frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 03 December 2025
DOI 10.3389/fcomp.2025.1694979

OPEN ACCESS

EDITED BY
Jacopo Soldani,
University of Pisa, Italy

REVIEWED BY
Stavroula Tsinorema,
University of Crete, Greece
Vasileios Alevizos,
Karolinska Institutet, Sweden

*CORRESPONDENCE
José Santisteban
✉ PCSILSAN@upc.edu.pe

RECEIVED 29 August 2025
REVISED 11 November 2025
ACCEPTED 20 November 2025
PUBLISHED 03 December 2025

CITATION
Romani G, Avendaño C and Santisteban J (2025) MESIAS: a web-based platform rooted in ethical principles for evaluating trustworthiness in AI projects. *Front. Comput. Sci.* 7:1694979. doi: 10.3389/fcomp.2025.1694979

MESIAS: a web-based platform rooted in ethical principles for evaluating trustworthiness in AI projects

Georgina Romani, Cesar Avendaño and José Santisteban*

Information Systems Engineering, Universidad Peruana de Ciencias Aplicadas (UPC), Lima, Peru

The accelerated growth of artificial intelligence (AI)-based projects has intensified the need for tools to assess their reliability, safety, and ethical alignment. In response to this challenge, the MESIAS initiative was developed. MESIAS is a web-based platform that provides a framework for evaluating AI systems through the lenses of ethical principles and international governance frameworks. The tool features a virtual assistant, adaptive forms, and a monitoring dashboard. The validation process comprised three steps: a preliminary investigation into operational efficiency, expert judgment validation with technological leaders, and a user satisfaction validation with 52 technology professionals. The operational assessment revealed a substantial 41.8% reduction in total assessment time and a 40% reduction in human resources required. Expert validation reflected a general acceptance of 85%. User validation revealed elevated satisfaction levels: 92% for usability, 94% for content, 91% for follow-up, and 95% for overall satisfaction. The study results indicate that the MESIAS strategy is a practical and effective approach to enhancing ethical governance in AI, particularly in public settings, fostering more responsible and informed decision-making processes.

KEYWORDS

artificial intelligence, ethical guidelines, evaluation tools, governance, risk assessment, security, usability, web platform

1 Introduction

For In recent decades, the field of AI has led to novel opportunities for value creation in various sectors, including business, industry, community, and society at large. As organizations have learned to harness the massive flow of information, multiple opportunities have arisen to apply AI solutions in sectors such as healthcare, education, finance, and public administration. The advent of this technology has precipitated a transformative paradigm within social, economic, and political systems on a global scale (Ansari et al., 2022; Tripathi and Rosak-Szyrocka, 2024). However, the rapid development of this technology has also given rise to a range of ethical, legal, and security challenges. AI systems, engineered to learn from vast quantities of data, may encounter scenarios that are outside their prior experience and exhibit behaviors that are not predicted, which could pose risks to users (Taeihagh, 2021; Mennella et al., 2024).

According to Vela (2022), institutions such as the Organization for Economic Cooperation and Development (OECD) and the United Nations Educational, Scientific, and Cultural Organization (UNESCO) emphasized the importance of establishing strong governance and regulatory frameworks of AI to ensure the moral and secure advancement. AI is currently employed in both the public and private sectors throughout Latin America, though concerns related to technical complexity and societal implications remain largely unaddressed (Español and Sylvan, 2023). In this context, governments and enterprises must respond proportionally to real challenges, avoiding severe restrictions that could stifle innovation or passivity that could damage technological superiority (Oladele et al., 2024).

While the advancement of AI has yielded numerous advantages, it has concurrently presented businesses with significant challenges that necessitate resolution (Daly et al., 2019). According to Marr (2023), the primary hazards associated with AI include algorithmic bias and discrimination, a paucity of transparency, ethical dilemmas surrounding automated decision-making, and cybersecurity threats. The absence of adequate procedures to anticipate and appropriately manage these issues from the onset of AI project development exacerbates them. Consequently, significant gaps persist in the realm of AI governance, impeding enterprises' capacity to discern and oversee potential hazards before their materialization (Prajapati, 2025; Boppiniti, 2022). The first underlying problem, therefore, lies in the limited adherence of Latin American organizations to ethical, regulatory, and social standards in AI system development. This phenomenon is primarily attributable to the absence of a robust culture of technological governance within many institutions in the region.

Globally, the rapid proliferation of AI underscores the urgent need for the establishment of ethical guidelines and principles to govern its use. Singla et al. (2024) assert that 72% of businesses and organizations worldwide currently utilize some form of AI, representing an increase from 50% the previous year. A significant proportion of these businesses now allocate more than 5% of their digital budgets to these technologies. Despite the advancements in technology, there is a paucity of moral considerations or effective methods to ensure that the development of AI is aligned with human and social values. A significant proportion of companies, specifically 92%, have expressed their intention to allocate increased financial resources to the field of AI within the subsequent three-year period. However, a mere 1% of leaders within these companies perceive their organizations to be "mature" in their implementation of AI, signifying the full integration of AI into their operational processes, to achieve substantial outcomes (Mayer et al., 2025). This absence of maturity underscores the urgent necessity to fortify the ethical management of AI. Furthermore, Benchaita (2024) notes that, according to the IBM Global AI Adoption Index 2023, 90% of IT professionals concur that consumers prefer services from companies that offer transparency and an ethical framework in the design and management of their AI systems, while 92% consider the ability to govern AI data and models throughout their lifecycle to be crucial. These figures demonstrate that while there is strong interest in adopting AI, a significant gap remains in ethical preparedness and risk management capacity—factors that may compromise public trust and the long-term sustainability of AI development.

Beyond regional and organizational shortcomings, a broader theoretical challenge persists. The central issue in the ethics of AI today lies not in what ethical principles to adopt, but in how to operationalize them—bridging the gap between abstract principles and concrete engineering practices (Morley et al., 2019). This theory–practice gap has become a defining concern for the field. Efforts to close this gap often fall into two traps. First, checklist-based approaches, while practical, risk oversimplifying complex moral concepts and fostering ethics-washing—the symbolic use of ethical language for reputational purposes without substantive accountability (Mittelstadt, 2019; Bietti, 2020). Second, comprehensive methodologies such as Value-Sensitive Design (VSD), while conceptually robust, are difficult to scale within the fast-paced development cycles typical of AI systems (Friedman et al., 2013).

The establishment of ethical auditing frameworks, such as the one proposed by Kiran et al. (2023), has enabled open-source tools to identify biases in classification models, particularly in the healthcare sector. This development has rendered the incorporation of transparency and algorithmic fairness methods a viable prospect. Furthermore, a human-centered artificial intelligence (HCAI) paradigm that prioritizes explainability, personalization, and human-AI collaboration was proposed (Usmani et al., 2023). Furthermore, they proffered moral standards to ensure trust and well-being. In a similar vein, Fedele et al. (2024) identified ethical and legal problems when creating the AI4SPP system, which predicts student achievement, using the Assessment List for Trustworthy Artificial Intelligence (ALTAI). Additionally, Buhmann and Fieseler (2021) discussed the significance of open, responsible, and multi-group AI governance. In a similar Buchholz et al. (2022) underscore the significance of integrating ethical and design principles into the development of AI. They propose an approach that ensures accountability is considered throughout all stages of the development process, emphasizing the necessity of addressing ethical issues from the earliest phases.

Despite the development of numerous conceptual frameworks, ethical principles, and governance guidelines for AI projects, the majority of these contributions have remained at a theoretical or general level, as indicated by the reviewed studies. However, there is a paucity of practical tools that enable a direct and quantitative assessment of an AI project's ethical status (Ortega-Bolaños et al., 2024). The landscape of available information is characterized by fragmentation and complexity, which poses a significant challenge for professionals in the field (De Borba et al., 2024). Thus, the field urgently requires hybrid models that are both conceptually sound and operationally scalable, ensuring ethical compliance without compromising innovation.

To address this persistent challenge—the gap between the articulation of ethical principles and their concrete implementation in AI projects—this study introduces the Model for Evaluating Security and Ethical Issues in Artificial Intelligence Systems (MESIAS). MESIAS is conceived as a practical and adaptive framework that operationalizes international principles of trustworthy AI through an automated and participatory web-based platform. Unlike checklist-oriented tools that risk promoting ethics-washing through superficial compliance, MESIAS integrates both top-down ethical frameworks—such as the European Commission's Trustworthy AI Assessment List—and bottom-up deliberative mechanisms that foster reflection, accountability, and continuous improvement within organizations. This dual approach seeks to transform ethics from a procedural obligation into an active dimension of governance and decision-making throughout the AI lifecycle.

In doing so, MESIAS acts as a deliberative and traceable scaffold, enabling developers, auditors, and decision-makers to evaluate ethical and security aspects in a transparent, dynamic, and context-sensitive manner. Beyond assessment, the platform promotes learning and dialogue, helping institutions internalize ethical awareness as part of their technological culture. While the challenge of trustworthy AI is global, it carries distinct implications in Latin America, where institutional capacities for ethical oversight and AI governance remain limited. Therefore, MESIAS not only adapts established frameworks to regional realities but also contributes to reducing the implementation gap between global standards and local practices.

The primary contributions of this study are as follows:

- The design and implementation of MESIAS, an interactive web platform that operationalizes abstract ethical AI principles into a concrete, automated assessment process tailored for the Latin American context.
- The development of a scalable, hybrid architecture that integrates a logic-based quantitative risk model with a virtual assistant, demonstrating a practical method for bridging the theory-practice gap in AI ethics.
- A multi-faceted validation of the platform (assessing operational efficiency, expert judgment, and user satisfaction), providing preliminary evidence of its value in significantly reducing evaluation time and strengthening ethical governance.

2 Related works

This section focuses on the practical ecosystem of tools, platforms, and methodologies developed to operationalize AI ethics, auditing, and trustworthy AI requirements. Recent research reveals that existing efforts can be grouped into three broad domains: ethics-focused procedural models, auditing-oriented governance frameworks, and multidimensional evaluation tools. Together, these illustrate the progress and limitations of current operationalization strategies.

2.1 Studies focused on ethical considerations

The literature on AI ethics consistently identifies a core challenge: the persistent gap between high-level ethical principles and their effective operationalization in real-world AI systems. This gap has driven a shift toward frameworks that provide procedural clarity and enable practical governance across the AI lifecycle.

One influential proposal is the Ethics as a Service (EaaS) paradigm, which argues that principle-based approaches lack the procedural detail required for consistent implementation (Morley et al., 2021). EaaS reframes ethics as an iterative organizational function—embedded in workflows, decisions, and oversight mechanisms—rather than a static set of rules. This perspective is directly relevant to our study because it highlights the demand for tools that can translate normative frameworks into repeatable, auditable, and context-responsive processes.

Empirical studies reinforce this need for contextualization. Kieslich et al. (2022), through survey-based analysis of tax fraud detection systems, show that ethical expectations vary significantly across user groups. Their results challenge the assumption of homogeneous societal values and underscore that ethical governance must accommodate diverse perceptions of fairness, accountability, and societal risk.

At the system governance level, recurring ethical issues—such as discrimination, opacity, and ambiguous responsibility—have motivated integrated lifecycle approaches to risk management (Guan et al., 2022). These works emphasize that ethical oversight cannot rely solely on technical fixes: it requires organizational traceability, documentation structures, and clear institutional roles. This aligns with the growing recognition that auditability is a foundational component of trustworthy AI.

Complementarily, Ashok et al. (2021) introduce an ontological framework that maps ethical concerns across different technological domains. By distinguishing issues related to intelligibility, autonomy, equity, and governance, their framework illustrates why broad ethical principles remain insufficient for domain-specific evaluation. Their work demonstrates the value of structured tools that convert abstract principles into domain-relevant assessment criteria.

Finally, transdisciplinary approaches highlight the importance of anticipating long-term consequences and extending ethical evaluation beyond system design (Trinkley et al., 2024). These studies argue that ethical governance must span deployment, monitoring, and post-implementation review—requiring collaboration across technical, social, and institutional expertise.

Collectively, these works converge on a central insight: the field lacks practical, iterative, and context-sensitive mechanisms to operationalize ethics in ways that support traceability, accountability, and real-world governance. This gap sets the foundation for the need addressed in our research.

2.2 Studies focused on AI auditing considerations

The auditing literature converges on the need to move from principle-based ethics toward systematic, evidence-driven, and institutionally anchored evaluation mechanisms capable of governing AI deployments at scale. Several studies highlight that ethics cannot be operationalized without structured forms of assessment and verification, and that existing governance practices lack the procedural depth required for effective oversight.

A foundational contribution in this space is the Ethics-Based Auditing (EBA) framework, which positions ethical auditing as a core instrument for translating normative principles into enforceable governance practices (Mökander and Floridi, 2021). EBA argues that ethical oversight must be continuous, embedded throughout the development lifecycle, and supported by independent entities empowered by public regulatory bodies. The authors identify sixteen structural limitations—ranging from unclear audit scopes to insufficient methodological guidance—that constrain current audit practices. Importantly, they caution that auditing cannot replace individual ethical deliberation, underscoring the need for hybrid models that integrate procedural audits with value-oriented reflection.

Building on the limitations of both purely technical and purely conceptual approaches, Brown et al. (2021) introduce a socio-technical auditing framework that integrates stakeholder analysis, metric-based assessment, and a relevance matrix. Their work emphasizes that algorithmic audits must incorporate societal implications—especially the effects on vulnerable populations—rather than focusing solely on performance or fairness metrics.

In parallel, Koshiyama et al. (2024) advocate for the establishment of a formal algorithm auditing industry, analogous to financial auditing, to manage legal, ethical, and operational risk. Their governance framework stresses the need for both internal and external audit functions and highlights the absence of standardized methods as a major barrier to organizational safety. By proposing risk-based classification schemes and sector-specific audit requirements, the authors underscore the urgency of infrastructure that enables reproducibility and traceability—capabilities directly.

Laine et al. (2024) contribute a comprehensive meta-analysis of ethical algorithm auditing, addressing the theoretical fragmentation that has slowed the field's maturation. Through the analysis of 93 academic works, they develop a consolidated typology of ethical principles and map the epistemic value produced by audits across stakeholder groups, including developers, regulators, auditors, and the public. Their classification into technical vs. social and process vs. outcome perspectives clarifies the multidimensional nature of auditing and reinforces the notion that practical auditing requires methodological unification.

Finally, the work of Lacmanovic and Skare (2025) illustrates the increasing regulatory formalization of auditing practices. By comparing the EU AI Act's conformity assessments with New York City's Local Law 144 bias audit mandates, the authors identify emerging best practices and persistent methodological challenges, including inconsistent audit procedures and the absence of interoperable standards. Their analysis highlights that regulatory efficacy depends on audits that not only ensure compliance but promote equity, transparency, and social justice—particularly in sectors such as healthcare and employment where algorithmic harms can have disproportionate societal impacts.

Together, these studies indicate a broad consensus: effective AI governance requires auditing infrastructures that are procedural, context-aware, institutionally standardized, and capable of integrating both quantitative metrics and qualitative reasoning. This provides strong conceptual grounding for the design choices of MESIAS, which operationalizes these principles through a hybrid evaluation logic, risk-based scoring, and semantically informed validation mechanisms.

2.3 Studies integrating multiple dimensions (ethical, security, legal, etc.)

Research that integrates ethical, technical, legal, and sociotechnical dimensions highlights the increasing need for governance mechanisms capable of evaluating AI systems holistically. This body of work underscores that isolated approaches—whether ethical checklists, technical robustness metrics, or regulatory compliance reviews—are insufficient for managing the complexity of contemporary AI deployments. Instead, multi-dimensional assessment tools are required to address transversal risks such as discrimination, opacity, security vulnerabilities, and misalignment with fundamental rights.

Radclyffe et al. (2023) offer one of the most comprehensive critiques of ALTAI, the European Commission's self-assessment tool for trustworthy AI. While ALTAI provides a structured consolidation of ethical requirements, the authors identify crucial limitations that restrict its operational value, including its weak connection to fundamental rights assessments, limited cross-organizational comparability, and the absence of risk-based scoring mechanisms. Their analysis emphasizes the importance of quantitative metrics, rating systems, and alignment with international regulatory frameworks. This reinforces a central theme in the literature: tools for AI governance require methodological precision and evaluative depth if they are to support real-world accountability.

Similarly, Díaz-Rodríguez et al. (2023) present a multidimensional framework for responsible and reliable AI that integrates technical specifications, ethical principles, and regulatory strategies. By operationalizing the European Commission's seven technical requirements, the authors elucidate how legal, technical, and ethical considerations must be interwoven across the entire AI lifecycle. Explainability and auditability emerge as foundational elements in this integration, serving as mechanisms that bridge system-level design with public expectations of transparency and accountability. Their study highlights that robust regulatory frameworks are essential to avoid both safety risks and innovation stagnation, particularly in high-stakes sectors.

Complementing these perspectives, Merhi (2022) examines cross-cutting barriers that hinder responsible AI adoption. Using expert evaluations and the Analytic Hierarchy Process, the study identifies eleven obstacles—spanning technological, organizational, and contextual dimensions—with data quality emerging as the most critical. This aligns with broader findings that trustworthy AI requires not only ethical principles but also enabling infrastructure, including robust data governance, traceability structures, and institutional mechanisms that support oversight.

The risks associated with explainable AI (XAI) further illustrate the need for integrated approaches. Nannini et al. (2024) show that technical explanations alone cannot address sociotechnical concerns related to safety, accountability, and public perception. Through a thematic analysis of the literature, they propose a multi-layered risk assessment framework that supports documentation, monitoring, and intervention strategies. Their work reinforces the notion that evaluation mechanisms must be sensitive to both system-level performance and societal context, particularly when explanations are used as proxies for trust.

Public perception also plays a significant role in determining the legitimacy of AI systems. Dorotic et al. (2023) demonstrate that acceptance of AI varies across application domains and is strongly influenced by concerns about privacy, intrusiveness, and autonomy. Their findings highlight the necessity of integrating user-centered considerations—such as transparency in data practices and communication of system benefits—into governance frameworks. These results mirror the broader consensus that multi-dimensional assessments must account for both institutional requirements and user expectations.

Despite these advances, a review of scientific articles indexed in Scopus and Web of Science reveals a persistent gap: the field lacks technological platforms capable of operationalizing the multi-dimensional evaluation of AI systems (See Table 1). Using the search strategy "(evaluation OR audit) AND ('artificial intelligence' OR AI) AND (ethics OR fairness OR reliability)," the analysis shows that most existing tools focus on isolated components—such as fairness metrics or privacy guarantees—without integrating ethical, security, and governance dimensions into a unified assessment workflow. In the Peruvian context, only one preliminary initiative has been identified that attempts to evaluate ethical principles through a web-based platform (Romani et al., 2025). The scarcity of such tools underscores an enduring disconnect between conceptual frameworks and their implementation in organizational settings.

Recent efforts such as Peru's Readiness Assessment Methodology for AI (UNESCO, 2024) signal progress toward nationally grounded

TABLE 1 Technological platforms for AI evaluation.

Characteristics	PADTHAI-MM (Cohen et al., 2025)	ETAPAS (Tsoumma et al., 2023)	Aequitas flow (Jesus et al., 2024)	ANN-ANFIS (Wankhade et al., 2025)
Main focus	User-centered trustworthy AI design	Ethical assessment of AI in the public sector	End-to-end fair ML experimentation	Ethical evaluation of AI projects using a computational approach
Type of evaluation	Principle-based trust evaluation	Assisted ethical self-assessment	Algorithmic fairness auditing	Multicriteria ethical evaluation using fuzzy logic
Application domain	AI systems for analytical tasks	AI applications in public administration	Machine learning models	AI projects in diverse organizations
Evaluation focus	Trust and transparency in AI	Identification of ethical, legal, and social risks	Fairness, reproducibility, transparency	Transparency, fairness, accountability, social impact
Methods used	MAST (Methodology for trust assessment)	Surveys and interviews with public users	Metrics, optimization, fair ML pipelines	Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), Fuzzy Logic (FL), and Multi-Criteria Decision-Making (MCDM)
Level of interaction	User-centered AI system design	Web platform with end-user feedback	Programmable interface for researchers and practitioners	Evaluation model incorporating expert judgment and data
Results classification	Trust (high, low)	Requires qualitative interpretation by users	Comparative metrics and audit reports	Project prioritization based on ethical principles

strategies for AI governance. However, the literature consistently indicates that operationalizing ethical, technical, and legal requirements remains a major challenge. Together, these studies reinforce a central insight across the related work: although multidimensional frameworks exist, the field still lacks practical, integrated mechanisms capable of translating normative principles into systematic, scalable, and context-aware evaluation processes.

3 Materials and methods

The development of MESIAS followed a meticulous three-phase process. First, a conceptual model of the proposed system was formulated, encompassing a comprehensive description of each module and a visual representation of the application's main components and intended functionality. This stage was guided by the need to operationalize ethical principles in concrete development contexts, ensuring that the framework moved beyond abstract guidelines toward actionable mechanisms. Second, the conceptual model was implemented through the definition of its architectural framework, development procedures, and constituent modules, integrating both automated functions and deliberative structures to support traceable and transparent decision-making. Finally, a usability evaluation was conducted through a user validation phase to assess how effectively MESIAS supported responsible and accountable decision-making in AI projects.

3.1 MESIAS model and evaluation process

The development of MESIAS followed a meticulous three-phase process. A conceptual model of the proposed system was initially formulated, encompassing a comprehensive explanation of each module (see Figure 1), along with a systematized and visual representation of the application's primary concept and features. Secondly, the conceptual model was implemented through the delineation of its architectural framework, developmental procedures, and constituent modules (see Table 2). A final evaluation of the system's usability was conducted following a user validation phase. The objective of this evaluation was to ascertain the extent to which the system supported responsible decision-making in AI projects.

The MESIAS architecture consists of four main modules: Administration (user management, system settings, access control), Repository (storing results, project history, risk levels), Virtual Assistant (guiding assessments, generating automated feedback), and User Module (project evaluation, profile management, alerts). Icons for "Administrator" and "Consultant" represent the respective end-users, while directional arrows illustrate the flow of information and feedback loops among modules.

The process begins when the user accesses the web-based platform. Depending on their registered role—Administrator or Consultant—they authenticate into the system, where they can manage profiles, consult FAQs, or initiate evaluations through the Virtual Assistant module. This module also includes a public-facing landing page that outlines the purpose of the tool and provides open access to essential guidelines on AI governance, ensuring transparency even for non-registered users.

Upon successful login, users initiate the evaluation by selecting their AI project and choosing between a comprehensive or sectional assessment (seven sections in total). After completing the questionnaire, MESIAS applies an internal computational logic to process the data. Responses (s_i) are numerically encoded on a 1-to-5 Likert scale, where 1 indicates minimal adherence or high ethical risk, and 5 represents strong adherence or low risk (see Table 3). The system does not rely solely on a linear question–answer sequence; instead, it employs a weighted multi-criteria assessment model to compute the

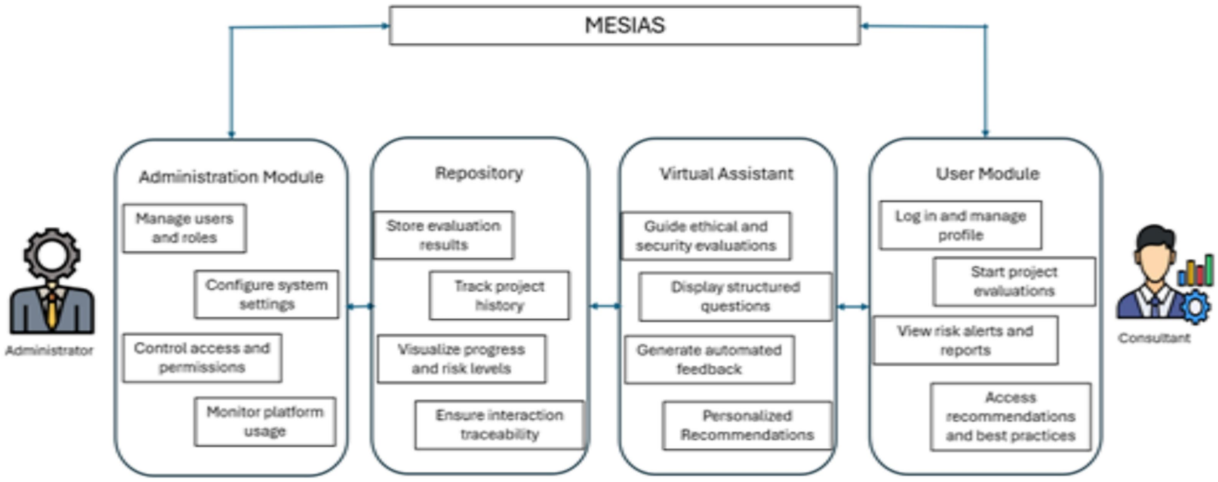

MESIAS
Administration Module
Manage users and roles
Configure system settings
Control access and permissions
Monitor platform usage
Repository
Store evaluation results
Track project history
Visualize progress and risk levels
Ensure interaction traceability
Virtual Assistant
Guide ethical and security evaluations
Display structured questions
Generate automated feedback
Personalized Recommendations
User Module
Log in and manage profile
Start project evaluations
View risk alerts and reports
Access recommendations and best practices
Administrator
Consultant

FIGURE 1
MESIAS model.

TABLE 2 Mesias modules.

Module	Description
Administrator	Manages users, roles, permissions, and the general configuration of the system.
Consultant	Conducts AI project evaluations, accesses Virtual Assistant, and reviews evaluation results.
Virtual Assistant	Provides the interface for conducting evaluations, guiding the user through structured questions and generating automated feedback. Includes a public landing page.
Repository	Stores evaluation results and enable monitoring and analysis of project progress.

TABLE 3 Risk ranges by dimension.

Dimension	Min.	Max.	High risk	Medium risk	Low risk
Human agency and oversight	22	110	22–51	52–81	82–110
Technical robustness and safety	36	180	36–84	85–132	133–180
Privacy and data governance	12	60	12–28	29–44	45–60
Transparency	13	65	13–30	31–47	48–65
Diversity and non-discrimination	25	125	25–58	59–91	92–125
Societal and environmental well-being	16	80	16–37	38–58	59–80
Accountability	12	60	12–28	29–44	45–60

ethical risk index R for each project dimension. Formally, the model calculates:

$$R_d = \frac{\sum_{i=1}^{n_d} w_i \cdot s_i}{\sum_{i=1}^{n_d} w_i}$$

where R_d represents the normalized risk score for dimension d, s_i is the user's response to question i on a scale from 1 to 5, and w_i is the assigned weight reflecting the relative importance of each ethical criterion as derived from the EU Trustworthy AI framework. Once all dimensional scores are computed, the system determines the total ethical risk index (R_T) for the project through an aggregation of weighted dimensions:

$$R_T = \frac{\sum_{d=1}^{7} \alpha_d \cdot R_d}{\sum_{d=1}^{7} \alpha_d}$$

where α_d denotes a scaling coefficient for dimension d. This coefficient is envisioned as a user-defined input (e.g., 1.0 for 'low context risk', 1.5 for 'high context risk') that allows auditors to manually increase the importance of specific dimensions based on the project's context (e.g., applying a higher α for 'Privacy' in a healthcare project).

Based on the resulting R_d values, MESIAS dynamically assigns each dimension a risk level according to proportional thresholds T_1 and T_2 derived from the theoretical range of possible scores (see Table 4).

Then, MESIAS visualizes the evaluation outcomes through a color-coded interface: red for high risk, yellow for medium, and green for low. The Repository module securely stores all evaluations, enabling longitudinal analysis of ethical performance across projects. Meanwhile, the dashboard integrates behavioral analytics, allowing administrators to detect trends, compare compliance evolution, and identify recurrent areas of ethical vulnerability within AI development practices.

The Virtual Assistant not only guides users throughout the process but also applies interpretative heuristics to contextualize risk levels, offering tailored recommendations based on aggregated patterns from prior evaluations. This dynamic feedback mechanism enhances the system's learning capacity and supports continuous improvement in ethical AI governance practices.

Unlike traditional checklists or static compliance frameworks, MESIAS integrates a hybrid evaluation logic that combines quantitative scoring and semantic interpretation through its Virtual Assistant module. To complement, MESIAS integrates semantic consistency validation through GPT-4. The model analyzes user explanations and detects potential contradictions or ethical inconsistencies by applying embedding-based similarity scoring and logical entailment checks.

3.2 MESIAS implementation

3.2.1 Technical architecture rationale

The architecture of MESIAS was conceived under a modular and service-oriented design paradigm to ensure scalability, transparency, and computational robustness (see Figure 2). The system is composed of a frontend interface and a backend infrastructure, which operate cohesively to deliver a seamless user experience through any web browser on desktop or laptop environments. The platform requires only a stable internet connection, making it accessible to auditors, consultants, and administrators of AI projects seeking to perform structured self-assessments of ethical and security compliance through the integrated Virtual Assistant. At its core, MESIAS embodies a three-layer logic model—presentation, application, and data management—that supports modular updates and algorithmic traceability.

The architecture illustrates MESIAS deployed on the PythonAnywhere cloud infrastructure, where users interact with the system via encrypted HTTPS sessions. The frontend and backend components are hosted jointly in the cloud and

TABLE 4 Risk levels and evaluation criteria.

Risk level	Description	Condition	Representative color
Low	Low risk, project considered safe	$S_{d_i} < T_1$	Green
Medium	Medium risk, requires attention	$T_1 < S_{d_i} < T_2$	Yellow
High	High risk, urgent action required	$S_{d_i} \geq T_2$	Red

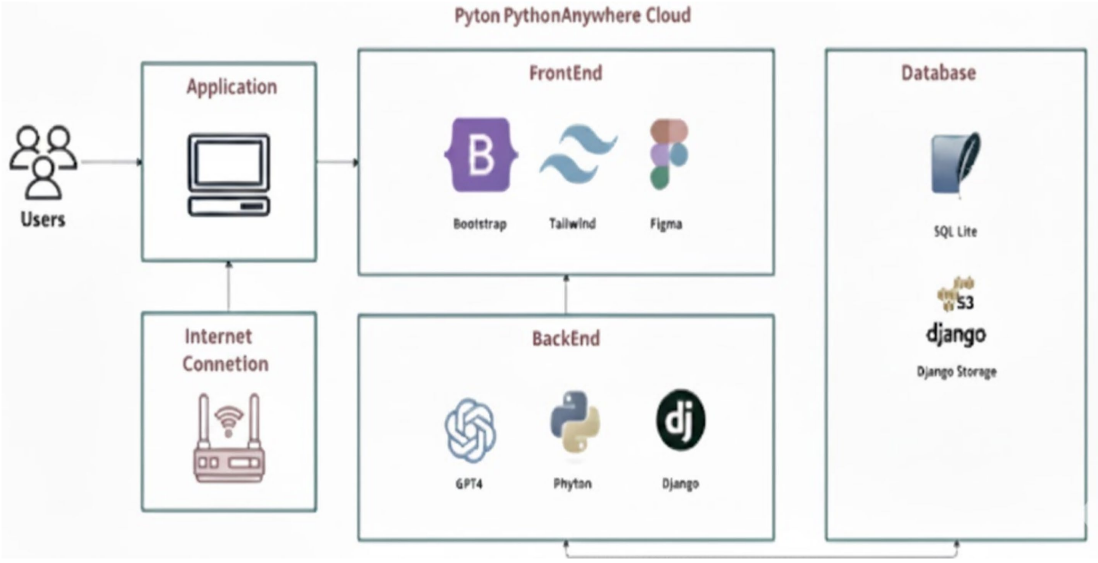


FIGURE 2
Physical architecture of MESIAS.

supported by a SQLite relational database and an S3-compatible object storage layer for persistent file management. Bidirectional arrows denote real-time data exchange between modules, while logic flows represent the orchestration of service requests, user queries, and automated responses generated by the Virtual Assistant.

The frontend layer, built with Bootstrap and Tailwind CSS, provides a responsive and adaptive interface that dynamically renders content across varying resolutions and operating systems. Figma was employed in the early design phase to prototype interactive components, allowing for rapid iteration and consistent user experience validation. The backend layer was implemented using the Django framework, leveraging Python's versatility to handle data models, access control, and logical workflows. Django's Model-View-Template (MVT) pattern ensures structural clarity and facilitates the traceability of ethical evaluation processes. Within this layer, GPT-4 is integrated as an intelligent agent responsible for dialog management, semantic interpretation of questions, and automated feedback generation. Through fine-tuned prompting strategies, GPT-4 aligns user responses with the evaluation model derived from the European Union's Ethics Guidelines for Trustworthy AI, ensuring consistency and interpretability in risk classification. The database layer relies on SQLite, chosen for its high I/O efficiency and suitability in agile research and pilot environments. It supports transactional integrity and facilitates the rapid querying of multi-dimensional risk data. Complementing this, Django Storage orchestrates secure document management for generated reports, project metadata, and trace logs. Deployment is executed on PythonAnywhere, a cloud-based platform optimized for Django hosting, which guarantees operational stability, automatic scaling, and continuous availability.

3.2.2 Development

The development of MESIAS adhered to an iterative prototyping methodology that prioritized adaptability, traceability, and empirical validation of each component. The frontend was crafted to maximize usability through visual modularity and consistent navigation flows. These were collaboratively defined and refined in Figma, integrating heuristic usability testing results from early prototypes. The backend development, implemented in Python using Django, centralized critical operations such as user authentication, project registration, dynamic routing, and evaluation management. A specific computational pipeline was established for the Virtual Assistant module, structured as follows:

- Input normalization: user responses are numerically encoded on a 1-to-5 Likert scale.
- Dimensional scoring: responses within each ethical dimension are aggregated via weighted arithmetic mean.
- Threshold calibration: dynamic cut-off points are computed based on the minimum and maximum attainable scores per dimension, establishing proportional intervals for high, medium, and low risk.
- Inference and feedback: GPT-4 processes the quantitative outputs to contextualize ethical vulnerabilities and generate natural-language insights for the user.

All assessment data, system logs, and results are securely persisted in the SQLite database, while Django Storage ensures encrypted file handling for reports and supporting documentation. This guarantees both the integrity and the reproducibility of evaluations, which are crucial in research-oriented contexts. Finally, PythonAnywhere was chosen for deployment due to its alignment with the project's operational philosophy: lightweight, cost-efficient, and capable of continuous integration cycles. Its architecture ensures that MESIAS remains accessible to global users while maintaining secure communication channels and consistent system uptime.

3.2.3 Modules

The MESIAS program was meticulously engineered around four interdependent core modules, each designed to fulfill a specific operational function while ensuring coherence across the overall architecture. These modules work synergistically to facilitate the secure, traceable, and analytically consistent evaluation of AI projects.

3.2.3.1 Administrator module

This module manages user authentication, role-based access control, and the overall configuration of the system's operational environment. It serves as the entry point for managing identities and permissions, ensuring data integrity and compliance with organizational hierarchies. The module performs two main functions: the registration of new user accounts and the validation of access credentials through a secure login interface. Authentication is handled using Django's built-in encryption framework, which protects credentials through salted hashing and session tokens. Once authenticated, users can access a profile management interface to modify or update their personal and contact information in real time. Figure 3A displays the new user registration form, Figure 3B depicts the secure login screen, and Figure 3C illustrates the editable user profile interface accessible after authentication.

3.2.3.2 Consultant module

This module is designed for subject-matter experts and auditors who evaluate AI projects. Following successful authentication, consultants are granted access to a data-driven dashboard that aggregates project information and operational metrics in real time. The dashboard provides analytical indicators such as the total number of assessments conducted, the count of pending and finalized evaluations, six-month activity trends, and the distribution of projects across evaluation phases and ethical dimensions. Interactive charts allow consultants to visualize progress and identify patterns of compliance or recurring risk factors across multiple AI initiatives. Figure 4A presents the main dashboard interface, while Figure 4B shows the detailed project evaluation summary, including metadata such as creation dates, active phases, and current evaluation status.

3.2.3.3 Virtual assistant module

The module constitutes the analytical core of MESIAS. It integrates GPT-4, which operates as a conversational agent that guides users through the self-assessment process based on the Ethics Guidelines for Trustworthy AI developed by the European Commission. The assessment instrument consists of 136 questions distributed across seven thematic dimensions, covering principles such as transparency, accountability, robustness, and human oversight. Questions are presented in an adaptive conversational

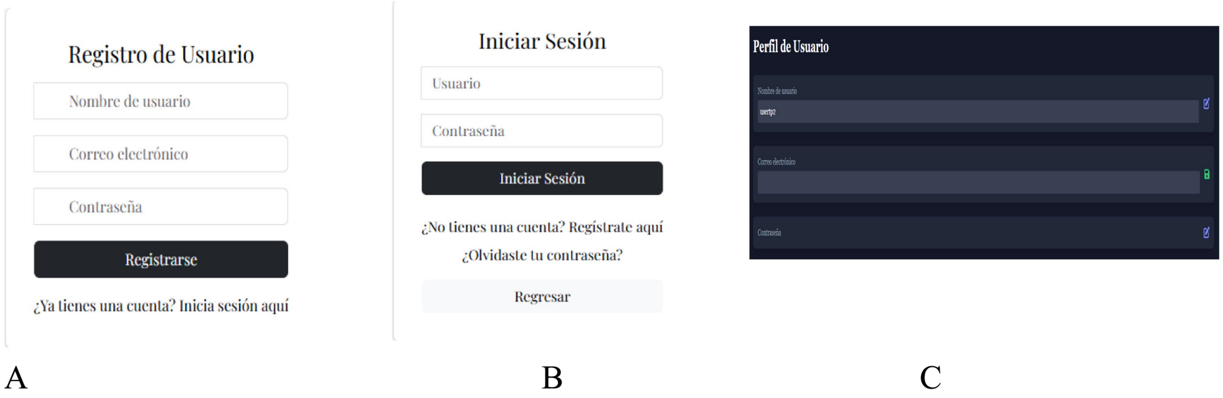


FIGURE 3
Administrator module: **(A)** registration of a new user; **(B)** user login screen; **(C)** user profile management interface.

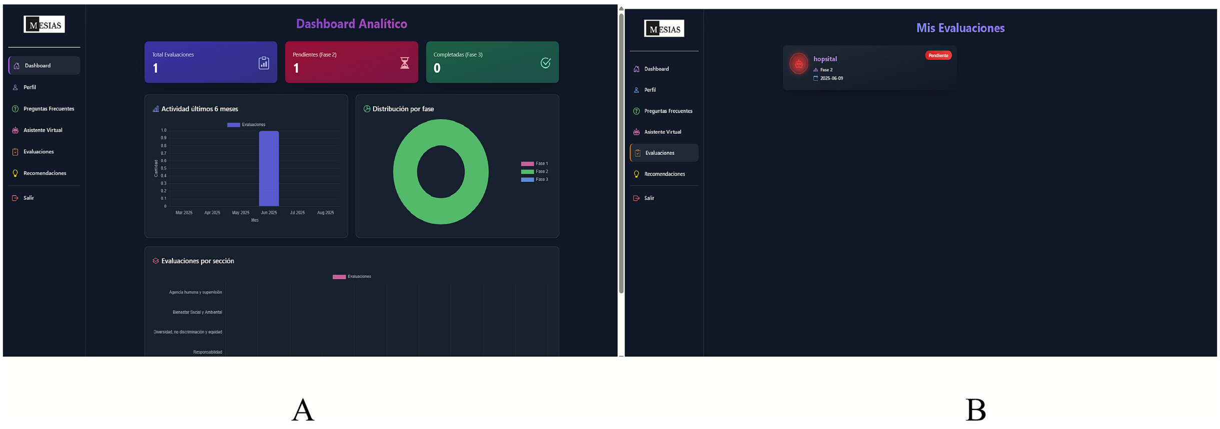


FIGURE 4
Consultant module: **(A)** interaction dashboard; **(B)** summary of project evaluations.

format, where GPT-4 dynamically interprets context to refine prompts and provide clarifications. Responses are encoded on a five-point Likert scale (1–5), enabling the quantification of ethical compliance levels. The module also includes a public landing page providing open access to the foundational guidelines underpinning the evaluation framework. Each entry displays key attributes such as the publication date, objectives, highlights, and references to related documentation. Once responses are submitted, the system computes risk scores and, through the GPT-4 engine, generates automated interpretive feedback, offering users personalized recommendations for risk mitigation and governance improvement. Figure 5A illustrates the user's interactive dialogue with the virtual assistant, Figure 5B depicts the structured evaluation interface, and Figure 5C displays the automated feedback visualization.

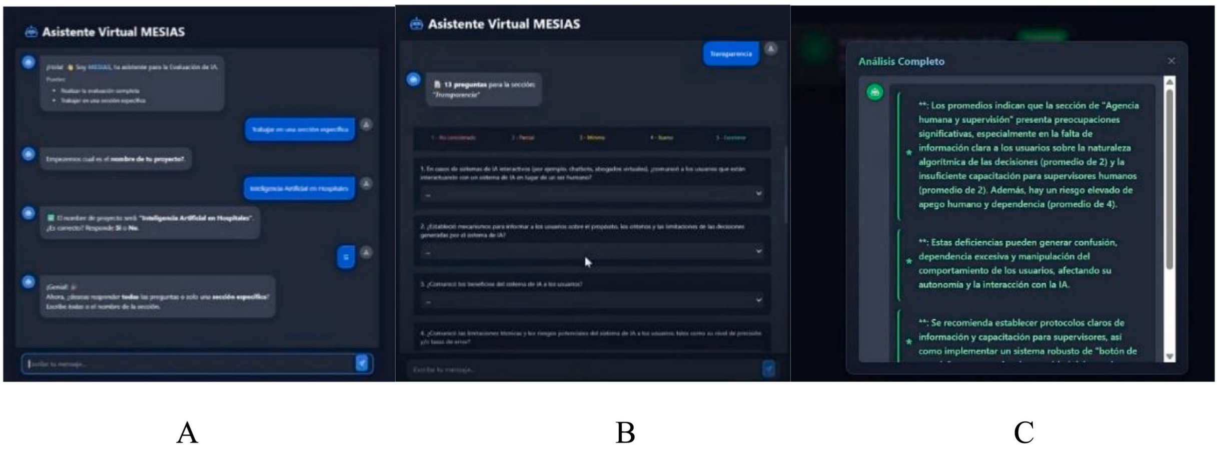
3.2.3.4 Repository module

The Repository Module ensures the long-term storage, traceability, and accessibility of all completed evaluations. It enables users to review project histories, monitor the evolution of ethical compliance, and access previously generated reports for comparative analysis. In addition to its archival function, this module provides auxiliary services, such as a dynamic repository of frequently asked questions and a feedback submission interface where users can propose enhancements to the system's design or suggest new evaluation criteria. As shown in Figure 6A, the landing page of this module displays the core guidelines available for consultation. Figure 6B presents the detailed view of a selected guideline, including its content hierarchy and associated documentation.

4 Results

To assess the initial implementation of the MESIAS tool, an exploratory case study was conducted in a private organization. This study compared two projects evaluated through the traditional manual procedure with one project assessed using MESIAS. The purpose was not to establish a causal validation, but rather to observe practical improvements in key performance indicators, such as the total time required to complete the evaluation process and the number of human resources involved (see Figure 7).

As illustrated in the above figure, under conventional methodologies, the mean operational duration for a given project



¡Hola! 👋 Soy MESIAS, tu asistente para la Evaluación de IA.
Puedes:
- Realizar la evaluación completa
- Trabajar en una sección específica

Trabajar en una sección específica

Empecemos: cual es el **nombre de tu proyecto?**

Inteligencia Artificial en Hospitales

✅ El nombre de proyecto será: **"Inteligencia Artificial en Hospitales"**. ¿Es correcto? Responde **Sí** o **No**.

Sí

¡Genial! ¡
Ahora, ¿deseas responder **todas** las preguntas o solo una **sección específica**?
Escribe todas o el nombre de la sección.

Transparencia

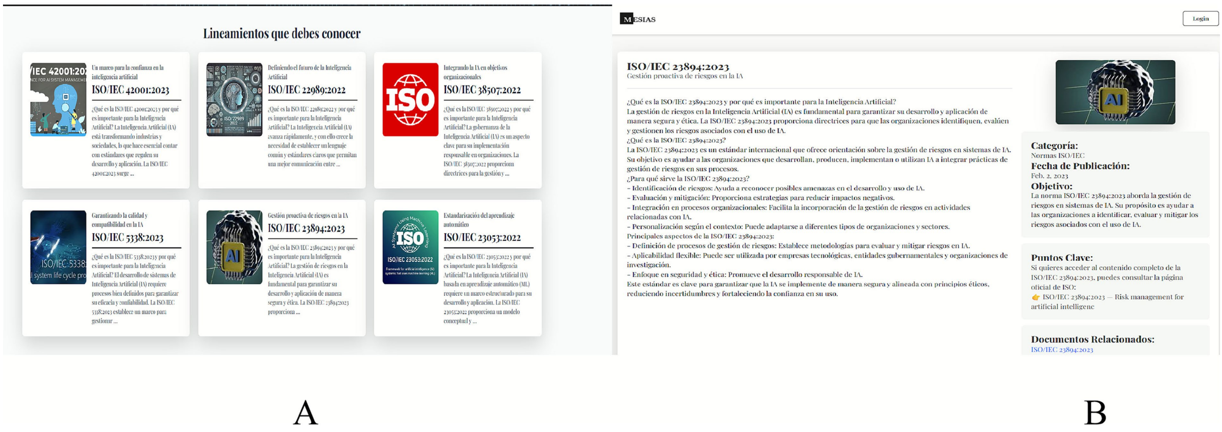
13 preguntas para la sección: *"Transparencia"*

1 - No considerado · 2 - Parcial · 3 - Mínimo · 4 - Bueno · 5 - Excelente

1. En casos de sistemas de IA interactivos (por ejemplo, chatbots, abogados virtuales), ¿comunicó a los usuarios que están interactuando con un sistema de IA en lugar de un ser humano?
2. ¿Estableció mecanismos para informar a los usuarios sobre el propósito, los criterios y las limitaciones de las decisiones generadas por el sistema de IA?
3. ¿Comunicó los beneficios del sistema de IA a los usuarios?
4. ¿Comunicó las limitaciones técnicas y los riesgos potenciales del sistema de IA a los usuarios, tales como su nivel de precisión y/o tasas de error?

- **: Los promedios indican que la sección de "Agencia humana y supervisión" presenta preocupaciones significativas, especialmente en la falta de información clara a los usuarios sobre la naturaleza algorítmica de las decisiones (promedio de 2) y la insuficiente capacitación para supervisores humanos (promedio de 2). Además, hay un riesgo elevado de apego humano y dependencia (promedio de 4).
- **: Estas deficiencias pueden generar confusión, dependencia excesiva y manipulación del comportamiento de los usuarios, afectando su autonomía y la interacción con la IA.
- **: Se recomienda establecer protocolos claros de información y capacitación para supervisores, así como implementar un sistema robusto de "botón de

FIGURE 5
Virtual assistant module: **(A)** virtual assistant conversation; **(B)** project evaluation; **(C)** automated feedback.


Lineamientos que debes conocer

Un marco para la confianza en la inteligencia artificial
ISO/IEC 42001:2023
¿Qué es la ISO/IEC 42001:2023 y por qué es importante para la Inteligencia Artificial? La Inteligencia Artificial (IA) está transformando industrias y sociedades, lo que hace esencial contar con estándares que regulen su desarrollo y aplicación. La ISO/IEC 42001:2023 surge ...

Definiendo el futuro de la Inteligencia Artificial
ISO/IEC 22989:2022
¿Qué es la ISO/IEC 22989:2022 y por qué es importante para la Inteligencia Artificial? La Inteligencia Artificial (IA) avanza rápidamente, y con ello crece la necesidad de establecer un lenguaje común y estándares claros que permitan una mejor comunicación entre ...

Integrando la IA en objetivos organizacionales
ISO/IEC 38507:2022
¿Qué es la ISO/IEC 38507:2022 y por qué es importante para la Inteligencia Artificial? La gobernanza de la Inteligencia Artificial (IA) es un aspecto clave para su implementación responsable en organizaciones. La ISO/IEC 38507:2022 proporciona directrices para la gestión y ...

Garantizando la calidad y compatibilidad en la IA
ISO/IEC 5338:2023
¿Qué es la ISO/IEC 5338:2023y por qué es importante para la Inteligencia Artificial? El desarrollo de sistemas de Inteligencia Artificial (IA) requiere procesos bien definidos para garantizar su eficacia y confiabilidad. La ISO/IEC 5338:2023 establece un marco para gestionar ...

Gestión proactiva de riesgos en la IA
ISO/IEC 23894:2023
¿Qué es la ISO/IEC 23894:2023 y por qué es importante para la Inteligencia Artificial? La gestión de riesgos en la Inteligencia Artificial (IA) es fundamental para garantizar su desarrollo y aplicación de manera segura y ética. La ISO/IEC 23894:2023 proporciona ...

Estandarización del aprendizaje automático
ISO/IEC 23053:2022
¿Qué es la ISO/IEC 23053:2022 y por qué es importante para la Inteligencia Artificial? La Inteligencia Artificial (IA) basada en aprendizaje automático (ML) requiere un marco estructurado para su desarrollo y aplicación. La ISO/IEC 23053:2022 proporciona un modelo conceptual y ...

ISO/IEC 23894:2023
Gestión proactiva de riesgos en la IA

¿Qué es la ISO/IEC 23894:2023 y por qué es importante para la Inteligencia Artificial?
La gestión de riesgos en la Inteligencia Artificial (IA) es fundamental para garantizar su desarrollo y aplicación de manera segura y ética. La ISO/IEC 23894:2023 proporciona directrices para que las organizaciones identifiquen, evalúen y gestionen los riesgos asociados con el uso de IA.
¿Qué es la ISO/IEC 23894:2023?
La ISO/IEC 23894:2023 es un estándar internacional que ofrece orientación sobre la gestión de riesgos en sistemas de IA. Su objetivo es ayudar a las organizaciones que desarrollan, producen, implementan o utilizan IA a integrar prácticas de gestión de riesgos en sus procesos.
¿Para qué sirve la ISO/IEC 23894:2023?
- Identificación de riesgos: Ayuda a reconocer posibles amenazas en el desarrollo y uso de IA.
- Evaluación y mitigación: Proporciona estrategias para reducir impactos negativos.
- Integración en procesos organizacionales: Facilita la incorporación de la gestión de riesgos en actividades relacionadas con IA.
- Personalización según el contexto: Puede adaptarse a diferentes tipos de organizaciones y sectores.
Principales aspectos de la ISO/IEC 23894:2023:
- Definición de procesos de gestión de riesgos: Establece metodologías para evaluar y mitigar riesgos en IA.
- Aplicabilidad flexible: Puede ser utilizada por empresas tecnológicas, entidades gubernamentales y organizaciones de investigación.
- Enfoque en seguridad y ética: Promueve el desarrollo responsable de IA.
Este estándar es clave para garantizar que la IA se implemente de manera segura y alineada con principios éticos, reduciendo incertidumbres y fortaleciendo la confianza en su uso.

Categoría:
Normas ISO/IEC
Fecha de Publicación:
Feb. 2, 2023
Objetivo:
La norma ISO/IEC 23894:2023 aborda la gestión de riesgos en sistemas de IA. Su propósito es ayudar a las organizaciones a identificar, evaluar y mitigar los riesgos asociados con el uso de IA.

Puntos Clave:
Si quieres acceder al contenido completo de la ISO/IEC 23894:2023, puedes consultar la página oficial de ISO:
👉 ISO/IEC 23894:2023 — Risk management for artificial intelligenc

Documentos Relacionados:
ISO/IEC 23894:2023

FIGURE 6
Repository module: **(A)** guidelines landing page; **(B)** guideline detail page. Two interfaces of the repository module.

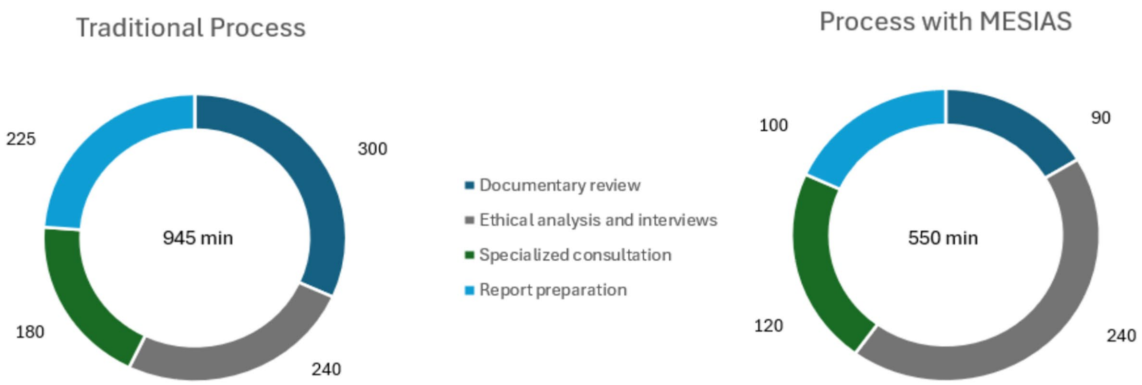


FIGURE 7
Comparison of phase-wise duration between the traditional process and the MESIAS-based process.

was approximately 945 min. Conversely, the implementation of the MESIAS approach yielded a 41.8% reduction, with an average operational time of 550 min being attained. This efficiency enhancement was accomplished by means of the automation of the analysis and the generation of reports. An analysis of the time allocation by phase indicated that the documentation review stage was reduced from 300 to 90 min, a 70% improvement due to the implementation of guided forms to streamline documentation input. While the duration of interviews and ethical analysis remained constant at 240 min across both methods, MESIAS provided a more structured process. The duration of specialist consultations was reduced from 180 to 120 min, representing a 33% improvement, which can be attributed to the MESIAS integration of normative criteria. The duration required for report preparation was reduced from 225 to 100 min, representing a 55% improvement. This enhancement can be attributed to the utilization of MESIAS, a system that automatically generates reports based on user responses.

With respect to the optimization of human resources, the conventional process necessitated an average of five individuals per evaluation. The individuals responsible for this review and signature included an IT risk analyst, an ethics and governance specialist, a legal advisor, a project coordinator, and a department head. In the MESIAS study, the process was efficiently executed by only three individuals: one user responsible for data entry and review, one ethics analyst, and one supervisor for final report validation. This represents a 40% reduction in human resources, enabled by the automation of diagnostics, automated reporting, and integrated regulatory repositories.

Concerning the roles involved, the MESIAS streamlined the workflow without compromising technical rigor. Consequently, the process was streamlined to encompass only essential roles for control and final validation. This eliminated the need for coordination with external departments during the preliminary phases. Moreover, the implementation of process standardization has been observed to evolve from a state of variability among teams to a state of uniformity, facilitated by the MESIAS framework, which provides a consistent structure and set of evaluation criteria. Traceability decisions has undergone significant enhancement, progressing from a partial state to a comprehensive, complete traceability. This enhancement can be attributed to the meticulous logging of each step of the process by the tool, thereby facilitating subsequent audits.

To obtain expert judgment regarding the usability of the MESIAS application, four leaders from key organizational areas were invited. The participants, each with over a decade of professional experience, were selected to ensure a mature, critical, and representative evaluation of the system's functional capabilities. The subjects' varied areas of expertise enabled feedback from multiple functional and strategic perspectives.

In the aftermath of their engagement with the MESIAS application, the participating leaders were requested to furnish responses to a series of inquiries devised to assess the application's usability across multiple dimensions, encompassing aspects such as ease of use, navigational clarity, content relevance, design appropriateness, and its impact on decision-making. Given that the Likert scale employed integer values and the potential for average results to include decimal fractions, the mean score for each question was calculated using all participant responses. Subsequently, these results were aggregated to identify general trends. For interpretation, the following classification was applied: The interval between 1.8 and 2.6 is designated as "low," the interval between 2.6 and 3.4 as "moderate," the interval between 3.4 and 4.2 as "high," and the interval between 4.2 and 5 as "very high" (Yaska and Nuhu, 2024).

The evaluation, which was conducted by leaders with extensive experience in key areas of technology and management, revealed an overall average score of 4.16 (83%). The aspects of MESIAS that received the highest ratings were its potential impact on governance (Q08, 5.00), which was considered exceptional, particularly its ability to transform ethical decision-making into AI projects. Question 04, which pertains to the efficacy of the virtual assistant, received a score of 4.50. This result indicates that the virtual assistant is effective in promoting ethical reflection and assuming a pivotal role in the evaluation process.

In the high-scoring categories, the interpretation of results (Q03, 4.25) was particularly noteworthy for its ability to communicate complex issues related to AI ethics and security in an accessible manner. However, the ease of initiating the ethical evaluation (Q01, 4.00) could be enhanced by the implementation of improved training and guidance resources, to reduce the learning curve. The ethical questionnaire questions (Q05, 4.25) were also positively rated in terms of relevance and clarity; however, some users noted that certain questions could be more detailed, specifically about ethical frameworks.

In areas where moderate scores were observed, the navigation between sections (Q02, 3.25) emerged as the paramount dimension, with users reporting challenges in transitioning between the evaluation form, the dashboard, and the history section. It was recommended that the navigation and visual elements be revised to enhance efficiency. The web interface design (Q06, 3.50) was identified as a significant area for improvement. Despite its functionality, the system was perceived as being unintuitive and offering a visual experience that did not support agile navigation. Despite the identified areas for improvement, the willingness to recommend MESIAS remained high (Q07, 4.50), which reinforces its strategic value and emphasizes the need for further work on optimizing the user experience to ensure its large-scale success in AI environments (Table 5).

In addition, the participating leaders offered recommendations for enhancing the MESIAS framework. These recommendations included the integration of additional international regulatory

TABLE 5 Results of the evaluation by leaders regarding MESIAS and its usability.

ID	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08
L1	5	4	5	4	5	4	5	5
L2	4	3	4	5	4	3	4	5
L3	4	3	4	5	4	4	5	5
L4	3	3	4	4	4	3	4	5
Average scores	4.00	3.25	4.25	4.50	4.25	3.50	4.50	5.00

frameworks and the establishment of internal policy repositories. Furthermore, they recommended the implementation of automatic report generation and the enhancement of interoperability with tools such as Jira or Trello. The proposed solution involved the optimization of result visualization through the implementation of real-time dashboards and the provision of contextual guidance tailored to the user's profile. The objective of these measures was to facilitate the adoption of digital tools within organizations characterized by a lower degree of digital maturity. The collective feedback indicates a strategic vision that acknowledges MESIAS transformative potential while delineating a clear roadmap for its scalability and integration in corporate environments.

The survey on satisfaction was administered to 52 professionals in the IT field, including roles such as quality analyst, programmer analyst, and information security analyst, with specialties ranging from AI security, AI solution development, IT governance, to technology ethics. In a manner consistent with the findings of the preceding survey, the mean scores for each question were computed by aggregating all responses without exclusion. The means by dimension were subsequently ascertained, computed as the mean of the five questions for each dimension (see Table 6). The results from this evaluation reflected a highly positive assessment of the tool across all dimensions evaluated. The mean score for all questions was 4.5 out of 5, indicating an exceptionally high and consistent favorable perception among the users.

The objective of this study is to assess the usability of the system under review. The usability dimension received a highly positive overall assessment, with a general average score of 4.58. The highest ratings were concentrated in the items related to ease of use (Q1 and Q2, both scoring 4.63), indicating a favorable perception regarding the system's intuitiveness and the support provided by the integrated virtual agent. Conversely, the system's communicative capacity, specifically the clarity of the feedback received (Q3), was also highly regarded by users, with a mean rating of 4.60. However, opportunities for enhancement were identified in specific components, such as the repository of international ethical guidelines (Q4) and the monitoring dashboard (Q5), both of which recorded slightly lower scores (4.52 each). These results suggest potential refinements in the presentation of information and the usability of these modules. The findings indicate an efficient and user-friendly interaction, with satisfaction levels reaching 92%, thereby positioning MESIAS as an accessible tool from a user experience design perspective.

Contents: The content evaluation yielded exceptional results, with an overall average score of 4.72, thus classifying it as one of the highest-rated dimensions in the study. It was noted by users that the content related to ethics and AI security was marked by clarity, relevance, and pertinence. There was particular emphasis on the alignment of the content with international standards (Q7, 4.73) and the usefulness of the information provided by the virtual assistant (Q8, 4.75). The logical structure of the questions (Q9, 4.71) and the appropriateness of the recommendations generated (Q10, 4.69) were also positively assessed, reinforcing MESIAS as a robust technical and regulatory resource that enhances users' evaluative capabilities. The recorded satisfaction level was 94%, reflecting significant acceptance of the proposed framework.

Tracking: About the monitoring dimension, the results indicate a favorable reception, with an average score of 4.54. The functionalities related to project progress visualization and traceability were highly valued, particularly the use of the monitoring dashboard (Q11, 4.50) and the display of evaluation progress (Q12, 4.56). Furthermore, the participants expressed appreciation for the functionality that enabled the review of previous evaluations (Q13, 4.52) and the resumption of ongoing processes (Q14, 4.52). This suggests that MESIAS actively supports continuous evaluation and knowledge management. Another highly rated feature was access to the detailed response history (Q15, 4.60). These data confirm that the tool offers effective mechanisms for tracking, registering, and retrieving information, although some functionalities could be further optimized to enhance the user experience. The satisfaction index in this dimension reached 91%.

Satisfaction: The general satisfaction dimension received the highest average score in the analysis, with a mean of 4.73, indicating a highly satisfactory user experience. The mean score for the tool's overall assessment (Q16, 4.79) and the mean score for its willingness to recommend its use (Q17, 4.69) reflect a consolidated perception of MESIAS effectiveness. Furthermore, the platform's contributions to ethical governance in project development (Q18, 4.77) and its expressed intent to continue using the tool in the future (Q19, 4.60) were acknowledged. The item that received the highest rating was the contribution to improved ethical decision-making (Q20, 4.81), thereby solidifying the position of MESIAS as a strategic instrument of significant value in organizational contexts. The overall satisfaction level for this dimension was 95%, positioning the tool as a solution with strong potential for sustained adoption and scalability in corporate environments.

5 Discussion

The findings of this study contribute to the ongoing debate on how to operationalize ethical principles in artificial intelligence governance, a challenge that continues to reveal both epistemological and institutional tensions (De Laat, 2021). In this regard, the MESIAS model provides a concrete response to one of the central gaps in the literature: the absence of formalized and auditable instruments that enable the translation of normative guidelines into organizational practice, particularly in regions with limited regulatory and technical capacities.

TABLE 6 Results of the evaluation by professionals regarding MESIAS and its usability.

Dimension	Usability					Content					Tracking					Satisfaction				
ID	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18	Q19	Q20
Average scores	4.63	4.63	4.60	4.52	4.52	4.69	4.73	4.75	4.71	4.69	4.50	4.56	4.52	4.52	4.60	4.79	4.69	4.77	4.60	4.81
	4.58					4.72					4.54					4.73				

Unlike frameworks such as ETAPAS, which focus on ethics-by-design through web interfaces for public-sector actors in Europe but struggle with issues of digital literacy and trust (Tsourma et al., 2023), MESIAS introduces a hybrid approach that combines conversational interaction with structured analytical logic. This design choice positions MESIAS not as a compliance tool, but as a deliberative reasoning system, capable of guiding users through context-sensitive ethical assessments. The sustained usability performance (above 4.5/5 across dimensions) supports the hypothesis that structured, dialogic mediation enhances both comprehension and adherence to ethical frameworks (Binns et al., 2018; Morley et al., 2019; Winfield et al., 2021).

From a technical standpoint, MESIAS departs from the dominance of data-intensive fairness tools such as Aequitas Flow or FATE, which rely on algorithmic auditing within machine learning pipelines. Instead, MESIAS employs a logic-based and semi-quantitative architecture, where ethical dimensions are decomposed into weighted subcomponents and integrated into a composite ethical risk index. Each indicator contributes to the overall assessment through a normalized aggregation function of the system, which represents the standardized score of the dimension and its relative weight, empirically calibrated by user consensus. This approach preserves mathematical transparency and traceability, offering an interpretable evaluative mechanism absent in large language models or purely heuristic frameworks. The system thus bridges symbolic reasoning and computational ethics, aligning with the emergent paradigm of "structured deliberation engines" in AI ethics research.

In operational terms, MESIAS demonstrated a 41.8% reduction in evaluation time, a 40% decrease in human resources, and a 47.8% drop in costs, indicating that ethical governance can be both rigorous and scalable. These outcomes contrast with systems that, while theoretically robust—such as ANN/ANFIS-based models for moral uncertainty (Wankhade et al., 2025)—are often impractical for institutions lacking large datasets or specialized AI ethics expertise. MESIAS's lightweight architecture thus expands the accessibility of ethical oversight mechanisms, making them feasible in low-resource environments typical of Latin American institutions.

Conceptually, MESIAS advances the field by reframing ethics not as a static checklist but as a dynamic evaluative process, where normative criteria are interpreted through human-machine collaboration. This echoes recent calls to move from "ethics as compliance" toward "ethics as practice" (Floridi et al., 2018; Jobin et al., 2019), aligning normative reflection with real-time decision processes. By embedding deliberation in an algorithmic structure, MESIAS provides empirical evidence that ethical reasoning can be operationalized without sacrificing interpretability or contextual sensitivity. Moreover, the tool introduces an innovative institutional layer absent in other evaluation systems. While Aequitas Flow and similar frameworks optimize fairness metrics at the model level, MESIAS extends evaluation to organizational governance, tracking accountability, documentation, and decision pathways. This expands the notion of AI ethics from technical compliance to systemic accountability, an approach particularly relevant for public-sector and consultancy environments.

Finally, from a theoretical standpoint, MESIAS contributes to the emerging body of work advocating for regional epistemologies of AI ethics. By translating global principles into a Latin American governance model, the system demonstrates that ethical AI is not only a universal aspiration but also a contextual construct that must adapt to socio-institutional realities. Its modular and auditable design enables reproducibility across diverse domains while maintaining normative coherence. This work thereby positions MESIAS as a reference model for future research seeking to reconcile ethical rigor, institutional feasibility, and computational interpretability in the design of AI oversight systems.

6 Conclusion

The objective of this study was to design, develop, and validate MESIAS, a digital tool designed to strengthen the ethical governance of AI projects. A validation strategy centered on criteria of operational efficiency, usability, and perceived satisfaction was implemented to demonstrate that MESIAS constitutes a comprehensive, accessible, and effective solution for evaluating ethical, normative, and security principles in AI-based technological projects.

The efficiency validation results, obtained through a comparative pre- and post-implementation study in a private sector organization, showed significant operational improvements. The platform achieved a significant reduction in the average evaluation time, from 945 to 550 min, representing a 41.8% decrease. Additionally, the number of human resources required per evaluation decreased from five to three individuals, marking a 40% reduction in the necessary workforce. These results indicate that MESIAS not only contributes to ethical concerns but also enhances resource-intensive processes, thereby increasing their efficiency and sustainability.

Regarding usability validation, this was conducted with leaders from the technological sector, who reported an overall acceptance rate of 85%. These strategic actors not only provided a favorable assessment of the value proposition of MESIAS but also furnished pivotal recommendations for its institutional scalability. It is important to note that the necessity of sector-specific versions of the tool, as well as its integration with existing risk management and compliance platforms, was emphasized. This input is critical for projecting MESIAS's adoption in organizational ecosystems. Furthermore, data obtained from surveys administered to 52 professionals in the technology field demonstrated consistently high average ratings across all evaluated dimensions: usability (4.58), content (4.72), monitoring (4.54), and overall satisfaction (4.73), all on a 5-point scale. These scores are indicative of the ease of use and clarity of the interface, the relevance of the ethical content provided, and the effectiveness of the system in guiding critical reflection and informed decision-making processes.

The findings of the study generally support the innovative contribution of MESIAS as a pioneering tool that translates international ethical frameworks into standardized, auditable, and replicable operational practices. The system's modular design, user-friendly interface, and evidence-based decision-making orientation position it as a valuable instrument for institutionalizing ethics in AI projects, particularly in contexts that require low-cost, highly adaptable solutions.

However, it is imperative to acknowledge the methodological limitations of the present study, which inform the direction of future research. First, the exploratory case study on efficiency was based on a pre/post design within a single organization and did not include a control group. Therefore, while the observed reductions in evaluation time (41.8%) and resource utilization (40%) are encouraging, potential confounding factors—such as the learning effect of the team or contextual organizational dynamics—cannot be ruled out. Consequently, these findings should be interpreted as indicative observations rather than causal evidence. Second, the 136-item evaluation instrument has not yet undergone a formal psychometric validation (e.g., factor analysis, reliability assessment). This is a significant limitation, as the validity of the risk scores depends on the validity of the instrument.

However, we argue that the instrument's primary function at this stage is not as a diagnostic measurement tool, but as a deliberative scaffolding mechanism. Its purpose is to operationalize abstract principles into concrete questions, guiding teams through a structured and comprehensive reflection process. This approach aligns with methodologies like Value-Sensitive Design (VSD) (Friedman et al., 2013), where the process of inquiry is a key outcome, not just the quantitative result. This conceptual approach, emphasizing deliberative over diagnostic evaluation, aligns with the theoretical framework introduced in Section 2. Nevertheless, establishing the instrument's validity is a critical next step. Future work will prioritize a formal psychometric validation, likely using a Delphi panel to establish content validity followed by exploratory and confirmatory factor analysis to ensure construct validity and reliability.

Finally, both the usability testing sample and the efficiency observations were limited to professionals from a single institution. Future studies should include more heterogeneous samples across multiple sectors and organizational maturity levels to strengthen generalizability and cross-context validation.

This study establishes a foundation for future research and development in this field. The MESIAS system has already been enhanced based on survey feedback, which identified key areas for improvement, including navigation, interface design, and report generation. Future work should focus on developing sector-specific versions tailored to regulatory frameworks in areas such as healthcare, finance, and justice. Additionally, the integration of adaptive machine learning modules capable of refining recommendations based on usage history—without compromising the ethical core of the system—represents a promising direction. Another relevant avenue concerns the assessment of MESIAS's contribution to ethical decision-making and AI risk mitigation in operational settings.

Finally, it is proposed that the institutional impact of MESIAS be analyzed in terms of cultural transformation and governance strengthening, particularly in organizations with low digital maturity or resistance to change. These findings demonstrate that MESIAS represents a significant advancement in the operationalization of ethics in AI, bridging the gap between normative discourse and practical implementation. Its scalability, adaptability, and traceability position it as a benchmark for future developments aimed at fostering transparent, inclusive, and responsible AI ecosystems.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants or participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

GR: Conceptualization, Data curation, Formal analysis, Software, Validation, Visualization, Writing – original draft. CA: Conceptualization, Data curation, Formal analysis, Software, Validation, Visualization, Writing – original draft. JS: Investigation, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors acknowledge the Dirección de Investigación de la Universidad Peruana de Ciencias Aplicadas (UPC) for the support provided, as well as the participants of the study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ansari, M. F., Dash, B., Sharma, P., and Yathiraju, N. (2022). The impact and limitations of artificial intelligence in Cyberse-curity: a literature review. doi: 10.17148/IJARCCE.2022.11912

Ashok, M., Madan, R., Joha, A., and Sivarajah, U. (2021). Ethical framework for artificial intelligence and digital technologies. *Int. J. Inf. Manag.* 62:102433. doi: 10.1016/j.ijinfomgt.2021.102433

Benchaita, S. "Data suggests growth in Enterprise adoption of AI is due to widespread deployment by early adopters". (2024) IBM Newsroom. Accessed May 24, 2025. [Online]. Available online at: https://newsroom.ibm.com/2024-01-10-Data-Suggests-Growth-in-Enterprise-Adoption-of-AI-is-Due-to-Widespread-Deployment-by-Early-Adopters

Bietti, E. (2020). "From ethics washing to ethics bashing" in Proceedings of the 2020 conference on fairness, accountability, and transparency (FAT* '20), 210–219. doi: 10.1145/3351095.3372860

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N., *It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions, ACM Conference on Human Factors in Computing Systems (CHI'18)*, pp. 1–14, 2018

Boppiniti, S. T. (2022). Ethical implications of artificial intelligence: a review of early research and perspectives. *IEJRD – Int. Multidiscip. J.* 7, 1–8.

Brown, Davidovic, J., and Hasan, A. (2021). The algorithm audit: scoring the algorithms that score us. *Big Data Soc.* 8:2053951720983865. doi: 10.1177/2053951720983865

Buchholz, J., Lang, B., and Vyhmeister, E. (2022). The development process of responsible AI: the case of ASSISTANT. *IFAC-PapersOnLine* 55, 7–12. doi: 10.1016/j.ifacol.2022.09.360

Buhmann, A., and Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technol. Soc.* 64:101475. doi: 10.1016/j.techsoc.2020.101475

Cohen, M. C., Kim, N., Ba, Y., Pan, A., Bhatti, S., Salehi, P., et al. (2025). Padthai-mm: principles-based approach for designing trustworthy, human-centered AI using the MAST methodology. *AI Mag.* 46. doi: 10.1002/aaai.70000

Daly, A., Hagendorff, T., Hui, L., Mann, M., Marda, V., Wagner, B., et al. (2019). Artificial intelligence, governance and ethics: global perspectives. *SSRN Electron. J.* doi: 10.2139/ssrn.3414805

De Borba, J. G. R., Canedo, E. D., and Filho, G. P. R. (2024). Bridging theory and practice: a tool for translating ethical Ai re-quirements into ethical user stories. *SSRN Electronic Journal*. doi: 10.2139/ssrn.5067796

De Laat, P. B. (2021). Companies committed to responsible AI: from principles towards implementation and regulation? *Philosophy Technol.* 34, 1135–1193. doi: 10.1007/s13347-021-00474-3

Díaz-Rodríguez, N., Javier, D. S., Coeckelbergh, M., López, D. P. M., Herrera-Viedma, E., and Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: from AI principles, ethics, and key requirements to responsible AI systems and regula-tion. *arXiv (Cornell University) Information Fusion*, 19:101855. doi: 10.48550/arxiv.2305.02231

Dorotic, M., Stagno, E., and Warlop, L. (2023). AI on the street: context-dependent responses to artificial intelligence. *Int. J. Res. Mark.* 41, 113–137. doi: 10.1016/j.ijresmar.2023.08.010

Español, A. G., and Sylvan, E. (2023). "AI: what should Latin American governments do?" CAF, [Online]. Available online at: https://cyber.harvard.edu/story/2023-05/generative-ai-what-should-governments-latin-america-do (Accessed May 14, 2025).

Fedele, A., and Punzi, C. y Tramacere, S., "The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education", *Comput. Law & Secur. Rev.*, vol. 53:105986, 2024. doi: 10.1016/j.clsr.2024.105986

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., et al. (2018). AI4People—An ethical framework for a good AI society: opportunities, risks, principles, and recom-mendations. Mind. Mach. 28, 689–707. doi: 10.1007/s11023-018-9482-5

Friedman, B., Kahn, P. H., Borning, A., and Huldtgren, A. (2013). "Value sensitive design and information systems" in *Philosophy of engineering and technology* (Dordrecht: Springer), 55–95.

Guan, H., Dong, L., and Zhao, A. (2022). Ethical risk factors and mechanisms in artificial intelligence decision making. *Behavioral Sciences* 12:343. doi: 10.3390/bs12090343

Jesus, S., Saleiro, P., Oliveira e Silva, I., Jorge, B. M., Ribeiro, R. P., Gama, J., et al. (2024). Aequitas flow: streamlining fair ML experimentation. arXiv (Cornell University). arXiv preprint arXiv:2405.05809. doi: 10.48550/arxiv.2405.05809

Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* 1, 389–399. doi: 10.1038/s42256-019-0088-2

Kieslich, K., Keller, B., and Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data Soc.* 9:205395172210929. doi: 10.1177/20539517221092956

Kiran, N., Sapna, F. N. U., Kiran, F. N. U., Kumar, D., Raja, F. N. U., Shiwlani, S., et al. (2023). Digital pathology: transforming diagnosis in the digital age. *Cureus* 15:e44620. doi: 10.7759/cureus.44620

Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., et al. (2024). Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. *R. Soc. Open Sci.* 11:230859. doi: 10.1098/rsos.230859

Lacmanovic, S., and Skare, M. (2025). Artificial intelligence bias auditing – current approaches, challenges and lessons from prac-tice. *Rev. Account. Finance* 24, 375–400. doi: 10.1108/raf-01-2025-0006

Laine, J., Minkkinen, M., and Mäntymäki, M. (2024). Ethics-based AI auditing: a systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Inf. Manag.* 61:103969. doi: 10.1016/j.im.2024.103969

Marr, B. (2023). The 15 biggest risks of artificial intelligence. *Forbes*.

Mayer, H., Yee, L., Chui, M., and Roberts, R. 2025 Superagency in the workplace: Empowering people to unlock AIâs full potential, McKinsey & Company. Available online at: https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Mennella, C., Maniscalco, U., De Pietro, G., and Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: a narrative review. Heliyon 10:e26297. doi: 10.1016/j.heliyon.2024.e26297

Merhi, M. I. (2022). An assessment of the barriers impacting responsible artificial intelligence. *Inf. Syst. Front.* 25, 1147–1160. doi: 10.1007/s10796-022-10276-3

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nat. Mach. Intell. 1, 501–507. doi: 10.1038/s42256-019-0114-4

Mökander, J., and Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds Mach.* 31, 323–327. doi: 10.1007/s11023-021-09557-8

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., and Floridi, L. (2021). Ethics as a service: a pragmatic operationalisation of AI ethics. *Mind. Mach.* 31, 239–256. doi: 10.1007/s11023-021-09563-w

Morley, J., Floridi, L., Kinsey, L., and Elhalal, A. (2019). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics* 26, 2141–2168. doi: 10.1007/s11948-019-00165-5

Nannini, L., Huyskes, D., Panai, E., Pistilli, G., and Tartaro, A. (2024). Nullius in explanans: an ethical risk assessment for explainable AI. *Ethics Inf. Technol.* 27. doi: 10.1007/s10676-024-09800-7

Oladele, I., and Orelaja, A. y Akinwande, O. T., "Ethical implications and governance of artificial intelligence in business Deci-sions: a deep dive into the ethical challenges and governance issues surrounding the use of artificial intelligence in making critical business decisions", *Int. J. Latest Technol. Eng., Manage. & Appl. Sci.*, vol. XIII, pp. 48–56, 2024. doi: 10.51583/ijltemas.2024.130207

Ortega-Bolaños, R., Bernal-Salcedo, J., Germán Ortiz, M., Galeano Sarmiento, J., Ruz, G. A., and Tabares-Soto, R. (2024). Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems. *Artif. Intell. Rev.* 57:110. doi: 10.1007/s10462-024-10740-3

Prajapati, S. B. (2025). Existing challenges in ethical AI. *World J. Adv. Res. Rev.* 25, 2130–2137. doi: 10.30574/wjarr.2025.25.1.0267

Radclyffe, C., Ribeiro, M., and Wortham, R. H. (2023). The assessment list for trustworthy artificial intelligence: a review and recommendations. *Front. Artificial Intell.* 6:1020592. doi: 10.3389/frai.2023.1020592

Romani, G., Avendaño, C., and Santisteban, J., *MESIAS: A web application for evaluating ethical and security considerations in AI project implementation," 2025 Joint International Conference on Digital Arts, Media and Technology With ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON)*, pp. 412–416, 2025

Singla, A., Sukharevsky, A., and Yee, L. y Chui, M. "El estado de la IA a principios de 2024: la adopciÃ3n de la IA generativa aumenta y comienza a generar valor, McKinsey & Company". (2024). Accessed April 23, 2025. [Online]. Available online at: https://www.mckinsey.com/locations/south-america/latam/hispanoamerica-en-potencia/el-estado-de-la-ia-a-principios-de-2024-la-adopcion-de-la-ia-generativa-aumenta-y-comienza-a-generar-valor/es-CL

Taeihagh, A. (2021). Governance of artificial intelligence. *Polic. Soc.* 40, 137–157. doi: 10.1080/14494035.2021.1928377

Trinkley, K. E., An, R., Maw, A. M., Glasgow, R. E., and Brownson, R. C. (2024). Leveraging artificial intelligence to advance im-plementation science: potential opportunities and cautions. *Implement. Sci.* 19:17. doi: 10.1186/s13012-024-01346-y

Tripathi, S., and Rosak-Szyrocka, J. (2024). Impact of artificial intelligence on society. doi: 10.1201/9781032644509

Tsourma, M., Carmeno, N. L., Codagnone, J. A., Mancini, S., Krognos, J., Drosou, A., et al. (2023). User experience of a web-based platform that enables ethical assessment of artificial intelligence in the public sector. *AHFE International.* 70. doi: 10.54941/ahfe1004047

UNESCO, "Perú: Colaboración en inteligencia artificial para la implementación de la metodología de evaluación de prepar-ación," 2024. Available online at: https://www.unesco.org/es/articles/peru-colaboracion-en-inteligencia-artificial-para-la-implementacion-de-la-metodologia-de-evaluacion

Usmani, A., Happonen, J., and Watada, J., "*Human-centered artificial intelligence: designing for user empowerment and ethical considerations," 2023 5th international congress on human-computer interaction, optimization and robotic Ap-plications (HORA)*, Istanbul, Turkiye, 2023, pp. 1–7

Vela, J. M. M. (2022). Retos, riesgos, responsabilidad y regulación de la inteligencia artificial: Un enfoque de seguridad física, lógica, moral y jurídica. Cizur Menor (Navarra), Spain: Aranzadi.

Wankhade, S., Sahni, M., León-Castro, E., and Olazabal-Lugo, M. (2025). Navigating AI ethics: ANN and ANFIS for transparent and accountable project evaluation amidst contesting AI practices and technologies. *Front. Artif. Intell.* 8:1535845. doi: 10.3389/frai.2025.1535845

Winfield, A. F. T., Booth, S., Dennis, L. A., Egawa, T., Hastie, H., Jacobs, N., et al. (2021). IEEE P7001: a proposed standard on transparency. *Front. Robot. AI* 8:665729. doi: 10.3389/frobt.2021.665729

Yaska, M., and Nuhu, B. M. (2024). Assessment of measures of central tendency and dispersion using Likert-type scale. *Deleted J.* 16, 33–45. doi: 10.62154/ajastr.2024.016.010379