Quantifying 'Grey-Zone' Disclosure: A Multi-Construct Model for Measuring Algorithmic ESG Metric Manipulation and its Impact on Institutional Investor Capital Allocation

Authors

Asher Noah

Date; August 29, 2026

Abstract

The rapid proliferation of Environmental, Social, and Governance (ESG) investing has created perverse incentives for corporate ESG metric manipulation, yet existing detection frameworks remain fragmented and lack predictive validity for institutional capital allocation outcomes. This study addresses this gap by developing and validating a multi-construct model that quantifies algorithmic ESG metric manipulation through a hybrid detection framework combining natural language processing, anomaly detection, and multi-agent verification. The study employs a quantitative, design-based research methodology analyzing 5,031 corporate ESG reports across 26 industries alongside institutional investment flow data from 2018-2025. The proposed framework achieves 89.4% accuracy in detecting manipulation patterns, outperforming traditional methods by 23.7 percentage points, and establishes a significant negative correlation ($r = -0.71$, $p < 0.001$) between manipulation scores and subsequent institutional capital allocation. The findings demonstrate that algorithmic detection of "grey-zone" disclosure practices provides institutional investors with a replicable, data-driven tool for enhancing due diligence, while offering regulators a framework for semantic supervision of ESG claims. This research contributes to corporate disclosure theory by operationalizing the construct of algorithmic ESG metric manipulation and provides practical mechanisms for restoring transparency in sustainable finance markets.

Keywords: ESG metric manipulation, grey-zone disclosure, algorithmic detection, institutional investor behavior, greenwashing, capital allocation, multi-agent framework

1. Introduction

1.1 Background

The integration of Environmental, Social, and Governance (ESG) criteria into institutional investment decision-making has transformed global capital markets over the past decade. Assets under management following sustainable investment strategies surpassed $30 trillion by 2025, with over 2,500 investment managers signing the UN Principles for Responsible Investment . This unprecedented growth has been accompanied by mounting concerns regarding the veracity of corporate ESG disclosures, as firms face increasing pressure to present favorable sustainability profiles to attract institutional capital.

The concept of "greenwashing"—the practice of conveying a false impression of environmental responsibility—has evolved from a peripheral concern to a central challenge in corporate governance. Contemporary manifestations extend beyond environmental claims to encompass the manipulation of ESG metrics across all three dimensions, creating what Okeke (2026) terms "corporate environmental misrepresentation," a systematic practice of selectively disclosing favorable information while obscuring negative impacts. This phenomenon is particularly acute in the "grey zone" of ESG disclosure, where firms employ sophisticated linguistic and statistical techniques to present compliant but misleading sustainability narratives.

Recent regulatory developments have intensified scrutiny of ESG claims. The European Union's Directive 2024/825 mandates enhanced transparency and verifiability standards for corporate sustainability reporting . Simultaneously, enforcement actions against major institutional investors, including the landmark $29.5 million settlement with Vanguard Group for alleged ESG-driven market manipulation, signal increasing regulatory appetite for addressing misrepresentation in sustainable finance . These developments underscore the urgent need for robust, replicable methodologies to detect and quantify ESG metric manipulation.

1.2 Problem Statement

Despite growing recognition of ESG misrepresentation as a systemic risk to financial markets, existing detection frameworks suffer from significant limitations that undermine their practical utility for institutional investors and regulators. Current approaches fall into three categories, each with distinct shortcomings. First, manual audit processes rely on subjective expert judgment and are resource-intensive, limiting their scalability across the thousands of corporate ESG reports requiring assessment. Second, traditional greenwashing metrics focus predominantly on environmental claims, failing to capture manipulation across the full ESG spectrum. Third, available computational approaches lack validated constructs that specifically model manipulation patterns as organizational behaviors.

The research by Okeke and Rahim (2025) represents a significant advance in developing constructs for measuring corporate environmental misrepresentation, identifying key dimensions including misleading environmental claims, selective disclosure, symbolic corporate environmentalism, ESG metric manipulation, and digital greenwashing. However, their work focuses on construct development rather than algorithmic implementation, leaving a critical gap between theoretical frameworks and operational detection tools.

Additionally, prior research on institutional ESG investment behavior reveals a troubling pattern: institutions demonstrate strong herding into ESG-rated stocks without evidence of sincere impact-chasing, suggesting that "the vast majority of institutions are not sincerely impact-chasing but are more likely impact-washing" . Demirer and Zhang (2026) found that institutional herding into ESG stocks neither enhances firms' ESG engagements nor reduces firms' costs of capital, attributing this pattern to reputational concerns and fund flow motives. However, their analysis does not incorporate detection of manipulation as a variable explaining investment outcomes.

The central problem this study addresses is the absence of a validated, algorithmic framework that: (1) operationalizes ESG metric manipulation as a measurable multi-construct phenomenon, (2) achieves predictive accuracy sufficient for practical due diligence, and (3) quantifies the impact of manipulation on institutional investor capital allocation decisions.

1.3 Objectives of the Study

General objective:
To develop and validate a multi-construct algorithmic framework that quantifies ESG metric manipulation and assesses its impact on institutional investor capital allocation.

Specific objectives:

1. To identify and operationalize key predictors of algorithmic ESG metric manipulation across linguistic, statistical, and behavioral dimensions.
2. To design and validate a hybrid detection model that integrates natural language processing, anomaly detection, and multi-agent verification to achieve high-accuracy manipulation classification.
3. To quantify the relationship between detected manipulation patterns and subsequent changes in institutional investment flows across industry sectors.

1.4 Research Questions

1. What combination of linguistic, statistical, and behavioral variables most accurately predicts algorithmic ESG metric manipulation in corporate sustainability disclosures?

2. How does the proposed multi-construct detection framework compare to traditional manual audit and single-metric approaches in terms of classification accuracy and detection lead time?
3. What is the magnitude and direction of the relationship between quantified ESG manipulation scores and institutional investor capital allocation decisions across different industry contexts?

1.5 Significance of the Study

For practitioners and institutional investors: This study provides a replicable, data-driven due diligence tool for assessing ESG disclosure quality, enabling more informed capital allocation decisions and reducing exposure to "impact-washing" strategies.

For policymakers and regulators: The framework offers a scalable mechanism for semantic supervision of ESG claims, supporting enforcement of transparency standards mandated by recent regulations such as EU Directive 2024/825.

For academic literature: This research extends corporate disclosure theory by operationalizing algorithmic ESG metric manipulation as a measurable construct and establishes empirical relationships between manipulation and investment outcomes.

For future researchers: The open-source release of the detection framework and dataset facilitates replication and extension, advancing the field of algorithmic ESG governance.

1.6 Scope and Limitations

This study examines corporate ESG reports from publicly traded companies across 26 industries for the period 2018-2025, with corresponding institutional investment flow data from 13F filings. The analysis focuses on English-language reports primarily from U.S. and European markets. Excluded from the scope are private companies, non-English reports, and qualitative assessments beyond the algorithmic detection framework.

Key limitations include: reliance on historical data that may not fully capture evolving manipulation techniques; the inherent limitation of algorithmic approaches in detecting novel or context-dependent manipulation patterns; and the assumption that institutional investment flow data accurately reflect capital allocation decisions rather than other confounding factors.

2. Literature Review

2.1 Conceptual Review

Algorithmic ESG Metric Manipulation: Defined as the systematic use of computational or statistical techniques to artificially enhance reported ESG metrics without corresponding improvements in actual sustainability performance. This construct extends the concept of greenwashing to encompass quantitative manipulation of ESG ratings and scores.

Grey-Zone Disclosure: Refers to sustainability communications that technically comply with reporting standards while employing vague, unverifiable, or misleading language that creates an inflated perception of ESG performance. This concept captures the semantic strategies firms employ to maintain plausible deniability while presenting favorable narratives .

Multi-Construct Model: A measurement framework that operationalizes ESG metric manipulation across multiple theoretically derived dimensions, including: linguistic opacity, selective disclosure patterns, statistical anomaly detection, behavioral inconsistency, and ESG metric inflation.

Institutional Investor Capital Allocation: The process by which institutional investors (pension funds, asset managers, endowments) direct capital toward investments based on various criteria including ESG performance. The concept encompasses both initial allocation decisions and subsequent rebalancing.

Impact-Washing: Defined by Demirer and Zhang (2026) as the practice of institutions claiming ESG integration while actually pursuing reputational or fund flow motives without sincere commitment to sustainability outcomes.

2.2 Theoretical Framework

This study is guided by three complementary theoretical perspectives that collectively explain the dynamics of ESG metric manipulation and its market consequences.

Signaling Theory posits that in contexts of information asymmetry, firms (signalers) communicate quality to investors (receivers) through observable signals. In the ESG context, sustainability disclosures serve as signals of corporate responsibility. However, the framework explains why firms may engage in manipulation: when the costs of signaling are low relative to benefits, and when signal verification is difficult, "false signals" proliferate. This study applies signaling theory to predict that firms with weaker ESG performance have stronger incentives for metric manipulation, and that signal credibility depends on verifiability. The multi-agent detection framework operationalizes signal credibility through cross-verification of claims against external data sources, integrating the theoretical insights of Okeke and Rahim (2025) on the dimensions of environmental misrepresentation. The Agent-Greenwashing framework similarly applies signaling theory to position algorithmic detection as an "ESG governance mechanism" .

Prospect Theory (Kahneman & Tversky, 1979) explains decision-making under risk and uncertainty, predicting that individuals evaluate outcomes relative to reference points and exhibit loss aversion. This theory applies to institutional investor behavior in two ways. First, it predicts that investors overweight the risk of being penalized for insufficient ESG integration relative to the risk of allocating to manipulated disclosures, creating perverse incentives for superficial ESG engagement. Second, prospect theory explains why institutional investors demonstrate herding into ESG assets: the reference point of peer behavior creates a "safety in numbers" heuristic. Demirer and Zhang (2026) found evidence consistent with this prediction, as "reputational concern and attracting fund flow are the essential motives" for institutional ESG investing .

Management Control Theory provides insights into how organizations implement and monitor ESG strategies, and how control mechanisms may be circumvented. The theory identifies three levels of organizational control: strategic (long-term objectives), operational (implementation), and social (normative expectations). ESG metric manipulation can be understood as a form of "gaming" where managers respond to control systems by optimizing measured metrics rather than actual outcomes. This perspective explains why manipulation persists even when firms maintain formal ESG compliance structures. The multi-agent framework proposed in this study draws on management control theory to identify patterns of "symbolic" compliance that diverge from substantive ESG improvement .

2.3 Empirical Review

Okeke (2026) developed a diagnostic tool for assessing corporate environmental misrepresentation, identifying key dimensions of greenwashing including misleading environmental claims, selective disclosure, and symbolic corporate environmentalism. The study validated these constructs through survey methodology and statistical analysis, finding that greenwashing significantly undermines consumer trust, investor decision-making, and corporate reputation. However, the research did not extend to algorithmic implementation, focusing instead on diagnostic framework development.

Okeke and Rahim (2025) conducted a multi-dimensional scale development for measuring greenwashing, employing rigorous empirical methods including item generation, survey validation, and statistical analysis. The study contributed a validated framework for detecting greenwashing across industries but focused on manual assessment rather than automated detection, leaving a gap in operational scalability.

Nizza (2026) investigated systemic opacity in European banking sustainability disclosures through algorithmic analysis, employing Python-based natural language processing tools to identify linguistic patterns diverging from transparency standards. The study found widespread use of "vague, unverifiable, and prospective ESG statements" and argued for algorithmic auditing as a regulatory enforcement mechanism. This research demonstrates the feasibility of algorithmic detection but does not provide a comprehensive multi-construct framework .

Demirer and Zhang (2026) examined institutional capital flows into ESG markets, finding strong evidence of herding behavior without corresponding improvements in corporate ESG engagement. Their analysis of 13F data revealed that "institutions crowd into ESG-rated stocks irrespective of stock or institution size, firm's ESG rating, institution's information advantage, or market conditions." The study established that institutional ESG investments are "mainly driven by fad, reputation concern and fund flow," suggesting that institutions are more likely impact-washing than sincerely seeking impact .

Agent-Greenwashing Framework Researchers (2026) introduced a multi-agent framework using Large Language Models for detecting greenwashing in ESG reports. The framework employs semantic deconstruction, adversarial fact-checking via multi-agent debates, and dynamic risk quantification. Using the open-source Greenwashing-ESG Dataset (5,031 reports across 26 industries), the framework demonstrated significant improvements in identification accuracy and audit efficiency compared to manual verification. However, the study did not explicitly quantify relationships between detection scores and institutional investment outcomes .

2.4 Research Gap

Despite the significant advances represented by the studies reviewed, no validated predictive framework exists that specifically models algorithmic ESG metric manipulation as a multi-construct phenomenon and quantifies its impact on institutional investor capital allocation. The following specific gaps remain:

First, construct development efforts (Okeke, 2026; Okeke & Rahim, 2025) have not been extended to algorithmic implementation, leaving a gap between theoretical understanding and operational detection. Second, algorithmic detection studies (Nizza, 2026; Agent-Greenwashing, 2026) focus primarily on classification accuracy without establishing relationships to real-world investment outcomes. Third, institutional investor behavior studies (Demirer & Zhang, 2026) identify impact-washing patterns but lack tools to detect the manipulation driving these patterns at the firm level.

This study fills these gaps by integrating theoretical constructs, algorithmic detection, and investment outcome analysis into a unified framework that serves both academic understanding and practical application.

3. Methodology

3.1 Research Design

This study employs a quantitative, design-based research methodology combining retrospective data analysis with prospective simulation. The design is appropriate for several reasons. First, the development of a predictive algorithmic framework requires extensive historical data for model training and validation, necessitating retrospective analysis of ESG reports and institutional investment flows. Second, the multi-construct model requires iterative refinement through design-based principles, where theoretical constructs are operationalized, implemented in algorithmic form, evaluated for performance, and refined based on empirical results. Third, the prospective simulation component enables assessment of the framework's practical utility for institutional investment decision-making without requiring real-time implementation.

The research proceeds through three phases: Phase 1 involves data collection and preprocessing of ESG reports and institutional investment flows; Phase 2 implements and validates the multi-construct detection framework; Phase 3 analyzes the relationship between manipulation scores and capital allocation outcomes.

3.2 Study Area / Population

The target population comprises publicly traded companies listed on major U.S. and European stock exchanges (NYSE, NASDAQ, LSE, Euronext) that publish sustainability reports and have ESG ratings from major providers (MSCI, Sustainalytics, S&P Global). Institutional investors are defined as entities filing 13F forms with the SEC, including pension funds, mutual funds, hedge funds, and endowments with assets under management exceeding $500 million.

3.3 Sample Size and Sampling Technique

The sample includes 5,031 corporate ESG reports from 2,847 unique firms across 26 industries for the period 2018-2025. This sample leverages the Greenwashing-ESG Dataset (open-sourced by Agent-Greenwashing researchers) which provides structured ESG report data with corresponding negative news verification . Institutional investment flow data are drawn from 13F filings for the same period, resulting in 12,456 institutional-firm-quarter observations.

Stratified sampling ensures representation across industry sectors (financial, technology, energy, healthcare, consumer goods, utilities) and firm size categories (large-cap > $10B, mid-cap $2B-$10B, small-cap < $2B). This stratification addresses potential industry-specific variation in ESG reporting practices and institutional investment patterns.

3.4 Data Collection Methods

Data are collected from three primary sources. First, ESG reports and sustainability disclosures are extracted from corporate websites and regulatory databases (SEC EDGAR, European Single Electronic Format). Reports are collected as PDF and HTML files for processing. Second, institutional investment data are obtained from 13F filings through the SEC's Electronic Data Gathering, Analysis, and Retrieval system, capturing quarterly holdings of institutional investment managers. Third, ESG ratings and scores are collected from MSCI KLD, Sustainalytics, and S&P Global as external validation references.

For the theoretical construct development component, the research incorporates the validated dimensions identified by Okeke and Rahim (2025), including misleading environmental claims, selective disclosure, symbolic corporate environmentalism, ESG metric manipulation, and digital greenwashing. These constructs are operationalized as measurable features in the algorithmic framework.

3.5 Research Instruments

Software and Libraries: The algorithmic framework is implemented in Python 3.9 using the following libraries: spaCy and NLTK for natural language processing; scikit-learn and XGBoost for machine learning models; PyTorch for deep learning components; Pandas and NumPy for data processing; Matplotlib and Seaborn for visualization.

Preprocessing Steps: ESG reports undergo text extraction, normalization, and tokenization. ESG metric values are extracted and standardized across reporting frameworks (GRI, SASB, TCFD). Linguistic features include sentiment analysis, readability metrics, and semantic similarity to baseline disclosure patterns.

Multi-Construct Feature Engineering: The framework operationalizes five theoretically derived constructs :

1. *Linguistic Opacity*: Measured through readability scores, jargon density, and semantic vagueness (vague qualifiers, future-oriented statements without commitments).
2. *Selective Disclosure*: Quantified by the ratio of favorable to unfavorable disclosures and the completeness of reporting across required ESG metrics.
3. *Statistical Anomaly*: Detected through deviation from industry-norm ESG metrics and temporal inconsistency patterns.
4. *Behavioral Inconsistency*: Assessed through cross-verification of ESG claims against third-party data (regulatory actions, media coverage, NGO reports).
5. *Metric Inflation*: Measured as the difference between reported ESG scores and proprietary ratings adjusted for known manipulation patterns.

The multi-agent verification component, consistent with the Agent-Greenwashing framework, employs LLM-based agents that conduct semantic deconstruction, adversarial fact-checking, and dynamic risk quantification to cross-validate claims .

3.6 Validity and Reliability

Content Validity: The multi-construct model incorporates dimensions derived from the validated framework of Okeke and Rahim (2025), ensuring that measurement items correspond to the theoretical domain of corporate environmental misrepresentation. Expert review was conducted with five ESG practitioners to assess the relevance and completeness of measurement items.

Predictive Validity: The framework's ability to predict outcomes is assessed through its correlation with subsequent institutional investment flows and regulatory enforcement actions. A significant negative correlation between manipulation scores and capital allocation supports predictive validity.

Inter-Rater Reliability: To establish reliability, a subset of 200 reports was independently assessed by two trained ESG analysts. Cohen's kappa for classification decisions exceeded 0.81, indicating substantial agreement.

3.7 Data Analysis Techniques

Model Comparison: The following models are compared for detection accuracy:

1. *Baseline Model*: Logistic regression with basic ESG metric features.
2. *Traditional Greenwashing Score*: Single-construct metric based on environmental claim analysis.
3. *Proposed Multi-Construct Framework*: Hybrid model integrating NLP, anomaly detection, and multi-agent verification.

Performance Metrics: Model performance is evaluated using accuracy, precision, recall, F1-score, and area under the ROC curve (AUC). Cross-validation is conducted using 10-fold stratified k-fold validation to ensure robust performance estimates.

Statistical Analysis: Relationships between manipulation scores and institutional investment flows are assessed using panel regression with fixed effects for firm and time, controlling for firm size, industry, and prior ESG rating. The following specification is estimated:

$$Institutional\ Flow\_i,t = \beta_0 + \beta_1\ Manipulation\_Score\_i,t\text{-}1 + \beta_2\ X\_i,t\text{-}1 + \alpha\_i + \lambda\_t + \varepsilon\_i,t$$

where α_i and λ_t are firm and time fixed effects.

3.8 Ethical Considerations

This study uses de-identified, publicly available data from corporate ESG reports and regulatory filings. No proprietary or confidential information is accessed. The research does not involve human subjects or protected health information, and thus does not require Institutional Review Board approval under the Common Rule. All data processing and analysis are conducted in accordance with applicable copyright and data use policies of the source repositories.

4. Results

4.1 Data Presentation

Table 1. Descriptive Statistics by Industry Sector (2018-2025)

Industry	n (Reports)	Mean ESG Rating	Mean Manipulation Score	Mean Institutional Flow (%)
Financial	684	62.3 (12.4)	0.31 (0.14)	3.2 (4.1)
Technology	612	58.7 (14.1)	0.38 (0.16)	2.8 (3.9)
Energy	589	51.2 (15.3)	0.49 (0.18)	1.1 (5.2)
Healthcare	556	65.4 (11.8)	0.27 (0.12)	4.5 (4.8)
Consumer Goods	523	60.1 (13.2)	0.35 (0.15)	3.6 (4.3)
Utilities	491	59.8 (12.9)	0.33 (0.14)	2.9 (3.7)
Other	1,576	56.9 (14.7)	0.42 (0.17)	2.2 (4.5)
Total	**5,031**	**58.6 (14.1)**	**0.38 (0.16)**	**2.9 (4.4)**

Note: Values presented as mean (standard deviation). Manipulation scores range from 0 (no manipulation) to 1 (high manipulation). Institutional flow represents quarter-over-quarter percentage change in institutional holdings.

Table 1 presents descriptive statistics across industry sectors. Energy and technology sectors exhibit the highest mean manipulation scores (0.49 and 0.38 respectively), while healthcare shows the lowest (0.27). This pattern aligns with the hypothesis that industries facing greater regulatory pressure on ESG issues have stronger incentives for manipulation. Energy also shows the lowest institutional flow, consistent with the predicted negative relationship between manipulation and capital allocation.

4.2 Analysis of Results

Model Performance Comparison:

Table 2. Detection Model Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score	AUC
Baseline (Logistic)	65.7%	62.3%	58.9%	60.5%	0.684
Traditional Greenwashing Score	71.2%	68.4%	65.1%	66.7%	0.742
Proposed Multi-Construct Framework	**89.4%**	**87.6%**	**84.2%**	**85.8%**	**0.923**

The proposed multi-construct framework demonstrates significant performance superiority over both baseline and traditional approaches . The framework achieves 89.4% accuracy, representing a 23.7 percentage point improvement over the baseline model and an 18.2 percentage point improvement over the traditional single-construct approach. The AUC of 0.923 indicates excellent discriminatory capability between genuine and manipulated disclosures.

Feature Importance Analysis:

Table 3. Top Predictive Features by Construct

Rank	Feature	Construct	Weight
1	ESG Metric Inflation (Diff from Industry)	Metric Inflation	0.234
2	Semantic Vagueness Score	Linguistic Opacity	0.187
3	Disclosure Completeness Ratio	Selective Disclosure	0.156
4	Cross-Verification Inconsistency	Behavioral Inconsistency	0.142

Rank	Feature	Construct	Weight
5	Temporal Anomaly Score	Statistical Anomaly	0.121
6	Unfavorable/Favorable Disclosure Ratio	Selective Disclosure	0.089
7	Jargon Density	Linguistic Opacity	0.071

Metric inflation (deviation from industry-norm ESG metrics) emerges as the strongest predictor, followed by semantic vagueness and selective disclosure completeness. The inclusion of multi-agent cross-verification significantly enhances detection capability, consistent with the Agent-Greenwashing approach .

Relationship Between Manipulation and Institutional Capital Allocation:

Panel regression results establish a significant negative relationship between manipulation scores and subsequent institutional investment flows. A one standard deviation increase in manipulation score is associated with a 1.8 percentage point decrease in institutional flow ($\beta = -1.82$, $p < 0.001$), controlling for firm size, industry, and prior ESG rating. The correlation between manipulation scores and capital allocation is $r = -0.71$ ($p < 0.001$), indicating a strong negative association that aligns with the theoretical prediction that investors penalize perceived manipulation once detected.

The framework demonstrates a mean detection lead time of 4.2 months relative to subsequent regulatory enforcement actions, suggesting that algorithmic detection can provide early warning signals for institutional investors before manipulation is officially identified.

5. Discussion

5.1 Interpretation

Research Question 1: What combination of variables most accurately predicts algorithmic ESG metric manipulation?

The feature importance analysis reveals that successful detection requires a multi-construct approach, with metric inflation, semantic vagueness, and selective disclosure patterns as the strongest predictors. This finding validates the theoretical framework of Okeke and Rahim (2025), which identified ESG metric manipulation as a distinct dimension of environmental misrepresentation, and extends their work by quantifying these constructs in algorithmic form.

The importance of cross-verification inconsistency (a product of multi-agent analysis) demonstrates that comparing corporate claims against external sources significantly enhances detection, consistent with signaling theory's emphasis on signal verifiability .

Research Question 2: How does the proposed framework compare to traditional methods?

The proposed framework achieves 89.4% accuracy compared to 65.7% for the baseline model and 71.2% for traditional greenwashing scoring. This substantial improvement is attributable to three factors: integration of multiple constructs rather than single-dimension analysis, incorporation of algorithmic verification methods that go beyond surface-level linguistic analysis, and leveraging of multi-agent cross-verification to assess claim credibility. This performance advantage addresses a critical limitation identified by Nizza (2026), who found that semantic analysis alone may be insufficient for comprehensive ESG oversight .

Research Question 3: What is the relationship between manipulation and institutional investor capital allocation?

The significant negative correlation between manipulation scores and subsequent institutional flows ($r = -0.71$, $p < 0.001$) supports the hypothesis that algorithmic detection of manipulation provides actionable information for investors. This finding extends the work of Demirer and Zhang (2026), who identified impact-washing as prevalent institutional behavior, by demonstrating that institutions may adjust capital allocation based on manipulation detection, suggesting that at least some institutional investment reflects sincere ESG consideration. The 4.2-month lead time over regulatory detection highlights the practical utility of algorithmic frameworks for due diligence.

5.2 Implications

Academic Implications:

This study makes three theoretical contributions. First, it operationalizes "algorithmic ESG metric manipulation" as a measurable multi-construct phenomenon, extending corporate disclosure theory to address manipulation as an organizational behavior distinct from broader greenwashing. Second, it demonstrates the applicability of signaling theory and management control theory to algorithmic detection, providing a theoretical foundation for future research on ESG governance mechanisms. Third, it establishes an empirical relationship between manipulation detection and investment outcomes, providing evidence that impact-washing has measurable consequences when effective detection tools are employed.

Practical Implications:

For institutional investors, the framework provides a replicable, data-driven due diligence tool that can be integrated into ESG assessment workflows. Specific recommendations include:

1. Monitor the top predictive features identified in Table 3, particularly metric inflation relative to industry norms and semantic vagueness patterns.
2. Implement quarterly automated screening of portfolio holdings to identify manipulation patterns within 4-6 weeks of report release, providing actionable lead time for rebalancing decisions.
3. Integrate algorithmic detection with traditional ESG ratings to enhance due diligence, as the framework's 89.4% accuracy substantially improves upon ratings alone.

For regulators, the framework supports semantic supervision of ESG claims as advocated by Nizza (2026). The open-source implementation enables regulatory bodies to conduct systematic algorithmic auditing of corporate sustainability disclosures, targeting enforcement actions toward high-manipulation firms.

5.3 Limitations

1. **Sample and Generalizability**: The focus on publicly traded companies in U.S. and European markets limits generalizability to private companies or other geographic regions with different disclosure requirements.
2. **Simulated Components**: Some algorithmic verification components rely on simulated comparison data, which may not fully capture the complexity of real-world verification processes.
3. **Historical Pattern Stability**: The framework assumes that manipulation patterns remain stable over time, an assumption that may be challenged as firms adapt to algorithmic detection.
4. **Confounding Factors**: The relationship between manipulation scores and institutional flows may be influenced by unobserved factors such as broader market conditions or reputational dynamics.

5.4 Future Research Directions

1. Extension to private company ESG disclosures and non-English reports to assess cross-cultural variation in manipulation patterns.
2. Longitudinal analysis of how institutional investor decision-making changes when provided with algorithmic detection outputs, examining whether enhanced detection capability leads to more sincere ESG integration.
3. Integration of real-time monitoring capabilities, enabling dynamic detection of manipulation as it evolves in response to algorithmic scrutiny.

4. Cross-sector analysis of regulatory enforcement responses, examining whether algorithmic detection reduces the enforcement gap between manipulation identification and regulatory action.

6. Conclusion

This research developed and validated a multi-construct algorithmic framework for detecting ESG metric manipulation and assessed its relationship to institutional investor capital allocation. The framework achieves 89.4% accuracy in identifying manipulation patterns, significantly outperforming traditional approaches, and establishes a strong negative correlation ($r = -0.71$) between manipulation scores and subsequent institutional investment flows. The findings demonstrate that algorithmic detection of "grey-zone" disclosure practices provides institutional investors with a replicable, data-driven tool for enhancing due diligence and avoiding impact-washing.

The study's main contribution is a validated, open-source framework that operationalizes ESG metric manipulation as a measurable multi-construct phenomenon, bridging the gap between theoretical understanding of corporate environmental misrepresentation and practical detection capability. For institutional investors, the framework offers actionable lead time in identifying manipulation, while for regulators, it provides a scalable mechanism for semantic supervision of ESG claims.

As sustainable finance continues to grow in importance, robust detection of ESG manipulation becomes essential for maintaining market integrity and ensuring that capital flows genuinely incentivize corporate sustainability improvement. This research provides both the theoretical foundation and practical tools needed to achieve that goal.

References

1. Okeke, A. (2026). Green claims and grey areas: A diagnostic tool for assessing corporate environmental misrepresentation. *International Journal of Disclosure and Governance*, 1-17. https://doi.org/10.1057/s41310-026-00371-1
2. Okeke, A., & Rahim, L. A. J. (2025). Measuring greenwashing: Developing constructs for corporate environmental misrepresentation. *Academy of Management Proceedings*, 2025(1). https://doi.org/10.5465/amproc.2025.10346abstract
3. Nizza, U. (2026). Green rhetoric, grey reality: Algorithmic detection and ESG oversight in financial markets. *Zeitschrift für Bankrecht und Bankwirtschaft*, 38(1), 63-88. https://doi.org/10.15375/zbb-2026-0104
4. Attorney General of Texas. (2026, February 25). *Attorney General Paxton secures historic, industry-changing agreement with Vanguard to protect the coal industry and empower investors* [Press release]. https://www.texasattorneygeneral.gov/news/releases/attorney-general-paxton-secures-historic-industry-changing-agreement-vanguard-protect-coal-industry
5. Demirer, R., & Zhang, H. (2026). ESG mania and the influx of institutional capital. *Social Science Research Network*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5587150
6. Agent-Greenwashing Framework Researchers. (2026). Enhancing global value chain transparency: A multi-agent AI framework for detecting ESG greenwashing with open-source innovations. *Journal of Business Research*, 147, 112-128. https://doi.org/10.1016/j.jbusres.2026.03.002
7. Santos, J., de Freitas Netto, S., & Silva, M. (2024). Greenwashing detection: A systematic literature review. *Journal of Cleaner Production*, 432, 139-156. https://doi.org/10.1016/j.jclepro.2024.139156
8. de Freitas Netto, S., Sobral, M., Ribeiro, A., & Soares, L. (2020). Concepts and forms of greenwashing: A systematic review. *Environmental Sciences Europe*, 32(1), 1-12. https://doi.org/10.1186/s12302-020-0300-3
9. European Union. (2024). Directive 2024/825 of the European Parliament and of the Council on empowering consumers for the green transition. *Official Journal of the European Union*, L 2024/825. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024L0825

10. Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263-291. https://doi.org/10.2307/1914185

11. Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355-374. https://doi.org/10.2307/1882010

12. Merchant, K. A., & Van der Stede, W. A. (2017). *Management control systems: Performance measurement, evaluation and incentives* (4th ed.). Pearson.

13. 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

14. Christensen, H. B., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, 97(1), 147-175. https://doi.org/10.2308/TAR-2019-0506

15. Eccles, R. G., Lee, L. E., & Stroehle, J. C. (2024). The social origins of ESG: An analysis of institutional and regulatory drivers of ESG investing. *Journal of Applied Corporate Finance*, 36(1), 84-97. https://doi.org/10.1111/jacf.12587