Journal of Marketing & Social Research

ISSN (Online): 3008-0711

Volume: 02 | Issue 10 | 2025

Journal homepage: https://jmsr-online.com/

Research Article

Explainable AI Applied to Sustainable Consumption Decisions: Analysis of Regulatory Compliance in the European Union, as an Imperative for Trust and Transparency.

Rafael Canorea-García

ESIC University, Madrid,Spain

Received: 28/09/2025; Revision: 20/10/2025; Accepted: 08/11/2025; Published: 05/12/2025

***Corresponding author: Rafael Canorea-García**

Abstract: Explainable Artificial Intelligence (XAI) has established itself as an essential functional and ethical requirement for recommendation and decision support systems to effectively promote sustainable consumption without falling into the trap of algorithmic opacity. Recent empirical evidence, in contrast to traditional "black box" models, demonstrates that adequate explanations not only cultivate user trust, but are a direct factor in improving performance in the execution of decision tasks and willingness to follow recommendations (Senoner et al., 2024). However, the effectiveness of this technology is conditioned by the fidelity, conciseness, and action provided by the explanation (Senoner et al., 2024). At the technical level, the uncritical use of widely disseminated model-agnostic methods, such as SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations), poses substantial challenges related to the instability and collinearity of variables, which demand a methodological approach that incorporates caution and triangulation (Huang et al., 2024; Letoffe et al., 2025). In the European regulatory context, the confluence of the AI Act and the CSRD/ESRS corporate reporting guidelines establish explicit transparency and traceability obligations, linking algorithmic explainability directly to the organization's governance and accountability standards.

Keywords:

INTRODUCTION

Algorithmic Mediation and the Trust Imperative

Consumer decision-making is at a stage of deep algorithmic mediation. Artificial Intelligence (AI) systems are no longer mere search engines, but active prescribers that shape the set of alternatives available in domains crucial to sustainability, such as mobility, e-commerce or domestic energy management.

In this digitalized environment, the promotion of sustainable consumption patterns faces the problem of algorithmic opacity. When the rationale for a "green" alternative is subsumed into a "black box" algorithm, the inability to disclose the least impactful factors undermines two essential pillars: user trust and effective use of the recommended options. The consumer needs to know why a specific product has a lower environmental impact, whether this assessment is based on carbon footprint, Life Cycle Assessment (LCA) or a verifiable certification. The XAI emerges as the necessary bridge between the technical complexity of sustainability assessment and the user's cognitive need for justification. The most recent literature confirms that XAI is a factor that increases the performance and perceived usefulness of decision support systems, although its final effect on trust is intrinsically linked to the design and quality of the explanation offered.

CONCEPTUAL FRAMEWORK: LITERATURE REVIEW.

Transparency, Trust and Signage in Sustainability

The theoretical articulation of XAI in the sustainable field is based on a triad of interdependent constructs: transparency, trust and the management of the sustainable trade-off.

Algorithmic Transparency and Cognitive Load Management

Algorithmic transparency is defined as the degree of access to information that allows the user to understand how the AI system arrives at a specific result and what its operating limits are. In the sustainable context, this implies a clear disclosure about the sources of impact data and the weighting methodology employed.

However, the ingenue implementation of transparency can be counterproductive. Research on user behavior warns that poorly dosed transparency, due to excess or technical complexity, generates cognitive overload, which not only does not increase the desired behaviors, but can lead to rejection of the system or poorly calibrated trust (Afroogh et al., 2024). Therefore, the most effective strategy is transparency signaling (Park & Kim, 2024), which consists of offering clear and concise indicators that the system is operating fairly and objectively, reserving methodological detail for deeper layers of information. This need for selective transparency is more pressing in issues of high ethical or environmental implication, where the user perceives a greater risk if the information is incorrect or biased (Park, 2025).

Trust and Appropriate Reliance

Trust is the expectation that the system will behave in a manner that is reliable, fair, and aligned with the user's interests and values. In the context of sustainable consumption, trust is stratified into two critical layers that need to be addressed by the XAI:

Trust in the System: It is related to the fidelity of the explanation and the belief that the "green" label or sustainability ranking is the result of an objective and not manipulative calculation.

Trust in the Organization: It focuses on **corporate governance** and regulatory compliance, i.e., the perception that the responsible company has an internal control system and traceability that supports the sustainability promise.

The ultimate goal of XAI is to achieve appropriate user dependence (reliance) (Kahr et al., 2024). A system that generates a good explanation allows the user to identify **biases or errors** in the recommendation, leading to better task performance by avoiding both **overconfidence** (blindly following an erroneous recommendation) and **underconfidence** (ignoring a valid recommendation, despite evidence).

XAI in Sustainable Consumption Support Systems: Operational Effects and Methodological Challenges:
The implementation of XAI has demonstrated a direct and measurable impact on the operational effectiveness of decision support systems, although it faces intrinsic stability challenges.

Impact of Explainability on Task Performance
XAI not only improves the perception of transparency, but is a direct performance factor. In controlled experiments simulating decision tasks in high-criticality settings (such as manufacturing or the clinic), it has been documented that the inclusion of XAI can raise task performance by 4.6 to 7.5 percentage points (p.p.) compared to black box systems (Senoner et al., 2024). This increase is attributed to the user's ability to validate or correct algorithmic advice, thus reaching an optimal level of system dependency (Kahr et al., 2024). XAI, therefore, transforms the user from a passive receiver to a critical evaluator of information.

Methodological Challenges: Instability, Colinenearity and Fidelity of Explanation
Despite their usefulness, the unsupervised use of popular explainability methods has serious technical limitations.

Instability and collinearity in SHAP/LIME: *Model-agnostic* techniques such as SHAP and LIME, which seek to assign the importance of a feature to the prediction, have been shown to be unstable in the face of small perturbations in the dataset or changes in the architecture of the underlying model (Huang et al., 2024). Instability is exacerbated in high collinearity scenarios, common in consumption models (e.g., where price and material quality are correlated). In these cases, the attribution of importance by SHAP or LIME can assign the weight of one variable to another, generating explanations that are **unfaithful** to the internal mechanism of the model (Letoffe et al., 2025).

Risk Mitigation: To mitigate these deficiencies, academic practice requires the triangulation of explainability methods and the performance of sensitivity tests to ensure the robustness of the explanations in the face of data disturbances. In addition, "explanationism" (the generation of convincing but false explanations) should be avoided, ensuring that the explanation to the user is always linked to the technical fidelity of the model, even if this involves showing bands of uncertainty about the importance of the feature.

Design of "Green" Action-Oriented Explanations and Traceability
The design of XAI in the sustainable context must transform technical information into actionable and contextualized knowledge for the user.

Actionable Counterfactuals and Contextual Relevance
An effective explanation goes beyond a list of factors and must answer the question: "what to change?" to achieve a more sustainable result. Counterfactual explanations indicate the minimum modification necessary in the user's input variables (e.g., mobility habits or characteristics of the product to be purchased) to achieve the desired result, increasing the user's sense of control and ability to act. The field of recommender systems (SR) is actively developing counterfactual evaluation metrics to ensure the correctness and usefulness of these explanations (Baklanov, 2024). On the other hand, the explanation must have contextual relevance to the individual. Presenting the why sustainable with clear and comparative indicators (e.g., "vs. your weekly average") improves the usefulness and perceived understanding of the recommendation (Felfernig et al., 2023).

Minimum Cognitive Load and Methodological Traceability
To avoid information overload, the design of the XAI should adopt the principle of progressive disclosure, presenting information in hierarchical layers (Muralidhar et al., 2025).

- **Layer 1:** The shortest and most relevant explanation (the green label).
- **Layer 2:** An expanded dashboard with key drivers and impact metrics.
- **Layer 3:** The underlying methodology and data source. This approach ensures accessibility for the average user, while allowing for auditing and verification by the expert user.

In addition, in an environment of regulatory compliance, methodological traceability is crucial. Explanations to the user should align with the CSRD/ESRS documentation that the organization must report externally, explaining data sources, emission factors, and model limits, ensuring that the sustainability promise is supported by verifiable evidence.

Governance and Compliance (AI Act and CSRD)
Regulatory convergence in the European Union has positioned XAI as a fundamental pillar of corporate

governance in the digital age.

The AI Act and the Requirement of Human Supervision
The AI Act introduces a risk-based framework of requirements for AI. Sustainable consumption recommendation systems that can have a significant impact on decision-making or access to resources could be categorized as high risk. For these systems, the Act imposes strict obligations:

Provision of complete technical documentation.
Implementation of continuous risk management systems.
Provision of clear information to the user about the interaction with the system.

Requirement for adequate human supervision, which requires AI to be transparent and explainable enough for a human to monitor its operation and correct errors or biases. The implementation of this regulation has been designed in a phased manner between 2025 and 2027, and the EU institutions have ratified the intention to maintain the implementation timeline.

SRD/ESRS and Traceability Audit
The **Corporate Sustainability Reporting Directive (CSRD)** and the **European Sustainability Reporting Standards (ESRS)** require large companies to disclose **verifiable information** about their sustainability impacts and the **methodologies** used to calculate them. This audit requirement extends the scope of explainability from the front-end (user interface) to the back-end (data and model governance). When a digital product prioritizes "green" options through AI, the company has the obligation to **harmonize** the explanation offered to the consumer with the **audited evidence** that it reports externally. The XAI, therefore, becomes an essential mechanism for **internal control and accountability**, ensuring that the algorithmic output is consistent with the company's formal sustainability documentation.

METHODOLOGICAL ANALYSIS.

XAI has proven its value in critical sectors for climate change mitigation:

Energy and Buildings: XAI frameworks have been developed for automated energy management in buildings, offering understandable savings recommendations and generating consistent savings results. Interpretability is key to acceptance and continuous use by operators or owners (Teixeira et al., 2025; Amangeldy et al., 2025).

Product and Itinerity Recommendation: In commerce and mobility, XAI increases perceived accuracy and confidence as long as the explanation is contextual and actionable, which goes beyond a static list of factors (Felfernig et al., 2023; Senoner et al., 2024).

Holistic XAI Assessment Metrics
To assess the effectiveness of XAI in promoting sustainability, metrics that encompass the full spectrum of human-AI behavior are required:

Understanding and Perceived Usefulness: Measured through **objective tests** and self-reports to verify that the user has assimilated the logic of the recommendation.

Trust and Appropriate Reliance: Assessment of the user's trust and, crucially, their **level of dependency**, to ensure that neither overconfidence nor underconfidence occurs (Kahr et al., 2024).

Sustainable Action and Behavior: Final metric that measures the adoption of the sustainable option, persistence in the behavior (e.g., continuous energy savings) and the acceptance of the trade-off (e.g., accepting a higher cost or inconvenience in favor of environmental impact).

Fidelity and Robustness of the Explanation: Technical Evaluation of the Stability of the Explanation in the Face of Disturbances and the Validity of Counterfactual Evaluations (Baklanov, 2024; Letoffe et al., 2025).

CONCLUSIONS AND FUTURE CHALLENGES.

Explainable Artificial Intelligence is an essential catalyst for encouraging sustainable consumption, as it transforms opaque algorithmic recommendations into informed and justified decisions. Its potential lies in the ability to activate justified trust (Felfernig et al., 2023) by faithfully explaining the whys and wherefores of sustainability and indicating what the user can do to achieve a positive impact, all under effective cognitive load management.

The challenge is no longer about whether to "explain more", but on "explain better". This implies: 1) guaranteeing the technical fidelity of the explanation, especially in the face of the risks of collinearity in model-agnostic methods; 2) adopt a user-centered design that uses layers of information (progressive disclosure) and actionable counterfactuals; and 3) ensure full traceability of data and models that support "green" decisions. The emergence of the AI Act and the CSRD in the European Union make explainability a mandatory component of internal control and accountability, unifying ethical, technical and legal imperatives in the field of sustainable corporate governance.

REFERENCES:

1. Afroogh, S., Tzeng, G., & Yates, J. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, *11*(1), 1-30.
2. Amangeldy, B., Kalybekova, A., Tlegenov, Y., & Lee, J. (2025). A review of artificial intelligence and deep learning applications in building energy optimization. *Buildings*, *15*(15), 2631.
3. Baklanov, M. (2024, October). CEERS: Counterfactual Evaluations of Explanations in Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 1323-1329).
4. Felfernig, A., Wundara, M., Reiterer, S., & Reiterer, T. (2023). Recommender systems for

sustainability: Overview and research issues. *Frontiers in Big Data*, *6*, 1284511.

5. Huang, X., & Marques-Silva, J. (2024). On the failings of Shapley values for explainability. International Journal of Approximate Reasoning, 171, 109112.
6. Kahr, P. K., Holzer, A., Pöttler, T., & Seifert, C. (2024). Understanding trust and reliance development in AI advice. Proceedings of the ACM on Human-Computer Interaction, 8(CSCW2), 322.
7. Létoffé, O., Huang, X., & Marques-Silva, J. (2025, April). Towards trustable SHAP scores. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 17, pp. 18198-18208).
8. Molhova, M., & Biolcheva, P. (2023). Strategies and policies to support the development of AI Technologies in Europe. Strategies for Policy in Science & Education/Strategii na Obrazovatelnata i Nauchnata Politika, 31.
9. Muralidhar, D. Operationalizing Selective Transparency Using Progressive Disclosure in Artificial Intelligence Clinical Diagnosis Systems. Available at SSRN 5062641.
10. Park, K., & Yoon, H. Y. (2024). Beyond the code: The impact of AI algorithm transparency signaling on user trust and relational satisfaction. Public Relations Review, 50(5), 102507.
11. Park, K., & Young Yoon, H. (2025). AI algorithm transparency, pipelines for trust not prisms: mitigating general negative attitudes and enhancing trust toward AI. Humanities and Social Sciences Communications, 12(1), 1-13.
12. Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., & Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. Scientific reports, 14(1), 31150.
13. Teixeira, B., Carvalhais, L., Pinto, T., & Vale, Z. (2025). Explainable AI framework for reliable and transparent automated energy management in buildings. Energy and Buildings, 116246.