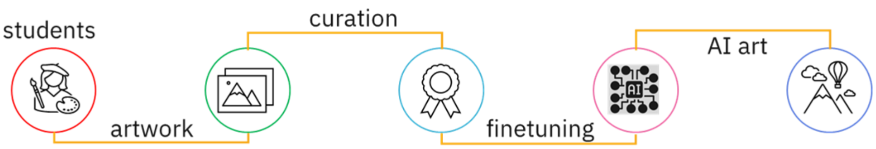


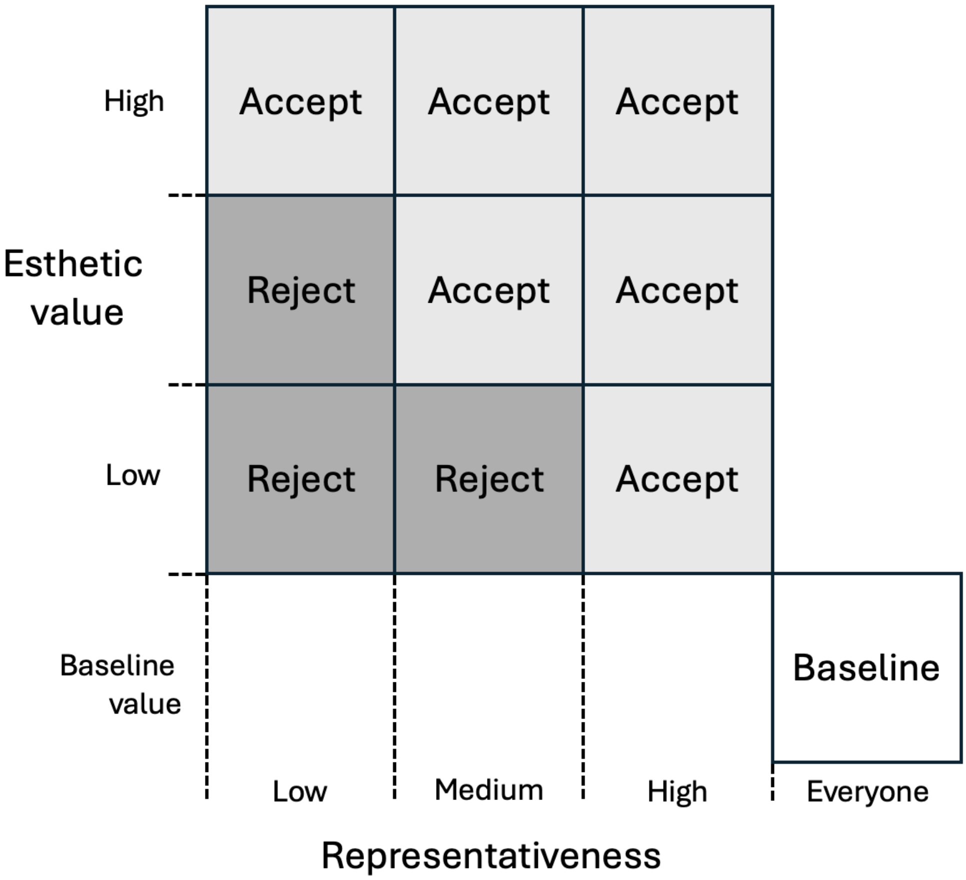
Figure 4 Steps needed to generate images by the *StableArtists* app involve collection and curation of student artworks which is used for fine-tuning a LORA model which produces AI art in the artistic style of the students.

4.2 Ethics-by-Design Approach

StableArtists allows the AI-based creation of artwork that reflects the diverse skills and styles of students from different age groups or backgrounds. The system is designed to be used in formal and non-formal educational settings. The user group consists of students under the guidance of a teacher or instructor. The development of the StableArtists system is motivated by an educational objective. Students shall learn to identify biases, acquire knowledge about ethical AI practices, and eventually become responsible citizens and remain independent actors in an increasingly AI-driven society. StableArtists provides the first use case to test the operability of the AI ethics assessment blueprint. We use the outcomes of the assessment for an ethics-by-design approach in the specification, technical realization of the diversity-sensitive AI system and its intended use. The following section is based on the results of the selected measures (c.f. appendix B) from the UNESCO ethical impact assessment [33] following the processes of the AI ethics blueprint.

4.3 Acceptability of the fairness-performance equilibrium

StableArtists has the dilemma to maximize two antagonistic metrics of the underlying AI model which are fairness and performance. Fairness is measured with respect to the representation of students who contribute artwork to the training data. Students are characterized by a set of features such as age, gender, ethnicity. The performance or quality of the model is measured by the mean esthetic value of the artwork composing the training data. For finetuned image-generation models we can fairly assume that the quality of the output is representative to the quality of the input training data. The quality, or synonymously the esthetic value, will be established by grading individual student's contributions by (i) grading by the teacher, (ii) consensus decisions by the students, or (iii) by a multi-modal AI model acting as a "judge". The students can vote on their preferred evaluation method. The finetuned model has the task to produce images of higher quality (mean grade) than a baseline model where all data would be included in the finetuning process. To fulfill this objective, artwork of lower grades must be removed from the training data which introduces selection bias. This exclusion bias will in turn lower the representativeness of the model with respect to students. The stakeholders of the system must agree on a bias mitigation strategy to select those images which will improve the esthetic value and still balance the representation of diverse student characteristics in the training data. Different bias mitigation strategies can be mapped by a materiality matrix assessment of the AI model as shown in figure 5. Students will understand how (cultural) bias may be inherent to generative AI systems as output bias is related to bias in the seen training data. StableArtists serves as a practical example through which students will gain insights into the ethical implications of AI technologies and developing a more responsible approach to their use.



	Low	Medium	High	Everyone
High	Accept	Accept	Accept	
	Reject	Accept	Accept	
Low	Reject	Reject	Accept	
Baseline value				Baseline

Figure 5 AI model materiality matrix assessment. Esthetic value measures the appreciation of art. The Esthetic value of the training data correlates to the generated output images of the finetuned model. Representativeness is the measure of selection bias for excluded images in the training dataset to achieve a higher esthetic value.

5 Conclusions

The paper proposes a framework for the ethics assessment along the AI lifecycle divided in phases and processes. This blueprint is based on the concept of adaptability; the framework is agnostic to specific ethics guidelines, regulatory approaches, industry sectors, business models, and technologies. It allows to choose use-case specific measures from the selected ethics framework (e.g. UNESCO) and to prioritize the most relevant ethics KPIs from the outcomes catalog. Conducting an AI ethics assessment according to the blueprint is not merely a compliance criterion; but adds value to the overall AI system by enhancing user adoption and trustworthiness, towards achieving Trusted AI. Trusted AI in practice requires two components, first an enforceable component to achieve compliance with regulatory standards on AI quality, and second an voluntarily component built on an AI assessment blueprint for ethics-by-design approach with selectable an adaptable AI Ethics KPIs.

5.1 Recommendations further research

While the Blueprint provides a promising foundation for AI ethics assessment, further research is needed to continue refining the framework. We propose three prospective directions:

1. Develop a process for applying the blueprint in ethical assessment of potential transitions of AI applications between different risk categories with respect to the classification defined by the EU AI Act.
2. Test the adaptability of the AI ethics assessment framework through use cases in: (i) different geographical zones with different interpretability of the ethics principles, and in (ii) sensitive areas like healthcare, recruiting, AI at the workplace, or collaborative AI systems.
3. Identify generalisation aspects of AI Ethics Assessment across different application field sectors, with respect to the harmonization of outcomes, and guiding the standardisation of AI Ethics Assessment.

References

1 African Union. *Continental Artificial Intelligence Strategy. Harnessing AI for Africa's Development and Prosperity.* African Union, July 2024.

2 BCLP. US state-by-state AI Legislation snapshot. *bclplaw.com*, 2024.

3 Anu Bradford. *The Brussels Effect: How the European Union Rules the World.* Oxford University PressNew York, 1 edition, February 2020.

4 Anu Bradford. *Digital Empires: The Global Battle to Regulate Technology.* Oxford University Press, Oxford, New York, September 2023.

5 Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR, January 2018. URL: http://proceedings.mlr.press/v81/buolamwini18a.html.

6 CIFAR. Pan-Canadian Artificial Intelligence Strategy. *cifar.ca*, 2017.

7 Nicholas Kluge Corrêa, Camila Galvão, James William Santos, Carolina Del Pino, Edson Pontes Pinto, Camila Barbosa, Diogo Massmann, Rodrigo Mambrini, Luiza Galvão, Edmund Terem, and Nythamar De Oliveira. Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance. *Patterns*, 4(10):100857, October 2023. doi:10.1016/j.patter.2023.100857.

8 Embassy of the People's Republic of China in Grenada. Global AI Governance Initiative. *gd.china-embassy.gov.cn*, October 2023.

9 European Commission. Directorate-General for Communications Networks, Content and Technology. *Ethics Guidelines for Trustworthy AI.* Publications Office, LU, 2019.

10 European Commission. Directorate-General for Communications Networks, Content and Technology. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self Assessment.* Publications Office, LU, 2020.

11 European Parliament and European Council. AI Act, Regulation 2024/1689. *Official Journal of the European Union*, June 2024.

12 Luciano Floridi. Soft Ethics and the Governance of the Digital. *Philosophy & Technology*, 31(1):1–8, March 2018. doi:10.1007/s13347-018-0303-9.

13 Government of Canada. Bill C-27: An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts. *justice.gc.ca*, November 2022.

14 Government of Canada. The Artificial Intelligence and Data Act (AIDA). *ised-isde.canada.ca*, 2023.

15 Government of Nigeria. National Artificial Intelligence Strategy. *ncair.nitda.gov.ng*, August 2024.

16 Ministry of Law and Justice. The Digital Personal Data Protection Act. *meity.gov.in*, August 2023.

17 Monetary Authority Singapore. Assessment Methodologies for Responsible Use of AI by Financial Institutions. *mas.gov.sg*, February 2022.

18 NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report, National Institute of Standards and Technology, Gaithersburg, MD, January 2023. `doi:10.6028/nist.ai.100-1`.

19 NITI Aayog. National Strategy for AI #AIForAll. *niti.gov.in*, 2018.

20 Ricardo Ortega-Bolaños, Joshua Bernal-Salcedo, Mariana Germán Ortiz, Julian Galeano Sarmiento, Gonzalo A. Ruz, and Reinel Tabares-Soto. Applying the ethics of AI: A systematic review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*, 57(5):110, April 2024. `doi:10.1007/s10462-024-10740-3`.

21 Personal Data Protection Commission Singapore and Infocomm Media Development Authority. Model AI Governance Framework (2nd edition). *pdpc.gov.sg*, January 2020.

22 PRC Cyberspace Administration. Interim Measures for the Management of Generative Artificial Intelligence Services (translated). *cac.gov.cn*, July 2023.

23 PRC Ministry of Science and Technology. Ethical Norms for New Generation Artificial (translated). *most.gov.cn*, October 2021.

24 Republic of South Africa. National Artificial Intelligence Policy Framework. *dcdt.gov.za*, August 2024.

25 Reuters. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women. *reuters.com*, October 2018.

26 The Hollywood Reporter. Scarlett Johansson's AI Legal Threat Sets Stage for Actors' Battle With Tech Giants. *hollywoodreporter.com*, May 2024.

27 The Indian Express. His 'jhakaas': HC issues order against misuse of Anil Kapoor's persona. *indianexpress.com*, May 2024.

28 The White House. Blueprint for an AI Bill of Rights. *whitehouse.gov*, October 2022.

29 The White House. Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI. *whitehouse.gov*, July 2023.

30 The White House. Delivering on the Promise of AI to Improve Health Outcomes. *whitehouse.gov*, December 2023.

31 The White House. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. *whitehouse.gov*, October 2023.

32 UNESCO. Recommendation on the Ethics of Artificial Intelligence. Technical report, UNESCO, 2022. URL: `https://unesdoc.unesco.org/ark:/48223/pf0000381137`.

33 UNESCO. Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial Intelligence. Technical report, UNESCO, 2023. `doi:10.54678/YTSA7796`.

34 UNESCO. AI competency framework for students. Technical report, UNESCO, 2024.

35 United Nations. Sustainable Development Goals. *sdgs.un.org*, 2024.

36 United Nations. AI Advisory Body. Governing AI for Humanity. Technical report, United Nations, September 2024.

37 Angela Huyue Zhang. The Promise and Perils of China's Regulation of Artificial Intelligence. *Columbia Journal of Transnational Law (forthcoming)*, January 2024. `doi:10.2139/ssrn.4708676`.

A Appendix 1: UNESCO Recommendation on Ethics of AI

UNESCO produced the first global standard on AI ethics – the Recommendation on the Ethics of Artificial Intelligence [8]. This framework was adopted by all 193 Member States in 2021. The Recommendation is centered around 10 principles:

I. Proportionality and Do No Harm. through which AI must be used only to achieve its legitimate purposes. It must not cause harm, discriminate, or manipulate. Risk assessments must ensure AI goals are appropriate, balanced, respect human rights, and are scientifically reliable.

II. Safety and Security. where AI actors should prevent and address unwanted harms (safety risks) and vulnerabilities to attacks (security risks).

III. Fairness and Non-Discrimination. by which AI actors must ensure fairness, inclusivity, and accessibility, address biases and digital divides. Member States should promote equity, and advanced countries should support less advanced ones. Measures for discrimination must be available.

IV. The Sustainability. principle states that AI technologies can either support or hinder sustainability goals, depending on their application. A continuous assessment of their human, social, cultural, economic, and environmental impact is required to align the system with the Sustainable Development Goals (SDGs).

V. Right to Privacy, and Data Protection. recommends that privacy (imperative for human dignity and autonomy) is protected throughout the AI lifecycle. Data handling must align with international and local laws, and strong data protection frameworks should be established considering societal and ethical aspects.

VI. Human Oversight and Determination. by which member states must ensure that ethical and legal responsibility for AI systems can always be attributed to individuals or entities. Human oversight should include both individual and public oversight. Ultimate responsibility and accountability are always ascribed to humans.

VII. Transparency and Explainability. support accountability, help individuals understand AI decisions, and promote democratic oversight. UNESCO recommends that the level of transparency and explainability should be appropriate to the context of use, as there may be tensions between these two and other principles such as privacy, safety and security.

VIII. Responsibility and Accountability. principle recommends developing AI systems that are auditable and traceable. Oversight, impact assessments, audits, and whistle-blower measures are needed to avoid conflicts with human rights and environmental standards.

IX. Awareness and Literacy. states that the public awareness of AI and data must be increased through open education, civic engagement, civil society actions, academia and the private sector involvement, etc. AI education needs to address its impact on human rights, freedoms, and the environment.

X. Multi-Stakeholder and Adaptive Governance and Collaboration. calls on data use to respect international law and national sovereignty, allowing states to regulate data within their territories and ensure data protection while upholding privacy rights. Stakeholder participation is needed to achieve inclusive AI governance, involving governments, organizations, the technical community, civil society, academia, media, policymakers, and others. Participation from marginalized groups and Indigenous Peoples is contributing to sustainable development and effective AI governance.

In alignment with the above-mentioned principles, following the need for an AI impact assessment, UNESCO developed a methodology for ethical impact assessment of AI systems in 2023. The methodology was published in the document Ethical Impact Assessment: A Tool of the Recommendation on the Ethics of AI [33]. The goal of the assessment is to ensure alignment of AI system with values and principles recognized by UNESCO in the Recommendation. However, there is still a step to go from endorsement of a recommendation by governments to an actual implementation of the ethical impact assessment by AI producers in practice.

B Appendix 2: Use Case *StableArtists*

Fig. 6 shows the selected UNESCO framework measures mapped to the blueprint for AI ethics assessment which we conducted on the generative AI use case *StableArtists*. A description of the measures is given below.

Questions for phase 1

- *Q-111:* Please provide an initial description of the AI system you intend to design, develop or deploy:
- *Q-112:* Please describe the aim or objective of this system. If the aim is to address a specific problem, please specify the problem you are trying to solve. Please also specify how this system may fit within broader schemes of work:
- *Q-1141:* Who will the users who interact with your system be (include their level of competency)?
- *Q-62214:* Have you developed a process to document how data quality issues can be resolved during the design process?

Questions for phase 2

- *Q-6232:* How has the principle of fairness been approached from a technical perspective? For example, are you able to specify what the technical notion of fairness is that the AI system is calibrated for? (e.g., individual fairness, demographic parity, equal opportunity, etc.)
- *Q-4245:* Which activities will help your team to identify potential impacts and ensure they are mitigated?

Questions for phase 3

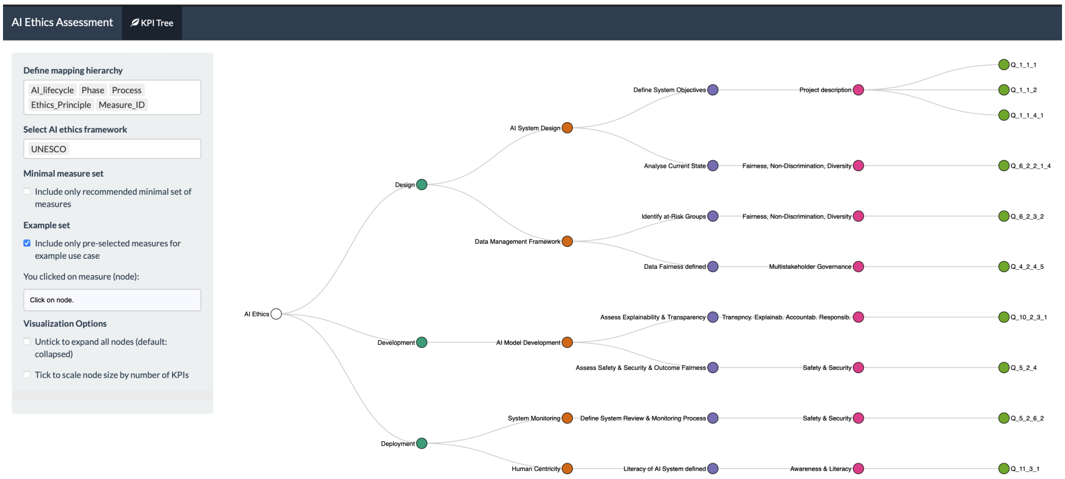
- *Q-10231:* Is the algorithm, including its inner-working logic, open to the public or any oversight authority? Is the code of the AI system in an open-source format?
- *Q-524:* If the training data or data being processed by the AI system were poisoned or corrupted, or if your system was manipulated, how would you know?

Questions for phase 4

- *Q-5262:* How often will the AI system be tested in the future and which components will be tested?

Questions for phase 5

- *Q-1131:* What are the prospective positive impacts of the system on AI awareness and literacy? How, if at all, could the deployment of this system increase awareness surrounding AI? Are there any other ways in which this system could increase awareness and literacy?



Define mapping hierarchy

AI_lifecycle Phase Process Ethics_Principle Measure_ID

Select AI ethics framework

UNESCO

Minimal measure set

□ Include only recommended minimal set of measures

Example set

☑ Include only pre-selected measures for example use case

You clicked on measure (node):

Click on node.

Visualization Options

□ Untick to expand all nodes (default: collapsed)

□ Tick to scale node size by number of KPIs

Figure 6 Selected framework measures for the use case *StableArtists*.