Governance-Aware Greenwashing Detection: A Multi-Agent Pipeline with Explainability

Chayma Sakrani
SMART Lab, ISG
Tunis, Tunisia
sakrani.chayma@gmail.com
https://orcid.org/0009-0009-0427-669X

Boutheina Jlifi
ESCT Manouba, SMART Lab, ISG
Tunis, Tunisia
CRECC, Paris School of Business
Paris, France
LITIS, Le Havre Normandie University
Normandie, France
boutheina.jlifi@esct.uma.tn
https://orcid.org/0000-0002-9835-4338

Lamjed Ben Said
SMART Lab, ISG
Tunis, Tunisia
bensaid$_l amjed@yahoo.fr$
https://orcid.org/0000-0001-9225-884X

***Abstract*—Greenwashing, the practice of making misleading environmental claims, threatens the integrity of Environmental, Social, and Governance (ESG) reporting, investor trust, and global sustainability goals. Existing AI detection methods treat this as a standalone classification problem, ignoring the governance ecosystem in which claims are produced, reviewed, and acted upon. This paper makes two contributions. First, we formalise greenwashing detection as a multi-agent decision problem, modelling six interacting actors and the feedback loops connecting corporate disclosure, regulatory action, and market response. Second, we evaluate the explainability of the AI screening component using SHapley Additive exPlanations (SHAP) and Integrated Gradients (IG), demonstrating statistically significant faithfulness for both methods. Together, these contributions position AI as an accountable intermediary in a critical governance process, advancing human-centred design for high-stakes regulatory decision-making.**

***Index Terms*—greenwashing detection; multi-agent systems; ESG; explainability; SHAP; integrated gradients; human-centred AI; critical systems**

I. Introduction

The expansion of Environmental, Social, and Governance (ESG) reporting has transformed sustainability claims into consequential signals for investors, regulators, and the public. Yet the same disclosure incentives that encourage transparency also create conditions for greenwashing, the practice of projecting a positive environmental image that is decoupled from, or actively conceals, poor environmental performance [1]. In corporate disclosures, greenwashing manifests across a spectrum of strategies: inflated emissions-reduction claims, selective indicator reporting that highlights favourable metrics while omitting material harms, and deliberately vague sustainability language designed to resist verification. Left undetected, such practices distort capital allocation, undermine environmental policy, and erode the credibility of sustainability reporting as an institution.

Detecting greenwashing is therefore not a narrow text-classification problem. It is a governance challenge shaped by the interplay of multiple actors, each with distinct roles, interests, and information asymmetries. Companies produce and strategically frame ESG claims, controlling the evidence that is disclosed. Regulators investigate suspected violations and must translate automated risk signals into defensible enforcement decisions. Investors condition capital allocation on compliance signals and bear financial exposure to misclassified disclosures. Civil society actors, including NGOs, journalists, and advocacy groups, exercise external oversight through public complaints and reputational pressure. Financial auditors and third-party certifiers provide independent verification that anchors the credibility of the entire system. The natural environment itself constitutes the ultimate stakeholder: inadequate enforcement allows harmful practices to persist with cumulative ecological consequences. Critically, these actors are not independent. A regulatory decision propagates back through the system, reshaping corporate disclosure behaviour, investor confidence, and public trust simultaneously.

Existing computational approaches to greenwashing detection have not kept pace with this governance complexity. Current NLP and machine-learning methods typically reduce the problem to isolated binary classification, flagging individual claims without contextual grounding in sector norms, verification evidence, or regulatory thresholds [5], [6]. These limitations have two consequences. First, misclassification costs are asymmetric: undetected greenwashing misleads markets and delays corrective action, while false accusations impose reputational and legal costs that deter genuine sustainability investment. Second, benchmark accuracy does not guarantee real-world utility. Detection does not end with a model output; it propagates through regulatory decisions, market responses, and public oversight, and AI design must reflect that multi-actor reality.

This paper addresses these gaps through two contributions: a multi-agent model of the greenwashing governance system that characterises agent roles and interdependencies; and an NLP screening approach built on that model, prioritising transparency and auditability for regulatory use. Together, they argue that effective greenwashing detection requires AI systems designed around their governance context, not merely

better classifiers.

This work is submitted to the special session on *"AI for Humanity: Toward Transparent and Robust Optimisation in Critical Systems."* Greenwashing detection exemplifies the critical systems addressed by that theme: its failures carry economic, environmental, and social consequences across multiple interdependent actors. Transparency and institutional fit must therefore be treated as first-class design requirements alongside predictive performance.

II. Contributions

This work makes two main contributions: a governance-level multi-agent model and an AI mechanism for supporting screening decisions.

- **Contribution 1—Multi-agent modelling of the greenwashing governance system:** We conceptualise greenwashing detection as a multi-agent critical system rather than a standalone text-classification task. The model specifies six interacting agents—Company, AI Screening System, Regulator, Investors, Society/NGOs, and Environment—with their objectives, decision roles, information exchanges, and feedback loops. It provides a conceptual basis for understanding how ESG claims, AI risk signals, regulatory actions, market responses, public oversight, and environmental consequences propagate through the governance ecosystem, including the asymmetric effects of false positives and false negatives.
- **Contribution 2—A governance-aware AI screening approach:** Grounded in the multi-agent model, we propose an AI-assisted screening component that integrates cost-sensitive greenwashing classification with explainability evaluation. The approach applies a Distilled Bidirectional Encoder Representations from Transformers (DistilBERT)-based classifier and augments its predictions with risk scores, SHAP explanations, and Integrated Gradients attributions subject to faithfulness tests, producing outputs regulators can inspect, audit, and use in human-centred enforcement workflows.

III. Related Work

Greenwashing detection has evolved from keyword-based sentiment analysis flagging gaps between stated commitments and corporate actions [2] to transformer-based classifiers applied to ESG disclosures [3]. The Environmental Claims Dataset [4] provides a benchmark for binary classification of legitimate versus potentially misleading claims, and recent work has extended this to sentence-level annotation and domain-adapted language models [5], [6]. Despite this progress, existing approaches remain classification-centred: they optimise predictive accuracy without addressing transparency for regulatory use or integration into enforcement workflows.

The Multi-Agent Systems (MAS) literature offers conceptual tools to move beyond this limitation by modelling heterogeneous actors with conflicting objectives in regulatory settings [7]. Applied to algorithmic decision-making, this tradition demonstrates that decisions cannot be assessed at the prediction point alone but must be traced across entire decision pipelines [8], [9]. Strategic regulation models further show how the design of detection mechanisms shapes the disclosure behaviour of regulated entities, an incentive effect that purely technical classifiers ignore [10].

Producing outputs fit for multi-stakeholder pipelines requires explainability guarantees. SHAP offers model-agnostic feature attributions grounded in cooperative game theory [11], while Integrated Gradients provide gradient-based attributions satisfying axiomatic completeness [12]. Faithfulness tests, including deletion and occlusion protocols, assess whether attributed features are genuinely decision-driving rather than post-hoc rationalisations [13], [14].

Our work integrates these three strands in a way that none does individually. We conceptualise greenwashing detection as a multi-agent governance problem, explicitly modelling the feedback loops between AI outputs, regulatory decisions, and corporate disclosure behaviour that classification-centred approaches overlook. On this foundation, we build a screening pipeline that embeds explainability as a first-class design requirement, producing outputs that regulators can inspect, challenge, and act upon within accountable enforcement workflows.

IV. Problem Definition and Multi-Agent Pipeline

We define greenwashing detection as the assessment of whether an environmental claim is credible, misleading, or insufficiently supported in its governance context. A claim becomes problematic not solely because of its wording, but when the statement, omitted evidence, company incentives, and likely stakeholder interpretation collectively create a risk of distorted environmental judgement. The operational input is a corporate ESG claim, optionally paired with metadata such as sector, region, and company type; the desired output is a governance-ready risk report comprising a greenwashing probability, salient textual explanations, uncertainty indicators, and sufficient evidence for human review—not merely a binary label.

This formulation carries two design imperatives. First, the objective must prioritise recall, since missed greenwashing misdirects capital and weakens environmental accountability. Second, explanations are mandatory, since regulators must justify why a claim is suspicious. Greenwashing detection is therefore modelled as a socio-technical decision problem embedded in a multi-agent governance system. We use the Environmental Claims Dataset [4] as the benchmark corpus: it contains approximately 3,000 sentence-level ESG statements annotated as legitimate or potentially misleading, with a positive (greenwashing) rate of roughly 40%, drawn from corporate reports across multiple sectors.

We define six agents in the pipeline. The **Company** generates sustainability disclosures and receives regulatory and market feedback. The **AI Screening System** produces predictions, risk scores, and explanations as an accountable intermediary.

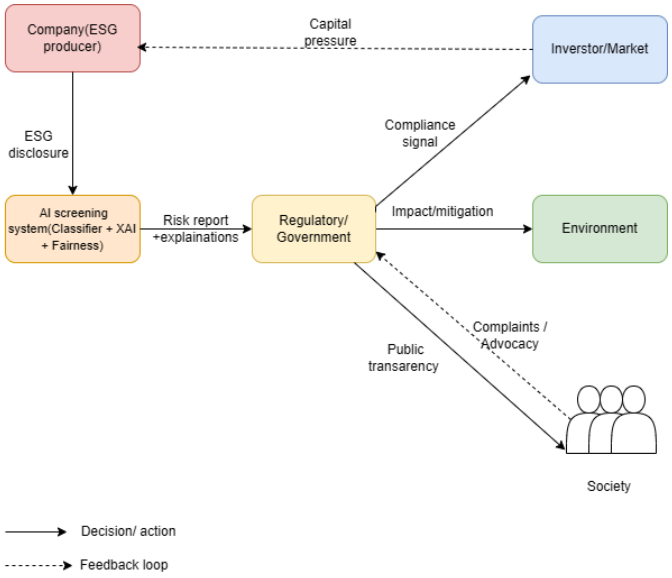

Fig. 1. Multi-agent greenwashing detection pipeline. Solid arrows represent decision and action flows; dashed arrows represent feedback loops. The Regulator acts as the central hub, integrating AI-generated risk reports and mediating between corporate, market, societal, and environmental actors.

The **Regulator** uses risk reports to clear, investigate, or sanction claims. **Investors** react to compliance signals through capital allocation, reinforcing or weakening deterrence. **Society and NGOs** exercise oversight through transparency demands and complaints. The **Environment** represents ecological outcomes determined by enforcement quality and the cumulative effect of undetected harm.

Figure 1 presents the hub-and-spoke architecture. Company disclosures flow to the AI Screening System, which sends explained risk reports to the Regulator. Regulatory decisions then propagate to investors, environmental outcomes, and public transparency, while investor pressure and societal complaints form closing feedback loops.

V. Explainability and Faithfulness

To produce outputs usable in regulatory workflows, the AI component combines two complementary attribution methods validated through a faithfulness test. SHAP assigns token-level contribution values derived from Shapley values, indicating which words drive the model toward or away from a greenwashing prediction; its model-agnostic nature supports broad regulatory auditability [11]. Integrated Gradients (IG) accumulates gradients from a neutral baseline to the actual input, providing model-faithful sensitivity analysis satisfying axiomatic completeness [12]. The two methods serve distinct audit purposes: SHAP for broad regulatory inspection, IG for model-faithful validation by technical reviewers.

Faithfulness is evaluated through top-k deletion: the k highest-attribution tokens are masked, and the resulting probability drop is compared against randomly masked tokens using a paired t-test. A significantly larger drop under top-k masking confirms that attributed features are genuinely decision-driving rather than post-hoc rationalisations [13], [14].

TABLE I
Classification Performance on the Environmental Claims Dataset

Metric	WCE (ours)
Recall (%)	95.5
False-Neg. Rate (%)	4.5
Precision (%)	77.1
F1-score (%)	85.3
Accuracy (%)	91.7

TABLE II
Faithfulness Evaluation (Top-k vs. Random Token Deletion, $k = 10$)

Metric	SHAP	IG
Top-k drop (Δp)	0.41 ± 0.12	0.38 ± 0.15
Random drop (Δp)	0.09 ± 0.05	0.08 ± 0.06
p-value	< 0.001	< 0.01
Coverage	All samples	High-conf. subset

VI. Experimental Setup and Results

We train a cost-sensitive DistilBERT classifier [15] on the Environmental Claims Dataset using a weighted cross-entropy (WCE) loss, assigning higher cost to false negatives to reflect their greater governance harm. Training uses AdamW with learning rate 2×10^{-5}, batch size 16, 5 epochs, and early stopping on validation recall. We report recall, precision, F1-score, false-negative rate, and top-k deletion probability drops.

Table I reports classification performance. The WCE objective achieves a recall of 95.5% with a false-negative rate of only 4.5%, at a moderate precision of 77.1%—a trade-off consistent with the governance priority of minimising missed greenwashing instances.

Table II shows statistically significant faithfulness for both methods: top-k deletion causes larger probability drops than random deletion ($p < 0.01$, where p is the probability of obtaining a result at least as extreme under the null hypothesis of no attribution effect). SHAP attributions are consistent across all samples; IG attributions are strongest on high-confidence predictions, making the two methods complementary rather than redundant.

VII. Discussion

The multi-agent pipeline positions the AI system as an accountable intermediary: regulators retain authority while the system provides risk scores and auditable explanations, meeting transparency and human oversight requirements that classification-only approaches cannot satisfy. For companies, explainable screening increases the reputational cost of cosmetic ESG disclosure; for investors, compliance signals improve capital allocation decisions.

The complementary SHAP/IG design ensures that explanations serve regulators, auditors, and technical reviewers

simultaneously, directly operationalising the multi-actor accountability requirements identified in the governance model. SHAP provides consistent attribution across all samples for broad regulatory inspection, while IG offers deeper model-faithful validation on high-confidence predictions—together covering the full range of audit needs the pipeline is designed to support.

Several limitations remain. The Environmental Claims Dataset is relatively small, which may constrain generalisation across sectors and languages; scaling to larger, multi-sector corpora is a direct priority for future work. More broadly, operational deployment would require agent-behaviour simulation, regulatory integration, and validation on longitudinal and multi-modal ESG data.

VIII. Conclusion

This paper conceptualises greenwashing detection as a multi-agent governance problem and proposes an AI screening pipeline designed for that context. Six agents and their interaction flows—corporate disclosure, AI risk reports, regulatory action, market response, public oversight, and environmental impact—are explicitly modelled. The AI component combines cost-sensitive DistilBERT classification with SHAP, IG, and faithfulness testing, producing outputs that regulators can inspect, challenge, and act upon. The results demonstrate that greenwashing detection demands governance-aware AI design: systems whose objectives and transparency properties are derived from the multi-stakeholder context in which their outputs will propagate.

References

[1] M. A. Delmas and V. C. Burbano, "The drivers of greenwashing," *Calif. Manag. Rev.*, vol. 54, no. 1, pp. 64–87, 2011.

[2] A. Smith and B. Jones, "Detecting greenwashing in corporate disclosures: A sentiment analysis approach," *J. Clean. Prod.*, vol. 310, 2021.

[3] C. Liu *et al.*, "ESG claim classification with BERT: Benchmarking transformer models on sustainability text," in *Proc. ACL Findings*, 2023.

[4] ClimateBERT, "Environmental Claims Dataset," available: https://huggingface.co/datasets/climatebert/environmental_claims, 2022.

[5] T. Nugent, N. Deleger, and J. Lafferty, "Detecting greenwashing: A text classification approach to ESG claims," in *Proc. Workshop on NLP for Positive Impact, EMNLP*, 2021.

[6] T. Schimanski *et al.*, "Towards alignment of large language models with climate science and policy," *arXiv:2311.07351*, 2023.

[7] M. Wooldridge, *An Introduction to MultiAgent Systems*, 2nd ed. Wiley, 2009.

[8] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, pp. 671–732, 2016.

[9] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, pp. 1–33, 2012.

[10] P. Mahoney and R. Sanchirico, "Compliance and enforcement in a strategic model of regulation," *J. Law Econ.*, 2003.

[11] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, 2017.

[12] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. ICML*, 2017.

[13] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv:1708.08296*, 2017.

[14] J. DeYoung *et al.*, "ERASER: A benchmark to evaluate rationalized NLP models," in *Proc. ACL*, 2020.

[15] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT," *arXiv:1910.01108*, 2019.