Revista de Științe Politice. Revue des Sciences Politiques • No. 92 • 2026: 71 – 85

ORIGINAL PAPER

Governing ESG Claims about Artificial Intelligence: Evidence and Institutional Accountability

Narcis Eduard Mitu[1)], George Teodor Mitu[2)]

Abstract:
Organisations may document the use of artificial intelligence (AI) without substantiating the environmental, social or governance (ESG) benefits attributed to it. This article examines how public claims about AI can be assessed and connected to institutions able to demand explanations and corrections. A purposive conceptual synthesis draws on sustainability reporting, ethical assurance and algorithmic accountability. The analysis distinguishes implementation, technical capability, attributed ESG outcomes and broader sustainability contributions. It specifies four evidential judgements: substantiated as stated, supportable only after qualification, undetermined and contradicted. The original assertion is assessed before any revised wording is considered, so that a defensible narrower statement cannot erase a contradiction in the initial claim. Three hypothetical cases examine environmental efficiency under changing deployment conditions, average social improvement alongside subgroup harm, and governance assurance based on confidential evidence. They illustrate why support depends on the outcome, population and period asserted, and why restricted public access can coexist with independent verification. The contribution is a claim-level adaptation of assurance approaches that connects evidential judgements to disclosure responses while assessing institutional answerability separately. Its practical usability remains untested. The framework provides a basis for examining both the justification of an ESG assertion and the distribution of authority to challenge its public use.
Keywords: *artificial intelligence; ESG disclosure; substantiation; institutional accountability; ethical assurance; greenwashing.*

JEL Classification: G34; M14; M42; O33; Q56.

[1)] Associate Professor, PhD, University of Craiova, Faculty of Economics and Business Administration, Public Finances specialization, E-mail: narcis.mitu@edu.ucv.ro, ORCID ID: https://orcid.org/0000-0003-0265-7658
[2)] PhD Student, Economics Field, Doctoral School of Economic Sciences ”Eugeniu Carada”, University of Craiova, E-mail: mitu.george.h8y@student.ucv.ro, ORCID ID: https://orcid.org/0000-0003-0551-2497

1. Introduction

An organisation can accurately report that it uses artificial intelligence to monitor energy consumption while leaving its environmental contribution unestablished. Monitoring, reliable prediction and reduced consumption are different achievements. A disclosure can move between them without explaining the additional inference. The same difficulty arises when a technical result is presented as evidence of fairer access, safer work or more effective oversight. Readers must assess what the evidence supports and whether anyone can require the organisation to reconsider its interpretation.

Sustainability reporting research distinguishes organisational activities from their measurement and communication (Christensen et al., 2021). Studies of AI-assisted reporting examine opportunities for analysis alongside the risk of persuasive but misleading disclosures (de Villiers et al., 2024; Moodaley & Telukdarie, 2023). Our concern is AI as the subject of an ESG claim, regardless of whether people or software prepared the disclosure. The problem is what an organisation is justified in saying about AI's contribution and who can demand an explanation when that justification is challenged.

Existing research provides substantial foundations for this inquiry. Sætra's (2023) AI ESG protocol connects impact assessment, information and responsibilities. Burr and Leslie (2023) develop ethical assurance through structured argumentation. Pittz's (2026) editorial places AI governance and ESG assurance within a shared research agenda. Building on these approaches, we examine how the scope of a particular public assertion affects the evidence it requires and the response warranted when that evidence is incomplete, qualified or contradictory.

The research question is: *How should AI-related ESG claims be assessed against the available evidence, and what institutional arrangements allow those assessments to be challenged and the claims to be corrected?*

The framework separates evidential status from institutional answerability. It distinguishes partial support, missing information and contradiction, including situations in which AI deployment changes after evaluation. Its contribution is to connect those distinctions to decisions about retaining, narrowing, suspending or correcting individual ESG assertions. This is an application and adaptation of assurance ideas, with the limits of the contribution established through comparison with neighbouring approaches.

The political significance concerns control over evaluation. Organisations, vendors, auditors and affected groups may have unequal access to evidence and unequal authority to demand revision. Sections 2 and 3 establish the conceptual basis and research procedure. Section 4 develops assessment criteria and examines three hypothetical cases. Section 5 considers the distribution of evidential and corrective authority, before Section 6 sets out limitations and conclusions.

2. Conceptual foundations and relevant literature

2.1. Defining AI-related ESG claims and their scope

An AI-related ESG claim is a public organisational assertion about the use, capabilities or effects of AI that is presented as relevant to an environmental, social or governance objective. Statements about implementation and technical capability fall within this definition when they form part of an ESG justification, although neither establishes a benefit on its own. A claim's meaning depends on the application, period, population, comparison and assessment boundary. Anticipated contributions also require a clear distinction between a forecast and an outcome already observed.

AI can produce information used to describe its own effects. This creates two distinct questions: whether the reporting process is reliable and whether the reported contribution is substantiated. De Villiers et al. (2024) examine generative AI across sustainability accounting, reporting, assurance and report use. Moodaley and Telukdarie (2023) review the relationship between artificial intelligence, sustainability reporting and greenwashing. These studies inform the assessment of reporting practices; they do not settle every claim about AI's effects. Even a well-controlled reporting process can communicate an inference requiring further justification.

The assessment must also consider relevant consequences of AI itself. Van Wynsberghe's (2021) distinction between AI for sustainability and the sustainability of AI prevents a beneficial application from defining the entire evaluation. Environmental life cycle analysis develops this point by considering the system's own burdens alongside those of its application (Ligozat et al., 2022). A statement about a net contribution consequently requires a wider evidential basis than a bounded claim about one operational benefit.

ESG indicators leave room for evaluative choices. Berg et al. (2022) identify scope, measurement and weighting as sources of divergence between ESG ratings. Their findings concern ratings rather than AI outcomes, but they clarify why an automated indicator still needs a defensible interpretation of improvement. We use evidential adequacy to describe the fit between that interpretation, the assertion's scope and the evidence offered. Relevant coverage, reliable information and a warranted inference matter whether the supporting material is quantitative or qualitative.

2.2. Existing approaches to substantiation and accountability

The relationship between claims and evidence is already central to several relevant frameworks. Sætra's (2023) protocol for assessing and disclosing the ESG implications of AI capabilities, assets and activities includes impact statements, information sources, knowledge gaps and responsible persons. Burr and Leslie (2023) connect ethical claims with evidence and argument while incorporating deliberation throughout a project's lifecycle. These approaches provide the basis for asking how far a particular public statement is justified and how its interpretation can be examined.

Auditing research locates this examination within organisational practice. Raji et al. (2020) organise internal algorithmic auditing across development stages, with documentation connecting decisions and review. Mökander (2023) treats technical evaluation and scrutiny of organisational processes as complementary. Batool et al. (2025) organise AI governance research around who governs, what is governed, when and how. This literature helps explain why assessing a disclosure requires attention to the decisions and responsibilities behind its wording.

Lam et al. (2024) provide a particularly close comparator. Their criterion audit framework concerns external compliance and assurance and examines the need for interpretable audit outcomes, including the limitations of a simple pass/fail result. We draw on this concern with differentiated judgements while assessing the support for individual public assertions rather than issuing an opinion on a professional audit engagement.

Recent ESG-specific research addresses related issues from different positions. Al-Zoubi et al. (2026) examine AI-driven reporting, greenwashing and legal accountability. Krasteva-Hristova and Georgieva (2026) consider AI-assisted sustainability assurance and the continuing role of professional judgement. Pittz (2026)

proposes an institutional research agenda linking AI governance and ESG assurance. Table 1 identifies the basis retained from neighbouring approaches and the more specific application developed here.

Table 1. *Relationship to neighbouring conceptual approaches*

Approach	Basis retained	Application developed here
AI ESG protocol Sætra (2023)	Impact statements, information sources and responsibilities.	Examine the inference supporting a particular public assertion.
Ethical assurance Burr and Leslie (2023)	Structured justification and participatory examination.	Differentiate evidential judgements and disclosure responses.
Assurance audits Lam et al. (2024)	Criteria-based examination and interpretable audit outcomes.	Distinguish judgements about a public assertion from professional audit opinions.
Algorithmic auditing Raji et al. (2020, 2022)	Lifecycle documentation and institutional conditions for review.	Locate evidence access and corrective authority for the claim.
AI in ESG assurance Krasteva-Hristova and Georgieva (2026)	Professional judgement when AI assists assurance.	Assess claims about AI's contribution, regardless of the reporting tool.

Note: *The final column specifies this article's application; it does not imply that earlier frameworks exclude these questions.*
Source: *Authors' comparison of the cited approaches.*

Within the literature reviewed, the task remaining for this article is one of analytical specification: how should support for an individual ESG assertion be classified when its wording extends beyond the evaluated task, population or period, and how does that judgement relate to the obligation to answer a challenge? Existing frameworks can accommodate these questions. We develop their application explicitly, including circumstances in which evidential support and institutional answerability diverge. The focused review does not establish the absence of comparable formulations elsewhere.

2.3. Evidence and accountability challenges in AI deployment

Evidence about AI is obtained under conditions that disclosure can obscure. Evaluation concerns a particular task, data distribution and deployment configuration. A model update or a change in users may affect the relevance of earlier findings. The NIST AI Risk Management Framework connects context, evaluation and continuing monitoring (Tabassi, 2023). Applied to ESG claims, this means that an earlier result may remain available without remaining applicable. Reassessment is warranted when a change affects the premises of the claim; it does not depend on assuming that every AI system learns during operation.

Different actors may hold different parts of the justification. A vendor may know the model's evaluation conditions, an employer may control deployment, and a reporting team may endorse the public statement. Raghavan et al. (2020) show why claims about

algorithmic hiring assessments require scrutiny of validation practices, data choices and prediction targets. Their study offers a sector-specific precedent for examining the relationship between a promised social benefit and the technical procedures said to deliver it, without establishing the quality of all current hiring systems.

These difficulties intensify familiar accountability problems. Bovens (2007) understands accountability as a relationship in which an actor explains and justifies conduct to a forum able to question, judge and attach possible consequences. Raji et al. (2022) show why external auditing also depends on institutional design, including access and independence. A statement may therefore be well supported while offering weak opportunities for challenge, just as an accessible forum may examine an unsupported assertion. The framework keeps these positions distinct instead of concealing them within an overall credibility score.

3. Research approach

The study used conceptual synthesis and adaptation to examine how the evidence supporting AI-related ESG claims relates to the institutional arrangements through which those claims can be challenged and corrected. Following Jaakkola (2020), the research design assigned a distinct role to each theoretical component: sustainability reporting specified the object of assessment; ethical assurance informed the examination of the relationship between claims and evidence; and accountability theory located this assessment within an actor–forum relationship. The unit of analysis was an individual public assertion, including separable propositions within a compound disclosure.

The bibliographic basis comprised the 23 sources cited in the article and listed in its final bibliography. Sources were selected purposively for their contribution to the research question through targeted searches and examination of reference lists, completed on 1 October 2026. Searches combined “AI” or “artificial intelligence” with “ESG”, “sustainability reporting”, “assurance”, “claims”, “accountability” and “auditing”. Additional searches on environmental assessment and algorithmic hiring helped clarify assessment boundaries and the substantiation of social benefits. Priority was given to publications from 2020 onwards. Earlier foundational work was retained where necessary, particularly Bovens (2007), whose account of institutional accountability informed the analysis of who must provide explanations and who can demand them.

Inclusion required an identifiable contribution to the argument, verifiable publication details and access to the relevant content. Verification distinguished bibliographic accuracy from substantive support. Publication details were checked against publisher or institutional records and DOI information, where available. The relevant content, including author-deposited versions where appropriate, was consulted to assess whether the source supported the claim or interpretation attributed to it. Citations used the journal issue year when it differed from the year of initial online publication.

The selected literature combined conceptual frameworks, empirical studies informing specific analytical distinctions, and institutional guidance. Pittz’s (2026) editorial was used to locate the discussion within a contemporary research agenda. Its arguments were treated as a scholarly position rather than empirical findings. The selection supported a focused critical synthesis, with no claim to exhaustive coverage or systematic-review completeness. Promotional accounts lacking a traceable evidential basis were excluded.

The analysis proceeded in three steps. First, it differentiated assertions about AI implementation, technical capability, attributed ESG outcomes and broader sustainability

contributions, identifying the additional justification required to move between them. Second, it examined how partial support, missing information and contradictory evidence affected the assessment of a defined assertion. Third, it connected these evidential judgements to the actors holding the evidence, endorsing the public claim and reviewing challenges. This sequence kept the question of whether a statement was supported separate from the question of who could require an explanation or correction.

Three hypothetical cases were then constructed to examine the resulting criteria: environmental efficiency under changing assessment boundaries and deployment conditions; an aggregate social improvement alongside harm to a subgroup; and governance assurance based on confidential evidence. The premises were stated explicitly, and changes to a premise or to the wording of the claim were examined to determine whether they warranted a different judgement. These cases tested conceptual distinctions that a binary assessment of credibility could obscure. They were not observations of sampled organisations, and no original performance data were collected. Their role was limited to examining the framework's reasoning; they did not establish its practical usability, agreement between independent reviewers or effects on reporting behaviour. Section 4 presents the resulting conceptual criteria and their application, while Section 5 considers their governance implications.

4. A framework for assessing and governing claims

4.1. From using AI to demonstrating ESG benefits

Consider an organisation that introduces an AI system to identify workplace hazards. Records can confirm that the system is operating, and tests may show that it detects certain risks reliably. A claim that employees are safer requires further evidence: what happened after the risks were identified, how the organisation responded, and whether exposure to danger or injuries declined. Detecting a problem is one part of the explanation for an improvement; action on the warning must also be examined.

This example illustrates why the wording of an ESG claim matters. An organisation may report that it uses AI, that the technology performs a particular task, or that its use has produced a benefit. It may also make a broader statement about its contribution to sustainability. These claims can appear together, but each requires different support. Table 2 distinguishes them and identifies the conclusions that evidence at each level leaves open.

Table 2. *Types of AI-related ESG claims and the evidence needed to support them*

Type of claim	Evidence needed	What remains to be established
Implementation	Records identifying the system, its task, the setting and the period of use.	Use alone does not demonstrate effectiveness or an ESG benefit.
Technical capability	Evaluation relevant to the task, users and system version, with errors and limitations reported.	Performance on a technical task does not establish an improvement in an ESG outcome.
ESG outcome attributed to AI	Evidence of the outcome, an appropriate comparison and an explanation of AI's contribution.	An observed change alone does not establish that AI produced it.

Broader or net contribution	Relevant benefits and burdens across the stated activities, period and affected groups.	An improvement in one area does not settle the wider consequences or the acceptability of trade-offs.

Note: *The four claim categories and their associated limits of inference are the authors' analytical formulation.*

Source: *Authors' conceptual synthesis. Christensen et al. (2021) inform the distinction between sustainability activities and their reporting; Mökander (2023) and Tabassi (2023) inform the treatment of technical assessment; van Wynsberghe (2021) and Ligozat et al. (2022) inform the consideration of the sustainability burdens associated with AI itself.*

Returning to the workplace example, a decline in injuries after the system's introduction would be relevant evidence. The organisation would still need to examine whether changes in training, staffing or equipment contributed to that decline. Records of employees' and managers' responses to warnings could help explain AI's role. The assessment can draw on qualitative evidence about decisions and working practices, provided the public claim reflects how convincingly that evidence supports the proposed explanation.

Extending a result beyond the setting in which it was obtained requires similar care. Evidence from one workplace may be relevant to another, but differences in tasks, users and working conditions need consideration. A later change to the AI system may also affect whether an earlier evaluation remains applicable. The result need not be discarded. The organisation should explain why it continues to support the claim now being made.

The period covered by a statement matters as well. Expected savings depend on assumptions about future use, whereas savings already achieved require evidence of what occurred. Keeping that distinction visible allows anticipated benefits to be communicated without presenting them as established results. Once the assertion is clear, reviewers can examine whether its evidence is sufficient, supports a narrower formulation, leaves the issue unresolved or conflicts with the statement.

4.2. Evidential judgements and their limits

The framework uses four evidential judgements, set out in Table 3. Each concerns a defined assertion and an identified body of available evidence. These are categories for assessing public claims, not professional audit opinions or organisational ratings. Their use requires an explanation of how the evidence bears on the wording under review.

Table 3. *Evidential judgements and corresponding disclosure responses*

Judgement	Decision criterion	Response indicated
Substantiated as stated	Sufficiently reliable evidence covers the assertion and its material assumptions, with relevant counterevidence addressed.	Retain with scope, uncertainty and review conditions visible.
Supportable only after qualification	Evidence supports a material restriction of the same assertion; its extension is unsupported, not contradicted.	Replace the wording; specify the unsupported extension.
Undetermined	Necessary evidence is missing, inaccessible or inconclusive; neither substantiation, a supported restriction nor contradiction is established.	Seek specified evidence; suspend categorical endorsement meanwhile.
Contradicted	The assessed evidence warrants rejecting the original assertion, even if different wording would be supportable.	Correct or withdraw the affected assertion and explain why.

Note: *Assess the original wording before considering a replacement. Contradiction takes priority over possible qualification. Responses concern disclosure, not legal liability; uncertainty does not establish falsehood.*
Source: *Authors' proposed criteria.*

Assessment begins by separating compound assertions and weighing favourable evidence alongside relevant counterevidence. Where that assessment warrants rejecting the original wording, the assertion remains contradicted even if a narrower statement would be defensible. Qualification applies where evidence supports a material, bounded part of the same assertion, while its extension remains unsupported rather than contradicted. If no such part can be established and rejection is unwarranted, the judgement is undetermined. Unresolved disagreement between credible sources may also warrant that judgement; an isolated adverse observation does not automatically refute an average or probabilistic claim.

Qualification must preserve the kind of assertion being assessed. Restricting a claim to the sites for which the stated outcome is supported may be justified when other sites remain unevaluated. Replacing a claim about total environmental impact with one about energy efficiency changes the proposition. That replacement may improve the disclosure, but it does not substantiate even a bounded part of the original net-impact assertion.

Adequacy depends on the outcome asserted and the methods appropriate to it. Material uncertainty should be disclosed, and reviewers should distinguish evidence they inspected from an assurance report they relied on or information they could not examine. Public availability alone establishes neither reliability nor completeness. Conversely, restricted public access does not show that verification is absent. The governance case below examines how a claim can be supported through independent scrutiny while its underlying records remain confidential.

4.3. Applying the framework to three hypothetical ESG cases

Environmental assessment provides the first case. Luccioni et al. (2024) examine energy use during AI inference across tasks and model configurations, making deployment choices relevant to assessment without establishing the net impact of any particular application. Dauvergne's (2022) political economy analysis questions the assumption that local efficiency gains necessarily yield wider sustainability gains. These perspectives support a distinction between an operational comparison and the broader consequences attributed to it.

Suppose a facility has a credible comparative evaluation supporting lower energy use per unit through AI-assisted scheduling under specified operating conditions. Production subsequently expands, while computing and equipment burdens remain unmeasured. The evaluated efficiency improvement can be substantiated. A claim that the organisation's total environmental burden fell remains undetermined on these premises: a different metric and a wider boundary are involved. Reporting the local result separately preserves useful information without treating it as a qualified version of the net-impact claim. If the model is replaced, the historical result may remain valid, but continuing savings require justification that the earlier evaluation still applies. Better technical performance in the replacement model does not supply that justification on its own.

The social case concerns a service whose average waiting time falls after an AI-supported allocation system is introduced. Assume that credible evaluation supports AI's contribution, but disaggregated evidence shows longer waits for a group requiring assisted access. A claim of lower average waiting time can be substantiated; a claim that waiting times fell for every eligible group is contradicted. The supported average does not change the status of the universal assertion. Nor does it establish more inclusive access, which concerns who can use the service and under what conditions, beyond the time spent waiting.

This case separates correction of a statement from resolution of a distributional conflict. The organisation may communicate the average improvement accurately while still needing to explain and address the adverse subgroup outcome. Affected users should be able to question the definition of improvement as well as data accuracy. In workplace settings, the International Labour Organization (2025) similarly brings privacy, monitoring-related stress and worker participation into its assessment of digital safety tools. A benefit on one measure does not decide the acceptability of a deployment across these dimensions. The framework identifies the unresolved judgement without assigning weights that would offset one group's losses against another's gains.

The governance case concerns a disclosure that transactions flagged by AI were reviewed consistently against stated criteria during a specified period. Assume that a competent independent reviewer examines evaluation records, samples operational files and checks follow-up, finding sufficient evidence to support that bounded claim. Transaction data remain confidential, but a public assurance statement identifies the criteria, period and limitations. Restricted public access can then coexist with substantiation through independent examination. Readers rely partly on the reviewer's competence, independence and reported findings. The conclusion concerns the operation of the review process; it does not establish that AI improved governance overall.

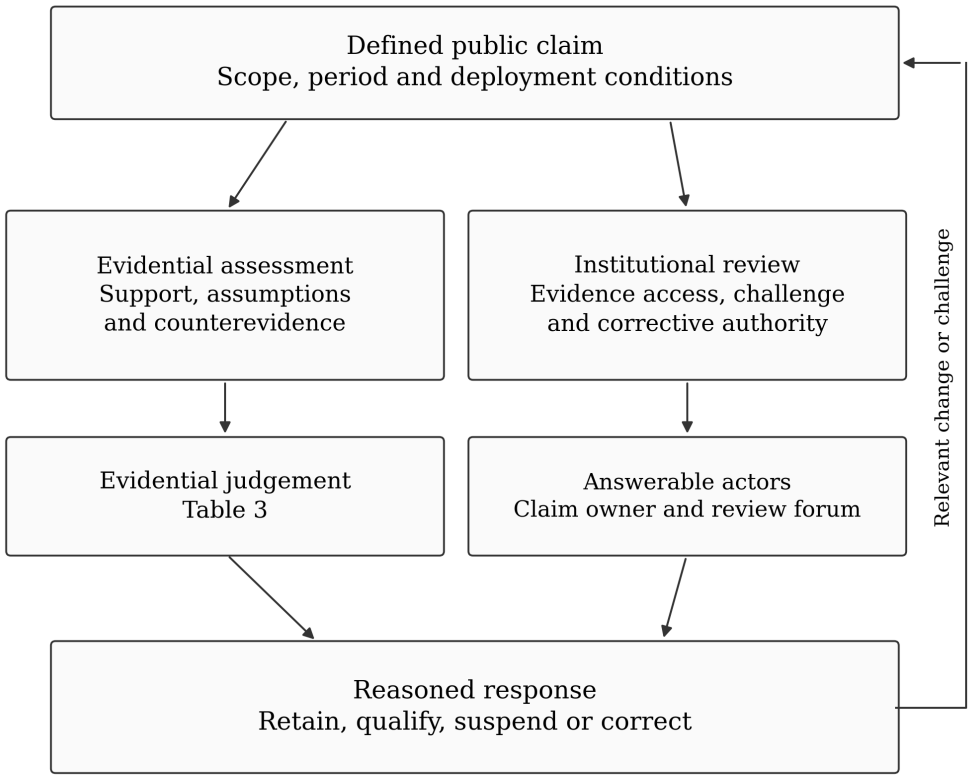
Now suppose that evidence necessary to assess the claim is unavailable to both the reviewer and the deploying organisation, no adequate alternative exists, and the disclosure relies only on an unexplained vendor assurance. The claim becomes

undetermined for that review, rather than disproved. A policy requiring human oversight cannot establish how oversight operated. This differs from a situation where inaccessible vendor records can be replaced by sufficient operational evidence. Even a substantiated operational claim may remain institutionally weak if no identified body must answer challenges or can secure a correction.

4.4. Institutional responsibility and corrective authority

Figure 1 connects evidential assessment and institutional review while preserving their separate conclusions. The first asks whether the assertion is supported. The second asks who can inspect its basis, contest the assessment and require an answer. Weak review arrangements do not make a factual statement false, but they restrict the opportunities to challenge its interpretation and continued public use.

Figure 1. *Separate assessments linked to a reasoned organisational response*



Note: *Arrows represent proposed review relationships, not measured effects. Evidential status and institutional answerability are assessed separately.*
Source: *Authors' framework.*

Responsibility must be located for holding the evidence, endorsing the claim and deciding on a response. Different actors may perform these roles, provided their relationship is explicit. A reporting team without access to technical records needs a route to a qualified assessment. An auditor who identifies an overstatement needs access to

someone authorised to amend it. Naming responsible actors is insufficient when the information or power necessary to act does not accompany the role.

The response follows from the evidential judgement and the forum's remit. Supported wording may be retained, a partly supported assertion narrowed, an undetermined benefit claim suspended pending examination, or a contradicted statement corrected or withdrawn. Where the forum lacks corrective authority, referral needs an identified destination. Changes in model version, use or population, and relevant counterevidence, warrant renewed assessment when they affect its premises. Retaining the earlier wording and reasons for revision prevents a correction from erasing the record of what the organisation previously communicated.

5. Discussion and governance implications

5.1. The contribution and its boundaries

The analytical gain lies in keeping the object of each judgement stable. The environmental case distinguishes a historical efficiency result from a present claim about total burden. The social case shows why a true average cannot rescue a false universal assertion. The governance case separates restricted public access from the absence of adequate verification. Each example changes a premise or the proposition under review, allowing the consequences for its assessment to be explained without assigning the organisation a single credibility label.

This application complements the attention to audit outcomes in Lam et al. (2024). Here, the judgement concerns the wording of an individual disclosure and does not certify an algorithmic system's compliance with a normative framework. The institutional distinction adds a separate question: who must answer for that wording and who can obtain a response? Adequate evidence can substantiate a statement without ensuring accountable governance. Equally, the existence of a complaints process cannot establish that the statement is true. The normative requirement for answerability follows from the actor–forum relationship developed in Section 2.3.

This separation also clarifies the treatment of competing ESG concerns. A narrowly defined environmental claim may be supported while the deployment raises unresolved social questions. Those questions remain relevant to oversight even if they fall outside the assertion being assessed. The framework therefore limits what a disclosure can establish while leaving room to challenge the organisation's choice of what to disclose. Judgements about the acceptability of trade-offs require a substantive evaluation of the affected interests, beyond the evidential status of one statement.

5.2. Evidential access and the distribution of authority

Deciding what counts as improvement distributes influence. An organisation's choice of comparison, population and time horizon shapes which effects are visible. Dauvergne (2022) connects accounts of AI's environmental benefits to corporate power and displaced costs. The question for governance is whether another actor can challenge the assessment boundary or merely check calculations within it. A review may accurately verify the selected indicator while leaving the central disagreement about excluded consequences untouched.

Meaningful access requires institutional arrangements as well as information. Raji et al. (2022) examine how selection, remuneration, access and subsequent action affect external auditing. A reviewer's independence should therefore not be inferred solely from the label "external". Commissioning terms should allow scrutiny of material

assumptions and communication of consequential limitations. Confidentiality can justify controlled access, but it should not automatically give the evidence holder final authority over what may be questioned. Public explanations should identify the basis and limits of the reviewer's conclusion.

The burden of substantiation raises a related asymmetry. In our view, organisations making benefit claims should remain responsible for supporting them. Affected groups should not have to reconstruct a proprietary system before a relevant challenge receives an answer. Their accounts may reveal excluded outcomes, inaccessible services or changes missed by reporting categories. Participation does not make every objection decisive. It should, however, entail a reasoned response and an opportunity for escalation; otherwise, an invitation to comment may leave the distribution of authority unchanged.

5.3. Proposals for implementation and empirical evaluation

One possible organisational measure is a dated claim record linking exact wording to evidence, assumptions, responsible actors and review conditions. It would distinguish publicly available material from evidence examined confidentially by a named reviewer, identify unresolved limitations and record changes requiring reconsideration. Its purpose is to make the justification traceable; keeping the record would not itself demonstrate improved sustainability. The extent of review should reflect the claim's scope and the consequences of the deployment, with its practical demands examined before wider adoption.

Standard setters and public oversight bodies could support comparable descriptions of assurance scope, access restrictions and corrective responsibilities. Such descriptions may help readers identify omissions without imposing one performance measure across ESG pillars. Their effects would depend on incentives, competence and enforcement, concerns already established in sustainability reporting research (Christensen et al., 2021). The proposal offers questions for oversight and does not state current legal obligations or replace applicable assurance standards.

Empirical examination could begin with a bounded sample of public claims and their supporting methodological documents. Independent reviewers could record the proposition assessed, the evidence accessible to them, their judgement and reasons for disagreement. Interviews could then examine whether named forums actually obtain answers and secure revisions. This would assess the criteria's interpretability separately from the operation of accountability arrangements. An unsupported claim would warrant scrutiny, but a finding of intentional greenwashing would require further evidence about knowledge, conduct and context.

6. Limitations and conclusions

The framework's limitations follow from its conceptual design and purposive literature selection. Its consistency across reviewers, organisational feasibility and effects on disclosure have not been tested. The hypothetical cases establish distinctions under stated premises, without showing their prevalence. A review of public documents would also face limits where relevant evidence remains private. Applying the framework would require sector-specific criteria and attention to the powers of the reviewing institution. It cannot resolve substantive disagreements about acceptable trade-offs merely by classifying the evidence.

The research question requires two connected answers. Evidence must be assessed against the scope, period and level of inference of the assertion actually made. Where it falls short, the reviewer should distinguish a supported restriction of that assertion from uncertainty or contradiction. Institutional arrangements must then identify an actor responsible for explaining the disclosure and a forum able to question it and secure an appropriate response. Neither answer can substitute for the other.

A correction is accountable only if its reasons and consequences remain open to examination. Replacing an overstatement with defensible wording should preserve the record of the earlier claim, particularly where credible evidence contradicted it. Organisations also need to reconsider continuing claims when deployment changes undermine their original justification. These requirements connect the assessment of evidence to the distribution of authority over public ESG information. Whether applying them produces more reliable disclosure is a question for subsequent empirical research.

Authors' Contributions:

The authors contributed equally to this work.

References:

Al-Zoubi, T. M. H., Al-Hailat, O., Alomar, A., & Al-Billeh, T. (2026). AI-driven sustainability reporting and corporate greenwashing: Legal accountability and governance challenges in the ESG era. *Sustainability, 18*(17), Article 8661. https://doi.org/10.3390/su18178661

Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. *AI and Ethics, 5*(3), 3265–3279. https://doi.org/10.1007/s43681-024-00653-w

Berg, F., Kölbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance, 26*(6), 1315–1344. https://doi.org/10.1093/rof/rfac033

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European Law Journal, 13*(4), 447–468. https://doi.org/10.1111/j.1468-0386.2007.00378.x

Burr, C., & Leslie, D. (2023). Ethical assurance: A practical approach to the responsible design, development, and deployment of data-driven technologies. *AI and Ethics, 3*(1), 73–98. https://doi.org/10.1007/s43681-022-00178-0

Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of Accounting Studies, 26*(3), 1176–1248. https://doi.org/10.1007/s11142-021-09609-5

Dauvergne, P. (2022). Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. *Review of International Political Economy, 29*(3), 696–718. https://doi.org/10.1080/09692290.2020.1814381

de Villiers, C., Dimes, R., & Molinari, M. (2024). How will AI text generation and processing impact sustainability reporting? Critical analysis, a conceptual framework and avenues for future research. *Sustainability Accounting, Management and Policy Journal, 15*(1), 96–118. https://doi.org/10.1108/SAMPJ-02-2023-0097

International Labour Organization. (2025). *Revolutionizing health and safety: The role of AI and digitalization at work.* https://www.ilo.org/publications/revolutionizing-health-and-safety-role-ai-and-digitalization-work

Jaakkola, E. (2020). Designing conceptual articles: Four approaches. *AMS Review, 10*(1–2), 18–26. https://doi.org/10.1007/s13162-020-00161-0

Krasteva-Hristova, R., & Georgieva, V. (2026). Artificial intelligence in sustainability assurance: Accounting challenges, audit risks and a conceptual framework for ESG verification. *Accounting and Auditing, 2*(3), Article 15. https://doi.org/10.3390/accountaudit2030015

Lam, K., Lange, B., Blili-Hamelin, B., Davidovic, J., Brown, S., & Hasan, A. (2024). A framework for assurance audits of algorithmic systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1078–1092). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658957

Ligozat, A.-L., Lefevre, J., Bugeau, A., & Combaz, J. (2022). Unraveling the hidden environmental impacts of AI solutions for environment life cycle assessment of AI solutions. *Sustainability, 14*(9), Article 5172. https://doi.org/10.3390/su14095172

Luccioni, A. S., Jernite, Y., & Strubell, E. (2024). Power hungry processing: Watts driving the cost of AI deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 85–99). Association for Computing Machinery. https://doi.org/10.1145/3630106.3658542

Mökander, J. (2023). Auditing of AI: Legal, ethical and technical approaches. *Digital Society, 2*(3), Article 49. https://doi.org/10.1007/s44206-023-00074-y

Moodaley, W., & Telukdarie, A. (2023). Greenwashing, sustainability reporting, and artificial intelligence: A systematic literature review. *Sustainability, 15*(2), Article 1481. https://doi.org/10.3390/su15021481

Pittz, T. G. (2026). The new accountability frontier: AI governance, ESG assurance, and the future of an accountability infrastructure [Editorial]. *Cogent Business & Management, 13*(1), Article 2685990. https://doi.org/10.1080/23311975.2026.2685990

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372828

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). 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* (pp. 33–44). Association for Computing Machinery. https://doi.org/10.1145/3351095.3372873

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. (2022). Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 557–571). Association for Computing Machinery. https://doi.org/10.1145/3514094.3534181

Sætra, H. S. (2023). The AI ESG protocol: Evaluating and disclosing the environment, social, and governance implications of artificial intelligence capabilities, assets, and activities. *Sustainable Development, 31*(2), 1027–1037. https://doi.org/10.1002/sd.2438

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics, 1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Article Info

Received: October 05 2026
Accepted: November 08 2026

How to cite this article:

Mitu, N.E., Mitu, G.T. (2026). Governing ESG Claims about Artificial Intelligence: Evidence and Institutional Accountability. *Revista de Științe Politice. Revue des Sciences Politiques,* no. 92, pp. 71 - 85.