

RESPONSIBLE AND ETHICAL AI: EXPLAINING THE DCO AI ETHICS EVALUATOR

GUIDANCE DOCUMENT

2025

DOCUMENT DISCLAIMER

The following legal disclaimer (“Disclaimer”) applies to this document (“Document”) and by accessing or using the Document, you (“User” or “Reader”) acknowledge and agree to be bound by this Disclaimer. If you do not agree to this Disclaimer, please refrain from using the Document.

This Document, prepared by the Digital Cooperation Organization (DCO). While reasonable efforts have been made to ensure accuracy and relevance of the information provided, the DCO makes no representation or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained in this Document.

The information provided in this Document is intended for general informational purposes only and should not be considered as professional advice. The DCO disclaims any liability for any actions taken or not taken based on the information provided in this Document.

The DCO reserves the right to update, modify or remove content from this Document without prior notice. The publication of this Document does not create a consultant-client relationship between the DCO and the User.

The designations employed in this Document of the material on any map do not imply the expression of any opinion whatsoever on the part of the DCO concerning the legal status of any country, territory, city, or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The use of this Document is solely at the User’s own risk. Under no circumstances shall the DCO be liable for any loss, damage, including but not limited to, direct or indirect or consequential loss or damage, or any loss whatsoever arising from the use of this Document.

Unless expressly stated otherwise, the findings, interpretations and conclusions expressed in this Document do not necessarily represent the views of the DCO. The User shall not reproduce any content of this Document without obtaining the DCO’s consent or shall provide a reference to the DCO’s information in all cases.

By accessing and using this Document, the Reader acknowledges and agrees to the terms of this Disclaimer, which is subject to change without notice, and any updates will be effective upon posting.

TABLE OF CONTENTS

01	Introduction	4
	1.1 Purpose of the Tool	7
02	Building the Tool	8
	2.1 The DCO Principles for Ethical AI	9
	2.2 Risk Categories in the Tool	11
	2.3 Recommendation Categories in the Tool	15
03	Using the Tool	18
	3.1 Registration and Characterization	19
	3.2 Risk Assessment Process	21
	3.3 Visualization and Analysis	23
	3.4 Recommendations	23
04	Frequency of risk evaluations	25
	4.1 Baseline Assessment Schedule	26



01

INTRODUCTION

Recognising AI's transformative potential, and the need for its ethical development, the DCO executes an extensive programme of work to help ensure that AI respects and upholds fundamental human rights.

The foundation of the DCO's ethical AI governance efforts lies in its rigorous research efforts, which aim to understand how AI technologies intersect with fundamental human rights. This research is such an example and was driven by a recognition that the transformative potential of AI must be harnessed responsibly to ensure that it does not undermine individual freedoms nor exacerbate societal inequalities.

This body of research has been consolidated under the Ethical AI Governance Toolbox, a framework aligned with global standards that incorporates the latest advancements in ethical AI practices. This toolbox is formed by the in-depth analyses presented in the DCO reports "Rights by Design: Embedding Human Rights Principles in AI Systems" and "Responsible AI Governance: Global Lessons and International Best Practices for DCO Member States," together with the DCO Principles for Ethical AI and the DCO AI Ethics Evaluator (the "Evaluator").



One of the primary areas of focus for the DCO research has been privacy, which has emerged as a cornerstone of ethical AI governance. The research explored the **multifaceted nature of privacy risks**, including issues related to data collection, storage, and usage. It highlighted the growing concerns around surveillance practices, the lack of transparency in data handling, and the potential for misuse of personal information. The findings revealed the need for robust data protection measures and transparent consent mechanisms to safeguard individuals' privacy rights.

Another critical area of investigation was **algorithmic bias** and its implications for non-discrimination. The research identified how biases embedded in training datasets could lead to unfair outcomes in AI systems, particularly in sensitive areas like recruitment, healthcare, and financial services. These biases not only perpetuate existing inequalities but also undermine public trust in AI technologies. The findings underscored the importance of incorporating diverse datasets, implementing fairness audits, and fostering collaboration between technologists and ethicists to mitigate these risks.

Lastly, the intersection of AI and **freedom of expression** also received significant attention. The research examined how automated content moderation systems could both promote and restrict free speech. While these systems have the potential to combat harmful content and misinformation, they also risk over-censoring legitimate expression or failing to address culturally sensitive nuances. The findings highlighted the need for transparent and accountable content moderation practices that respect freedom of expression while addressing the challenges posed by harmful material.

To inform its governance framework, the DCO analysed a wide range of international standards and principles. This research laid the groundwork for a rights-based approach to AI governance, ensuring that the organisation's initiatives are firmly rooted in the principles of fairness, accountability, and transparency.

The Toolbox has been designed to assist the DCO Member State governments, developers, and deployers in navigating the intricate and dynamic task of implementing an ethical and human rights-based approach to AI governance. Rooted in the **DCO Principles for Ethical AI**, it provides practical guidance to ensure AI systems adhere to ethical standards, uphold human rights, and deliver meaningful benefits to society. The **DCO AI Ethics Evaluator** ("the tool") is a comprehensive tool designed to assist individuals and organizations to systematically assess and address ethical considerations related to their AI systems with a focus on human rights risks. The tool provides tailored guidance for both developers building AI systems and users deploying them based on the outcome of the assessment.

This innovative digital tool helps individuals and organizations identify potential human rights impacts, align their practices with ethical standards, and implement practical mitigation strategies. Its development was informed by the DCO's research on AI governance and extensive stakeholder consultations.



1.1 PURPOSE OF THE TOOL

The DCO AI Ethics Evaluator is a practical tool for managing specific risks in AI systems. It translates complex ethical considerations into concrete assessment criteria and actionable recommendations, helping the users implement effective safeguards for ethical risks with a focus on human rights.

The tool's distinct guidance for AI developers¹ and deployers² recognizes the different challenges at each stage of the AI lifecycle. While developers receive guidance on building in protections during system design and creation, deployers get practical advice on implementing robust controls and monitoring impacts of the AI tools they employ within their systems. This role-based approach, combined with clear risk assessment criteria, enables users to systematically evaluate and address ethical and human rights concerns in their AI systems. The DCO AI Ethics Evaluator has broad applicability across different AI use cases and organizational roles. For example, it could help:



1. The developer of an AI recruitment system

Evaluate potential risks around fairness and discrimination related to training data composition and algorithmic bias. Based on an assessment of these risks, the tool would then provide targeted guidance on, for example, implementing bias testing protocols, establishing transparent evaluation criteria, and building in appeal mechanisms for automated screening decisions.



2. The deployer of an AI healthcare diagnostic system

Evaluate risks related to data protection practices and system accuracy across different patient groups. Depending on the assessed level of risk, the tool would offer recommendations such as implementing robust data encryption, establishing comprehensive testing protocols, and ensuring meaningful human oversight of diagnostic recommendations.



3. The designer of an AI content moderation platform

Evaluate potential transparency and cultural sensitivity issues. If the risks were substantive, the tool would suggest designing clear flagging criteria, incorporating human review checkpoints, and implementing culturally-aware training parameters that could help create a more fair and accountable system.

¹ Developers are entities or individuals involved in the creation, training, and provision of AI systems. They are responsible for problem/activity definition, data collection and pre-processing, model training, testing and evaluation, and placing the AI system on the market. Developers may include providers of general-purpose AI (GPAI) models and those who create specific AI applications based on these models (Engler, A & Renda, A. 2022. Reconciling the AI Value Chain with the EU's Artificial Intelligence Act. Available at: <https://cdn.ceps.eu/wp-content/uploads/2022/09/CEPS-In-depth-analysis-2022-03-Reconciling-the-AI-Value-Chain-with-the-EU-Artificial-Intelligence-Act.pdf>)

² Deployers are individuals or entities that use an AI system for different purposes (within their professional scope, and also personal and non-professional activities). This includes businesses or governments that utilize AI as part of their core operations or for ancillary activities such as organizational management or recruitment. Deployers encompass organizations and individuals using AI for both inward-facing applications (such as internal processes, employee management, and operational efficiency) and outward-facing products and systems (such as customer-facing services, public applications, and market-facing solutions). In the context of this tool, we will focus on the use of AI systems for professional activities at all levels of implementation.

It is important to mention here that the DCO AI Ethics Evaluator is a unique tool that reflects international best practice and principles, as well as DCO research. All recommendations provided are supportive of, and should be used in conjunction with, local regulations and requirements. Users must conduct their own analysis of applicable local laws and regulations, as this tool does not substitute for legal compliance obligations in any jurisdiction. Implementation of any recommendations should be reviewed against current legal requirements in your region.³ The tool is specifically designed to assess AI systems where human rights and ethical considerations are paramount – including those that make or influence decisions about individuals, process personal data, or impact social interactions or the environment. While AI systems may still require their own rigorous safety and performance assessments, the tool's risk scoring methodology focuses specifically on evaluating the severity and likelihood of potential human rights impacts.

This focused scope aligns with the DCO's mission to promote ethical AI development that respects fundamental human rights, while acknowledging that different types of AI systems may require different forms of assessment and oversight depending on their intended use, ethical implications, and potential impact on human rights.



³ DCO does not guarantee the accuracy, completeness, reliability, or suitability of the assessment results, recommendations, or any content available on the Portal. The generated outcomes are intended for informational purposes and should not be considered binding guidance, official policy, or a definitive measure of AI risks. DCO bears no responsibility for any decisions, actions, investments, or policies formulated by Users based on the assessment results. Users acknowledge that the Portal does not provide legal, financial, or regulatory advice, and any reliance on its content is solely their own responsibility. For official AI policy recommendations, strategic planning, or the development of an AI adoption roadmap, users are advised to seek expert consultation and refer to authoritative national and international sources.



02

BUILDING THE TOOL

The DCO AI Ethics Evaluator was conceptualized as a bridge between AI innovation and ethical governance, with a strong foundation in human rights principles. Its development began with research to understand how AI technologies potentially impact fundamental rights like privacy, equality, and freedom of expression.⁴ This section outlines the core principles guiding the tool, the structured risk assessment framework, and the tailored recommendations designed to support responsible AI deployment, which together form the overall framework for the DCO AI Ethics Evaluator.

2.1 THE DCO PRINCIPLES FOR ETHICAL AI

The development of the DCO Principles for Ethical AI followed a rigorous process grounded in international best practices and benchmarks. The methodology applied aims to ensure the resulting framework is both globally aligned and locally relevant. The process began with comprehensive analysis of existing ethical AI frameworks from leading multilateral organizations and advanced AI nations, involving systematic review of principles adopted by organizations like the OECD, UNESCO, and the G20, as well as examining national frameworks established by DCO Member States and other countries with mature AI governance systems.

Through comparative analysis of global frameworks, the DCO identified recurring ethical foundations that transcend cultural and regional boundaries. These consistently emphasized principles include human-centricity and well-being, transparency and explainability, accountability and responsibility, fairness and non-discrimination, and respect for human rights and autonomy. While maintaining alignment with global standards, the process incorporated methodologies to accommodate regional variations and cultural contexts through consultations with diverse stakeholders across member states, analysis of regional priorities and values, and evaluation of different interpretations across cultural, religious, and historical contexts.

The DCO Principles for Ethical AI, which provide the Member States with shared foundations for responsible AI development, deployment, and governance, were unanimously adopted by the 16 DCO Member States in February 2025 during DCO's 4th General Assembly. The seven principles are described below.

1

Accountability

Accountability establishes clear responsibility for the development, deployment, and consequences of AI systems. It requires transparent answerability for performance, impacts, and potential risks to individuals and society. This includes ownership frameworks that assign responsibility at every stage of AI system development and mechanisms for tracking and addressing system performance and impacts.

2

Transparency and Explainability

Transparency refers to providing clear and comprehensive disclosure about AI system usage, including data processing, operational mechanisms, and intended purpose. Explainability complements transparency by focusing on communicating the reasoning behind AI-driven decisions in accessible and understandable terms, ensuring individuals can comprehend how and why specific outcomes are reached.

⁴ See DCO (2024) Rights by Design: Embedding Human Rights Principles in AI Systems

3

Fairness and Non-discrimination

Fairness refers to the equitable treatment of all individuals and groups in AI system outcomes, ensuring benefits, risks, and costs are justly distributed across society. Non-discrimination means AI systems must not create or contribute to unjust impacts based on protected characteristics, requiring proactive measures to prevent, identify, and mitigate both direct and indirect forms of discrimination.

4

Privacy

Privacy in AI systems encompasses the protection of individuals' physical, decisional, mental, and associational privacy, while addressing critical cybersecurity concerns that could compromise these protections. This holistic approach requires robust data protection frameworks and explicit consent mechanisms, coupled with strong cybersecurity measures to prevent unauthorized access, data breaches, and malicious exploitation of AI capabilities.

5

Sustainability and Environmental Impact

This principle demands a holistic approach balancing the environmental costs of AI technologies with their capacity to drive climate action and sustainable development. It requires implementing concrete strategies for energy efficiency and sustainable computing while leveraging AI's potential for environmental protection.

6

Human-centered Development and Social Benefit

This principle prioritizes human well-being and societal benefits by aligning AI innovations with human rights, ethical standards, and social values. This requires mechanisms to assess compliance with ethical guidelines, evaluate potential social impacts, and gather feedback from users.

7

Human Autonomy and Oversight

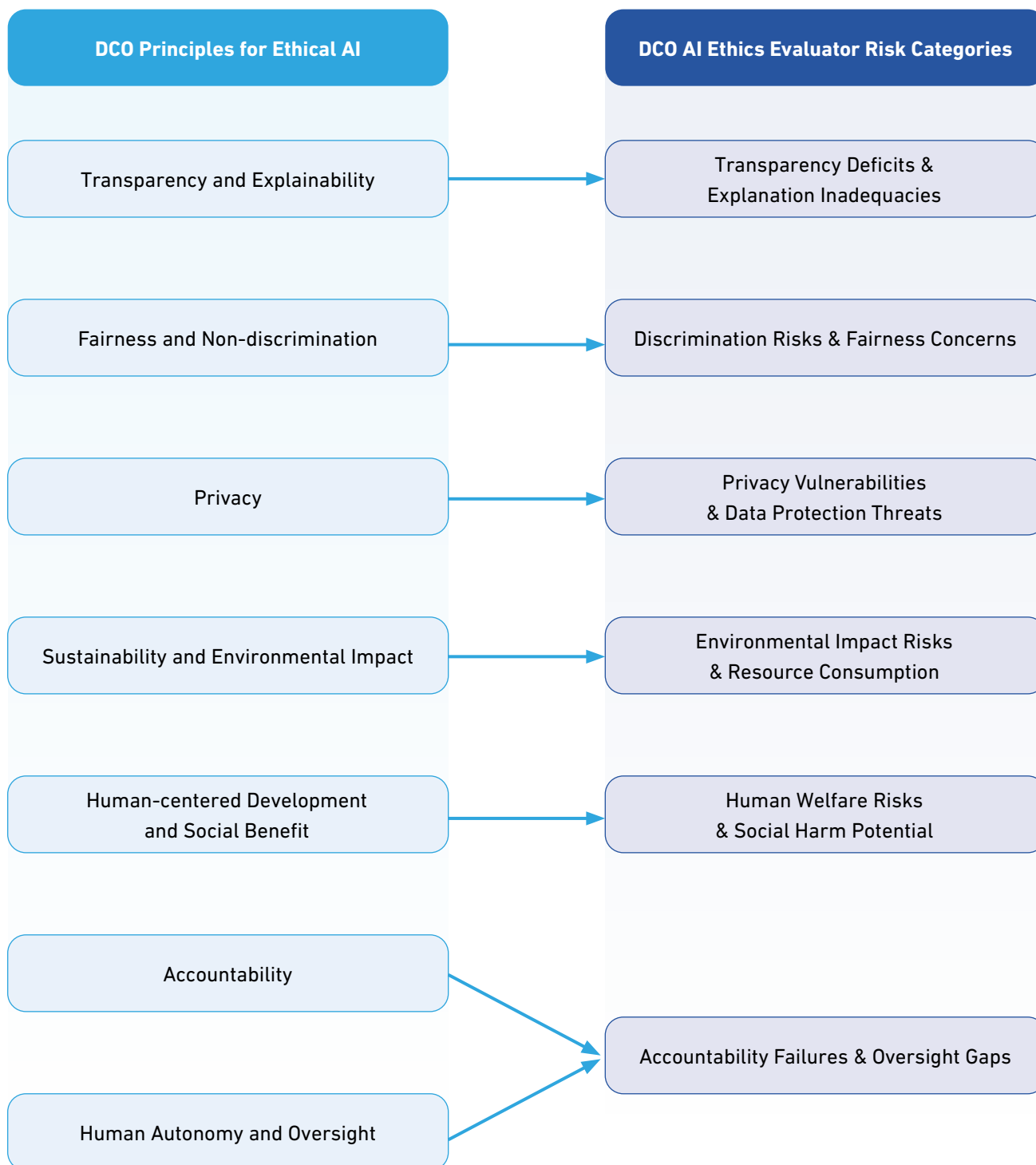
The principle of human autonomy in AI systems emphasizes maintaining human control and decision-making authority while leveraging AI to enhance rather than replace human capabilities. This concept is implemented through supervisory control and human-machine teaming, requiring comprehensive frameworks that enable transparent AI decision-making processes, meaningful human review, and clear intervention mechanisms.

Together, these principles support DCO Member States in fostering ethical AI governance that balances innovation with equity, accountability, and human dignity, establishing a global benchmark for responsible AI development and deployment.

2.2 PURPOSE OF THE TOOL

The seven DCO Principles for Ethical AI form the basis for six risk categories covered by the tool – see Figure 1 below. This mapping transforms abstract ethical principles into concrete, assessable risk areas that can be systematically evaluated through tailored questionnaires for both AI developers and deployers, creating a practical framework for identifying and mitigating human rights impacts throughout the AI lifecycle.

Figure 1: Mapping Principles to Risk Categories



To simplify the questionnaires that form the core element of the Evaluator tool, and make them more user-friendly, and because of similarities in their practical measurement, risks related to Accountability and Human Oversight are covered in one category Accountability Failures & Oversight Gaps.

Risk categories

1

Accountability Failures & Oversight Gaps

Risks arising from unclear responsibility frameworks, insufficient human verification of critical decisions, and inadequate mechanisms to respond to system failures or track decision pathways.

2

Transparency Deficits & Explanation Inadequacies

Risks stemming from insufficient clarity about how AI systems operate, make decisions, or process data, preventing users from understanding or challenging outcomes.

3

Discrimination Risk & Fairness Concerns

Risks of AI systems creating or amplifying biased outcomes across different demographic groups due to data representation issues, algorithmic design choices, or inadequate fairness testing.

4

Privacy Vulnerabilities & Data Protection Threats

Risks associated with improper handling of personal information, including unauthorized access, excessive data collection, insufficient protection measures, and potential data breaches.

5

Environmental Impact Risks & Resource Consumption

Risks related to AI systems' ecological footprint, including excessive energy usage, infrastructure capacity limitations, and unsustainable resource consumption patterns.

6

Human Welfare Risks & Social Harm Potential

Risks that AI systems may negatively impact human capabilities, autonomy, or social well-being by replacing rather than enhancing human functions or misaligning with genuine user needs.

Figure 2 -below- summarises the framework for the DCO AI Ethics Evaluator. As it is presented, each category addresses specific risks that could emerge during development and deployment of AI systems, connecting these risks to fundamental human rights that might be impacted. The framework enables organizations to identify, evaluate, and mitigate ethical concerns systematically before they cause harm, by applying the recommendations suggested by the Evaluator at the end of the process.

Figure 2: Framework for DCO AI Ethics Evaluator

Risks						
	Accountability Failures & Oversight Gaps	Transparency Deficits & Explanation Inadequacies	Discrimination Risk & Fairness Concerns	Privacy Vulnerabilities & Data Protection Threats	Environmental Impact Risks & Resource Consumption	Human Welfare Risks & Social Harm Protection
Specific Risks	<ul style="list-style-type: none"> Inadequate human verification of critical decisions Insufficient response to system failures Gaps in audit trails and responsibility tracking 	<ul style="list-style-type: none"> Inadequate human verification of critical decisions Insufficient response to system failures Gaps in audit trails and responsibility tracking 	<ul style="list-style-type: none"> Discriminatory impact on vulnerable groups Performance disparities across demographics Use of discriminatory proxy variables 	<ul style="list-style-type: none"> Insufficient data protection and security controls Unauthorized collection and processing of sensitive information 	<ul style="list-style-type: none"> Excessive energy consumption Infrastructure capacity limitations Unnecessary resource usage 	<ul style="list-style-type: none"> Degradation of human expertise/capabilities Misalignment with user needs and wellbeing
Human rights	<ul style="list-style-type: none"> Right to effective remedy Right to equality before the law 	<ul style="list-style-type: none"> Right to information Right to participate in public affairs 	<ul style="list-style-type: none"> Right to non-discrimination Right to equality before the law Right to freedom of opinion and expression 	<ul style="list-style-type: none"> Right to privacy 	<ul style="list-style-type: none"> Right to health Right to adequate standard of living Right to a healthy environment 	<ul style="list-style-type: none"> Right to work Right to participate in cultural life Right to liberty Right to education Right to self-determination
Recommendations						
	1. System Architecture Development:	2. Validation & Testing:	3. Data Management:	4. Operational Controls:	5. Documentation & Reporting:	6. Continuous Evolution:
	<ul style="list-style-type: none"> Building/ implementing robust safety mechanisms Ensuring appropriate human oversight capabilities 	<ul style="list-style-type: none"> Comprehensive testing of system behaviour Monitoring performance across different groups 	<ul style="list-style-type: none"> Implementing data minimization practices Maintaining proper data governance controls 	<ul style="list-style-type: none"> Creating effective incident response procedures Implementing monitoring and control processes 	<ul style="list-style-type: none"> Tracking decisions and their impacts Creating clear audit trails and reports 	<ul style="list-style-type: none"> Updating based on performance monitoring Improving safety and effectiveness measures

For example, under the Discrimination Risk & Fairness Concerns category, a specific risk might be “discriminatory impact on vulnerable groups” which could directly impact the human right to non-discrimination. Similarly, in the Privacy Vulnerabilities & Data Protection Threats category, “insufficient data protection and security controls” could violate an individual’s right to privacy.

With these risk categories defined, the tool development process focused on creating structured self-assessment questionnaires:

- This included developing questions to evaluate both severity and likelihood of risks, with distinct versions for developers and deployers reflecting their different roles and responsibilities.
- For each risk area, based on how the user answers questions, average scores are calculated for both potential severity of impact (on a scale of 1-5) and the likelihood of occurrence (also on a scale of 1-5).
- These two dimensions are then multiplied to yield an overall risk score ranging from 1-25, which is categorized into three risk levels across the 6 risk categories:
 - a. low (1 or 2)
 - b. medium (greater than 2 but less than or equal to 9)
 - c. high (greater than 9 but less than or equal to 25)

The tool employs this nuanced assessment approach for each of the six risk-categories independently, recognizing that AI systems may demonstrate varying levels of risk across different dimensions. Under this framework, an organization might, for instance, have a low-risk rating when it comes to Accountability Failures & Oversight Gaps while simultaneously struggling with Privacy Vulnerabilities & Data Protection threats, resulting in a high-risk rating.

This granular assessment methodology enables targeted improvement strategies. The final assessment presents a detailed risk profile visualization that maps performance across all six risk categories individually.

This approach acknowledges that excellence in one area cannot compensate for deficiencies in another, as each category addresses distinct and crucial aspects of ethical AI implementation. While an overall risk assessment is provided for general guidance, the risk category-specific ratings and corresponding recommendations serve as the primary drivers for targeted improvement actions. These categories maintain clear distinctions between different types of mitigation measures while ensuring comprehensive coverage of necessary controls.

The tool’s value lies in this ability to highlight specific areas requiring attention while acknowledging existing strengths, allowing users to prioritize their improvement efforts effectively across all categories independently.

2.3 RECOMMENDATION CATEGORIES IN THE TOOL

The DCO AI Ethics Evaluator not only identifies and assesses risks but also provides targeted recommendations structured around six key operational areas namely i. System Architecture & Development, ii. Validation & Testing, iii. Data Management, iv. Operational Controls, v. Documentation & Reporting and, vi. Continuous Evolution. These recommendations are calibrated to address the specific risk levels identified during assessment, offering more intensive interventions for high-risk areas while providing lighter guidance for lower-risk scenarios.

Figure 3. Recommendations

1	System Architecture & Development	<ul style="list-style-type: none">• Building/implementing robust safety mechanisms• Ensuring appropriate human oversight capabilities• Establishing clear system boundaries and fail-safes
2	Validation & Testing	<ul style="list-style-type: none">• Comprehensive testing of system behaviour• Monitoring performance across different groups• Detecting and addressing potential issues
3	Data Management	<ul style="list-style-type: none">• Securing sensitive data processing/storage• Implementing data minimization practices• Maintaining proper data governance controls
4	Operational Controls	<ul style="list-style-type: none">• Establishing clear oversight mechanisms• Creating effective incident response procedures• Implementing monitoring and control processes
5	Documentation & Reporting	<ul style="list-style-type: none">• Maintaining comprehensive system documentation• Tracking decisions and their impacts• Creating clear audit trails and reports
6	Continuous Evolution	<ul style="list-style-type: none">• Integrating stakeholder feedback• Updating based on performance monitoring• Improving safety and effectiveness measures



The recommendation framework follows the same operational structure as the risk assessment, creating a coherent path from risk identification to practical mitigation. Each category focuses on a distinct aspect of AI system development and deployment, covering the full lifecycle from initial architecture through continuous improvement. Importantly, there is no strict one-to-one mapping between risk categories and recommendation categories, as a single risk often requires mitigations across multiple operational areas.

Recommendation categories

1

System Architecture & Development

Recommendations in this category focus on building ethical considerations directly into AI system design. They cover implementing robust safety mechanisms, establishing clear boundaries, and creating appropriate fail-safes to prevent harm. For developers, this might include guidance on designing system features with built-in protections, while deployers receive recommendations on proper system configuration and customization.

2

Validation & Testing

This category provides recommendations for comprehensive testing regimes that evaluate system performance across different scenarios and user groups. Recommendations address how to build effective testing frameworks, create appropriate validation tools, and implement monitoring capabilities that can identify potential issues before they impact users.

3

Data Management

Recommendations here address the proper collection, processing, storage, and protection of data throughout the system lifecycle. Guidance covers data structure design, processing system implementation, protection mechanisms, and appropriate policies for data retention and access control.

4

Operational Controls

This category focuses on maintaining effective human oversight and control throughout system operation. Recommendations address building control mechanisms, creating monitoring tools, and implementing procedures that allow humans to effectively monitor, manage, and intervene in system operations when necessary.

5

Documentation & Reporting

Recommendations in this area ensure comprehensive tracking of system behaviour, decisions, and incidents to maintain transparency and accountability. Guidance covers creating documentation systems, building reporting tools, and maintaining comprehensive audit capabilities.

6

Continuous Evolution

This final category addresses how systems should evolve over time, providing recommendations on monitoring performance, gathering feedback, and adapting practices based on stakeholder input and emerging risks. Recommendations cover building update mechanisms, creating feedback systems, and implementing improvement processes.

Figure 4: Tailored recommendations structured around 6 core recommendation categories

			Examples of recommendations for developers	Examples of recommendations for deployers
1	System Architecture & Development	Ensuring the AI system is built and implemented with robust safety mechanisms, clear boundaries, and appropriate fail-safes to prevent harm	<ul style="list-style-type: none"> • Design and build system features • Implement technical controls • Create architectural safeguards 	<ul style="list-style-type: none"> • Configure system settings • Enable/disable features • Customize implementation
2	Validation & Testing	Rigorous testing and validation of system performance, bias, and reliability across different scenarios and user groups	<ul style="list-style-type: none"> • Build testing frameworks • Create validation tools • Design monitoring capabilities 	<ul style="list-style-type: none"> • Run tests • Conduct validation • Monitor performance
3	Data Management	Ensuring proper collection, processing, storage, and protection of data throughout the system lifecycle to maintain privacy and security	<ul style="list-style-type: none"> • Design data structures • Build data processing systems • Create data protection mechanisms 	<ul style="list-style-type: none"> • Manage data flows • Implement retention policies • Control data access
4	Operational Controls	Maintaining effective human oversight and control mechanisms to monitor, manage, and intervene in system operations	<ul style="list-style-type: none"> • Build control mechanisms • Create monitoring tools • Design override capabilities 	<ul style="list-style-type: none"> • Monitor system rohacodeure • Monitor system behavior • Manage oversight processes
5	Documentation & Reporting	Ensuring comprehensive documentation and tracking of system behaviour, decisions, and incidents to maintain transparency and accountability	<ul style="list-style-type: none"> • Create documentation systems • Build reporting tools • Design audit capabilities 	<ul style="list-style-type: none"> • Maintain documentation • Generate reports • Track system behavior
6	Continuous Evolution	Actively monitoring system performance, gathering feedback, and evolving practices based on stakeholders imput and emerging risks	<ul style="list-style-type: none"> • Build update mechanisms • Create feedback systems • Design improvement capabilities 	<ul style="list-style-type: none"> • Implement update • Collect and act on feedback • Manage system improvements

A woman with dark hair and glasses is looking at a computer screen. She has her hand resting on her chin, appearing thoughtful. The background is dark, and the lighting is focused on her face and the screen. A semi-transparent, dark, triangular shape is overlaid on the left side of the image, containing the text '03' and 'USING THE TOOL'.

03

USING
THE TOOL

The tool operates through a structured process, as laid out below:

Figure 4: Workflow



3.1 REGISTRATION AND CHARACTERIZATION

Users begin by registering and identifying their role as either AI developers or deployers, determining their specific assessment pathway. At this stage, country and system characterization (e.g. the type of technology, data processed, application domain, potential impact scale) information is collected for statistical purposes, but does not affect the questionnaires that the user will receive or the subsequent analysis.

Below are options for the drop-down menus:

Type of Technology (Primary):

- Machine Learning (Traditional)
- Deep Learning
- Natural Language Processing
- Computer Vision
- Expert/Rule-Based Systems
- Reinforcement Learning
- Generative AI
- Hybrid/Multi-modal AI
- Other AI Technology

Data Processed (Primary Type)

- Personal Data - Sensitive (health, biometric, financial, etc.)
- Personal Data - Non-sensitive (basic identifiers, preferences)
- Business/Commercial Data (non-personal)
- Environmental/Sensor Data (non-personal)
- Public Domain Data
- Synthetic/Generated Data (no real-world data)
- No Data Processing
- Other Data Type

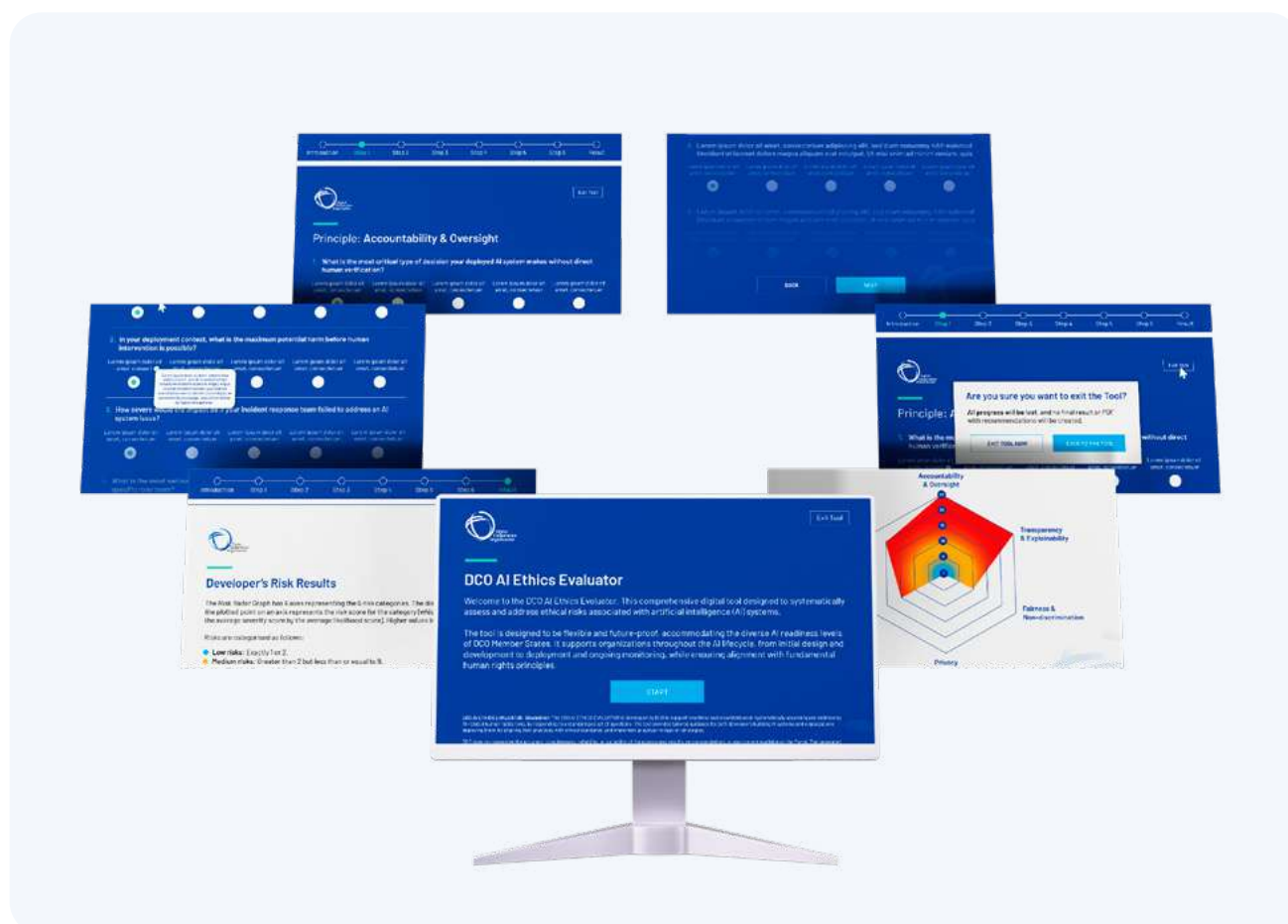
Application Domain (Primary)

- Internal Business Operations
- Consumer/Client Services
- Government/Public Services
- Healthcare Applications
- Financial Services
- Human Resources Management
- Education/Training
- Safety & Security
- Media & Content
- Research & Development
- Infrastructure & Logistics
- Other Application Domain

Potential Impact Scale

- System-specific (affects only direct users of the system)
- Organization-wide (affects operations across the organization)
- Market-facing (affects customers/clients outside the organization)
- Industry-level (affects practices across an industry or sector)
- Society-level (affects social or public infrastructure)

Instructions for users: Please select the furthest-reaching category that applies to your AI system. For example, if your system affects both direct users and customers, select "Market-facing." If it affects customers across an entire industry, select "Industry-level."



3.2 RISK ASSESSMENT PROCESS

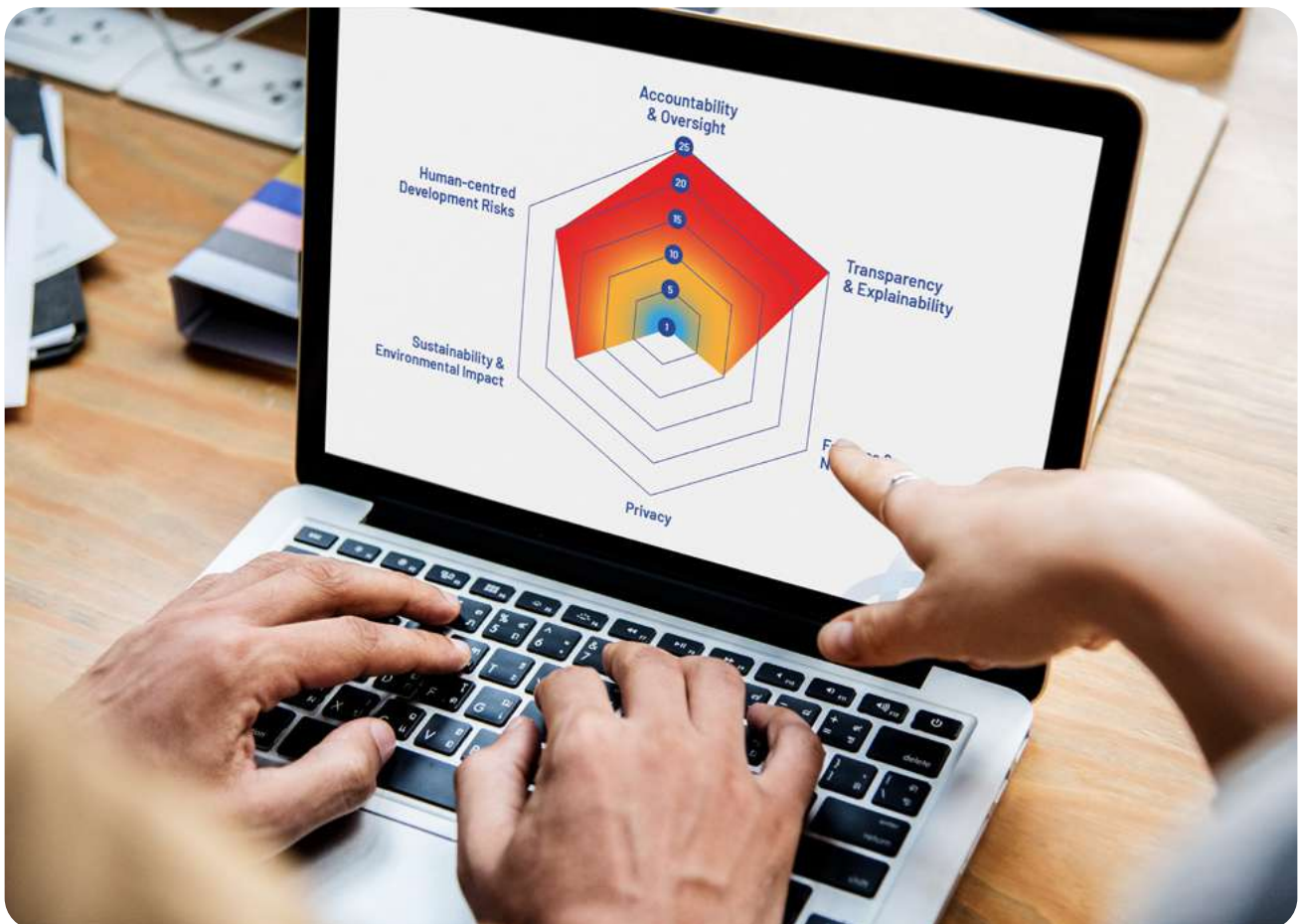
The tool's core function is its risk assessment mechanism. Users complete questions evaluating both severity and likelihood of potential risks across six categories. The questions are tailored to the user's role as developer or deployer, with options to indicate where risks may not be applicable to their context.

Framework Structure

The DCO AI Ethics Evaluator focuses on six primary risk categories that comprehensively cover all relevant risks for developers and deployers:

- Transparency Deficits & Explanation Inadequacies
- Discrimination Risks & Fairness Concerns
- Privacy Vulnerabilities & Data Protection Threats
- Environmental Impact Risks & Resource Consumption
- Human Welfare Risks & Social Harm Potential
- Accountability Failures & Oversight Gaps

Each category examines specific ethical and human rights dimensions of AI systems, ensuring comprehensive evaluation throughout the system lifecycle.



Risk Scoring Methodology

For each risk category, users evaluate both severity and likelihood using a 1–5 scale:

- Severity ratings range from minimal impact (1) to critical impact (5)
- Likelihood ratings range from rare occurrence (1) to almost certain (5)

If a question does not apply to the AI system being assessed, it can be marked as “Not Applicable.”

To determine the overall impact score, the severity and likelihood scores are directly multiplied. This results in a final score ranging from 1 to 25, categorizing risks as:

- **Low risks:** Exactly 1 or 2.
- **Medium risks:** Greater than 2 but less than or equal to 9.
- **High risks:** Greater than 9 but less than or equal to 25.

Example 1: Privacy

Factor	Rating	Explanation
Severity	4	High impact
Likelihood	3	Possible some security measures in place but vulnerabilities exist
Risk Score	12	$4 \times 3 = 12$
Risk Category	High Risk	Score > 9, requires immediate action

Example 2: Human Autonomy and Oversight

Factor	Rating	Explanation
Severity	2	Minor impact
Likelihood	3	Possible gaps may occur occasionally
Risk Score	6	$2 \times 3 = 6$
Risk Category	Medium Risk	Score > 2 but ≤ 9 , requires monitoring and certain actions

Differentiated Assessment

The tool provides distinct assessments for developers and deployers, reflecting their different responsibilities and challenges. Developer assessments focus on design decisions and technical implementation, while deployer assessments address operational risks and actual impacts on users and stakeholders.

3.3 VISUALIZATION AND ANALYSIS

Responses are processed through a scoring system combining severity and likelihood measurements to generate a risk profile. The analysis appears in visual format, including radar charts that highlight priority areas. These visualizations enable users to quickly identify their most significant risks.

Results appear through risk matrix heat maps and radar graphs, providing clear visualization of risk profiles. These tools help users identify priority areas requiring immediate attention and track changes over time.

This structured approach enables users to:

- Systematically evaluate AI risks across different dimensions
- Prioritize areas requiring immediate intervention
- Track changes in risk profiles over time

3.4 RECOMMENDATIONS

Based on the risk analysis, the tool provides targeted recommendations for implementing appropriate safeguards and controls. These align with international human rights standards and scale according to risk level. Users receive a downloadable report facilitating integration of these recommendations into their existing processes.

Structure

The DCO AI Ethics Evaluator generates recommendations through an integrated framework that maps identified risks to practical controls. Recommendations are organized into distinct operational categories:

- System Architecture & Development
- Validation & Testing
- Data Management
- Operational Controls
- Documentation & Reporting
- Continuous Evolution

For example, managing privacy risks requires coordinated implementation across recommendation categories:

- System Architecture: Implementing privacy-preserving features
- Validation & Testing: Conducting privacy impact assessments
- Data Management: Establishing data handling protocols
- Operational Controls: Maintaining access controls
- Documentation & Reporting: Creating audit trails
- Continuous Evolution: Updating privacy measures

Risk Response and Scaling

Recommendations scale in three levels based on assessed risk:

- **Low Risk:** Focus on verifying and maintaining existing controls
- **Medium Risk:** Enhanced controls and additional monitoring mechanisms
- **High Risk:** Comprehensive controls with rigorous oversight requirements

For example, addressing privacy risks might scale from basic data handling procedures at low risk, to enhanced protection mechanisms at medium risk, to comprehensive privacy-by-design implementation at high risk.

Role-Based Implementation

The tool provides different guidance for developers and deployers while maintaining consistency in control objectives. For example, when addressing fairness risks:

Developers receive recommendations for:

- Building bias detection into system architecture
- Creating testing frameworks for demographic performance
- Implementing data quality controls

Deployers receive guidance on:

- Configuring and using bias detection features
- Running regular fairness assessments
- Monitoring demographic impacts

This structured approach ensures that developers and deployers receive practical guidance matched to their role, risk level, and capabilities, while maintaining comprehensive risk coverage through coordinated controls across operational areas.

The questionnaire and recommendations for the tool were inspired by a range of credible sources that represent the cutting edge of AI ethics frameworks globally. These include:

- A range of leading international organizations have developed comprehensive frameworks and guidelines to promote ethical AI development and deployment. **UNESCO's "Recommendation on the Ethics of Artificial Intelligence" (2021)** provides a global standard on AI ethics, emphasizing human rights, fairness, transparency, and sustainability. The **European Union's "Ethics Guidelines for Trustworthy AI" (2019)** and the **EU AI Act** set out requirements for lawful, ethical, and robust AI, including provisions on human oversight and accountability. The **OECD Principles on Artificial Intelligence (2019 and 2024)** provide globally recognized recommendations for responsible AI, focusing on inclusive growth, human-centered values, and transparency.
- Industry and technical standards have also been foundational to our approach. **IEEE's "Ethically Aligned Design"** offers detailed guidance for embedding ethical considerations in AI and autonomous systems, covering topics like privacy, bias, and human well-being. The **U.S. NIST AI Risk Management Framework (2024)** provides practical guidelines for identifying and mitigating risks associated with AI systems, including fairness, transparency, and accountability.
- Additional resources that informed our recommendations include research from academic institutions such as the **Stanford Institute for Human-Centered AI** and the **Oxford Internet Institute**.

Together, these diverse sources form the foundation of our tool, ensuring it reflects both established standards and emerging best practices in responsible AI development and deployment

The background is a dark blue field filled with out-of-focus bokeh lights in shades of orange, yellow, and light blue. Overlaid on this are faint, glowing binary digits (0s and 1s) in a light blue color, some of which are sharp while others are blurred, creating a sense of depth and digital movement.

04

FREQUENCY OF RISK EVALUATIONS

The risks analysed in this tool are not static but, on the contrary, are continuously evolving due to multiple factors. Therefore, the recommended assessment frequency may vary for different users and cases, as suggested below.

4.1 BASELINE ASSESSMENT SCHEDULE

For **High-risk AI systems**, users should conduct comprehensive risk assessments every 6 months, with rapid review checkpoints every 2 months. Any significant system changes should trigger an immediate assessment.

Medium-risk AI systems require full assessments every 12 months, with rapid review checkpoints every 4 months. Following significant system changes, an assessment should be conducted within 1 month.

Low-risk AI systems should undergo full assessments every 18 months, with rapid review checkpoints every 6 months. After significant system changes, an assessment should be completed within 2 months.





Beyond the baseline schedule, several circumstances necessitate additional assessments. External factors may also necessitate reassessment, such as new regulatory requirements or legal frameworks, emerging societal concerns or public discourse, identified incidents or near-misses, and stakeholder feedback indicating potential issues.

Implementation progress also affects assessment frequency as more frequent assessments are typically needed during early deployment phases, with gradual adjustment as mitigation measures prove effective. Users should regularly review whether the assessment frequency remains appropriate.

Organizational context plays a crucial role in determining appropriate assessment frequency. This includes consideration of resource availability and technical capacity, complexity of deployment environment, stakeholder requirements and expectations, and industry-specific risk factors.



Follow us on

   @dcorg |  www.dco.org

© 2025, The Digital Cooperation Organization, all rights reserved.