



ARTICLE

Ethical-technological governance of generative artificial intelligence: toward a human-centered integrity framework in contexts of digital innovation

Gobernanza ético-tecnológica de la inteligencia artificial generativa: hacia un marco de integridad centrada en el humano en contextos de innovación digital

Wacner Galindez 

Universidad Nacional Experimental de los Llanos Occidentales Ezequiel Zamora, Barinas, Venezuela

Corresponding Author: wacner@unellez.edu.ve

Submitted: 25 April 2026 Accepted: 12 July 2026 Published online: 15 July 2026

Abstract

The expansion of generative artificial intelligence has revealed that the abundance of ethical principles contrasts with the scarcity of verifiable mechanisms to apply them. In response to this gap, the study designs and evaluates an ethical-technological integrity framework that steers digital innovation toward algorithmic transparency and the safeguarding of fundamental rights without inhibiting technical development. Following the design science paradigm, the framework was built from five theory-derived requirements and a document analysis of Regulation (EU) 2024/1689, the NIST framework, and China's provisions on generative AI, coded through a matrix of eight indicators. The results show that none of the three regimes fully guarantees binding human veto or the participation of affected groups. In response, the framework proposes four axes, namely auditability, proactive accountability, participatory bias mitigation, and a risk-segmented escalation protocol. Its demonstration in a social protection public service indicates that the framework is operable and turns ethics into a strategic asset by generating auditable evidence.

Keywords: Generative artificial intelligence, algorithmic governance, technological jurisprudence, digital integrity, human-centered ethics, algorithmic transparency

Resumen

La expansión de la inteligencia artificial generativa ha puesto en evidencia que la abundancia de principios éticos contrasta con la escasez de mecanismos verificables para aplicarlos. Ante esta brecha, el estudio diseña y evalúa un marco de integridad ético-tecnológica que orienta la innovación digital hacia la transparencia algorítmica y la salvaguarda de derechos fundamentales sin inhibir el desarrollo técnico. Siguiendo el paradigma de la ciencia del diseño, el marco se construyó a partir de cinco requisitos derivados de la teoría y de un análisis documental del Reglamento (UE) 2024/1689, del marco del NIST y de las disposiciones chinas sobre IA generativa, codificados mediante una matriz de ocho indicadores. Los resultados muestran que ninguno de los tres regímenes garantiza plenamente el veto humano vinculante ni la participación de los afectados. En respuesta, el marco propone cuatro ejes, que comprenden la auditabilidad, la responsabilidad proactiva, la mitigación participativa de sesgos y un protocolo de escalamiento segmentado por riesgo. Su demostración en un servicio público de protección social indica que el marco es operable y convierte la ética en un activo estratégico al generar evidencias auditables.

Palabras clave: Inteligencia artificial generativa, gobernanza algorítmica, jurisprudencia tecnológica, integridad digital, ética centrada en el humano, transparencia algorítmica

This is an open-access article and is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

Cite as: W. Galindez, "Ethical-technological governance of generative artificial intelligence: toward a human-centered integrity framework in contexts of digital innovation," *CISAI*, vol. 2, no. 2, pp. 63–81, July 2026, doi: 10.64439/cisai.v2i2.46.

1. Introduction

The proliferation of large language models reshaped the architecture of digital innovation during the 2022-2024 period, in a process driven by the sustained growth of the computational resources devoted to training, whose doubling time has shortened markedly over the last decade [1]. This scaling has given rise to capabilities that emerge abruptly and are difficult to anticipate, to the extent that developers themselves acknowledge that the behavior of these systems combines aggregate predictability with specific surprises in concrete tasks [2]. As a consequence, their adoption in strategic sectors, from legal services to education and health, has advanced faster than the understanding of their organizational, ethical, and regulatory implications [3]. This expansion, however, has not been free of negative externalities, among which stand out the generation of false content with a plausible appearance [4], the reproduction of stereotypes present in training corpora [5], and a broader set of risks encompassing misinformation, the exposure of private data, and the harms arising from interaction between humans and conversational systems [6].

In this scenario, a critical asymmetry can be observed between the speed of technical iteration of generative artificial intelligence (GenAI) and the responsiveness of regulatory frameworks. Smuha [7] documented how geopolitical competition shifted from a race to AI toward a race to AI regulation, in which the European Union, the United States, and China deploy divergent normative architectures that can hardly be synchronized with software development cycles. This divergence is also visible in the fragmentation of governance modes, which combine ethics committees, industry self-regulation, technical standards, and legislation with extraterritorial effects, without a common coordinating logic [8]. In the European case, Veale and Zuiderveen [9] warned, when analyzing the draft AI Act, that the risk-based categorization had gaps regarding general-purpose models, whose versatility hinders a static classification. With the adoption of Regulation (EU) 2024/1689, this tension did not disappear, since the notion of trustworthiness underpinning the regulation tends to be conflated with the acceptability of risk, which shifts decisions of a markedly political nature onto providers [10].

Faced with this problem, the specialized literature has identified the absence of comprehensive frameworks that articulate technological jurisprudence, ethics by design, and algorithmic governance into a coherent system applicable to GenAI. Jobin et al. [11] mapped eighty-four AI ethics documents and found that most converge on principles such as transparency, justice, and non-maleficence, although they differ in their interpretation and rarely specify how to apply them. Along the same lines, Mittelstadt [12] argued that principles alone do not guarantee ethical AI unless they are linked to accountability mechanisms and verification processes, and Hagendorff [13] showed that most current guidelines lack reinforcement mechanisms capable of changing development practices. This distance between normative abstraction and technical concreteness, which Morley et al. [14] characterized as the pending transition from what to how, constitutes the specific gap addressed by the present study.

Accordingly, the general objective of the research is to design and evaluate an ethical-technological integrity framework for GenAI that channels digital innovation under principles of technological jurisprudence, algorithmic transparency, and the safeguarding of fundamental rights, without generating inhibitory effects on technical development. This objective is broken down into four specific aims, namely to systematize the evolution of algorithmic accountability in the face of models with operational autonomy, to identify the ethics-by-design mechanisms applicable to the GenAI life cycle, to build an analytical matrix integrating transparency, accountability, bias mitigation, and human-AI interaction, and to discuss, on an exploratory basis, the implications of the framework for competitiveness and trust in the digital market. The first three aims translate into evaluable design requirements, whereas the fourth is addressed argumentatively in the discussion and remains open to subsequent empirical validation. The distinction between guiding principles and implementable practices proposed by Floridi and Cowls [15] provides the conceptual foundation for this operationalization, which is approached as original design science research [16], since the framework is conceived as an artifact that is built, demonstrated, and evaluated against explicit requirements, rather than as a synthesis of the existing literature.

Building on these aims, the work makes three contributions. The first is an eight-indicator evaluation matrix that makes it possible to compare heterogeneous regulatory regimes in a common language and that, when applied to the three reference frameworks, reveals shared gaps in binding human veto and in

the participation of affected groups. The second is a four-axis integrity framework that translates each identified gap into operational mechanisms and third-party verifiable evidence, and the third is a risk-segmented human escalation protocol, formalized as an algorithm and as a sequence diagram, which seeks to reconcile effective oversight with the operational scale of these systems, and whose applicability is demonstrated in a public-service scenario and evaluated against five design requirements. The article is organized as follows: Section 2 examines related work, Section 3 sets out the theoretical foundations, Section 4 describes the methodological design, Section 5 presents the results and the proposed framework, and Sections 6 and 7 discuss its implications and conclusions.

2. Related work

The literature on AI governance has advanced along several paths that, although complementary, rarely converge in a single proposal. A first path has focused on translating ethical principles into practical tools, and within it Morley et al. [14] reviewed the instruments available for each phase of the machine-learning life cycle, concluding that most are at early stages of validation and that their distribution across phases is uneven. This diagnosis, while valuable for guiding developers, predates the emergence of generative models and does not consider the challenges associated with general-purpose systems whose end use escapes the control of those who train them.

A second, legally oriented path has examined how existing legal orders respond to GenAI. Hacker et al. [17] analyzed the regulation of large generative models in the European context and proposed distinguishing between developers, deployers, and users in order to distribute obligations proportionately, while Laux et al. [10] questioned the European regulation's equating of trustworthiness with the acceptability of risk. For their part, Veale et al. [8] broadened the focus toward global governance, identifying the coexistence of regulatory modalities that respond to different rationales. These works provide rigorous legal analysis, although their scope stops at normative critique and does not reach down to the technical mechanisms that would make the proposed obligations enforceable.

From a third path, centered on auditing, Mökander et al. [18] formulated a three-layered approach, distinguishing governance, model, and application audits, as a mechanism for identifying and managing the risks of language models. This proposal represents a significant operational advance, but its emphasis lies on verification rather than on the allocation of responsibilities or the design of human oversight during operation. In a complementary way, Novelli et al. [19] clarified the concept of accountability in AI and its relational components, while Papagiannidis et al. [20] organized responsible governance into structural, relational, and procedural dimensions from an information-systems perspective. Both contributions enrich the conceptual vocabulary of the field, although they are not specifically oriented to GenAI and do not offer a comparison between regulatory regimes. Table 1 summarizes this analysis on the basis of five criteria that, taken together, define the space the present study seeks to occupy. As can be seen, none of the works examined combines specificity regarding GenAI, legal grounding, technical operationalization, comparison across regimes, and a binding human oversight mechanism, which justifies the need for a framework that articulates these dimensions within a single structure.

Table 1. Comparison of related work according to the criteria that define the contribution of the study

Work	GenAI-specific	Legal grounding	Operationa-lization	Regulatory comparison	Human veto
Morley et al. [14]	–	–	✓	–	–
Hacker et al. [17]	✓	✓	◦	–	–
Veale et al. [8]	–	✓	–	✓	–
Laux et al. [10]	–	✓	–	–	–
Novelli et al. [19]	–	◦	◦	–	–
Mökander et al. [18]	✓	◦	✓	–	◦
Papagiannidis et al. [20]	–	–	◦	–	–
Present study	✓	✓	✓	✓	✓

✓ addressed centrally ◦ partially addressed – not addressed

3. Theoretical framework: the governance triad

The background reviewed shows that GenAI governance rests on three pillars whose interaction, rather than their isolated consideration, determines the effectiveness of any framework. For this reason, this section successively examines technological jurisprudence, ethics by design, and the relationship between governance and digital innovation, in order to then contrast the regulatory architectures that embody these pillars in three jurisdictions.

3.1 Technological jurisprudence and GenAI

It is first necessary to clarify the scope with which this expression is used. In the Spanish-speaking legal tradition, the term *jurisprudencia* usually designates the body of judicial decisions that establish interpretive criteria. In this study, by contrast, the expression technological jurisprudence is used in a broader sense, close to that of technology law, to refer to the set of norms, doctrine, and interpretive criteria governing the attribution of responsibility and the protection of rights in relation to digital technologies. Consequently, the analysis is not based on court rulings but on normative texts and the doctrine that interprets them, and the pillar is understood as the legal dimension of GenAI governance.

The evolution of law in the face of the intangible has moved from a conception of responsibility centered on the programmer toward a paradigm that must address the operational autonomy of deep-learning models. Matthias [21] anticipated that systems capable of learning would generate a responsibility gap, since neither the manufacturer nor the operator fully controls their behavior, and Santoni de Sio and Mecacci [22] later specified that this gap manifests itself in four distinct forms, relating to culpability, moral accountability, public accountability, and active responsibility. In the field of data protection, Wachter et al. [23] argued that the General Data Protection Regulation does not guarantee a general right to an explanation of the internal logic of algorithms, but only to information about the existence of automated processing and its consequences, whereas Kaminski [24] offered a broader reading, according to which the regulation does impose an obligation of meaningful explanation that operates in combination with other safeguards.

This insufficiency is magnified in GenAI, where the distance between training data, model architecture, and generated content hinders the causal attribution of harms. Bryson [25] proposed that intelligent systems may be objects of consideration without this implying that they be granted moral agency, which requires keeping the chain of responsibility within the human actors of the ecosystem, that is, developers, deployers, and end users. This perspective is complemented by the finding of Cath et al. [26], who compared the approaches of the United States, the European Union, and the United Kingdom and found that jurisdictional fragmentation creates gray areas when models operate transnationally. Hence, the central legal question is no longer only who is liable, but how to distribute the burden of proof when a foundation model produces defamatory or discriminatory outputs in contexts of use not foreseen by its original developers, an uncertainty that Hacker et al. [17] identify as a systemic risk factor for investment in innovation.

3.2 Ethics by design

The second pillar concerns the integration of ethical principles into the development life cycle. The AI4People framework [27] synthesized previous proposals into five principles, beneficence, non-maleficence, autonomy, justice, and explicability, which later inspired the requirements for trustworthy AI adopted in the European sphere. Nevertheless, Schiff et al. [28] observed, when reviewing documents from the public, private, and non-governmental sectors, that there is rhetorical convergence accompanied by substantial divergence in implementation, and Floridi [29] warned of risks such as ethics washing, which consists of projecting an image of responsibility without effective changes in practice. When these considerations are transferred to GenAI, additional structural limitations arise. Alignment mechanisms based on human feedback constitute, strictly speaking, a form of ethics by design, although their effectiveness depends on the representativeness of the judgments incorporated and on the conception of value adopted in each communicative context [30]. Moreover, the emergence of unanticipated capabilities during training [2] implies that no finite set of coded rules can foresee all the behaviors of the system. Consequently, ethics by design in GenAI requires a reconceptualization that goes beyond rule coding and addresses the dynamics of continuous learning and deployment of the model, which leads naturally to

the relationship between governance and innovation.

3.3 Digital innovation and governance

Following on from the above, the third dimension examines whether governance acts as a brake on innovation or as an enabling condition for it. Zuboff [31] characterized the logic of surveillance capitalism as a model in which data accumulation takes precedence over the rights of the people affected, which illustrates how the absence of rules can steer innovation toward extraction rather than the creation of shared value. Birhane [32] complemented this critique by showing that systems deployed without attention to their relational context perpetuate injustices that erode the social legitimacy of technology. Conversely, the literature on organizational governance holds that responsible governance structures and processes allow organizations to capture the value of AI while managing its risks [20], and that such governance still requires a research agenda connecting principles with business practices [33]. This view must be qualified, however, by acknowledging inherent tensions. Whittaker [34] documented how the technology industry has partially captured AI ethics research, with the effect of neutralizing regulatory pressure, and Taddeo et al. [35] warned that trusting AI for security purposes is a double-edged sword, since the same systems can be attacked or manipulated. Therefore, any governance framework that aspires to function as trust infrastructure must incorporate independent verification mechanisms, since only in this way can ethics be prevented from being reduced to a public-relations strategy.

4. Methodology

The study is framed within design science research, a paradigm oriented toward creating and evaluating artifacts that solve relevant problems while also providing prescriptive knowledge about how to solve them [16, 36]. Under this approach, the integrity framework constitutes the research artifact, so its contribution does not lie in synthesizing the published literature but in proposing a new solution, grounded in kernel theories and primary documentary evidence, whose utility is explicitly demonstrated and evaluated. Following the methodology of Peffers et al. [37], the process was organized into six activities that Figure 1 links to the sections of the article, and in which the literature serves as kernel theory for deriving the design requirements, while the normative texts constitute the primary data of the analysis.

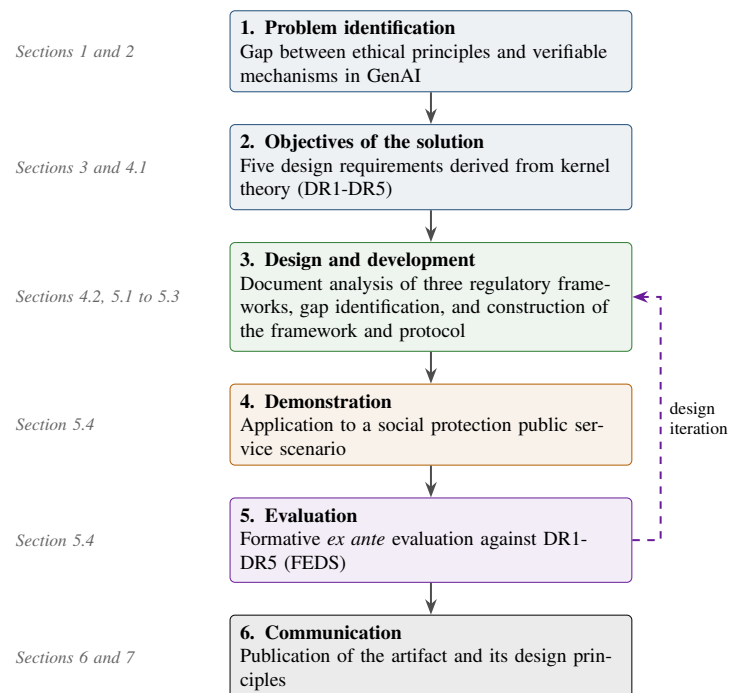


Figura 1. Design science research process followed in the study, adapted from Peffers et al. [37].

4.1 Problem identification and design requirements

The problem motivating the artifact, described in Sections 1 and 2, is that GenAI governance has widely shared principles but lacks mechanisms for verifying compliance with them and assigning responsibility when they fail [12, 14]. In order to transform this problem into solution objectives, five design requirements were derived from kernel theory that any satisfactory response should meet and that, in turn, serve as evaluation criteria for the artifact. Table 2 presents these requirements together with their theoretical foundation.

Tabla 2. Design requirements of the artifact derived from kernel theory

Code	Requirement	Statement	Foundation
DR1	Auditability	Mechanisms must produce evidence verifiable by third parties, beyond the disclosure of information	[38, 12]
DR2	<i>Ex ante</i> allocation of responsibility	Roles and obligations must be defined before deployment, even in the face of unanticipated capabilities	[21, 22, 2]
DR3	Participation of affected groups	Affected groups must take part in defining harm and validating fairness metrics	[39, 40]
DR4	Effective human oversight	The operator must have authority, competence, and a workload compatible with real control	[41, 42]
DR5	Proportionality	The intensity of controls must be adjusted to risk so as not to inhibit innovation	[43, 20]

4.2 Document analysis of regulatory frameworks

The design activity was supported by a qualitative document analysis [44] of three regulatory frameworks, which constitute the primary data of the study. The analysis covered Regulation (EU) 2024/1689 [45], the NIST AI Risk Management Framework in its 2023 version 1.0 [46], and the set of Chinese provisions on recommendation algorithms, deep synthesis, and generative AI services adopted between 2021 and 2023. Unlike the other two instruments, the NIST framework is voluntary, and it was selected because it is the most widely used federal technical reference in the United States in the absence of a general federal AI law. This asymmetry is deliberately assumed, since the rating scale distinguishes between recommended and binding provisions, so that the lower US score partly reflects the voluntary nature of the instrument and not only its content. For the same reason, the generative AI profile published by NIST in 2024 as a companion to the framework [47], which is also voluntary, would not alter the maximum level attainable by the United States on any indicator, although it could enrich the content of the recommended provisions, an aspect included among the limitations of the study. The Chinese provisions were analyzed on the basis of the unofficial English translations published by China Law Translate, without cross-checking against the original Chinese text, a circumstance also included among the limitations of the study.

The analytical instrument was an *ad hoc* matrix built from the four axes of the framework, each broken down into two indicators aligned with the design requirements, as detailed in Table 3. Each indicator was rated on a three-level ordinal scale, in which a value of 0 indicates the absence of a normative provision, a value of 1 indicates a partial or recommended provision or one limited to certain actors, and a value of 2 indicates a full, binding, and verifiable provision. Thus, each axis admits a score between 0 and 4, and each framework an overall score between 0 and 16.

Tabla 3. Operationalization of the document analysis matrix

Axis	Code	Indicator	Criterion for the full level (2)
I. Transparency and explainability	T1	Technical documentation	Mandatory documentation of the model, the data, and their versions, accessible to the competent authority
	T2	Explainability to affected persons	Enforceable right of affected persons to receive meaningful information about the output
II. Proactive accountability	R1	Role delimitation	Differentiated obligations for provider, deployer, and user along the value chain
	R2	<i>Ex ante</i> mechanisms	Pre-deployment due diligence with contingency plans and incident-reporting channels
III. Bias mitigation	S1	Prior assessment	Mandatory assessment of bias and risks before deployment
	S2	Monitoring and participation	Post-deployment monitoring with the participation of affected groups
IV. Human-AI	H1	Human oversight	Mandatory human oversight proportionate to risk
	H2	Binding veto	Effective authority of the operator to override the output, with a defined escalation protocol

The coding was carried out entirely by the author, since this is a single-author study, so it was not possible to estimate inter-coder agreement. To mitigate the risk of subjectivity that this entails, three safeguards were adopted. First, each level of the scale was defined through explicit criteria established before coding (Table 3), so that ratings would rest on rules rather than on general impressions. Second, each score was linked to the specific provisions supporting it (Table 5), which allows any reader to verify or challenge the rating assigned. Third, the complete matrix and the coded normative excerpts are available upon request to facilitate independent replication. Subsequently, the coded excerpts were subjected to a thematic analysis [48] aimed at identifying recurring patterns, contradictions, and shared gaps, which constituted the direct input for the design of the framework's mechanisms.

4.3 Demonstration and evaluation strategy

Once built, the artifact was demonstrated through its application to an illustrative public social-protection service scenario, chosen because it combines low-impact outputs and rights-affecting decisions within the same system, and because public administration is one of the areas where human oversight requirements have shown the greatest limitations [41]. For the evaluation, a formative, *ex ante*, and artificial strategy was adopted within the FEDS framework [49], appropriate for conceptual artifacts in their first iteration, which was implemented through the descriptive methods of informed argument and scenario analysis proposed by Hevner et al. [16]. In this way, each design requirement was contrasted with the framework's mechanisms and with the evidence they generate in the scenario, while naturalistic evaluation in real organizations is reserved for a later iteration. Since the research did not involve human subjects or personal data, it did not require approval from an institutional ethics committee. Even so, it should be made explicit from the outset that the analysis was limited to documents available in Spanish and English and to three predominant jurisdictions, which omits perspectives from the Global South, and that there is a possible Western bias in the selection of kernel theory, which was mitigated by deliberately incorporating critical perspectives into the discussion.

5. Results

This section follows the same order as the design process and successively answers four questions. The first is what shortcomings current regulatory frameworks present (Section 5.1), the second is how the proposed framework responds to those shortcomings (Section 5.2), the third is what exactly it contributes compared with what already exists (Section 5.3), and the fourth is whether it works when applied to a concrete case (Section 5.4). In this way, each subsection builds on the previous one and prepares the next, which makes it possible to follow the reasoning from diagnosis to application.

5.1 Gaps in current regulatory frameworks

The starting point consisted of examining, with the matrix described in Section 4.2, the extent to which the regulations of the European Union, the United States, and China provide for the eight mechanisms that the study considers necessary for GenAI with integrity. Each mechanism received 0 points if the regulation does not provide for it, 1 point if it provides for it partially or only as a recommendation, and 2 points if it requires it in a mandatory and verifiable manner. Table 4 presents the result, and Table 5 lists the provisions supporting each score, so that the rating can be replicated or discussed by other researchers.

Table 4. Evaluation matrix scores by indicator and regulatory framework

Indicator	EU	US	China
T1 Technical documentation	2	1	1
T2 Explainability to affected persons	1	1	0
R1 Role delimitation	2	1	1
R2 <i>Ex ante</i> mechanisms	1	1	1
S1 Prior assessment	2	1	1
S2 Monitoring and participation	1	1	0
H1 Human oversight	2	1	1
H2 Binding veto	1	0	0
Total (0-16)	12	7	5

0 = absence, 1 = partial provision, 2 = full and binding provision.

Table 5. Normative provisions supporting the scores in Table 4

Indicator	EU (Regulation 2024/1689)	US (NIST AI RMF 1.0)	China (2021-2023 provisions)
T1 Technical documentation	(2) Art. 11 and Annex IV; Art. 53 for general-purpose models	(1) Documentation recommended in the MAP and MEASURE functions	(1) Algorithm filing and explanations of training data upon request of the authority
T2 Explainability to affected persons	(1) Art. 86, limited to decisions based on Annex III high-risk systems; Art. 50	(1) Characteristics of explainable, interpretable, and transparent systems, on a recommended basis	(0) General information about the service, with no right to an explanation of the individual output
R1 Role delimitation	(2) Arts. 16, 25, and 26 (provider, value chain, and deployer)	(1) GOVERN 2.1 and GOVERN 6 (roles and third parties, recommended)	(1) Obligations concentrated on the service provider, without differentiation along the chain
R2 <i>Ex ante</i> mechanisms	(1) Arts. 9, 20, and 73, limited to high risk and without a compensation fund	(1) MANAGE 2.4 and MANAGE 4 (recommended)	(1) Prior security assessment and complaint mechanisms, oriented toward the authority
S1 Prior assessment	(2) Art. 10(2), points (f) and (g), and Art. 9	(1) MEASURE 2.11 (recommended)	(1) General obligation to prevent discrimination in design and data
S2 Monitoring and participation	(1) Arts. 72 and 27, without binding participation of affected persons	(1) GOVERN 5.1 and MANAGE 4.1 (recommended)	(0) Monitoring oriented toward content control, without participation of affected persons
H1 Human oversight	(2) Arts. 14 and 26(2)	(1) GOVERN 3.2 (recommended)	(1) Content review required of the provider, oriented toward state control
H2 Binding veto	(1) Art. 14(4), points (d) and (e), without a defined escalation protocol	(0) No provision for an operator veto over individual outputs	(0) No provision

The results reveal three clearly differentiated regulatory profiles. The European framework achieves the greatest coverage, with 12 out of 16 points, because it makes mandatory the technical documentation, the division of duties between those who develop and those who use the system, the pre-deployment assessment, and human oversight in high-risk uses. However, it loses points precisely where the regulation should hand decision-making capacity over to people, that is, in the right of affected persons to receive explanations, in their participation during monitoring, and in the operator's power to override an output, which is consistent with the critique of Laux et al. [10]. The US framework, with 7 points, recommends almost all the mechanisms but mandates none of them and does not provide for human veto. The Chinese regime, with 5 points, concentrates its coverage on state control and lacks provisions oriented toward affected persons, in line with the priority it gives to social stability [50]. Figure 2 groups these values by axis and shows that human-AI interaction is the axis with the lowest aggregate score across the three regimes, with 5 out of 12 possible points, compared with 6 for transparency and bias and 7 for accountability.

Moreover, its binding veto indicator (H2) is the least covered in the entire matrix, with 1 out of 6 possible points, since only the European framework provides for it, and only partially.

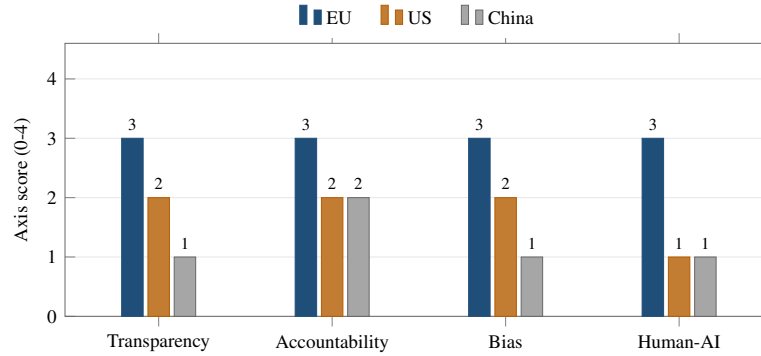


Figure 2. Normative coverage profile by axis in the three regulatory frameworks analyzed

Beyond these differences in intensity, however, the analysis revealed four shortcomings common to the three regimes. First, transparency is directed at the authority and does not allow a third party to reconstruct how a specific output was produced. Second, responsibility is assigned by type of actor, but nothing establishes what to do if the model develops unforeseen behaviors. Third, bias is treated as a technical problem in which affected communities do not take part, and, finally, human oversight is required in generic terms, without ensuring that the operator has the authority, time, and training to reject a decision. These results therefore indicate that current regulations describe what must be done but do not ensure that it can be verified that it was done, and it is precisely this verification gap that gives rise to the framework presented below.

5.2 Integrity framework as a response to the gaps

This subsection presents the proposed integrity framework, whose purpose is to respond to the four gaps identified in the previous analysis. To this end, the framework transforms each gap into an axis of action and organizes the four axes around a common core, human-centered integrity, which is in turn supported by the three theoretical pillars set out in Section 3, as illustrated in Figure 3. The logic that articulates the whole is simple, since each axis incorporates concrete mechanisms and each mechanism generates evidence that a third party can verify, so that the integrity of the system ceases to be a declaration of good intentions and becomes a verifiable property. Likewise, the dashed arrows in the figure show that the axes feed back into one another, as occurs when a case escalated by an operator reveals a bias that requires the training data to be reviewed.

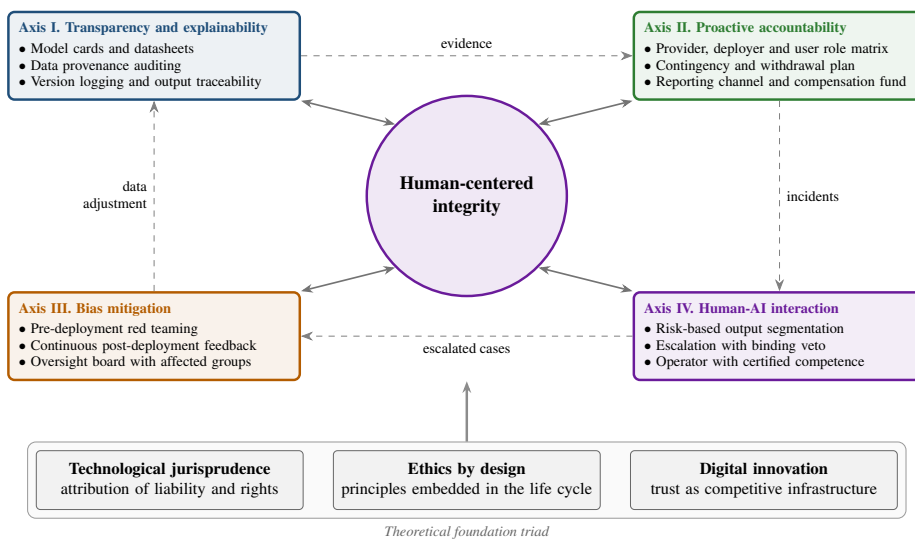


Figure 3. Ethical-technological integrity framework for generative AI.

For ease of reading, each axis is described below following the same path, which starts from the problem it addresses, continues with what the framework proposes, and closes with the support underpinning the proposal.

5.2.1 Axis I: algorithmic transparency and explainability

The first axis starts from the observation that publishing information about a system is not equivalent to being able to understand it or to hold anyone accountable for it [38], and in language models this difficulty is aggravated because they generate text from statistical patterns without understanding its meaning [5]. In response to this problem, the axis replaces mere disclosure with auditability, understood as the possibility for a third party to reconstruct how an output was obtained. To this end, it requires three instruments, namely a model card describing its performance, limitations, and intended uses [51], a datasheet documenting the origin and composition of the training data [52], and a version log linking each output to the model and the prompt that produced it. This explainability by design draws on explainable AI techniques [53] and responds to the legal obligation to provide meaningful information about automated decisions that Kaminski [24] identifies in European law.

5.2.2 Axis II: proactive accountability

The second axis deals with a different problem. Several actors take part in the GenAI chain, including those who train the model, those who adapt it, and those who use it, and when harm occurs, responsibility tends to be diluted among them. Moreover, these models may develop capabilities that no one anticipated during training [2]. To prevent this, the axis defines responsibilities before deployment rather than after the harm. Following Hacker et al. [17], risks originating in training correspond to the model provider, and those arising from its adaptation or use correspond to the deployer. This role matrix is complemented by a contingency plan, required every time a model is released, which specifies how to withdraw it, through which channel harms are reported, and with which fund affected persons are compensated. This proposal is justified because the early allocation of responsibilities throughout the life cycle prevents their dispersion [54], the decision on how to release a model is itself a governance decision [43], and external audits ensure that evaluation does not remain in the hands of those who benefit from it [55, 19].

5.2.3 Axis III: bias mitigation and fundamental rights

The third axis recognizes that no algorithm is neutral, since the data from which it learns reproduce existing inequalities [56, 57], and many studies on bias do not even define what harm they intend to measure [39]. To address this problem, the axis combines technical and social measures. On the technical side, it incorporates adversarial testing, which consists of subjecting the system to tests designed to elicit harmful responses before putting it into operation [58], as well as data provenance auditing and the continuous collection of user reports. On the social side, it creates an oversight board with representatives of affected groups, responsible for deciding what is considered an unfair output and how it is measured. The social dimension is indispensable, given that risks such as misinformation or stereotypes are not resolved by technical adjustments alone [6], and the participation of affected persons is only effective when it grants them real decision-making power and is not limited to formal consultation [40, 59].

5.2.4 Axis IV: human-AI interaction

The fourth axis addresses an idea that is often taken for granted, since including a person in the process does not guarantee that this person controls the system. Green [41] showed that human overseers are often unable to perform the function assigned to them, and the literature on automation bias indicates that people tend to accept machine recommendations without verifying them, especially when overloaded with work [60, 42]. For oversight to be real, the axis establishes an escalation protocol that classifies each output according to its risk to people. Low-risk outputs are delivered directly and a sample is reviewed afterward, while intermediate-risk outputs are delivered with a warning and reviewed as a priority. High-risk outputs, as well as decisions that by their nature can never be delegated to the machine, require the prior approval of a certified operator whose veto is binding, that is, one the system cannot ignore. In this way, human oversight is concentrated where it really matters and is not diluted across thousands of trivial responses. Algorithm 1 formalizes the procedure, in which the thresholds θ_l and θ_h delimit the low, intermediate,

and high risk levels. The ASSESSRISK function can be implemented through explicit rules or through a classification model, and both its criteria and the thresholds must be set in each context of use, which is why Section 5.4 offers an indicative example of its configuration. In this way, the protocol embodies the idea of human-centered AI, which combines high automation with high human control [61], and follows the logic of the layered auditing proposed by Mökander et al. [18].

Algorithm 1 Risk-segmented human escalation protocol

Require: Candidate output y generated for prompt x , context of use u , thresholds $\theta_l < \theta_h$, sampling rate ρ

Ensure: Final decision d and auditable record ℓ

```

1:  $r \leftarrow \text{ASSESSRISK}(x, y, u)$  ▷ Rights-impact classifier
2: if  $u$  belongs to a non-delegable decision category or  $r \geq \theta_h$  then
3:    $v \leftarrow \text{VALIDATEBYOPERATOR}(y, \text{justification})$  ▷ Mandatory prior validation
4:   if  $v = \text{veto}$  then
5:      $d \leftarrow \text{rejection}$  ▷ Human veto is binding
6:      $\text{NOTIFYBOARD}(x, y, \text{reason})$ 
7:   else
8:      $d \leftarrow y$ 
9:   end if
10: else if  $\theta_l \leq r < \theta_h$  then
11:    $d \leftarrow y$  with a warning to the user, flagged for priority review
12: else
13:    $d \leftarrow y$ 
14:   Include  $(x, y)$  in the post-hoc review sample with probability  $\rho$ 
15: end if
16:  $\ell \leftarrow \text{RECORD}(x, y, r, d, \text{operator}, t)$ 
17: return  $(d, \ell)$ 
  
```

Figure 4 shows the same protocol from the perspective of the actors involved. The upper box traces the path of a high-risk output, which passes through the operator before reaching the user, and the lower box that of a low-risk output, which is delivered immediately and recorded for review. The figure also shows that the protocol connects the four axes, since the record feeds transparency (Axis I), the notification to the board triggers accountability and bias review (Axes II and III), and the veto makes human oversight effective (Axis IV).

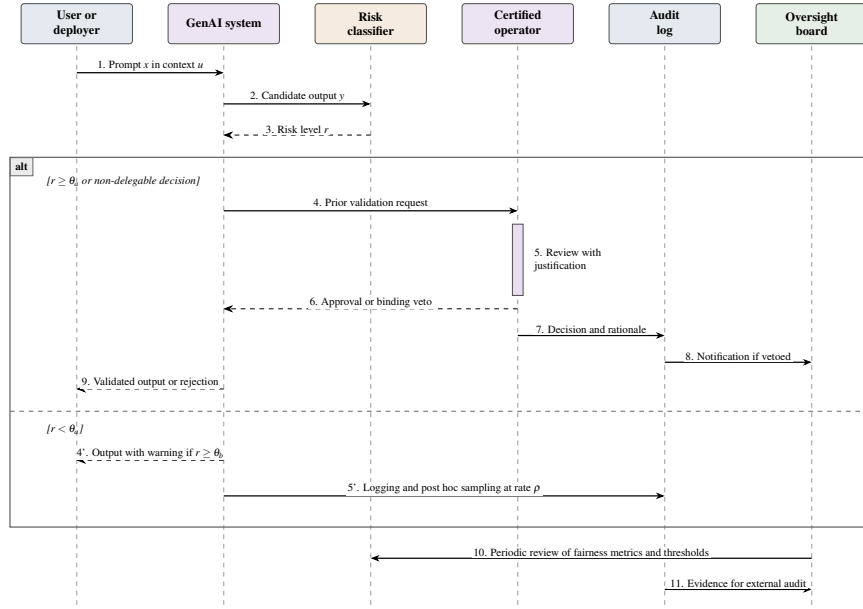


Figure 4. Sequence diagram of the human escalation protocol.

The foregoing thus highlights that the four axes do not function as isolated pieces but as a system in which the evidence generated by one feeds the others, a feature that is central to assessing the contribution of the framework.

5.3 Contribution of the framework compared with existing approaches

Having described the axes and their articulation, it is worth making the contribution of the framework explicit. Table 6 reads from left to right the complete path of the study, which goes from each gap detected in the regulations, through the mechanism the framework proposes to cover it, to the concrete evidence that makes it possible to verify compliance.

Table 6. Correspondence between identified gaps, framework mechanisms, and auditable evidence

Identified gap	Axis	Framework mechanism	Verifiable evidence
Transparency oriented toward the authority rather than toward output traceability	I	Explainability by design with version logging and provenance auditing	Model cards and datasheets, version log, and reproducible output trace
Accountability without prior mechanisms for unforeseen capabilities	II	Role matrix and contingency plan associated with each release decision	Withdrawal protocol, incident channel, and established compensation fund
Bias mitigation reduced to a technical assessment	III	Adversarial testing and oversight board with affected groups	Adversarial testing reports, board minutes, and validated fairness metrics
Generic human oversight without effective authority	IV	Risk-segmented escalation protocol with binding veto	Escalation log, veto rate, and operator certification

As Table 6 shows, the contribution does not consist of adding new ethical principles to an already long list, but of linking each gap to a mechanism and each mechanism to verifiable proof, which allows regulators, auditors, and developers to verify compliance without relying on declarations of good faith.

5.4 Application of the framework to a concrete case

To verify that the framework can be put into practice, it was applied to an illustrative scenario. A public social-protection agency uses a GenAI assistant, built on a model from another company, for two tasks. The first consists of answering general inquiries about requirements and deadlines, and the second of drafting resolutions on benefit applications, which are then signed by an official. The case is demanding because it combines low-impact responses and decisions affecting social rights within the same system. In this scenario, Axis II is applied first, because it defines who is responsible for what. The model provider is responsible for documenting training, the technology integrator for fine-tuning and safety filters, and the agency for deciding which decisions cannot be delegated and for appointing the operators, while a contingency plan provides for suspending the drafting of resolutions if discriminatory patterns are detected. On this basis, Axis I requires each draft to be linked to the model version and the prompt that originated it, so that any contested resolution can be reconstructed. Axis III, in turn, subjects the system before its launch to tests with simulated applicants from different territories, literacy levels, and linguistic varieties, and during its use a board with beneficiary representatives reviews, for example, whether proposed denials are concentrated in any group. Finally, Axis IV establishes that every denial resolution requires the prior approval of the official, who may veto it, that complex personal inquiries are delivered with a warning, and that general inquiries are reviewed by sampling, at a rate that could illustratively be set at 5 %.

To operationalize the AssessRisk function of Algorithm 1, the scenario proposes, on an indicative basis, a rule-based classifier that assigns each output a score $r \in [0, 1]$ on the basis of three criteria, namely the type of task (general inquiry or draft resolution), the presence of the applicant's personal data, and the potential effect of the output on access to a benefit. With illustrative thresholds of $\theta_l = 0.3$ and $\theta_h = 0.7$, general inquiries without personal data fall into the low level, complex personal inquiries into the intermediate level, and draft resolutions into the high level, while denial resolutions are treated as a non-delegable category regardless of their score. The final calibration of these values is the responsibility of the agency, with validation by the oversight board, and must be reviewed periodically on the basis of the escalation log, as provided for in step 10 of Figure 4.

With this scenario as a reference, it was then verified whether the framework meets the five requirements defined in Section 4.1. As summarized in Table 7, two requirements are fully satisfied, audibility and responsibility defined in advance, and three are conditionally satisfied, because their fulfillment depends

on factors that the framework can enable but not guarantee by itself, such as the board being truly representative, operators not being overloaded, and the organization having resources for documentation.

Tabla 7. Analytical evaluation of the artifact against the design requirements in the demonstration scenario

Requirement	Framework mechanism	Evidence in the scenario	Assessment
DR1 Auditability	Model cards and datasheets, version log, and traces	Each draft resolution can be reconstructed from the version and prompt that originated it	Satisfied
DR2 <i>Ex ante</i> responsibility	Role matrix and contingency plan	Obligations of the provider, integrator, and agency defined before deployment, with a rollback protocol	Satisfied
DR3 Participation of affected groups	Oversight board with affected groups	Validation of denial metrics by population group	Conditional on the board's representativeness
DR4 Effective oversight	Escalation protocol with binding veto	Denial resolutions validated by an official with authority to override them	Conditional on the operator's workload
DR5 Proportionality	Segmentation by risk levels	General inquiries supervised by sampling and rights-related decisions with prior validation	Conditional on the organization's capacity

These results allow us to conclude that the framework is applicable in a high-impact context and covers the verification and accountability requirements by design, whereas participation, effective oversight, and proportionality also require organizational conditions that will need to be verified during implementation, an issue taken up in the discussion.

6. Discussion

The results confirm that the gap between principles and practices described by Mittelstadt [12] and by Morley et al. [14] persists in the most influential regulatory regimes, including the European one, which obtained the greatest coverage. This finding qualifies the reading according to which the approval of Regulation (EU) 2024/1689 would have solved the problem of operationalization, since the matrix reveals that its strengths are concentrated in formal documentation and assessment obligations, while its weaknesses appear precisely in the indicators that require transferring decision-making capacity to affected persons or human operators. Thus, the results converge with the critique of Laux et al. [10] and refine it, as they identify the specific mechanisms in which declared trustworthiness does not translate into enforceable safeguards.

In terms of the typology of Gregor and Hevner [36], the contribution falls into the improvement category, since it offers a new solution to an already recognized problem, and its value lies both in the artifact and in the design principles underpinning it, which can be transferred to other contexts. From this perspective, the proposed framework engages with contributions on language-model auditing. Mökander et al. [18] showed that coordinated three-layered audits make it possible to identify and manage risks, and the present study extends this logic to operation time, since the escalation protocol generates during use the evidence that auditors subsequently examine. Similarly, the incorporation of oversight boards responds to the warning of Birhane et al. [40] about the risk of merely symbolic participation, by linking these boards to concrete decisions such as validating metrics and reviewing thresholds. For its part, the design of the protocol takes seriously the evidence of Green [41] and Goddard et al. [60], since it does not assume that human presence is sufficient, but rather concentrates oversight where risk justifies it and provides the operator with the authority and competence to exercise it.

As regards the fourth specific aim, concerning the implications for digital innovation, its treatment is exploratory, since it was not translated into an evaluable design requirement in this iteration. With this caveat, the framework can operate as a mechanism for preventing reputational and legal risks that reduces the uncertainty associated with GenAI projects. Following the perspective of responsible governance as an organizational capability [20], the availability of auditable evidence allows organizations to demonstrate their diligence to users, investors, and regulators, which turns ethics into a source of competitive trust. At the same time, the framework addresses the tension between security and technological sovereignty that

Smuha [7] and Veale et al. [8] describe as regulatory fragmentation, insofar as its indicators offer a common language that allows each jurisdiction to adjust risk thresholds to its sociopolitical context without giving up the interoperability of digital markets. The empirical verification of these effects, for example through case studies with organizations that adopt the framework, remains a line of future work.

It should also be acknowledged that there is a dimension the framework integrates only partially. Strubell et al. [62] showed that training large-scale models entails considerable energy and environmental costs, and Cowls et al. [63] warned that the potential of AI to tackle climate change must be weighed against its own footprint. Although the framework requires these costs to be documented in model cards, the full incorporation of sustainability as a fifth axis requires specific research that goes beyond the scope of this study. Finally, the study has limitations that delimit its scope. The evaluation carried out is formative and artificial, so it demonstrates the coherence and operability of the artifact in an illustrative scenario, but not its effectiveness in real organizations, which only a naturalistic evaluation will be able to establish [49]. In addition, the matrix scores reflect the coding of normative texts and not their effective application, so a regime with high formal coverage could show lower actual compliance, and the analysis is limited to three jurisdictions, which leaves out the perspectives of the Global South. Added to this, the coding was carried out by a single researcher, without the possibility of estimating inter-coder agreement, a limitation that was offset by explicit criteria, traceability of each score, and availability of the matrix upon request; the comparison combines binding instruments with a voluntary framework, whose score is bounded by its very nature; the NIST generative AI profile [47] was not coded independently; and the Chinese provisions were analyzed through unofficial translations, which may introduce interpretive nuances. In relation to the Global South, the framework assumes an institutional auditing capacity that many Latin American countries do not yet have, which could generate exclusionary effects if it is not accompanied by international technical cooperation mechanisms. Likewise, the critique by Hagendorff [13] and Floridi [29] regarding the risk of reducing moral deliberation to a checklist should be considered, a risk that the framework seeks to mitigate by requiring substantive evidence and the participation of affected persons, although it does not eliminate it completely. These limitations do not invalidate the proposal, but they indicate that the next iteration of the design cycle should consist of its evaluation in specific organizations.

7. Conclusions

This study designed, demonstrated, and analytically evaluated, from a design science approach, an ethical-technological integrity framework for generative AI structured around four axes, namely transparency and explainability, proactive accountability, bias mitigation, and human-AI interaction. The document analysis of Regulation (EU) 2024/1689, the NIST framework, and the Chinese regime showed that, despite their differences, the three regimes share gaps in explainability to affected persons, in social participation, and, above all, in binding human veto. The main contribution of the work consists of linking each of these gaps to an operational mechanism and to third-party verifiable evidence, and of incorporating a risk-segmented escalation protocol that makes effective human oversight compatible with the operational scale of these systems, so that ethics is integrated as a design variable rather than as an after-the-fact control. The demonstration in a social-protection scenario confirmed that the framework satisfies the requirements of auditability and *ex ante* responsibility, although participation, effective oversight, and proportionality depend on organizational conditions that must be guaranteed during implementation. Since the evaluation carried out was analytical, future work will aim to validate the framework in real organizations, empirically calibrate the risk classifier and its thresholds, verify its effects on competitiveness and trust, extend the matrix to other jurisdictions, including Latin American ones, and incorporate environmental sustainability as an additional dimension.

Author Contribution. W.G. carried out the conceptualization, methodology, investigation, formal analysis, writing of the original draft, review and editing of the text, and visualization, and also reviewed the results and approved the final version of the manuscript.

Acknowledgments. To the Universidad Nacional Experimental de los Llanos Occidentales Ezequiel Zamora for the institutional support provided for the development of this research.

Funding. This study was self-funded by the author.

Informed Consent Statement. Not applicable, since the study did not involve human participants.

Data Availability Statement. The coding matrix of the regulatory frameworks, the coded normative excerpts, the provisions supporting each score, and the documentation of the demonstration scenario are available upon request from the corresponding author.

Conflicts of Interest. The author declares no conflicts of interest.

Statement on the Use of Artificial Intelligence. Generative artificial intelligence tools were used exclusively to support text editing and $\LaTeX$ typesetting. The analysis, interpretation of the results, and conclusions are the exclusive intellectual responsibility of the author, who validated and approved the final version.

References

- [1] J. Sevilla, L. Heim, A. Ho, T. Besiroglu, M. Hobbhahn, and P. Villalobos, “Compute trends across three eras of machine learning,” in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8. doi: 10.1109/IJCNN55064.2022.9891914.
- [2] D. Ganguli, D. Hernandez, L. Lovitt, A. Askell, Y. Bai, A. Chen *et al.*, “Predictability and surprise in large generative models,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1747–1764. doi: 10.1145/3531146.3533229.
- [3] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar *et al.*, ““So what if ChatGPT wrote it?” Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy,” *International Journal of Information Management*, vol. 71, p. 102642, 2023. doi: 10.1016/j.ijinfomgt.2023.102642.
- [4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023. doi: 10.1145/3571730.
- [5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623. doi: 10.1145/3442188.3445922.
- [6] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor *et al.*, “Taxonomy of risks posed by language models,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 214–229. doi: 10.1145/3531146.3533088.
- [7] N. A. Smuha, “From a ‘race to AI’ to a ‘race to AI regulation’: Regulatory competition for artificial intelligence,” *Law, Innovation and Technology*, vol. 13, no. 1, pp. 57–84, 2021. doi: 10.1080/17579961.2021.1898300.
- [8] M. Veale, K. Matus, and R. Gorwa, “AI and global governance: Modalities, rationales, tensions,” *Annual Review of Law and Social Science*, vol. 19, pp. 255–275, 2023. doi: 10.1146/annurev-lawsocsci-020223-040749.
- [9] M. Veale and F. Zuiderveen Borgesius, “Demystifying the Draft EU Artificial Intelligence Act: Analysing the good, the bad, and the unclear elements of the proposed approach,” *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021. doi: 10.9785/cr-2021-220402.
- [10] J. Laux, S. Wachter, and B. Mittelstadt, “Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk,” *Regulation & Governance*, vol. 18, no. 1, pp. 3–32, 2024. doi: 10.1111/regg.12512.
- [11] A. Jobin, M. Ienca, and E. Vayena, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019. doi: 10.1038/s42256-019-0088-2.
- [12] B. Mittelstadt, “Principles alone cannot guarantee ethical AI,” *Nature Machine Intelligence*, vol. 1, no. 11, pp. 501–507, 2019. doi: 10.1038/s42256-019-0114-4.

- [13] T. Hagendorff, "The ethics of AI ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020. doi: 10.1007/s11023-020-09517-8.
- [14] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices," *Science and Engineering Ethics*, vol. 26, no. 4, pp. 2141–2168, 2020. doi: 10.1007/s11948-019-00165-5.
- [15] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," *Harvard Data Science Review*, vol. 1, no. 1, 2019. doi: 10.1162/99608f92.8cd550d1.
- [16] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004. doi: 10.2307/25148625.
- [17] P. Hacker, A. Engel, and M. Mauer, "Regulating ChatGPT and other large generative AI models," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1112–1123. doi: 10.1145/3593013.3594067.
- [18] J. Mökander, J. Schuett, H. R. Kirk, and L. Floridi, "Auditing large language models: A three-layered approach," *AI and Ethics*, vol. 4, no. 4, pp. 1085–1115, 2024. doi: 10.1007/s43681-023-00289-2.
- [19] C. Novelli, M. Taddeo, and L. Floridi, "Accountability in artificial intelligence: What it is and how it works," *AI & Society*, vol. 39, no. 4, pp. 1871–1882, 2024. doi: 10.1007/s00146-023-01635-y.
- [20] E. Papagiannidis, P. Mikalef, and K. Conboy, "Responsible artificial intelligence governance: A review and research framework," *The Journal of Strategic Information Systems*, vol. 34, no. 2, p. 101885, 2025. doi: 10.1016/j.jsis.2024.101885.
- [21] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics and Information Technology*, vol. 6, no. 3, pp. 175–183, 2004. doi: 10.1007/s10676-004-3422-1.
- [22] F. Santoni de Sio and G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philosophy & Technology*, vol. 34, no. 4, pp. 1057–1084, 2021. doi: 10.1007/s13347-021-00450-x.
- [23] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation," *International Data Privacy Law*, vol. 7, no. 2, pp. 76–99, 2017. doi: 10.1093/idpl/idx005.
- [24] M. E. Kaminski, "The right to explanation, explained," *Berkeley Technology Law Journal*, vol. 34, no. 1, pp. 189–218, 2019. doi: 10.15779/Z38TD9N83H.
- [25] J. J. Bryson, "Patience is not a virtue: The design of intelligent systems and systems of ethics," *Ethics and Information Technology*, vol. 20, no. 1, pp. 15–26, 2018. doi: 10.1007/s10676-018-9448-6.
- [26] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial intelligence and the 'good society': The US, EU, and UK approach," *Science and Engineering Ethics*, vol. 24, no. 2, pp. 505–528, 2018. doi: 10.1007/s11948-017-9901-7.
- [27] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018. doi: 10.1007/s11023-018-9482-5.
- [28] D. Schiff, J. Borenstein, J. Biddle, and K. Laas, "AI ethics in the public, private, and NGO sectors: A review of a global document collection," *IEEE Transactions on Technology and Society*, vol. 2, no. 1, pp. 31–42, 2021. doi: 10.1109/TTS.2021.3052127.
- [29] L. Floridi, "Translating principles into practices of digital ethics: Five risks of being unethical," *Philosophy & Technology*, vol. 32, no. 2, pp. 185–193, 2019. doi: 10.1007/s13347-019-00354-x.

- [30] A. Kasirzadeh and I. Gabriel, "In conversation with artificial intelligence: Aligning language models with human values," *Philosophy & Technology*, vol. 36, no. 2, p. 27, 2023. doi: 10.1007/s13347-023-00606-x.
- [31] S. Zuboff, "Big other: Surveillance capitalism and the prospects of an information civilization," *Journal of Information Technology*, vol. 30, no. 1, pp. 75–89, 2015. doi: 10.1057/jit.2015.5.
- [32] A. Birhane, "Algorithmic injustice: A relational ethics approach," *Patterns*, vol. 2, no. 2, p. 100205, 2021. doi: 10.1016/j.patter.2021.100205.
- [33] T. Birkstedt, M. Minkkinen, A. Tandon, and M. Mäntymäki, "AI governance: Themes, knowledge gaps and future agendas," *Internet Research*, vol. 33, no. 7, pp. 133–167, 2023. doi: 10.1108/INTR-01-2022-0042.
- [34] M. Whittaker, "The steep cost of capture," *Interactions*, vol. 28, no. 6, pp. 50–55, 2021. doi: 10.1145/3488666.
- [35] M. Taddeo, T. McCutcheon, and L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword," *Nature Machine Intelligence*, vol. 1, no. 12, pp. 557–560, 2019. doi: 10.1038/s42256-019-0109-1.
- [36] S. Gregor and A. R. Hevner, "Positioning and presenting design science research for maximum impact," *MIS Quarterly*, vol. 37, no. 2, pp. 337–355, 2013. doi: 10.25300/MISQ/2013/37.2.01.
- [37] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45–77, 2007. doi: 10.2753/MIS0742-1222240302.
- [38] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media & Society*, vol. 20, no. 3, pp. 973–989, 2018. doi: 10.1177/1461444816676645.
- [39] S. L. Blodgett, S. Barocas, H. Daumé III, and H. Wallach, "Language (technology) is power: A critical survey of "bias" in NLP," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5454–5476. doi: 10.18653/v1/2020.acl-main.485.
- [40] A. Birhane, W. Isaac, V. Prabhakaran, M. Díaz, M. C. Elish, I. Gabriel, and S. Mohamed, "Power to the people? Opportunities and challenges for participatory AI," in *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*, 2022, pp. 1–8. doi: 10.1145/3551624.3555290.
- [41] B. Green, "The flaws of policies requiring human oversight of government algorithms," *Computer Law & Security Review*, vol. 45, p. 105681, 2022. doi: 10.1016/j.clsr.2022.105681.
- [42] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, no. 3, pp. 381–410, 2010. doi: 10.1177/0018720810376055.
- [43] I. Solaiman, "The gradient of generative AI release: Methods and considerations," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 111–122. doi: 10.1145/3593013.3593981.
- [44] G. A. Bowen, "Document analysis as a qualitative research method," *Qualitative Research Journal*, vol. 9, no. 2, pp. 27–40, 2009. doi: 10.3316/QRJ0902027.
- [45] Parlamento Europeo y Consejo de la Unión Europea, "Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial)," *Diario Oficial de la Unión Europea*, Serie L, 12 de julio de 2024, 2024. [Online]. Available: <http://data.europa.eu/eli/reg/2024/1689/oj>
- [46] National Institute of Standards and Technology, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST AI 100-1, 2023. doi: 10.6028/NIST.AI.100-1.

- [47] —, “Artificial intelligence risk management framework: Generative artificial intelligence profile,” National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep. NIST AI 600-1, 2024. doi: 10.6028/NIST.AI.600-1.
- [48] V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006. doi: 10.1191/1478088706qp063oa.
- [49] J. Venable, J. Pries-Heje, and R. Baskerville, “FEDS: A framework for evaluation in design science research,” *European Journal of Information Systems*, vol. 25, no. 1, pp. 77–89, 2016. doi: 10.1057/ejis.2014.36.
- [50] H. Roberts, J. Cows, J. Morley, M. Taddeo, V. Wang, and L. Floridi, “The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation,” *AI & Society*, vol. 36, no. 1, pp. 59–77, 2021. doi: 10.1007/s00146-020-00992-2.
- [51] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, “Model cards for model reporting,” in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229. doi: 10.1145/3287560.3287596.
- [52] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, “Datasheets for datasets,” *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021. doi: 10.1145/3458723.
- [53] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.
- [54] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 33–44. doi: 10.1145/3351095.3372873.
- [55] I. D. Raji, P. Xu, C. Honigsberg, and D. Ho, “Outsider oversight: Designing a third party audit ecosystem for AI governance,” in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 557–571. doi: 10.1145/3514094.3534181.
- [56] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, “A survey on bias and fairness in machine learning,” *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. doi: 10.1145/3457607.
- [57] A. Birhane and V. U. Prabhu, “Large image datasets: A pyrrhic win for computer vision?” in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1536–1546. doi: 10.1109/WACV48630.2021.00158.
- [58] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, “Red teaming language models with language models,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 3419–3448. doi: 10.18653/v1/2022.emnlp-main.225.
- [59] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, “Algorithmic impact assessments and accountability: The co-construction of impacts,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 735–746. doi: 10.1145/3442188.3445935.
- [60] K. Goddard, A. Roudsari, and J. C. Wyatt, “Automation bias: A systematic review of frequency, effect mediators, and mitigators,” *Journal of the American Medical Informatics Association*, vol. 19, no. 1, pp. 121–127, 2012. doi: 10.1136/amiajnl-2011-000089.
- [61] B. Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy,” *International Journal of Human-Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020. doi: 10.1080/10447318.2020.1741118.

- [62] E. Strubell, A. Ganesh, and A. McCallum, “Energy and policy considerations for modern deep learning research,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 693–13 696. doi: 10.1609/aaai.v34i09.7123.
- [63] J. Cows, A. Tsamados, M. Taddeo, and L. Floridi, “The AI gambit: Leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations,” *AI & Society*, vol. 38, no. 1, pp. 283–307, 2023. doi: 10.1007/s00146-021-01294-x.