




1 Towards Trusted AI: A Blueprint for Ethics

2 Assessment in Practice

3 **Christoph Tobias Wirth**¹   

4 German Research Center for Artificial Intelligence (DFKI GmbH), Smart Data & Knowledge
5 Services, Kaiserslautern, Germany

6 **Mihai Maftai** 

7 German Research Center for Artificial Intelligence (DFKI GmbH), Ethics Team, Saarbrücken,
8 Germany

9 **Rosa Esther Martín-Peña**

10 German Research Center for Artificial Intelligence (DFKI GmbH), Educational Technology Lab,
11 Berlin, Germany

12 **Iris Merget**

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

15 — Abstract —

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

29 **2012 ACM Subject Classification** Computing methodologies → Artificial intelligence; Social and
30 professional topics → Codes of ethics; Human-centered computing → Collaborative and social
31 computing; Applied computing → Arts and humanities

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

35 **Digital Object Identifier** 10.4230/OASICS.SAIA.2024.12

36 **Category** Academic Track

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

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

¹ corresponding author



42 **1** Introduction

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

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

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

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

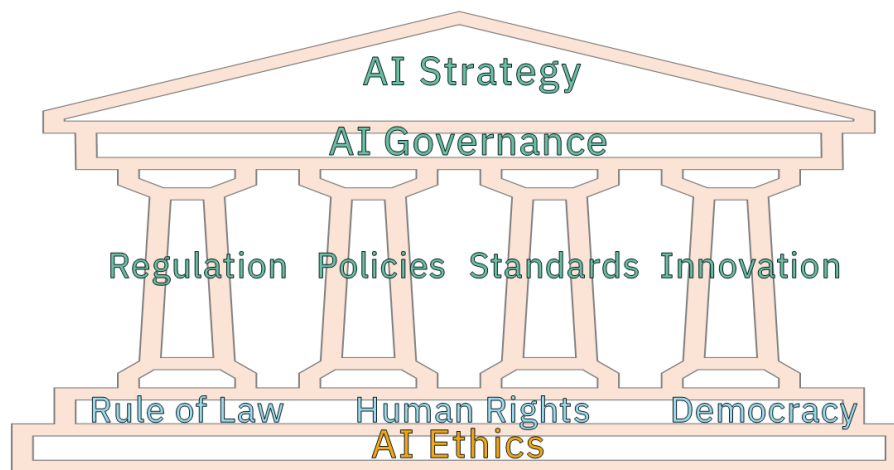
66 The concept of *Trusted AI* can be explored from multiple distinct perspectives. From the
67 multiplicity of definitions, we understand the term "Trusted AI" as the evaluation of artificial
68 intelligence concerning its reliability and effectiveness in individual applications from the
69 user's perspective,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

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

77 **2** Current global state of AI Ethics implementation

78 Countries around the world define national AI strategies to leverage the rapid advancement of
79 AI technology. Executing an AI strategy needs governance that includes oversight mechanisms
80 to address risks such as bias, privacy infringement and misuse, but also to build and maintain
81 trust in AI, while at the same time enables AI innovation and research. On international level,
82 the United Nations laid out foundations for the first global architecture for AI governance
83 based on international cooperation [36]. An effective AI governance framework provides a
84 structured approach based on the pillars of regulation, sound AI policy, supporting standards
85 for compliance, and innovation measures. Figure 1 illustrates the building blocks of an AI
86 governance framework. This structure highlights how every element depends on a strong
87 ethical foundation. The AI Strategy represents, in the context of a state, a government's

88 approach to the development, deployment, and regulation of AI technologies and from a
 89 corporate perspective, represents the enterprise AI roadmap. Below, the concept of AI
 90 Governance defines the structural support required to operationalize the pillars (regulation,
 91 policies, standards, and innovation), aligning them under a unified framework. The four
 92 pillars are grounded on a structural basis represented by the foundational aspects of rule
 93 of law, human rights, and democracy. At the very bottom, ethics serves as a fundamental
 94 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



■ **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.

95 ethics, examining how different Digital Empires are responding to the challenges posed by
 96 AI. Different regulatory approaches, ethical guidelines, and policy initiatives that have been
 97 implemented to ensure that AI technologies are developed and deployed responsibly will
 98 be explored. The Digital Empires create a pull effect on other countries in adapting their
 99 regulatory approach commonly denoted as Brussels, Beijing, and California effect. The
 100 following overview only presents the current point-in-time snapshot of the operationalization
 101 potential for AI ethics by selected global digital powers. The choice of geographies is not
 102 meant to be biased and presented in alphabetic order.

104 Africa

105 The African Union's (AU) "Continental AI Strategy" prioritizes "economic growth, social
 106 progress, and cultural renaissance" [1]. with the help of AI systems. The principles focus on
 107 local first and people-centeredness as well as ethics and transparency, inclusion and diversity,
 108 human rights and dignity, peace and prosperity, cooperation and integration, and skills
 109 development, public awareness and education. This strategy puts forward an Africa-centric
 110 and development-oriented and inclusive approach around five focus areas notably: harnessing
 111 AI's benefits, building AI capabilities, minimizing risks, stimulating investment and fostering
 112 cooperation. It is part of the AU Agenda 2063 which aims to further peace, prosperity,

12:4 Towards Trusted AI: A Blueprint for Ethics Assessment in Practice

113 self-governance, and international cooperation. The strategy is divided into 5 areas of actions
114 which should be implemented between 2025 and 2030, they are the following: Maximizing AI
115 Benefits, Building Capabilities for AI, Minimizing AI Risks, African Public and Private Sector
116 Investment in AI, and Regional and International Cooperation and Partnerships. Additionally,
117 South Africa has published the "National Artificial Intelligence Policy Framework" [24] and
118 Nigeria its corresponding "National Artificial Intelligence Strategy" [15], both in August 2024.

119 Canada

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

132 China

133 The National Governance Committee for the New Generation Artificial Intelligence published
134 the "Ethical Norms for the New Generation Artificial Intelligence" [23] in September 2021.
135 The norms for the AI life cycle include fairness, justice, harmony, and security, preventing bias,
136 discrimination, and privacy/information leaks. China has launched the Global AI Governance
137 Initiative (GAIGI) [8] as part of its Belt and Road Initiative, promoting international
138 cooperation in AI governance. Unlike the EU AI Act, China has been regulating specific
139 AI applications individually, such as internet recommendation algorithms, deep synthesis
140 technology, and generative AI. This approach allows China to address specific issues with
141 correspondent rules, building new policy tools and regulatory expertise with each regulation.
142 After the release of ChatGPT the Cyber Space Administration of China (CAC) reacted within
143 6 months with Draft Measures for Generative AI [37]. China's AI regulations are designed to
144 be iterative, allowing for quick updates in response to rapid AI developments. The "Interim
145 Administrative Measures for Generative AI Services" [22] exemplify this iterative approach,
146 with the expectation that AI regulation remains highly adaptive.

147 Europe

148 In August 2024 the world's first regulation on AI, the EU AI Act, went into force. This
149 Regulation shall support the EU objective of being a "global leader in the development
150 of secure, trustworthy and ethical AI" [11] and it shall "ensure the protection of ethical
151 principles" [11]. Recognition on the international level of the European legislation reflects the
152 global interest and adaptiveness to the EU regulatory framework, generating the Brussels
153 effect [3]. The AI Act's binding rules are built on a risk-based approach. However, the
154 implementation of ethics principles for providers and deployers of AI is left on a voluntary basis.
155 The AI Act suggests that for voluntary ethics codes of conduct to be effective, they should
156 be based on clear objectives and key performance indicators to measure the achievement

157 of those objectives. The AI Act does not explicitly mention that an ethics assessment
158 framework for trustworthy AI must be applied. The AI Act encourages to implement ethics
159 processes in AI system development. In this regard, the EU issued both independently and in
160 collaboration with international bodies multiple ethics principles, guidelines, and assessment
161 frameworks, such as: (i) The High-Level Expert Group on Artificial Intelligence (HLEG)
162 Ethics Guidelines for Trustworthy AI [9], (ii) the Assessment List for Trustworthy Artificial
163 Intelligence (ALTAI) [10],(iii) UNESCO Ethical impact assessment [33].

164 **India**

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

176 **Singapore**

177 Though Singapore does currently not have any binding regulation on AI, the Singaporean
178 government has developed variety of sector-specific and voluntary frameworks to guide the
179 responsible use of AI and to safeguard public interest in AI ethics and governance. In
180 the following two frameworks are introduced, one for financial institutions and the other
181 one for the deployment of generative AI. In 2022 the Monetary Authority of Singapore
182 published assessment methodologies for the fairness, ethics, accountability and transparency
183 (FEAT) principles, to guide the responsible use of AI by financial institutions [17]. The
184 fairness assessment methodology ensures that the AI-assisted decision-making process does
185 not systematically disadvantage individuals or groups of individuals, without appropriate
186 justification. The fairness principle is checked throughout the lifecycle of the AI system's
187 development process based on the key concepts such as selection of personal attributes,
188 types of bias and their mitigation methods, and fairness objectives and their metrics. In
189 2024 Singapore released the "Model AI Governance Framework for Generative AI" [21]
190 which addresses risks related to Generative AI and provides guidance on practices for safety
191 evaluation of Generative AI models. The framework is based on the core principles of
192 accountability, transparency, fairness, robustness and security and it extends the previous
193 version from 2019 developed for Traditional AI.

194 **U.S.A.**

195 In October 2023 the White House released the Executive Order on the Safe, Secure and
196 Trustworthy Development and Use of Artificial Intelligence. The Biden Administration fo-
197 cuses on eight principles, such as: Safety/Security, Robustness, Reliability, and Repeatability.
198 AI must be standardized and testable before its use to diminish risks. Furthermore, constant
199 monitoring is necessary to ensure ethical development, resilience against misuse, and compli-
200 ance with Federal laws [31]. The next step is the Blueprint for an AI Bill of Rights, with the

201 principles: safe and effective systems, algorithmic discrimination protection, data privacy,
202 notice and explanation, and human alternatives, consideration and fallback [28]. Although
203 this is a voluntary framework Amazon, Anthropic, Google, Inflection, Meta, Microsoft, and
204 OpenAI have offered their commitment [29]. Additionally, 28 healthcare providers and payers
205 have committed to the responsible use of AI in healthcare [30]. The different states can also
206 make their own laws to regulate AI [2]. The Artificial Intelligence Risk Management Frame-
207 work was published in January 2023 by the National Institute of Standards and Technology
208 (NIST). NIST uses a modified version of the AI lifecycle from the OECD Framework for the
209 Classification of AI systems. After the release of ChatGPT NIST has published the Artificial
210 Intelligence Risk Management Framework: Generative Artificial Intelligence Profile in July
211 2024.

212 **2.1 Implications for AI ethics assessment – The need for a process** 213 **blueprint**

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

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

232 In summary, a global ethical framework, with the objective to guide the deployment of
233 trusted AI and to promote the responsible use of AI, implies the need of a process blueprint.
234 The blueprint for an AI ethics assessment must fulfill two acceptance criteria. The first
235 criterion refers to its high level of independence, which implies it is agnostic to the underlying
236 regulatory model, to the deployed AI algorithm, to the technology in which the AI model
237 is embedded in, and it is agnostic to the needs of the industry sector or to the business
238 model or scale of business. The second criterion of the blueprint allows for adaptivity to
239 varying comprehension of ethical principles and values. As already been pointed out, the
240 interpretation or choice of ethical principles depends not only on the cultural perspective,
241 but it is also tailored to specific industry needs and it also aims to maximize the space for
242 AI innovation for which most national AI strategies of countries define a leading position.
243 Lastly, the AI ethics process blueprint that fosters a trusted AI ecosystem cannot be static.
244 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,

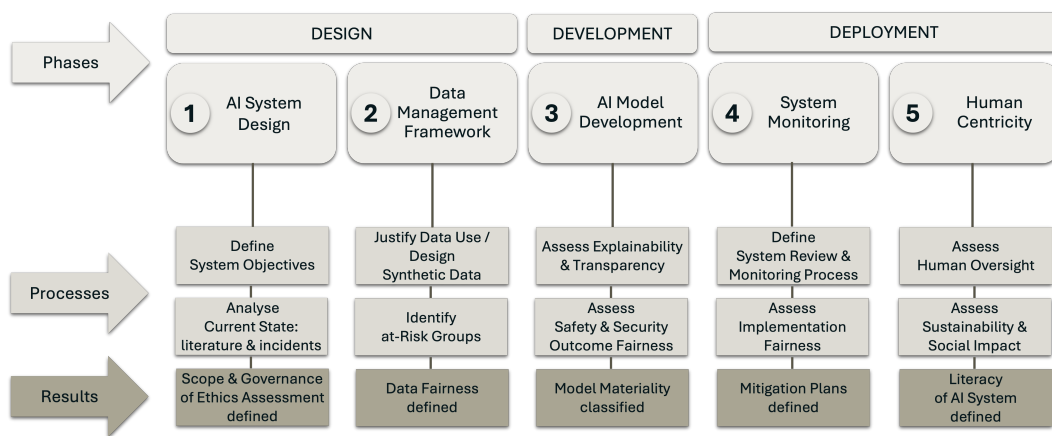
12:8 Towards Trusted AI: A Blueprint for Ethics Assessment in Practice

288 focusing on concrete phases such as system design, data management, model development,
289 system monitoring, and human-centric evaluation, the Blueprint can ensure that ethical
290 issues like fairness, transparency, and accountability are assessed in relation to the actual
291 system development process.

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

301 3.4 Specification: Detailing out the five phases of the Blueprint

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



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

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

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

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

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

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

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

347 **3.5 Applicability of the AI Ethics Assessment Blueprint**

348 The blueprint is designed to enable the operationalization of AI ethics assessment. The
349 applicability of the blueprint is manifold. It addresses AI systems working alongside human
350 subjects, for example in robotic-assistance or in AI-supported decision-making. It assesses
351 impact along the AI supply chain where downstream AI-driven products or solutions are
352 built around a (generative) AI model offered by an upstream provider. To allow for diverse
353 applications to be assessable, we established a harmonized terminology by mapping assessment
354 questions of a chosen ethics framework to our phases and processes of the AI ethics assessment
355 blueprint. Our approach focused initially on measures of the UNESCO ethical impact
356 assessment with about 160 questions or measures, but it can easily be extended to other
357 frameworks, for example to the European Commission's Assessment List for Trustworthy AI
358 (ALTAI). The mapping of measures is visualized in the dendrogram in figure 3. Answers to the

12:10 Towards Trusted AI: A Blueprint for Ethics Assessment in Practice

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.

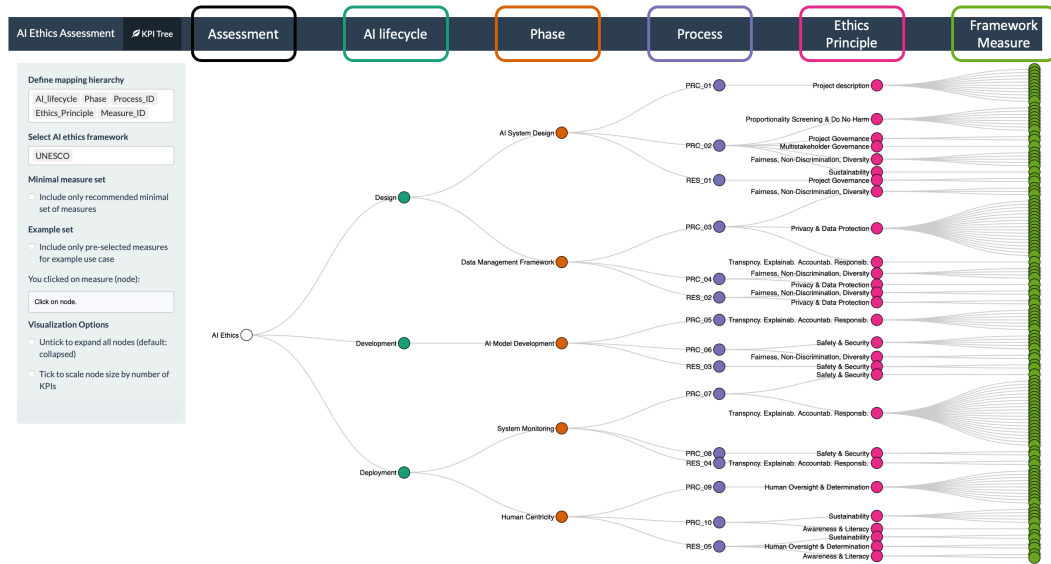


Figure 3 Selection tool for AI ethics assessment blueprint with hierarchical mapping of ethics framework measures to ethics principles, processes, phases and stages of the AI lifecycle.

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"

- 385 ■ **Identify at-risk-groups:** Identify at-risk groups which may be systematically disad-
386 vantaged by the AI system.
- 387 ■ **Define fairness:** Select fairness objectives and associated fairness metrics to measure
388 consequences of biased or unfair data on model outputs with respect to harms and benefits
389 which (at-risk) individuals may receive by use of the AI system.
- 390 ■ **Implement bias detection:** Implement bias detection processes throughout the AI
391 lifecycle at defined bias checkpoints based on selected fairness metrics. Definition of
392 processes for mitigation of bias present in data or outputs that impact fairness of the AI
393 system.
- 394 ■ **Implement fairness audit:** Define regular screening and data audits to ensure compli-
395 ance with fairness data guidelines and ethics principles.

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

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

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

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

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

428 3.11 AI Ethics KPIs for outcomes category of phase 5: "Literacy of AI 429 System defined"

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

456 4 Use Case: *StableArtists* - Generative Art in Education

457 4.1 Objectives of the AI system *StableArtists*

458 We propose an AI system, generative AI for Arts Education with the acronym *StableArtists*.
459 The technical realization of the system is based on a custom-trained text-to-image AI model
460 that generates images based on a given prompt or textual description. The custom-trained
461 model is obtained through a fine-tuning process where a pre-trained base model is trained
462 further on curated data of digitalized artwork which was previously created by students.
463 The fine-tuned model adjusts the weights of the base model so that it can now produce
464 images in the artistic style of the peer group of students who contributed with their artwork.
465 The workflow for image generation by the *StableArtists* app is presented in figure 4. The
466 main goal of this system is to build AI literacy by helping students to acquire the knowledge
467 necessary to understand AI from a technical, ethical and user or business needs perspective
468 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

504 AI technologies and developing a more responsible approach to their use.

	High	Accept	Accept	Accept
Esthetic value	---	Reject	Accept	Accept
	Low	Reject	Reject	Accept
Baseline value	---			Baseline
		Low	Medium	High
				Everyone
		Representativeness		

■ **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.

505 **5 Conclusions**

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

- 564 15 Government of Nigeria. National Artificial Intelligence Strategy. *ncair.nitda.gov.ng*, August
565 2024.
- 566 16 Ministry of Law and Justice. The Digital Personal Data Protection Act. *meity.gov.in*, August
567 2023.
- 568 17 Monetary Authority Singapore. Assessment Methodologies for Responsible Use of AI by
569 Financial Institutions. *mas.gov.sg*, February 2022.
- 570 18 NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical report,
571 National Institute of Standards and Technology, Gaithersburg, MD, January 2023. doi:
572 10.6028/nist.ai.100-1.
- 573 19 NITI Aayog. National Strategy for AI #AIForAll. *niti.gov.in*, 2018.
- 574 20 Ricardo Ortega-Bolaños, Joshua Bernal-Salcedo, Mariana Germán Ortiz, Julian Galeano Sarmi-
575 ento, Gonzalo A. Ruz, and Reinel Tabares-Soto. Applying the ethics of AI: A systematic
576 review of tools for developing and assessing AI-based systems. *Artificial Intelligence Review*,
577 57(5):110, April 2024. doi:10.1007/s10462-024-10740-3.
- 578 21 Personal Data Protection Commission Singapore and Infocomm Media Development Authority.
579 Model AI Governance Framework (2nd edition). *pdpc.gov.sg*, January 2020.
- 580 22 PRC Cyberspace Administration. Interim Measures for the Management of Generative
581 Artificial Intelligence Services (translated). *cac.gov.cn*, July 2023.
- 582 23 PRC Ministry of Science and Technology. Ethical Norms for New Generation Artificial
583 (translated). *most.gov.cn*, October 2021.
- 584 24 Republic of South Africa. National Artificial Intelligence Policy Framework. *dcdt.gov.za*,
585 August 2024.
- 586 25 Reuters. Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women.
587 *reuters.com*, October 2018.
- 588 26 The Hollywood Reporter. Scarlett Johansson's AI Legal Threat Sets Stage for Actors' Battle
589 With Tech Giants. *hollywoodreporter.com*, May 2024.
- 590 27 The Indian Express. His 'jhakaas': HC issues order against misuse of Anil Kapoor's persona.
591 *indianexpress.com*, May 2024.
- 592 28 The White House. Blueprint for an AI Bill of Rights. *whitehouse.gov*, October 2022.
- 593 29 The White House. Biden-Harris Administration Secures Voluntary Commitments from Leading
594 Artificial Intelligence Companies to Manage the Risks Posed by AI. *whitehouse.gov*, July 2023.
- 595 30 The White House. Delivering on the Promise of AI to Improve Health Outcomes. *whitehouse.gov*,
596 December 2023.
- 597 31 The White House. Executive Order on the Safe, Secure, and Trustworthy Development and
598 Use of Artificial Intelligence. *whitehouse.gov*, October 2023.
- 599 32 UNESCO. Recommendation on the Ethics of Artificial Intelligence. Technical report, UNESCO,
600 2022. doi:10.54678/YTSA7796.
- 601 33 UNESCO. Ethical impact assessment. A tool of the Recommendation on the Ethics of Artificial
602 Intelligence. Technical report, UNESCO, 2023. doi:10.54678/YTSA7796.
- 603 34 UNESCO. AI competency framework for students. Technical report, UNESCO, 2024.
- 604 35 United Nations. Sustainable Development Goals. *sdgs.un.org*, 2024.
- 605 36 United Nations. AI Advisory Body. Governing AI for Humanity. Technical report, United
606 Nations, September 2024.
- 607 37 Angela Huyue Zhang. The Promise and Perils of China's Regulation of Artificial Intelligence.
608 *Columbia Journal of Transnational Law (forthcoming)*, January 2024. doi:10.2139/ssrn.
609 4708676.

610 **A** Appendix 1: UNESCO Recommendation on Ethics of AI

611 UNESCO produced the first global standard on AI ethics – the Recommendation on the
612 Ethics of Artificial Intelligence [8]. This framework was adopted by all 193 Member States in
613 2021. The Recommendation is centered around 10 principles:

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

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

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

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

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

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

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

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

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

647 **X. Multi-Stakeholder and Adaptive Governance and Collaboration** calls on data use to
648 respect international law and national sovereignty, allowing states to regulate data within
649 their territories and ensure data protection while upholding privacy rights. Stakeholder
650 participation is needed to achieve inclusive AI governance, involving governments, organiz-
651 ations, the technical community, civil society, academia, media, policymakers, and others.

652 Participation from marginalized groups and Indigenous Peoples is contributing to sustainable
653 development and effective AI governance.

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

662 **B** Appendix 2: Use Case *StableArtists*

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

666 Questions for phase 1

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

676 Questions for phase 2

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

683 Questions for phase 3

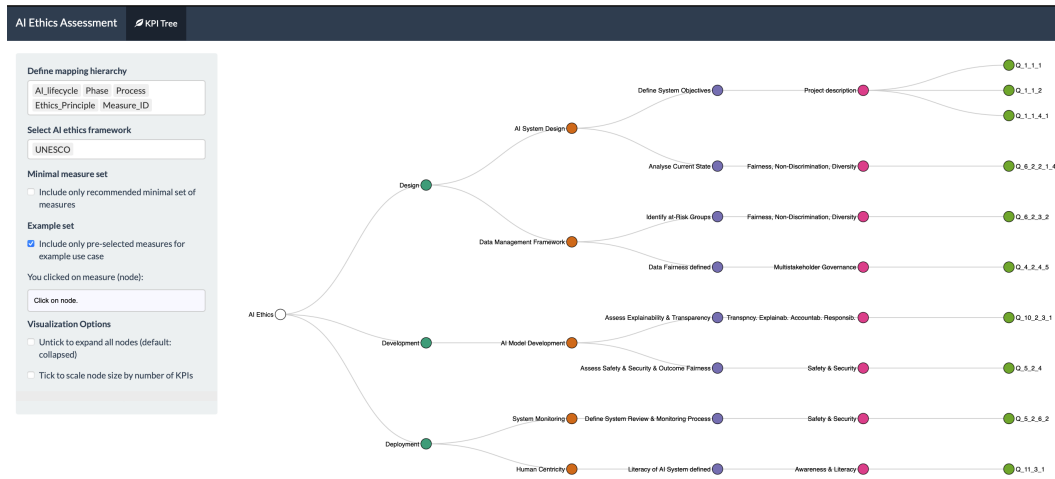
- 684 ■ *Q-10231*: Is the algorithm, including its inner-working logic, open to the public or any
685 oversight authority? Is the code of the AI system in an open-source format?
- 686 ■ *Q-524*: If the training data or data being processed by the AI system were poisoned or
687 corrupted, or if your system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

688 Questions for phase 4

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

691 **Questions for phase 5**

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



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