Department of Business and Management
Master's in Management
Chair of Organizational Design

"Trust in AI-Driven vs. Expert-Generated Sustainability Metrics: A Quantitative Experimental Study of Biases in Managerial ESG Decisions"

Prof. Cinzia Calluso
SUPERVISOR

Prof. Fabian Kurt Falk Homberg
CO-SUPERVISOR

Emanuele Pujia
ID: 790431
CANDIDATE

Academic Year 2024-2025

Table of Contents

1. Introduction 4
2. Literature Review 13
   2.1 ESG and Corporate Sustainability 13
   2.2 AI in Sustainability Reporting 16
   2.3 Cognitive Biases in Decision-Making 20
   2.4 Trust and Decision-Making in Managerial Contexts 27
   2.5 Research Gap 32
3. Experimental Study 34
   3.1 Setting 34
   3.2 Participants 34
   3.3 Materials and Procedure 35
   3.4 Statistical Analysis 40
   3.5 Results 42
4. Discussion 60
   3.1 Main Findings 60
   3.1 Theoretical and Managerial Implications 61
   3.3 Limitations 63

Abstract

As AI tools increasingly shape sustainability reporting, understanding how managers trust and use AI-driven versus expert-generated ESG (Environmental, Social, and Governance) metrics is key. This study looks at whether and how managerial trust differs based on the source of ESG performance information and the cognitive biases and individual attitudes that may influence these perceptions. Using theories of algorithm aversion, automation bias and confirmation bias we designed a between-subjects online experiment with 60 professionals experienced in ESG or managerial decision making. Participants were randomly assigned to evaluate an identical ESG report attributed either to a human expert or to an AI system. They then completed a budget allocation task and rated their confidence, perceived accuracy, perceived influence and trust (credibility) in the report. Measures of individual attitudes towards AI and ESG were also collected.

We used a moderated mediation model to test 8 models across 4 dependent variables (similarity to suggested allocation, confidence, accuracy, influence) and 2 moderators (AI attitude and ESG attitude). Results showed no significant effects of the report's source label (AI vs expert) on decisions or trust and no evidence of moderated mediation. But strong and consistent patterns emerged: participants' individual attitudes towards AI and ESG predicted their perceived credibility of the information which in turn predicted their confidence, accuracy and influence ratings.

These findings suggest that managerial trust in ESG information is driven more by individual predispositions than by whether the metrics are labelled as AI- or expert-generated. This has implications for AI adoption in sustainability reporting: simply introducing AI tools won't change decision behaviour unless managers' attitudes and trust are aligned. Limitations include the small sample size and single-exposure design but the results provide a good foundation for future research on responsible AI use in sustainability contexts.

1. Introduction

In recent years, corporate leaders and stakeholders have put Environmental, Social, and Governance (ESG) performance at the heart of corporate strategy and reporting. Companies worldwide are under pressure to deliver financial results and show positive ESG outcomes in areas like carbon footprint, labour practices, diversity and ethical governance. This is driven by growing public awareness of sustainability issues and investors integrating ESG into their decision making. Surveys show that over 85% of investment professionals now consider ESG factors when investing (Mazzacurati, 2021) and most companies track ESG metrics to inform their strategic decisions (Mazzacurati, 2021). In practice, this has led to widespread adoption of sustainability reporting standards and frameworks and the rise of ESG rating agencies that evaluate companies beyond financial metrics (Giudici & Wu, 2025).

Traditionally, ESG performance has been the domain of human experts and analysts. Companies have relied on expert generated sustainability metrics, for example ESG scores and ratings produced by analysts at rating agencies, consulting firms or internal sustainability teams. These expert-generated metrics synthesize data (e.g., emissions data, board diversity stats, employee surveys) through human judgement and established methodologies to produce scores. They are often the basis for sustainability reports, investment analysis and benchmarking against industry peers. However, there is significant variability and subjectivity in expert ESG ratings. Studies have found low correlation between ESG ratings from different providers (Giudici & Wu, 2025) and critics argue that rating methodologies can be *opaque and inconsistent* (Mazzacurati, 2021). One well known example is the divergence in ratings for a high-profile company like Tesla: some agencies give it a top ESG grade while others give it a poor rating (Mazzacurati, 2021). This has raised questions around which expert ratings to trust, and it highlights a broader challenge: ESG performance measurement is complex, context dependent and open to interpretation.

Alongside these developments in sustainability reporting, the past decade has seen rapid progress in artificial intelligence (AI) and data analytics, opening new possibilities for ESG measurement. Companies and researchers are increasingly using AI to collect, process and analyse large amounts of sustainability related data (Xiao & Xiao, 2025). AI driven sustainability metrics refer to scores or evaluations produced by machine learning algorithms, models that analyse both reported and alternative data (news articles, social media sentiment, satellite imagery) to assess a company's ESG profile. These AI systems can scan and aggregate information at a speed and scale beyond human capability, enabling more real time and granular ESG monitoring (Sustainability Directory, 2025). Industry observers note that AI powered ESG data platforms are transforming reporting by streamlining data collection and analysis (Centre for Sustainability and Excellence, 2025). Even traditional rating agencies have started to integrate AI tools: for example, some ratings firms use natural language processing to evaluate company disclosures or use algorithms to ensure consistency in scoring. The broader organizational trend of digital transformation is intersecting with sustainability management: as companies adopt AI for finance, marketing and operations they are also exploring AI for ESG analytics and reporting.

This sets the stage for synergies and tensions between expert generated and AI driven sustainability metrics. On one hand, AI driven metrics promise greater objectivity, breadth of coverage and timeliness. AI can uncover patterns and insights that human analysts might miss and can help standardise measurements across companies (Giudici & Wu, 2025). On the other hand, the introduction of AI into ESG evaluation raises important questions around transparency, accountability and trust. The algorithms used are often complex "black boxes" and their decision processes may not be completely explainable (White, 2023). Without explainability, managers may find it difficult to trust and act on AI generated ESG assessments even if the data processing is rigorous. Meanwhile expert generated metrics benefit from human judgement and domain expertise but are not free from bias or inconsistency (Mazzacurati, 2021). In summary, companies are now navigating two paradigms for sustainability metrics, one driven by human expertise and one by AI,

each with its own strengths and limitations. How these paradigms are perceived and trusted within managerial decision making is increasingly important in these technological and organisational shifts.

As companies use ESG metrics in decision making a critical question has arisen: *Do managers trust AI driven sustainability metrics as much as those generated by human experts?* This is key because trust will determine whether and how managers use the information from these metrics in strategic decisions. If managers don't trust AI generated ESG data, they may not use the insights and make suboptimal decisions or continue to rely on traditional analysis. If managers over trust AI metrics without proper scrutiny they could be misled by algorithmic biases or errors. The problem is that differing trust levels in AI based vs expert based ESG metrics will lead to different decision outcomes, and these trust levels themselves can be influenced by various cognitive biases.

Research in related fields suggests that humans are sceptical of algorithmic decisions. For example, in areas like hiring or financial forecasting people show "algorithm aversion" and tend to avoid or undervalue advice from algorithms after seeing them making mistakes (Logg, Minson, & Moore, 2019). In a recent experimental study on decision makers' reactions to algorithmic versus human outputs in a work setting, participants reported significantly higher trust and acceptance for human generated decisions compared to algorithmic ones (Wesche et al., 2022). People prefer human judgment even when algorithms outperform humans in accuracy (Dietvorst, 2017). Psychologists argue that people are uncomfortable with machines having control and are quick to lose confidence in an algorithm after an error, whereas they might be more forgiving of human error (Dietvorst, Simmons, & Massey, 2015). This could cause managers to default to expert assessments simply because they come from a human source, even if an AI metric is equally or more reliable. Indeed, algorithm aversion has been documented to the extent that companies sometimes disguise algorithmic recommendations as human to encourage acceptance (Logg, Minson, & Moore, 2019).

On the other hand, recent work also shows this is not a universal rule; under certain conditions, decision makers may exhibit "algorithm appreciation". When tasks are seen as data intensive, objective or time

pressured individuals can prefer algorithmic advice over human advice (Logg, Minson, & Moore, 2019). For example, Logg et al. (2019) found that across various forecasting tasks people prefer advice from algorithms to advice from people, suggesting that trust in AI can be high when the algorithm's superior data handling capability is salient (Logg, Minson, & Moore, 2019). This means that managers' trust in AI driven ESG metrics might increase if the metrics are seen as highly analytical (e.g., crunching large data for carbon emissions) or if managers have prior positive experience with data analytics. So, a nuanced view is required; managers' trust in AI vs expert metrics will depend on context and is mediated by psychological biases and perceptions. Notably, automation bias (the tendency to rely too much on automated decisions) could cause some managers to overweight AI generated metrics, especially if they assume the AI is objective and error-free. Conversely, confirmation bias might cause a manager to trust whichever source (AI or expert) aligns with their pre-existing belief about a company's ESG performance rather than evaluating each on merit.

The intersection of cognitive bias, trust and ESG decision making is underexplored in academic literature. While there is a rich body of research on cognitive biases in organizational decisions and a growing literature on trust in AI systems (Glikson & Woolley, 2020), few studies have examined these issues in the specific context of ESG metrics. ESG decision making has unique characteristics, it involves ethical considerations, uncertainty and often a mix of quantitative and qualitative information, which could influence how managers respond to AI vs human inputs. Some recent conceptual works note that biases can significantly hinder managers from making sustainable choices (Palmucci & Ferraris, 2023), but these works do not directly address the role of AI. Moreover, corporate practice is advancing faster than research in this area: organizations are deploying AI tools for sustainability reporting, yet we lack empirical evidence on how managers perceive and trust the outputs. This study addresses that gap by focusing explicitly on managerial trust in AI driven vs expert generated sustainability metrics and by investigating the cognitive biases that may lead to differential trust or decision biases in using those metrics. The research problem can thus be summarized as follows: *Managers may trust AI generated ESG metrics differently than expert*

generated ones and cognitive biases could be influencing these trust levels and consequently ESG related decisions. Understanding this problem is crucial as misaligned trust (either too little or too much) in either source of metrics could bias strategic decisions about sustainability initiatives, investments or disclosures.

Given the above problem statement and identified gap, this thesis sets out clear objectives and research questions. The primary objective of the study is to examine whether there are significant differences in trust and acceptance of AI driven versus expert generated sustainability metrics among managers, and to determine how any such differences affect decision making in the ESG domain. A closely related objective is to identify the cognitive biases that may mediate or moderate managers' trust in these metrics, thereby influencing their decisions. By achieving these objectives, the study aims to contribute both theoretical insights and practical guidance on the use of AI in ESG reporting.

To achieve these goals the study will address the following research questions (RQs):

RQ1 (Main Research Question): *Do managers trust AI driven ESG performance metrics more or less than expert generated ESG metrics when making sustainability related decisions?*

This question investigates the baseline trust differential. It will be tested by measuring managers' trust (or confidence) in a given sustainability metric, varying the source (AI vs human expert) in an experimental setting. The expectation based on literature is that there might be an inherent trust gap, potentially favouring expert metrics due to algorithm aversion (Wesche et al., 2022), or in some cases favouring AI metrics if they are seen as more data driven (Logg, Minson, & Moore, 2019).

RQ2 (Moderation & Decision-Making Question): *How do managers' attitudes toward ESG and toward AI moderate the effects of cognitive biases (e.g., algorithm aversion, automation bias, confirmation bias) on their trust in and use of AI-driven versus expert-generated sustainability metrics, and how do these moderated relationships influence ESG decision outcomes?*

This question will explore how cognitive biases (algorithm aversion, automation bias and confirmation bias) interact with managers' personal views on ESG and on AI to shape their trust in and use of AI-driven versus expert-generated sustainability metrics and how those interactions play out in real decisions. For example, we will test whether managers with a more positive view of AI have reduced algorithm aversion, i.e. do they trust an AI metric after it makes a small mistake, while those with a lower AI affinity punish the AI more harshly than they would a human expert's mistake. We will also look at whether a strong pro-ESG stance amplifies confirmation bias, so managers favour the metric source (AI or expert) that confirms their prior beliefs about a company's sustainability. And by comparing choices, for example the percentage of a sustainability budget or the risk rating assigned to an investment across different attitude profiles, we will see how these bias effects play out in managers' ESG decisions.

These research questions together address the objectives of the study. By answering RQ1 the study will quantitatively measure the trust differential between AI and expert sustainability metrics. RQ2 will uncover the psychological biases and decision-making implications behind that differential, so we will know *why* the trust gap matters and how it might lead to biased outcomes. Each question is designed to advance academic understanding of technology trust and decision biases while also yielding findings that can inform management practice.

This study is important for several reasons, across AI ethics, organizational behaviour and ESG management. Academically, the study will add to the literature on trust in artificial intelligence by providing empirical evidence from a new context: sustainability metrics in corporate decision-making. Trust in AI has been identified as a key factor in human-AI collaboration (Glikson & Woolley, 2020) and scholars have called for more domain specific studies to understand its determinants (Wesche et al., 2022). By focusing on the ESG domain, which involves high stakes and value laden decisions, this research will provide insights into how trust in AI might work differently (or similarly) compared to other domains like finance, healthcare or human resources. It will contribute to theory by linking concepts from decision science (cognitive biases like aversion or automation bias) with technology acceptance in a sustainability context.

Moreover, the study's quantitative experimental approach can strengthen causal inferences about bias and trust, so we can validate or challenge findings from prior surveys or observational studies on algorithm aversion (Wesche et al., 2022) and algorithm appreciation (Logg, Minson, & Moore, 2019).

This theoretical contribution sits at the intersection of organizational behaviour (how managers make decisions and the biases therein), technology management (adoption of AI tools) and sustainability accounting/reporting (use of ESG metrics) so it bridges multiple fields that don't often overlap. From an ethical and governance perspective, the research looks at the human factor in trusting and overseeing AI outputs. Frameworks for "Trustworthy AI" emphasize principles like transparency, fairness and accountability (Sustainability Directory, 2025). But even a well-designed AI system must be accepted and used correctly by human decision makers to be effective. If managers systematically mistrust AI driven ESG metrics, they might ignore valuable warnings or insights (and therefore ethical lapses or missed opportunities in sustainability initiatives). If they blindly trust AI metrics (perhaps due to automation bias) they may overlook AI errors or biases, for example if the AI's training data had gaps or prejudices it might underrate certain social issues, and over-reliance on it could reinforce those biases (Sustainability Directory, 2025). This study's exploration of trust and bias will inform AI governance practices: it will guide how organizations implement AI for ESG in a responsible way and ensure human oversight is effective. For example, the findings could show the need for explainable AI (XAI) in ESG reporting so managers have clarity on how an AI derived metric was generated and aligns with ethical AI use recommendations (Sustainability Directory, 2025).

Practically the study is highly relevant to corporate managers, investors and policy makers involved in sustainability and technology adoption. For managers (the direct subjects of this research) understanding their own biases in interpreting ESG metrics is the first step to improving decision quality. If the research finds that managers undervalue AI metrics due to lack of familiarity or undue scepticism, companies might invest in training programs to increase data literacy and comfort with AI tools. If the opposite is found (that managers are overconfident in AI outputs) then guidelines or checks (like requiring human review of AI

generated reports) could be put in place to prevent errors. Ultimately, increasing the right level of trust in ESG information will lead to better informed strategic decisions, such as more accurate risk management, more effective allocation of resources to sustainability projects, and greater credibility in external ESG communications. For investors and financial analysts, the study will shed light on the reliability of ESG data they increasingly use. It will show whether the source of ESG metrics (AI vs human) might inadvertently influence managerial behaviour, which in turn affects corporate performance and risk, knowledge that investors can feed into their engagements with companies. Also, the research has implications for the design of ESG reporting systems and tools. Developers of AI driven analytics platforms will get insights into user trust factors. Tool designers and ESG data providers can use such insights to improve their products (e.g. incorporating explainability features or combining human expert oversight with AI - a "hybrid" approach - to maximize user confidence (Logg, Minson, & Moore, 2019)). Standard-setters and regulators concerned with corporate sustainability disclosures may also find this work useful. As regulatory regimes (like the EU's Corporate Sustainability Reporting Directive) consider the role of technology in compliance, understanding trust dynamics will help ensure that any AI enabled reporting is done in a way that company directors and auditors trust and verify. In summary the relevance of this study lies in its potential to contribute to safer and more effective integration of AI in organizational decision processes, so that advances in technology actually lead to better ESG outcomes rather than unintended biases or resistance.

Finally, this research is timely and socially important. With climate change and social inequalities at the top of the global agenda, businesses are expected to make evidence based, unbiased decisions to improve ESG performance. By probing the trust and biases in ESG metrics, this study helps in *"getting the metrics right"*, so the information guiding ESG decisions is used optimally. This will ultimately support responsible management practices in line with academic calls for better understanding of human-AI trust and the practical imperative of achieving sustainability goals.

This thesis consists of five chapters. Chapter 1 (this chapter) has introduced the research topic, established the context and defined the problem, research questions and significance of the study. Chapter 2 reviews the literature and theoretical frameworks. This includes a review of ESG performance measurement and reporting and AI driven methodologies. It also reviews prior research on trust in technology and AI and cognitive bias theory. Chapter 3 describes the research design and methods used to answer the research questions and outlines the results. Chapter 4 interprets the results in the light of the research objectives and existing literature and their implications on theory and practice.

2. Literature Review

2.1 ESG and Corporate Sustainability

ESG (Environmental, Social and Governance) has become the way modern companies define and measure sustainability. ESG refers to the non-financial factors (environmental impact, social responsibility and governance practices) that together indicate a company's long-term performance and ethical impact. In practical terms, ESG are the three pillars of corporate sustainability, evaluating how companies manage issues like carbon emissions, labour relations, community impact, board oversight and shareholder rights (Jamali et al., 2008). Originally popularized in investor circles (notably the 2004 United Nations report *Who Cares Wins*) ESG has since moved from a niche concern of socially responsible investors to a mainstream component of corporate performance evaluation (United Nations, 2004). Over the past two decades research and industry reports have shown a sharp increase in ESG disclosures and analyses, as there is a growing consensus that strong performance on ESG can be linked to stable financial performance and lower risk (Friede et al., 2015; Gillan et al., 2021). Companies are no longer judged just on quarterly profits but on how sustainably and responsibly those profits are earned, making ESG a key metric in managerial decision making and stakeholder assessment (Gillan et al., 2021).

Theoretical frameworks underlying sustainability reporting provide a lens to understand why companies engage with ESG and how they should communicate it. Several key perspectives include:

Stakeholder Theory: Freeman's work posits that a company's success depends on creating value for all its stakeholders, not just shareholders but also customers, employees, communities and the environment (Freeman, 1984). In the context of ESG stakeholder theory suggests that companies have obligations to various stakeholder groups whose interests align with the E, S and G dimensions. For example, environmental stewardship responds to community and ecological stakeholders, social responsibility

responds to employees and society, and governance responds to investors and regulators. Sustainability reporting from this view is a tool to demonstrate accountability to stakeholder concerns and balance those interests in corporate strategy (Freeman, 1984; Donaldson & Preston, 1995). Stakeholder theory underpins many ESG frameworks by arguing that transparent disclosure of ESG performance is critical to managing stakeholder relationships and securing long term support (Donaldson & Preston, 1995).

Triple Bottom Line (TBL): Elkington's TBL framework expanded the notion of corporate success by arguing that companies should have three bottom lines (profit, people and planet) instead of one (Elkington, 1997). This concept introduced the idea that environmental and social outcomes should be measured with the same rigor as financial results. It directly led to the rise of ESG metrics by formalizing the idea that sustainable businesses must perform well financially and contribute positively to society and the environment (Elkington, 1997). Many early sustainability reports in the 2000s were structured around the triple bottom line, with equal weight given to economic, environmental and social performance indicators (Milne & Gray, 2013). The TBL framework has been instrumental in getting companies to collect data on carbon footprints, diversity and labour practices, community impacts, etc., and report these alongside financial data.

Legitimacy Theory: This theory argues that organizations continually seek to ensure their actions are seen as legitimate by society's norms and values (Suchman, 1995). With respect to sustainability reporting, legitimacy theory suggests companies disclose ESG information as a way to demonstrate that their operations are in line with social and environmental expectations, thereby justifying their existence and avoiding public disapproval (Suchman, 1995; Deegan, 2002). For example, when a company in a high pollution industry publishes detailed environmental performance data and targets it may be closing a "legitimacy gap" (the difference between societal expectations and the company's perceived behaviour) (Deegan, 2002). Research has found that companies often increase voluntary sustainability disclosures after incidents or periods of public criticism, consistent with a legitimacy-seeking motive for ESG reporting

(Deegan, 2002). In short legitimacy theory explains why transparency in ESG matters: it helps secure ongoing support from stakeholders by aligning the company's image with societal values.

Institutional Theory: Institutional theory argues that companies adopt certain practices (like ESG reporting) because of pressures from the institutional environment (laws, industry norms and mimicking peers) rather than purely economic motives (Di Maggio & Powell, 1983). Under this view, the increase in ESG reporting can be seen as a response to coercive pressures (regulations, listing requirements for sustainability disclosure), normative pressures (professional standards and expectations from bodies like the Global Reporting Initiative) and mimetic pressures (firms imitating industry leaders' sustainability practices) (Di Maggio & Powell, 1983). As sustainable business practices become institutionalized, companies may feel compelled to report ESG metrics simply because it is seen as the modern "best practice", or to maintain legitimacy within their field. Institutional theory thus complements stakeholder and legitimacy theories by showing how external pressures and the desire to conform drive the adoption of sustainability metrics and reporting standards.

Agency Theory: Agency theory typically focuses on the relationship between shareholders (principals) and managers (agents), often highlighting issues of incentive misalignment. In the ESG space, agency theory has been used in two ways. On one hand some argue that managers may engage in excessive ESG initiatives that cater to their personal values or reputational benefits at the expense of shareholder returns: an agency cost if those initiatives do not add value (Jensen & Meckling, 1976; Barnea & Rubin, 2010). On the other hand, agency theory can support ESG integration by noting that neglecting material ESG issues could actually harm shareholders in the long run (e.g. ignoring climate risks could destroy future value) (Eccles et al., 2014). Hence aligning managerial incentives with long term sustainable performance (through ESG targets, executive compensation tied to ESG outcomes etc.) can reduce agency conflict and improve firm value (Eccles et al., 2014). Agency theory requires robust measurement of ESG performance so that boards and shareholders can hold managers accountable for non-financial as well as financial outcomes (Gillan et al., 2021). In practice, this has led to mechanisms like independent ESG audits and linking CEO pay to

ESG metrics to ensure managers do not underinvest in sustainability or conversely overspend on immaterial ESG activities.

In addition to these theories, numerous reporting frameworks and standards have guided the practical implementation of ESG measurement. The Global Reporting Initiative (GRI) introduced in the late 1990s was one of the first comprehensive frameworks for sustainability reporting, providing standardized indicators for topics ranging from greenhouse gas emissions to labour practices. The GRI and similar standards (e.g., Sustainability Accounting Standards Board and more recently the International Sustainability Standards Board) operationalize the theoretical principles by telling companies what to report to satisfy stakeholders and maintain legitimacy (GRI, 2016). Integrated reporting frameworks encourage firms to combine financial and ESG information into one report, reinforcing the triple bottom line by showing how financial capital is linked with social, human and natural capital. The proliferation of these frameworks shows the growing importance of ESG in corporate performance evaluation: investors, regulators and other stakeholders increasingly expect quantifiable, credible sustainability metrics. Research evidence suggests that companies with higher ESG transparency enjoy lower cost of capital and higher investor trust, because comprehensive ESG disclosure reduces information asymmetry around non-financial risks and opportunities (Dhaliwal et al., 2011).

In summary, ESG and corporate sustainability have moved to the top of the performance agenda, underpinned by rich theoretical foundations and guided by evolving reporting frameworks. Together they form the basis for how managers today consider not just “profit”, but also “people” and “planet” in decision making and how those considerations are reported through formal metrics and reports.

2.2 AI in Sustainability Reporting

Artificial intelligence has been increasingly intersecting with sustainability reporting and offering new tools to generate, analyse and interpret ESG metrics. AI in sustainability reporting means using algorithms,

including machine learning, natural language processing and data analytics, to gather information on a company's environmental, social and governance performance and produce insights or scores based on that data. Over the past decade we've seen a big evolution: what started as simple automated data collection (like software pulling energy usage data from sensors) has moved to sophisticated AI systems that can write entire sustainability reports or derive ESG ratings from massive datasets. This evolution is driven by the growing volume and complexity of sustainability data. Companies must track everything from greenhouse gas emissions across global supply chains to employee satisfaction scores, regulatory compliance events and beyond. AI is great at handling large complex datasets, so it's a natural fit for sustainability where the data can be both quantitative (e.g. emissions numbers, accident rates) and qualitative (e.g. text reports, social media sentiment).

Applications of AI in generating sustainability metrics are many and growing. A key application is ESG scoring and analytics: AI models can aggregate data from multiple sources (company disclosures, news articles, satellite images, social media) to assess a company's performance on various ESG criteria and produce a composite score or risk rating (Berg et al., 2020). For example, natural language processing (NLP) algorithms can read through thousands of pages of annual reports and sustainability reports to detect mentions of key issues (like climate targets or diversity initiatives) and even gauge the tone or specificity of those disclosures. This helps develop metrics such as a "disclosure quality" score for ESG.

Machine learning models have also been used to predict future ESG performance or identify companies likely to have ESG related controversies, by finding patterns in historical data that human analysts might miss. Another use of AI is real-time sustainability monitoring. Traditional ESG reporting is periodic (annual or quarterly), but AI systems now enable more continuous tracking. For example, AI-driven platforms can harvest data on a company's environmental outputs (via IoT sensors measuring emissions or waste in real time) and flag deviations or improvements instantly (Correia & Água, 2024). This meets the growing demand for timely and proactive sustainability management; instead of waiting for an annual report, managers can get AI alerts about, say, a spike in water usage at a facility and act immediately. Similarly on

social issues, AI tools can analyse employee feedback or social media posts to detect emerging issues in company culture or reputation (for example, detecting a trend of employee complaints on forums might signal a social risk). These applications show how AI can generate sustainability metrics and insights at a speed and scale that human analysis alone could not.

Despite the potential, the integration of AI in ESG reporting comes with big challenges and debates, particularly around the reliability and transparency of AI-generated outputs. Reliability issues arise because AI models are only as good as the data and algorithms behind them. ESG data is notoriously unstandardized and sometimes unreliable; companies may use different methodologies for calculating metrics, or there may be gaps and errors in self-reported data. If an AI model is trained on flawed or biased data, its outputs (the sustainability metrics) will also be flawed or biased. For example, an AI might underestimate a company's carbon risk if that company has not reported indirect emissions (Scope 3) fully; the metric would then give a false sense of security. There is also the risk of algorithmic bias: if the AI's model places heavy weight on certain proxies that are more available for some companies than others, it could systematically favour or disfavour certain companies without truly reflecting performance.

One noted issue in ESG ratings is *"rating divergence"*: even human rating agencies often disagree on ESG scores for the same firm due to different methodologies (Berg et al., 2020). An AI could either help by finding a more data-driven consensus or exacerbate divergence if different AI systems use different training sets and criteria. Ensuring reliability therefore requires careful model design, extensive testing and often a human-in-the-loop to validate AI outputs against common sense and known benchmarks (Sulkowski, 2024). Transparency is the other big issue. Many AI models, especially complex machine learning (like neural networks), are "black boxes": they produce an output (say, an ESG risk score) without an easily interpretable explanation of how they arrived at that number. In the context of sustainability, this opacity is a problem. Managers and stakeholders might justifiably ask, *"Why did the AI rate our company's social performance as 4 out of 10?"* Without clear reasoning, the metric can be met with scepticism or confusion. A lack of transparency can directly erode trust in AI-driven metrics. If executives do not understand or

cannot explain the basis of an ESG score to their board or investors, they may be reluctant to use it for decision-making. Recognizing this, a recent strand of research and practice has focused on Explainable AI (XAI) for ESG and finance: developing models that can provide understandable justifications for their outputs. For instance, an explainable ESG scoring AI might output not just a score but also a list of contributing factors (e.g., "Score was lowered due to high water usage relative to peers and lack of diversity policy disclosure") to increase transparency. Ensuring interpretability and traceability of AI decisions in sustainability reporting is seen as key for managerial acceptance and for regulatory compliance, especially as authorities like the EU emphasize algorithmic transparency in AI governance frameworks.

The debates around reliability and transparency feed into a broader question: can AI outputs be trusted as much as (or more than) traditional expert analysis in ESG? Proponents argue that AI can be more accurate by processing more data free of human biases, potentially catching issues that a human analyst might miss and doing so consistently across companies. They also note that AI can increase transparency *in some ways*: for example, by avoiding the "black box" of individual expert judgment that might be influenced by personal biases or limited information. Indeed, AI could increase confidence in ESG metrics if it shows high predictive validity (e.g., AI ESG scores that better predict future incidents or financial outcomes than human ratings). But critics warn that AI is not the solution to ESG measurement. One concern is that companies might use AI as a symbolic tool rather than a substantive one, a kind of "AI washing" akin to greenwashing, to show they are modern, without actually improving data quality or accountability (Sulkowski, 2024). Additionally, heavy reliance on AI could lead to automation bias (discussed in the next section) where managers accept AI metrics without proper scrutiny, assuming the machine is always objective and correct. If the AI system has blind spots, that blind trust could be dangerous. There is also an ethical dimension: AI algorithms can reflect the values of their designers. For instance, how an AI weighs environmental vs. social criteria in a single ESG score involves value judgments (Is carbon footprint more important than workforce diversity? By how much?). Different stakeholders might disagree on those

weights, and an AI's "neutral" mathematical approach might hide what are essentially subjective priorities built into the algorithm.

In response to these debates, scholars and practitioners recommend best practices for using AI in sustainability reporting: combine AI efficiency with human expertise. For example, Correia and Água (2024) suggest using AI to collect data and do initial analysis but then have sustainability experts review AI findings for plausibility and context before finalizing reports (Correia & Água, 2024). This hybrid approach can harness AI's speed and breadth while human judgment addresses nuance and value-based interpretations. Another recommendation is to implement robust data governance and validation procedures (i.e., continuously check AI-generated metrics against other indicators or audit samples of its output against expert assessments to gauge accuracy) (Correia & Água, 2024). As ESG disclosure regulations strengthen (such as mandatory climate risk reporting in many jurisdictions), there is also a push for standardized metrics and data which would, if realized, improve AI reliability by providing clearer parameters to train models on. In short, AI will transform sustainability reporting by providing more data and insights on ESG performance. From data collection to predictive analytics to real-time monitoring, AI can help companies and investors make sense of sustainability factors. But the power of AI also raises questions of trust: managers must be sure these metrics are accurate (reliability) and know how they are calculated (transparency). The next sections will look at the human side of this equation: how cognitive biases affect the trust managers have in AI metrics versus expert judgments, and what research says about that.

2.3 Cognitive Biases in Decision-Making

When bringing in any new source of information into a decision, whether it's an AI system or expert advice, managers are affected by many cognitive biases. Biases are systematic errors in judgment and can subtly influence how information is perceived and acted upon.

A growing body of research, including experiments and surveys, has looked at how people trust algorithmic decisions versus human judgments and how cognitive biases come into play in these situations.

Trust in AI vs. Human Expert: Contrasting Results. Early research in this area found what was called *algorithm aversion*. In a notable study by Dietvorst et al. (2015), participants were asked to make predictions (e.g. forecast student performance) and were given a choice between using their own estimate, an algorithm's estimate or an average of both. Despite the algorithm being more accurate overall than humans, participants who saw the algorithm make even a small mistake would quickly lose confidence in it and prefer to rely on human judgment (their own or an expert's) afterwards (Dietvorst et al., 2015). This shows a reluctance to trust algorithms after seeing them make a mistake, whereas people often give humans a second chance. The researchers called this algorithm aversion; people are less tolerant of machine mistakes than human mistakes. One reason is that when an algorithm makes a mistake people assume the whole model is flawed (since it's opaque to them), whereas a human makes a mistake and people can rationalize it as a one off or attribute it to situational factors (Dietvorst et al., 2015).

But more recent research has shown it's not one sided. In fact, under many conditions people exhibit *algorithm appreciation*, a tendency to trust algorithmic advice more than human advice. Logg, Minson, and Moore (2019) did six experiments with various estimation tasks (e.g. counting objects in a photo or general knowledge questions) and manipulated whether the advice given to participants was labelled as coming from a human or an algorithm. They found that, contrary to the notion of general aversion, participants followed advice more when they thought it came from an algorithm than when the same advice was said to come from another person (Logg et al., 2019). This effect was strongest in tasks perceived as objective or mathematical. The authors argue that people assume algorithms are better in quantitative, data-driven domains, they are seen as unbiased number-crunchers, whereas a random person's advice is seen as less reliable. In these experiments participants showed a kind of inherent trust in the algorithm's competence (even though in reality the advice was the same, only the source changed). This algorithm appreciation means context matters: when a decision is seen as a technical optimization problem people may actually

prefer an algorithm's input, whereas if judgment and subjective factors are involved, they might lean towards human insight.

These seemingly contradictory findings (aversion vs. appreciation) have been resolved by looking at moderating factors. One key factor is whether people have seen the algorithm perform. If users directly see algorithm mistakes (as in Dietvorst's study), aversion sets in. If they haven't seen obvious errors and the task is one where algorithm strengths are salient, appreciation occurs. Task domain is important, for example, in creative or ethical decisions people might distrust AI, but in analytical ones they trust it. Another factor is the availability of a human expert alternative. If a known expert with credibility is present, people will favour the expert; absent that, they'll lean on AI. Personal experience and expertise also play a role: some studies show that novices are more likely to trust algorithmic recommendations than experts are. Experienced professionals have established heuristics and may be more sceptical of an AI that challenges their expertise (because they are overconfident in themselves and have a higher bar for trusting a machine in their domain) (Logg et al., 2019). In contrast, those less knowledgeable will happily defer to an AI, assuming it knows better.

Automation Bias: *Automation bias* is the tendency to favour suggestions or outputs from an automated system, sometimes at the expense of ignoring contradictory information or not using your own judgment (Mosier & Skitka, 1997). In essence, when automation bias occurs, individuals become over reliant on a computer or AI, assume it's correct and as a result may not take action on cues the automation didn't flag (errors of omission) or act on the automation even if it's wrong (errors of commission) (Parasuraman & Manzey, 2010). This bias has been found in many high-stakes fields. For example, in aviation, pilots have trusted an autopilot or cockpit warning system even when it malfunctioned and ignored their training or external signals that something was wrong (Parasuraman & Riley, 1997). In healthcare, studies have shown that clinicians using clinical decision support systems can fall victim to automation bias by following a diagnostic suggestion from software without noticing signs that contradict the suggested diagnosis, especially under time pressure. The bias is more pronounced when the automation has a history of accuracy

or when users lack confidence in themselves relative to the machine. In managerial decision making, automation bias could manifest as a manager giving too much weight to an AI generated sustainability metric or recommendation, perhaps implementing a strategy because "the model said so" even if there are red flags that a human would pick up. The implication is that managers might not think critically or double check when an "expert system" is involved, potentially leading to blind spots. Previous research suggests automation bias is not laziness per se, but *trust miscalibration*: users calibrate their trust too high for the automation, assume it's infallible and correspondingly lower their own vigilance (Parasuraman & Manzey, 2010).

Research specifically with corporate managers or decision-makers is still emerging, but one study in the public sector context shows automation bias and over-trust in AI. In an experimental study simulating a public policy decision, Alon-Barkat & Busuioc (2023) found that government officials were more likely to follow a recommendation labelled as coming from an algorithm even when there were clear "warning signals" that the recommendation was flawed, compared to when the same recommendation came from a human expert (Alon-Barkat & Busuioc, 2023). In other words, participants over-trusted the algorithmic advice, often choosing to implement it despite contradictory evidence that in a scenario with a human advisor would have made them pause or reject the advice. This is classic automation bias: the presence of a computer-generated solution seemed to induce a level of deference that overrode normal critical judgment. The officials likely assumed the algorithm had crunched more data than they could and so gave it the benefit of the doubt. Such findings mean in managerial settings the aura of AI objectivity can sometimes lend too much credibility to AI outputs, a manager might think, "The computer analysed more data than a human could, so I'll go with its suggestion," even if an analytical review would show reasons to question that suggestion.

Overconfidence Bias: *Overconfidence bias* is the tendency to overestimate the accuracy of your knowledge, predictions or abilities. In other words, people often think they are righter than they actually are. Overconfidence can take forms such as overestimation (thinking you are better than you are), over-precision

(excessive certainty in your beliefs), and over-placement (believing you're better than others) (Moore & Healy, 2008). In managerial settings, overconfidence is a well observed phenomenon: successful leaders might develop an inflated sense of their judgment, given past successes, and start to discount advice or evidence that contradicts their instincts (Malmendier & Tate, 2005). For example, an overconfident CEO might pursue a risky merger believing in their own vision even when the analysis and advisors say caution; studies have shown that overconfident CEOs are more likely to make high value acquisitions and overpay, often to the detriment of shareholder value (Malmendier & Tate, 2005). At a more day-to-day level, a manager might be overconfident in forecasting, assume their projections are correct and thus ignore forecast ranges or risk assessments provided by analysts. Previous experimental studies have repeatedly found overconfidence in decision making tasks; a classic finding is that when people say they are "99% confident" in an answer, they are correct far less than 99% of the time, indicating mis-calibrated confidence (Alpert & Raiffa, 1982). The presence of overconfidence bias means some managers might underweight both AI and expert input, trust their own judgment above all. Such managers might be resistant to suggestions, whether from a machine or a colleague, because they genuinely believe they know best. Overconfidence can thus reduce the uptake of new information and lead to slower adjustment of beliefs in the face of evidence. It also ties into risk-taking: overconfident managers perceive less risk, because they are too sure of positive outcomes, which can lead to decisions that a more calibrated (less confident) person would avoid or hedge (Moore & Healy, 2008).

Confirmation Bias: *Confirmation bias* is the tendency to look for, interpret and remember information in a way that confirms your existing beliefs or hypotheses, and give less consideration to alternative possibilities or opposing information (Nickerson, 1998). In other words, people see what they expect to see. This bias leads individuals to favour information that supports their prior views and to be overly critical or dismissive of information that challenges those views. Confirmation bias has been demonstrated in many decision-making experiments; for example, when evaluating a hypothesis, people tend to seek evidence that could confirm it rather than evidence that could disprove it (a tendency first noted in Wason's classic selection

task) (Nickerson, 1998). In organizational contexts, confirmation bias might occur when a manager has a favourite strategy or project: they may highlight any data that shows the project is doing well but rationalize away data that shows problems. For instance, if a manager believes "Investing in solar energy will be good for our company," they might accept market research that shows increasing solar demand (confirming evidence) but question the credibility of reports that highlight regulatory hurdles or past failures in solar investments (disconfirming evidence). In essence, confirmation bias acts as a filter that shapes your reality to fit your expectations. The implications for managerial decision making are significant: this bias can lead to reinforcing cycles where managers become more and more convinced of their initial positions, regardless of the facts, because they continuously cherry pick supportive information. When evaluating technical vs. expert information, confirmation bias might manifest as follows: a manager might trust whichever source (AI or human) that tells them what they *want* to hear and distrust the other. For example, if an AI system's ESG assessment aligns with the manager's belief that their company is a sustainability leader, the manager may praise the AI's objectivity and insight; but if the AI assessment suggests shortcomings, the manager might say "the AI doesn't understand our context" and prefer a human consultant's more favourable opinion, or vice versa.

Bashkirova and Krpan (2024) ran an experiment with mental health practitioners where an AI system provided diagnostic recommendations for patient cases. They found that practitioners were more likely to trust and accept the AI's recommendation when it matched their initial judgment, but when the AI's suggestion deviated from their own preliminary assessment, the practitioners became sceptical and often stuck to their original conclusion (Bashkirova & Krpan, 2024). In effect, these experts exhibited confirmation bias: they welcomed the AI as a second opinion when it confirmed their thinking but treated it with suspicion (or ignored it) when it challenged their view. Interestingly, those practitioners who rated themselves as more expert were the most likely to dismiss conflicting AI advice, indicating a possible interaction between confirmation bias and overconfidence. While this was in healthcare, the parallels to managerial ESG decisions are clear – an executive might be open to an AI's sustainability analysis if it

matches what they already believe about the company, but if the AI reveals an inconvenient truth (say, poor performance on a cherished initiative), they may find ways to discredit the AI or justify not acting on its findings. This shows that trust in AI is not just a function of the AI's accuracy, but also how its outputs fit with the decision-maker's existing frame of reference.

Each of these biases has been studied before in management or related fields. Automation bias has been researched mostly in domains like aviation and medicine but has been noted as a potential problem in business analytics: for instance, a study by Dzindolet et al. (2003) found that people using automated aids stopped looking for other information even when the aid was known to be imperfect (Dzindolet et al., 2003). Overconfidence bias in managers has been documented in behavioural corporate finance literature, linking overconfidence to actual outcomes like higher leverage and more aggressive investment policies in firms led by overconfident executives (Malmendier & Tate, 2005). Confirmation bias is common in strategic decision making; in one study, when given an initial hypothesis about a business problem, managers mostly looked for confirming evidence and had trouble integrating disconfirming evidence, which hurt decision quality. These biases operate subtly and unconsciously, so can't be eliminated completely. They surface especially under conditions of complexity and ambiguity, exactly the conditions of many ESG related decisions where data can be interpreted in many ways and outcomes are uncertain. The implications for managerial decision making are that awareness and mitigation of these biases are key, especially when introducing new decision support tools like AI. If a manager is prone to automation bias, they may need training to stay vigilant and cross-check AI outputs with other sources. If overconfidence is an issue, calibration tools (like getting feedback on past prediction accuracy) or decision protocols that force consideration of external input can help. To combat confirmation bias, managers can use practices like devils-advocate discussions or structured decision analysis that forces examination of contrary evidence. Without addressing these biases, even the best information, whether from AI or experts, will be misused or ignored. For example, a perfectly accurate AI sustainability metric is useless if an overconfident executive ignores it, and a highly experienced human advisor's warning will be overlooked if it doesn't match what

the decision maker already believes. In the context of this study, these biases set the stage for how managers might differentially trust AI-driven vs. expert-generated metrics. Do they default to the automated metric (automation bias)? Do they stick to their own initial judgment (overconfidence)? Do they favour the input that confirms their expectations (confirmation bias)?

In light of the evidence presented in this paragraph, we formulate the following hypothesis:

H1: Managers will trust AI more than human experts (automation bias) leading to (a) allocating budget choices following the ESG suggestion (i.e., high similarity), (b) displaying higher perceived confidence (c) perceived accuracy and (d) recognizing higher influence of the ESG reporting in their choices (i.e., confirmation bias).

2.4 Trust and Decision-Making in Managerial Contexts

Trust is key to decision making in organisations. It's a psychological state that means you're willing to accept vulnerability based on positive expectations of another's intentions or behaviour (Mayer et al., 1995). To trust means to rely on something or someone in a situation of uncertainty and risk, believing the outcome won't be averse to your interests. In a management context, trust operates at multiple levels: trust between individuals (e.g. a manager and an analyst), trust in teams or departments, and trust in systems or information sources. High levels of trust can make coordination and delegation smoother and more effective as managers feel confident in the information and analysis provided by others (Dirks & Ferrin, 2001). Low trust can lead to verification behaviours, duplication of work or paralysis in decision making as managers second guess the data or advice they receive (Dirks & Ferrin, 2001).

One of the most relevant areas is trust in the information used for decisions, especially when that information is complex or technical. In modern organisations, managers often rely on technical analysis (e.g. statistical models, data analytics) and expert opinion (e.g. reports from specialists or consultants) to make informed decisions. Trust plays a big role in whether and how managers use those inputs. If a manager

trusts the source of information, they'll use it; if they don't trust it, they'll ignore or discount it despite its value. For example, a CFO might trust the financial forecasts produced by an AI driven analytics system and base strategic plans on them, or if trust is low, override those forecasts with their own judgement or insist on a human audit of the numbers. Thus, trust is the filter through which information is accepted or rejected in organisational decision making (Mayer et al., 1995). When evaluating technical information versus expert information, managers face different considerations for each. Expert generated information (e.g., a report from a seasoned ESG analyst or a consulting firm) has the benefit of human judgement and context; managers will trust it if they perceive the expert as having high expertise, a good track record and aligned interests. AI driven or technical information (e.g., an algorithm's risk assessment) is often seen as data rich and objective but its trustworthiness is harder for a manager to gauge, especially if the algorithm's workings are opaque. The source credibility theory from communication research provides a useful lens here: it says that the credibility of a source in the eyes of the information receiver depends on two main dimensions: expertise and trustworthiness (Hovland et al., 1953). For a human expert, perceived expertise comes from credentials or experience and trustworthiness from reputation or demonstrated integrity. For an AI system, perceived expertise might come from accuracy or the sophistication of the analytics and trustworthiness from the reliability and objectivity of the outputs (and perhaps the reputation of the system's provider) (Glikson & Woolley, 2020). In both cases a manager's willingness to rely on the information will stem from an assessment (conscious or subconscious) of these credibility factors.

The importance of trust in handling technical vs expert information is highlighted by the potential consequences of mis-calibration of trust. If trust is too low (distrust in a reliable source) managers may ignore valuable insights, for example dismissing an accurate AI warning about a sustainability risk due to general scepticism about algorithms. On the other hand, if trust is too high (over-trusting) managers might accept information at face value without due critical thinking, for example taking an expert's claims as gospel even if the expert is biased or the analysis flawed. Either extreme can lead to suboptimal decisions. Prior management research suggests that optimal decision making often requires a balance: enough trust to

use others' expertise combined with a healthy degree of verification or critical analysis to guard against errors (Lewicki et al., 1998). Achieving this balance is hard as it requires managers to judge when their information sources are reliable and when they are not. In the context of ESG decision making, trust is even more important. Sustainability metrics can be complex, covering scientific data (e.g. emissions levels), social surveys and regulatory guidelines. Many managers are not domain experts in climate science or social impact assessment; thus, they rely on either in-house specialists, third party consultants or AI driven analytics to interpret ESG performance. Trust in the expert might be influenced by the expert's independence and credibility, for example managers may trust an independent auditor's assurance of ESG data more than the company's own sustainability report if they suspect internal bias (Hodge et al., 2009). Trust in an AI driven metric might depend on the clarity of how the metric was generated; managers are more likely to trust an AI system that provides transparent explanations for its ESG ratings than a "black box" that spits out a score without context (Kirkpatrick, 2016). Indeed, recent research highlights the need for explainable AI in business analytics to build user trust; explainability helps bridge the gap by providing rationale that managers can understand, just as a human expert would explain their reasoning. Researchers have developed tools to measure trust and credibility perceptions in these contexts. For example, in corporate communications Lock and Seele (2017) introduced the *PERCRED scale* (Perceived Credibility) to measure how credible stakeholders find a company's CSR or sustainability report (Lock & Seele, 2017). This scale measures dimensions such as truthfulness, sincerity and appropriateness of information, so managers can quantify trust in reported data. The implication for decision making is that the more credible the sustainability information is perceived (by managers or stakeholders) the more it will be trusted and weighted in decisions. Another relevant model is Fogg's framework of information credibility, which states that users assess the credibility of a system or content based on aspects of trustworthiness (e.g., honesty, unbiasedness) and expertise (e.g. knowledgeable, competent) (Fogg, 1999). Originally developed to understand why people trust certain websites or technologies, Fogg's model is very applicable to AI systems providing ESG metrics: managers will implicitly judge if the system seems expert (does it know what it's doing?) and trustworthy (is it likely to mislead?). High marks on both lead to higher trust.

Because perceived credibility has been shown to mediate the link between source and decision use (Lock & Seele, 2017; Söllner et al., 2016), the following hypothesis is formulated:

H2: The presence/absence of the automation bias on (a) budget allocation, (b) perceived confidence (c) perceived accuracy and (d) influence of the ESG reporting (i.e., confirmation bias) is mediated by perceived credibility of the ESG reporting.

Survey-based research also provides insight into general attitudes and trust levels. Broadly, surveys of managers and executives show a guarded optimism towards AI in decision-making. For example, a global survey by McKinsey (2021) found that while most executives are investing in AI tools, only a minority *fully trust* the analytics those tools provide without reservation (McKinsey, 2021). Common concerns were lack of transparency (the "black box" problem) and doubts about contextual understanding (essentially the transparency and reliability issues mentioned earlier). Many respondents said they prefer a hybrid approach: AI-generated metrics or forecasts should be reviewed by human experts and final decisions should combine automated analysis with managerial judgment. In academic research, Söllner et al. (2016) in an MIS context developed models showing that initial trust in an information system is key to its adoption; if users (managers) have initial distrust, they will ignore the system's recommendations no matter how good, a finding supported by surveys and usage data of business intelligence systems (Söllner et al., 2016). These surveys and models show that trust-building measures (like providing explanations, showing performance track records or aligning with user values) are necessary to get managerial buy-in for AI-driven insights. When it comes to biases like overconfidence affecting trust, some experimental evidence exists in the advice-taking literature. Generally, decision-makers who are more confident in their own initial answers give *less weight* to external advice (regardless of source). This phenomenon, sometimes called *egocentric advice discounting*, has been shown in lab studies: the more confident someone is in their estimate, the less they adjust it after seeing advice from others. Translating this to AI vs. expert, an overconfident manager might not differentiate much between AI and human input: they might be equally likely to downplay both, relying on their own analysis instead. However, one can speculate (and qualitative evidence suggests) that

if such a manager *had* to pick one, they might choose the source that more closely aligns with their self-image. For example, a tech-savvy overconfident manager might think, "I'm pretty much an expert myself, but the AI is data-driven like me" and thus lean slightly more on AI, whereas a more traditional overconfident manager might say "I trust my gut and the seasoned experts I know, not a machine." There's no one-size-fits-all; these tendencies are exactly why empirical research like the present study is needed to parse out.

Individual differences among managers also play a role in trust. Some managers have a general predisposition to trust technology, often called technology trust or AI attitude. A recently developed measure, the *AI Attitude Scale (AIAS-4)*, provides a brief assessment of a person's general attitude towards artificial intelligence and automation (Grassini, 2023). A manager with a positive attitude towards AI (scoring high on AIAS-4) will be more likely to give the benefit of the doubt to an AI driven sustainability metric, whereas a manager with a sceptical attitude will scrutinize it more heavily or favour human judgment instead. On the other hand, a manager with strong trust in traditional expertise might default to an expert's opinion even when an AI system provides a different analysis. So, trust is not only about the source in isolation but also about the trustor's disposition: their prior experiences, biases and attitudes shape how they evaluate information sources (Mayer et al., 1995).

Moreover, given evidence that a positive attitude toward AI increases trust in automated systems and that a strong ESG orientation boosts trust in sources that confirm one's sustainability values (Bashkirova & Krpan, 2024), we formulate the following hypothesis:

H3: the relationship between the source of the ESG reporting (AI vs Human) and the perceived credibility of the ESG report is moderated by (a) attitude toward AI and attitude toward ESG, such that manager with a more positive attitude will have higher perceived credibility of the ESG source.

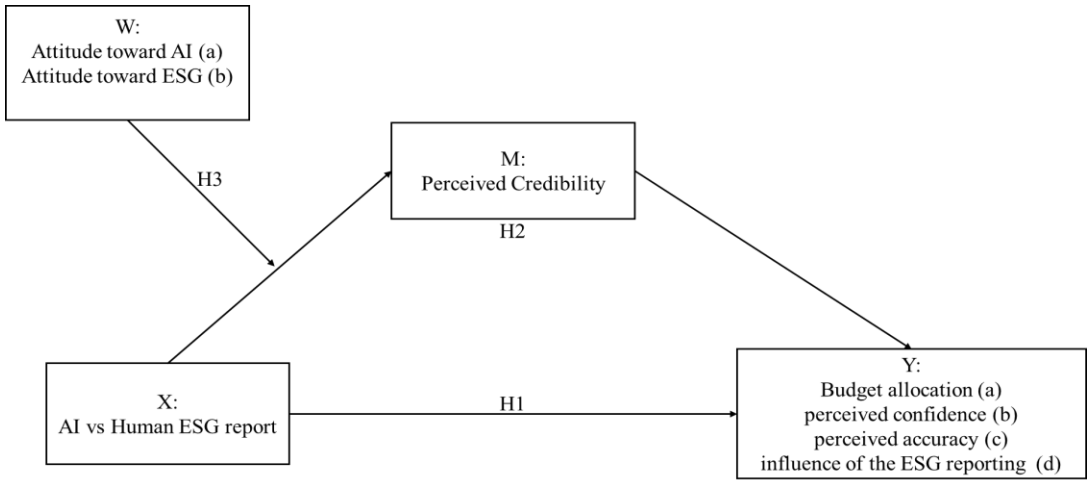


Figure 1. *Conceptual model summarizing the formulated hypothesis.*

In short, trust is a key moderator in how managers interpret and use both expert generated and AI generated information. In organizational decisions, especially those as complex as ESG evaluations, managers need to navigate the trust dynamics carefully. Building trust in reliable systems and experts (through credibility, transparency and track record) while keeping a degree of professional scepticism is often recommended as best practice (Söllner et al., 2016). Given the fast-emerging role of AI in providing decision inputs and the enduring role of human expertise, understanding the intricacies of trust in these contexts is crucial.

2.5 Research Gap

Despite all the research on algorithm aversion and appreciation in judgment tasks and the emerging evidence that perceived credibility mediates the impact of automation bias on decision outcomes, there is a glaring lack of studies that bring these insights into the world of ESG decision making. While scholars have shown that managers' pre-existing attitudes towards AI and towards sustainability can influence how they evaluate new information sources, no one has yet tested these moderating effects in a controlled experiment where the source of ESG metrics, AI versus human expert, is systematically varied.

In other words, while we know people sometimes distrust algorithms after one mistake or, conversely, defer to them in data-rich contexts, we don't have quantitative evidence on whether, and why, managers trust AI-driven sustainability metrics more or less than expert-generated ones, and how cognitive biases and individual attitudes jointly influence both that trust differential and its downstream impact on ESG investment or reporting decisions. This is especially pressing given the rapid adoption of AI tools in corporate sustainability reporting: without understanding these trust dynamics, organizations will under- or over-rely on AI outputs in ways that will systematically distort their ESG strategies.

Organizations need practical answers: Can managers actually use AI for ESG decisions, or will cognitive biases undermine its benefits? How can training or processes be designed to mitigate these biases? By filling these gaps, this study provides both academic and practical contributions.

Having grounded our three hypotheses in the literature above, the next chapter will describe the methodology of our experimental study to test these hypotheses and contribute to this interdisciplinary field.

3. Experimental Study

This chapter describes the methodology of the experimental study on trust in AI-driven versus expert-generated sustainability metrics. It outlines the research design, participants, materials and procedures, experimental manipulation, measures, ethical considerations, data analysis plan and methodological limitations. The aim is to explain how the quantitative experiment was conducted to compare managerial trust and decision-making under two conditions (AI vs. expert ESG metrics).

3.1 Setting

The study used a quantitative between-subjects design to investigate how sustainability (ESG) performance information affects trust and decision-making when it is attributed to an AI system versus a human expert. Each participant was randomly assigned to one of two conditions: an AI-generated ESG report condition or an expert-generated ESG report condition. This design allows for a controlled comparison of how the source of ESG metrics influences managerial judgments, eliminating potential learning or carryover effects by ensuring each participant experiences only one condition. The independent variable in this design is the stated source of the ESG metrics (AI vs. human expert), and the primary dependent variables are trust/credibility perceptions of the information and decision outcomes (see below). By holding the ESG report constant and only varying the source label (see Experimental Manipulation), the design isolates the effect of source credibility on participants' trust and resource allocation decisions. This is an experimental best practice for causal inference in behavioural research (Dietvorst et al., 2015).

3.2 Participants

The target population for this experiment was ESG/sustainability decision-makers in organizations. Participants were recruited through purposive sampling: the researcher personally contacted individuals in

his professional network and via industry groups who met the inclusion criteria. Inclusion criteria were that participants had experience or current responsibility in sustainability or ESG-related decision-making (e.g. sustainability managers, ESG analysts, corporate social responsibility directors). All participants were adults (18 years or older) and were proficient in English or Italian, as the survey was available in both languages. Participation was voluntary and anonymous, and no financial compensation was offered (participants were motivated by professional interest in the research topic). To ensure a relevant and knowledgeable sample, recruitment focused on professionals who would use or be familiar with ESG metrics in their work. Basic demographic and professional background data were collected to characterize the sample.

A total of 60 participants took part in the study. The survey was designed and administered via Qualtrics. The gender distribution is relatively balanced with 42% female and 57% male participants (1% preferred not to say). 55% of them are married, 23% are in a relationship, 18% are single and only 3% are separated or divorced. Most of the participants have children (83%). 27% hold bachelor's degrees, 65% master's degrees and 8% PhD or other postgraduate training. In terms of income range, most of them earn between 50 thousands and 100 thousands euros (28%) or between 35 thousands and 50 thousands (22%), while in terms of job position, 43% are Senior Managers or similar and 27% are Junior Managers or similar, predominantly in the tertiary sector (73%). The size of the companies they work for are mostly large (43%), medium (20%) or multinationals (18%), with 15% working in public administration.

3.3 Materials and Procedure

Each participant completed the study individually on their computer or device. The procedure consisted of several sections presented in a fixed order, as follows:

Introduction and Consent: Participants were told the purpose of the study in general terms (e.g. "to understand how managers make decisions using ESG information"), the expected duration and that their responses would be anonymous and used for research purposes only. An informed consent statement was

provided, in line with The Declaration of Helsinki and the APA ethical standards for the treatment of human samples.

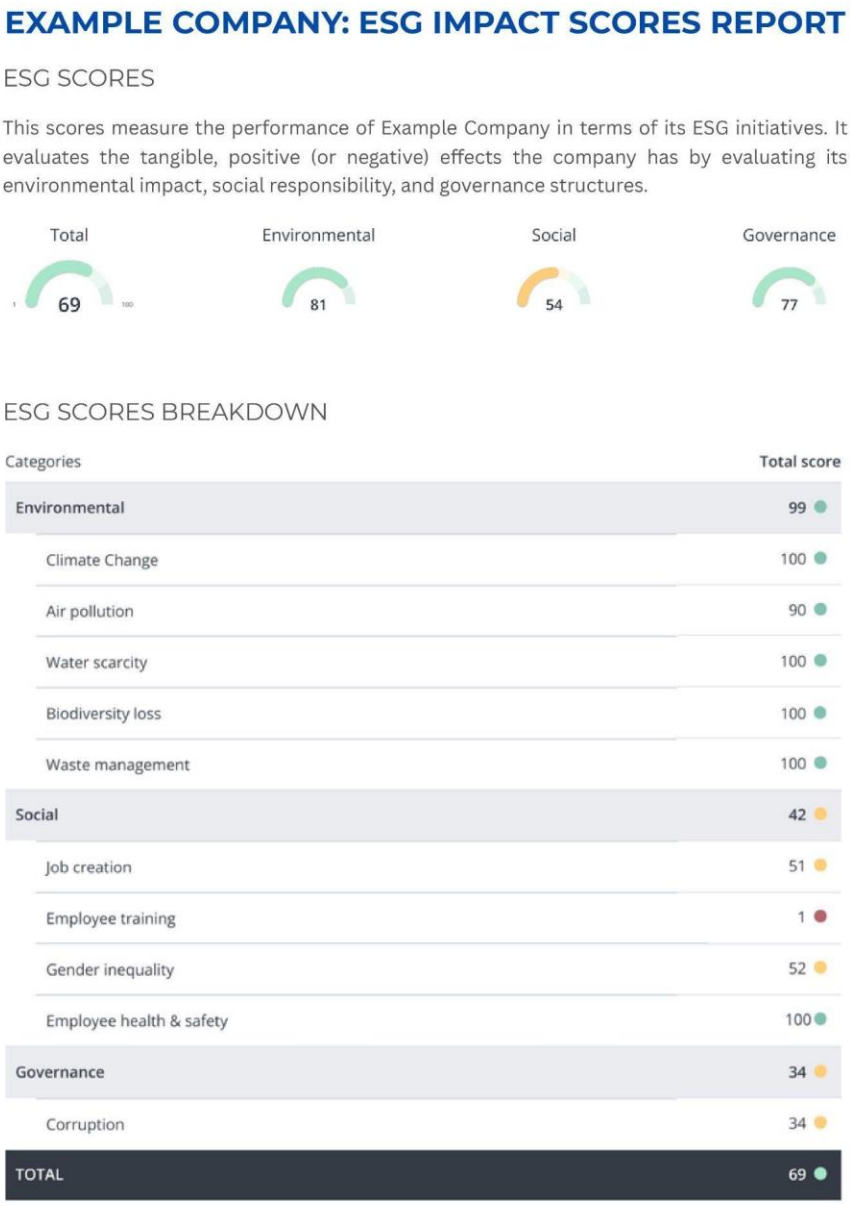
Demographic and Control Variables: Next, participants completed a demographic questionnaire and related background items. These questions asked age, gender, marital status, children, education, work experience, income range, current job position, sector and company size. The survey was anonymous (no names or employer details were asked, only broad information was collected to contextualise the sample).

Attitude Measures (Pre-Experiment): Before the experimental stimulus, the survey measured participants' baseline attitudes that might influence how they would interpret the ESG information. Two sets of attitude questions were administered:

- ESG Attitude Scale: A short questionnaire to assess the participant's general attitude towards environmental, social and governance issues in business. Participants rated statements such as "I consider environmental, social and governance (ESG) aspects whenever I am choosing an investment fund/company" on a Likert scale. These items measured how much the individual values ESG principles in a managerial context.
- AI Attitude Scale (AIAS-4): The Artificial Intelligence Attitude Scale was used to measure general attitudes toward AI (Grassini, 2023). This is a validated four-item scale where participants rated statements about AI (e.g. "I believe that AI will improve my life"). A higher score means a more positive attitude toward AI. This is an individual difference variable that will be tested as a potential moderator of trust in the AI-generated report. The AIAS-4 has shown good internal reliability in previous research (Grassini, 2023). In this study, the AIAS-4 was administered before the ESG information to capture participants' pre-existing attitudes toward AI.

ESG Report Stimulus (Experimental Condition): Participants were presented with an ESG performance report (Figure 1) for a hypothetical company (called "Example Company" in the stimulus materials). The report contained information about the company's performance on environmental, social and governance

criteria, including quantitative scores and their respective breakdowns into sub-scores. The content was the same for all participants; however, the source of the report was manipulated at this stage (see Experimental Manipulation below for details). Specifically, a brief note was added to the report. Participants in the AI condition saw a note saying the report (and sustainability metrics) were generated by an AI-driven analytics tool, while participants in the expert condition saw a note saying the report was prepared by human ESG experts. Everything else in the report was the same. The report included an overall ESG impact score, sub-scores for Environmental, Social and Governance. For example, it detailed the company's carbon footprint and environmental initiatives (E), labour practices and community engagement (S), and board governance structure and ethics (G). The information was presented in a professional but simplified report format. Participants were told to read the ESG report carefully as they would need to make a decision based on it in the next section.



EXAMPLE COMPANY: ESG IMPACT SCORES REPORT

ESG SCORES

This scores measure the performance of Example Company in terms of its ESG initiatives. It evaluates the tangible, positive (or negative) effects the company has by evaluating its environmental impact, social responsibility, and governance structures.

Total	Environmental	Social	Governance
69	81	54	77

ESG SCORES BREAKDOWN

Categories	Total score
Environmental	**99**
Climate Change	100
Air pollution	90
Water scarcity	100
Biodiversity loss	100
Waste management	100
Social	**42**
Job creation	51
Employee training	1
Gender inequality	52
Employee health & safety	100
Governance	**34**
Corruption	34
TOTAL	**69**

Figure 2 *- Example Company: ESG Impact Scores Report*

Budget Allocation Decision Task: After the ESG report, participants completed a decision-making task designed to simulate a managerial decision influenced by the ESG information. In this task, participants were told to assume the role of a manager at the company and were asked to allocate a fixed budget across

different initiatives based on the company's needs. They were given a hypothetical budget (100 allocation points) and several investment options. The options included the categories for which sub-scores were detailed in the report. Participants had to decide how much of the budget to allocate to each option, effectively revealing their priorities. The key outcome of interest is how much they chose to follow the scores detailed in the report for their allocation decisions. This budget allocation pattern is a behavioural measure of decision-making potentially influenced by trust in the provided ESG metrics. The task was framed as if there was no right or wrong answer, but rather it assessed their managerial judgment. All participants, regardless of condition, faced the same allocation options and total budget; only the source of the preceding report differed. The survey recorded the amount allocated to each option.

Decision Reflection (Post-Experiment): After completing the budget allocation task, participants rated three statements assessing their subjective evaluation of the decision they had just made (decision confidence, perceived decision accuracy and report influence on the allocation decisions). These items capture the manager's subjective assurance in their decision quality, which may be affected by how much they trusted the underlying data. For instance, a participant who doubts the accuracy of the ESG report might feel less confident that their allocation was the right one.

Trust and Credibility Perception Measures: Participants answered a series of questions about their trust in and perceived credibility of the ESG report they had just seen. This section used an adapted Perceived Credibility scale (PERCRED) by Lock and Seele (2017). We modified the PERCRED scale to fit our context of an ESG performance report. Participants rated statements on a 5-point Likert scale about the trustworthiness and credibility of the report. These items capture multiple facets of source credibility (e.g., perceived accuracy/truthfulness of the content, the expertise and trustworthiness of the source, and the completeness/clarity of the information). The credibility items were presented after the decision task to gauge how much the participant trusted the ESG metrics they saw, which we expect to be influenced by whether they thought an AI, or an expert provided them. Higher scores on this scale indicate higher perceived credibility of the ESG report.

Throughout the procedure, the online survey platform handled the random assignment and ensured that each participant only saw the materials corresponding to their condition. The flow was identical for both groups apart from the stimulus labelling. Every participant completed the survey in one sitting, and the software recorded response times and answers for analysis. Overall, the materials and procedure were designed to simulate a realistic decision scenario while systematically manipulating the source of ESG metrics to observe its influence on trust and behaviour.

3.4 Statistical Analysis

We conducted a reliability test to check the internal consistency of the scales used in the survey, namely ESG Attitude Scale, AI Attitude Scale and Credibility Scale. For ESG Attitude Scale, the initial reliability test showed a low Cronbach's alpha ($\alpha = 0.596$) for the 5-item scale. Looking at items' statistics, item number 3 had the lowest corrected correlation (0.149); further, the analysis indicated an increase in Cronbach's alpha ($\alpha = 0.663$) if the item was deleted. Hence, in order to increase reliability, the item was removed and not further considered in the analysis.

The AI Attitude Scale (AIAS-4) had high reliability with Cronbach's alpha ($\alpha = 0.925$) for 4 items. All items had high corrected correlations ranging from 0.783 to 0.896, confirming that all items were reliably measuring the same underlying construct.

The Credibility Scale (PERCRED) with 11 items had high internal consistency with Cronbach's alpha ($\alpha = 0.922$). Corrected items' correlations were robust ranging from 0.510 to 0.788, so the scale was reliable to capture the perceived credibility construct.

All variables were standardized to allow comparison across different scales, by converting them into z-scores prior to statistical testing. To test our hypotheses efficiently, we used a single moderated mediation model (Model 7 from the PROCESS macro for SPSS) and ran a total of 8 models. These varied across 4

different dependent variables and 2 moderators, allowing us to explore the conditional effects and interactions proposed in our hypotheses.

Model 1. The first model was conducted using the allocation similarity as dependent variable, the experimental manipulation (AI- vs Human-generated report) as independent variable, Credibility as mediator, and Attitude toward AI as moderator. The variables: gender, age, marital status, parental status, income, job position, sector and company's size were used as covariate in the analysis.

Model 2. The second model was identical to the first, with the only exception of the moderating variable; indeed, in this second analysis attitude toward ESG was included as moderator.

Model 3. The third model used decision confidence as the dependent variable, with condition as the independent variable, Credibility as mediator, and Attitude toward AI as moderator, controlling again for gender, age, marital status, parental status, income, job position, sector and company size.

Model 4. The fourth model mirrored Model 3 but substituted Attitude toward ESG for Attitude toward AI as the moderator; decision confidence, Credibility, condition and the same covariates were unchanged.

Model 5. The fifth model tested perceived accuracy as the dependent variable, with condition as the independent variable, Credibility as mediator, and Attitude toward AI as moderator, including the same set of eight covariates.

Model 6. The sixth model replicated Model 5's structure (perceived accuracy, condition, Credibility and covariates) but replaced Attitude toward AI with Attitude toward ESG as the moderator.

Model 7. The seventh model focused on perceived influence as the dependent variable, again using condition as the independent variable, Credibility as mediator, and Attitude toward AI as moderator, with gender, age, marital status, parental status, income, job position, sector and company size as covariates.

Model 8. The eighth model was identical to Model 7 except that Attitude toward ESG served as the moderator instead of Attitude toward AI; all other components remained the same.

3.5. Results

The inspection of the results obtained from the model 1 highlighted that the model was overall non-significant ($R^2 = 0.16$, $F_{11,48} = 0.83$, $p = 0.61$). In particular, the experimental manipulation (AI- VS. Human-generated rating) did not produce a statistically significant main effect ($\beta = -0.07$, $p = 0.60$) on the similarity of the budget allocation task. Similarly, the direct effect of credibility ($\beta = 0.20$, $p = 0.20$) was found non-significant. Across the covariates, the company's sector produces a marginally significant effect ($\beta = 0.30$, $p = 0.08$) indicating higher similarity in the primary and secondary sector as compared to the tertiary sector. Overall, these results point at the rejection of hypothesis H1a. As per the indirect effects – namely the credibility as a mediator of the relationship between the experimental manipulation of the similarity of the budget allocation task – the results suggest that the mediation was non-significant (Index = 0.04, LLCI = -0.07, ULCI = 0.16), hence pointing at the rejection oh hypothesis H2a (see table 1 for detailed results).

Table 1 - Results of Model 1. Dependent variable: Budget allocation similarity

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	-0.07	0.14	-0.52	0.60	-0.34	0.20
Credibility	0.20	0.15	1.30	0.20	-0.11	0.51
Gender	0.03	0.24	0.13	0.89	-0.45	0.51
Age	-0.01	0.02	-0.49	0.63	-0.04	0.02
Marital status	0.16	0.19	0.85	0.40	-0.22	0.55
Parental status	-0.65	0.41	-1.56	0.12	-1.48	0.18
Education	-0.12	0.25	-0.47	0.64	-0.62	0.39
Income	0.12	0.11	1.11	0.27	-0.10	0.35
Job Position	-0.13	0.15	-0.86	0.40	-0.44	0.18
Sector	0.30	0.17	1.76	0.08	-0.04	0.64
Size	0.07	0.13	0.56	0.58	-0.19	0.34
Credibility (indirect effect)	0.04	0.06	-	-	-0.07	0.16

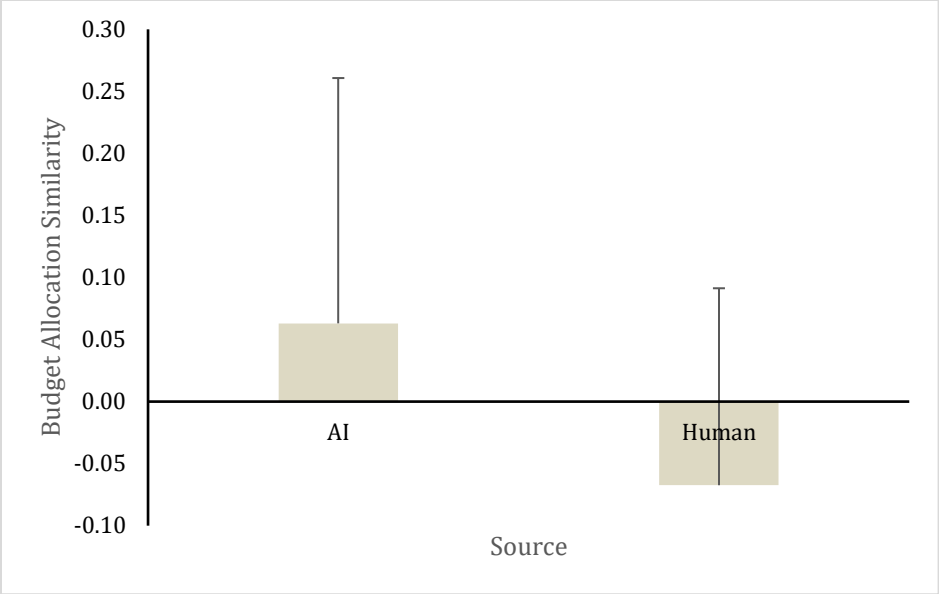


Figure 3 - Budget Allocation Similarity by Report Source (AI vs. Human)

Finally, regarding the moderation effect exerted by attitude toward AI on the relationship between the experimental manipulation and the perceived credibility of the report, the results indicate that the model was overall significant (R^2= 0.35, $F_{12, 47}$ = 2.11, p = 0.03). In particular, the main effect of the experimental manipulation (AI- VS. Human-generated rating) was found non-significant (β = 0.07, p = 0.58); conversely, the main effect of attitude toward AI on the credibility of the report was found significant (β = 0.42, p = 0.01), nonetheless, the interaction effect did not reach statistical significance (β = 0.22, p = 0.14, see figure xxx). Hence hypothesis H3a was rejected. As per the covariates, the company's size yielded a marginally significant effect (β = -0.22, p = 0.06), indicating higher perceived credibility of the report in smaller firms. The other covariates did not produce any significant effect (see table 2 for detailed results).

Table 2 - Results of Model 1. Moderator: Attitude toward AI

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.07	0.12	0.56	0.58	-0.18	0.31
Attitude toward AI	0.42	0.16	2.68	0.01	0.11	0.74
Experimental Manipulation by Attitude toward AI	0.22	0.15	1.48	0.14	-0.08	0.52
Gender	-0.10	0.22	-0.46	0.65	-0.55	0.34
Age	0.01	0.01	0.75	0.46	-0.02	0.04
Marital status	0.07	0.17	0.41	0.69	-0.28	0.42
Parental status	0.26	0.42	0.61	0.55	-0.59	1.10
Education	-0.23	0.23	-1.00	0.32	-0.70	0.24
Income	-0.05	0.10	-0.46	0.65	-0.24	0.15
Job Position	0.19	0.14	1.40	0.17	-0.08	0.47
Sector	-0.17	0.15	-1.17	0.25	-0.48	0.13
Size	-0.22	0.11	-1.95	0.06	-0.44	0.01

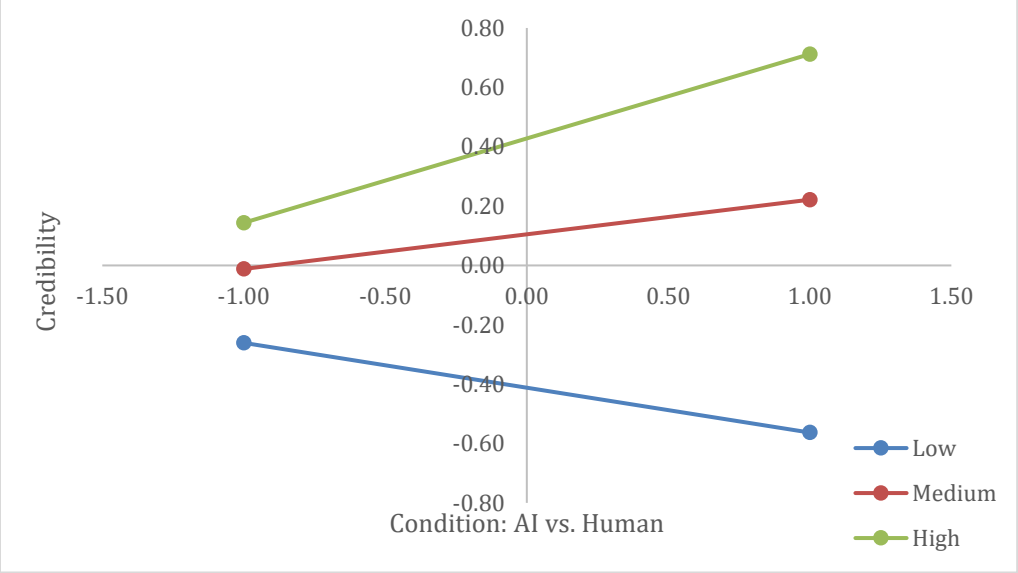


Figure 4 - Perceived Credibility by Report Source and AI-Attitude (Model 1)

The inspection of the results obtained from Model 2 highlighted that the model predicting budget-allocation similarity was overall non-significant ($R^2 = 0.16$, $F_{11,48} = 0.83$, $p = 0.61$). In particular, the experimental manipulation (AI- vs. Human-generated report) did not produce a statistically significant main effect on

allocation similarity (β = –0.07, p = 0.60). Similarly, the direct effect of Credibility on similarity was non-significant (β = 0.20, p = 0.20). Of the covariates, only sector showed a marginal effect (β = 0.30, p = 0.08), indicating slightly higher similarity in primary/secondary versus tertiary industries. These findings lead to the rejection of H1b. The index of moderated mediation for Credibility (H2b) was also non-significant (Index = –0.03, LLCI = –0.18, ULCI = 0.05), so H2b is rejected (see table 3 for detailed results).

Table 3 - Results of Model 2. Dependent variable: Budget allocation similarity

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	-0.07	0.14	-0.52	0.60	-0.34	0.20
Credibility	0.20	0.15	1.30	0.20	-0.11	0.51
Gender	0.03	0.24	0.13	0.89	-0.45	0.51
Age	-0.01	0.02	-0.49	0.63	-0.04	0.02
Marital status	0.16	0.19	0.85	0.40	-0.22	0.55
Parental status	-0.65	0.41	-1.56	0.12	-1.48	0.18
Education	-0.12	0.25	-0.47	0.64	-0.62	0.39
Income	0.12	0.11	1.11	0.27	-0.10	0.35
Job Position	-0.13	0.15	-0.86	0.40	-0.44	0.18
Sector	0.30	0.17	1.76	0.08	-0.04	0.64
Size	0.07	0.13	0.56	0.58	-0.19	0.34
Credibility (indirect effect)	-0.03	0.06	-	-	-0.18	0.05

When predicting Credibility itself, with ESG Attitude as moderator, the model was overall significant (R^2 = 0.38, $F_{12,47}$ = 2.39, p = 0.02). The main effect of condition on Credibility was not significant (β = 0.14, p = 0.25), whereas ESG Attitude had a strong positive effect (β = 0.40, p < 0.001). The interaction (condition × ESG Attitude) did not reach significance (β = –0.14, p = 0.29), leading to rejection of H3b. Company size again showed a marginal effect on Credibility (β = –0.22, p = 0.05); all other covariates were non-significant (see table 4 for detailed results).

Table 4 - Results of Model 2. Moderator: Attitude toward ESG

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.14	0.12	1.17	0.25	-0.10	0.39
Attitude toward ESG	0.40	0.14	2.95	0.00	0.13	0.67
Experimental Manipulation by Attitude toward ESG	-0.14	0.13	-1.07	0.29	-0.40	0.12
Gender	-0.37	0.20	-1.84	0.07	-0.78	0.04
Age	0.00	0.01	-0.20	0.84	-0.03	0.03
Marital status	0.10	0.17	0.60	0.55	-0.24	0.44
Parental status	0.45	0.36	1.28	0.21	-0.26	1.17
Education	0.03	0.23	0.14	0.89	-0.43	0.50
Income	-0.05	0.10	-0.47	0.64	-0.24	0.15
Job Position	0.08	0.13	0.62	0.54	-0.18	0.35
Sector	-0.11	0.15	-0.73	0.47	-0.41	0.19
Size	-0.22	0.11	-1.98	0.05	-0.44	0.00

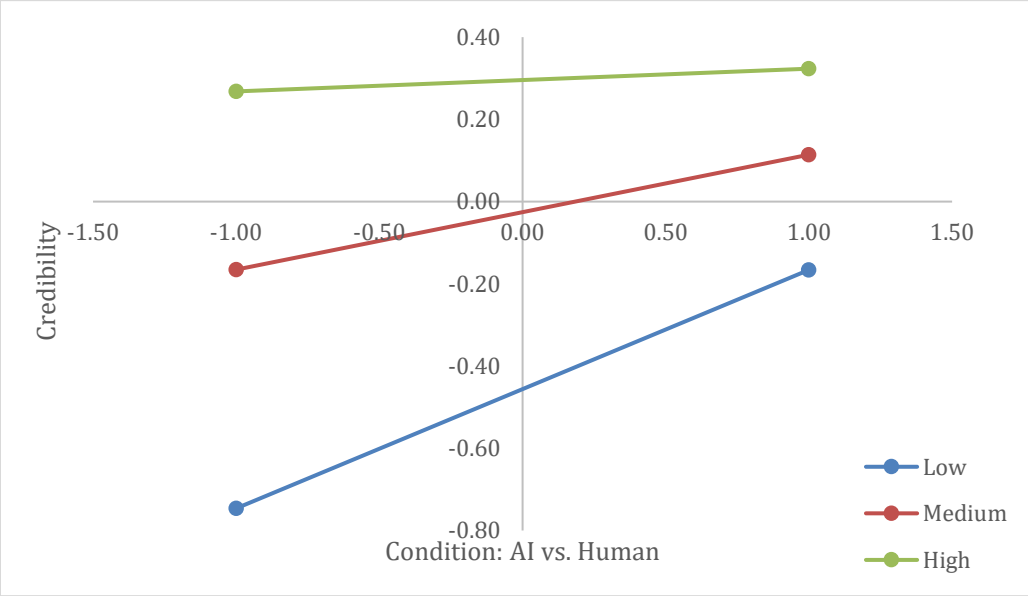


Figure 5 - Perceived Credibility by Report Source and AI-Attitude (Model 2)

The inspection of results from Model 3 showed that the overall model predicting decision confidence was significant ($R^2 = 0.42$, $F_{11,48} = 3.17$, $p < 0.001$). The experimental manipulation had no significant direct effect on confidence ($\beta = –0.10$, $p = 0.40$), but Credibility emerged as a strong predictor ($\beta = 0.57$, $p <$

0.001). None of the demographic or organizational covariates reached significance. Thus, H1c is rejected, while the direct effect of Credibility supports the underlying mechanism. The index of moderated mediation (H2c) was non-significant (Index = 0.13, LLCI = –0.11, ULCI = 0.34), so H2c is rejected (see table 5 for detailed results).

Table 5 - Results of Model 3. Dependent variable: Decision confidence

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	-0.10	0.11	-0.85	0.40	-0.32	0.13
Credibility	0.57	0.13	4.53	0.00	0.32	0.83
Gender	-0.10	0.20	-0.52	0.61	-0.50	0.29
Age	0.02	0.01	1.17	0.25	-0.01	0.04
Marital status	-0.08	0.16	-0.50	0.62	-0.40	0.24
Parental status	-0.14	0.34	-0.40	0.69	-0.83	0.55
Education	0.04	0.21	0.18	0.86	-0.38	0.46
Income	0.08	0.09	0.82	0.41	-0.11	0.26
Job Position	0.13	0.13	1.03	0.31	-0.12	0.38
Sector	0.06	0.14	0.43	0.67	-0.22	0.34
Size	0.03	0.11	0.29	0.77	-0.19	0.25
Credibility (indirect effect)	0.13	0.12	-	-	-0.11	0.34

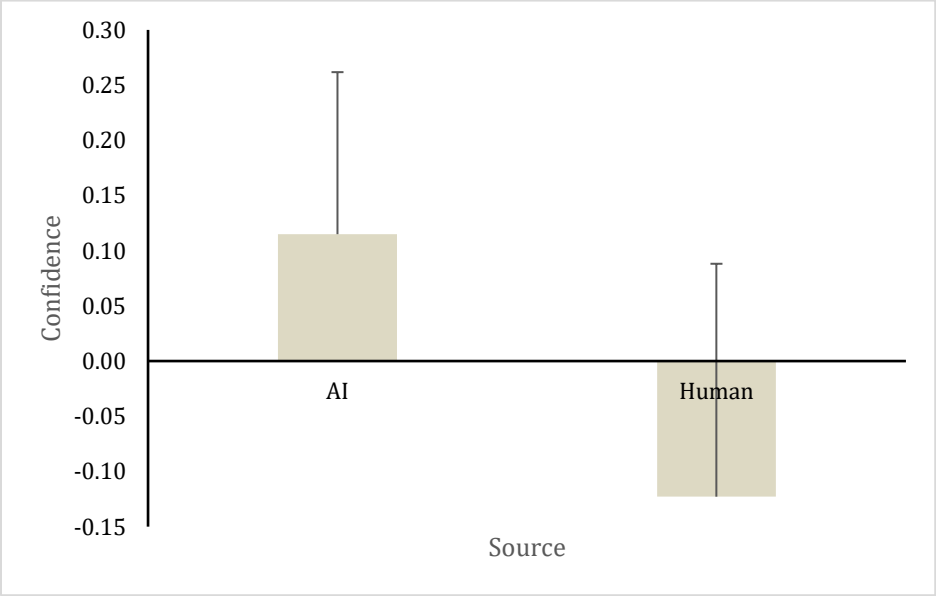


Figure 6 - Decision Confidence by Report Source (AI vs. Human)

Turning to the Credibility model with AI Attitude as moderator, the overall model was significant (R^2 = 0.35, $F_{12,47}$ = 2.11, p = 0.03). Condition again had no main effect on Credibility (β = 0.07, p = 0.58), AI Attitude was a significant positive predictor (β = 0.42, p = 0.01), and their interaction was non-significant (β = 0.22, p = 0.14), leading to rejection of H3c. Company size showed a marginal effect on Credibility (β = –0.22, p = 0.06); other covariates were non-significant (see table 6 for detailed results).

Table 6 - Results of Model 3. Moderator: Attitude toward AI

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.07	0.12	0.56	0.58	-0.18	0.31
Attitude toward AI	0.42	0.16	2.68	0.01	0.11	0.74
Experimental Manipulation by Attitude toward AI	0.22	0.15	1.48	0.14	-0.08	0.52
Gender	-0.10	0.22	-0.46	0.65	-0.55	0.34
Age	0.01	0.01	0.75	0.46	-0.02	0.04
Marital status	0.07	0.17	0.41	0.69	-0.28	0.42
Parental status	0.26	0.42	0.61	0.55	-0.59	1.10
Education	-0.23	0.23	-1.00	0.32	-0.70	0.24
Income	-0.05	0.10	-0.46	0.65	-0.24	0.15
Job Position	0.19	0.14	1.40	0.17	-0.08	0.47
Sector	-0.17	0.15	-1.17	0.25	-0.48	0.13
Size	-0.22	0.11	-1.95	0.06	-0.44	0.01

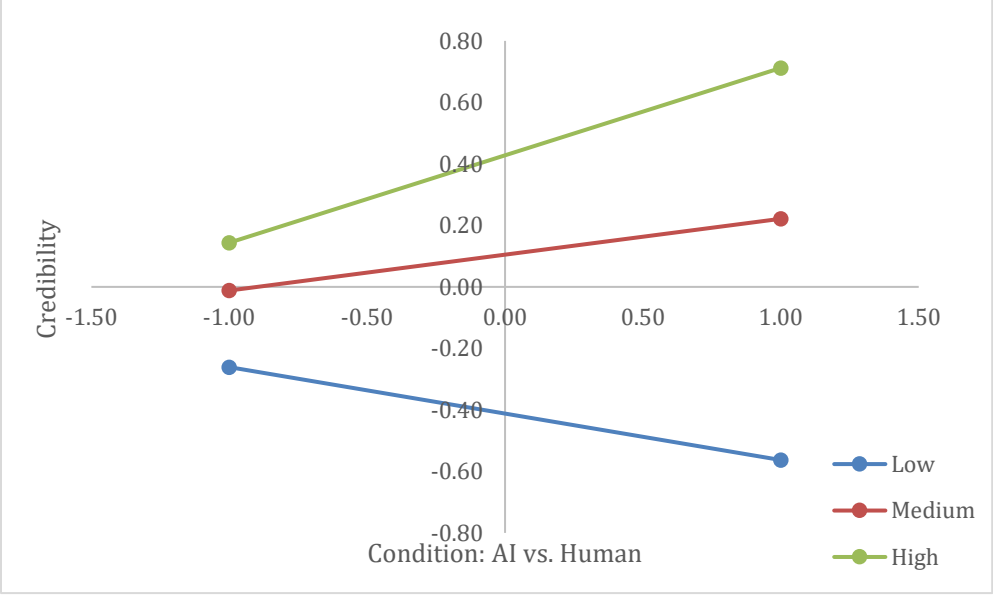


Figure 7 - Perceived Credibility by Report Source and AI-Attitude (Model 3)

In Model 4, the model predicting decision confidence was again significant ($R^2 = 0.42$, $F_{11,48} = 3.17$, $p < 0.001$). The manipulation had no direct effect ($\beta = –0.10$, $p = 0.40$), but Credibility remained a strong positive predictor of confidence ($\beta = 0.57$, $p < 0.001$). No covariates were significant. Therefore, H1d is rejected, and the direct credibility effect aligns with the theorized pathway. The moderated mediation index for ESG Attitude (H2d) was non-significant (Index = –0.08, LLCI = –0.31, ULCI = 0.12), so H2d is rejected (see table 7 for detailed results).

Table 7 - Results of Model 4. Dependent variable: Decision confidence

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	-0.10	0.11	-0.85	0.40	-0.32	0.13
Credibility	0.57	0.13	4.53	0.00	0.32	0.83
Gender	-0.10	0.20	-0.52	0.61	-0.50	0.29
Age	0.02	0.01	1.17	0.25	-0.01	0.04
Marital status	-0.08	0.16	-0.50	0.62	-0.40	0.24
Parental status	-0.14	0.34	-0.40	0.69	-0.83	0.55
Education	0.04	0.21	0.18	0.86	-0.38	0.46
Income	0.08	0.09	0.82	0.41	-0.11	0.26
Job Position	0.13	0.13	1.03	0.31	-0.12	0.38
Sector	0.06	0.14	0.43	0.67	-0.22	0.34
Size	0.03	0.11	0.29	0.77	-0.19	0.25
Credibility (indirect effect)	-0.08	0.11	-	-	-0.31	0.12

In the corresponding Credibility model, with ESG Attitude as moderator, the overall fit was significant ($R^2 = 0.38$, $F_{12,47} = 2.39$, $p = 0.02$). Condition did not affect Credibility ($\beta = 0.14$, $p = 0.25$), ESG Attitude had a significant positive effect ($\beta = 0.40$, $p < 0.001$), and the interaction was non-significant ($\beta = –0.14$, $p = 0.29$), leading to rejection of H3d. Company size again approached significance ($\beta = –0.22$, $p = 0.05$); all other covariates were non-significant (see table 8 for detailed results).

Table 8 - Results of Model 4. Moderator: Attitude toward ESG

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.14	0.12	1.17	0.25	-0.10	0.39
Attitude toward ESG	0.40	0.14	2.95	0.00	0.13	0.67
Experimental Manipulation by Attitude toward ESG	-0.14	0.13	-1.07	0.29	-0.40	0.12
Gender	-0.37	0.20	-1.84	0.07	-0.78	0.04
Age	0.00	0.01	-0.20	0.84	-0.03	0.03
Marital status	0.10	0.17	0.60	0.55	-0.24	0.44
Parental status	0.45	0.36	1.28	0.21	-0.26	1.17
Education	0.03	0.23	0.14	0.89	-0.43	0.50
Income	-0.05	0.10	-0.47	0.64	-0.24	0.15
Job Position	0.08	0.13	0.62	0.54	-0.18	0.35
Sector	-0.11	0.15	-0.73	0.47	-0.41	0.19
Size	-0.22	0.11	-1.98	0.05	-0.44	0.00

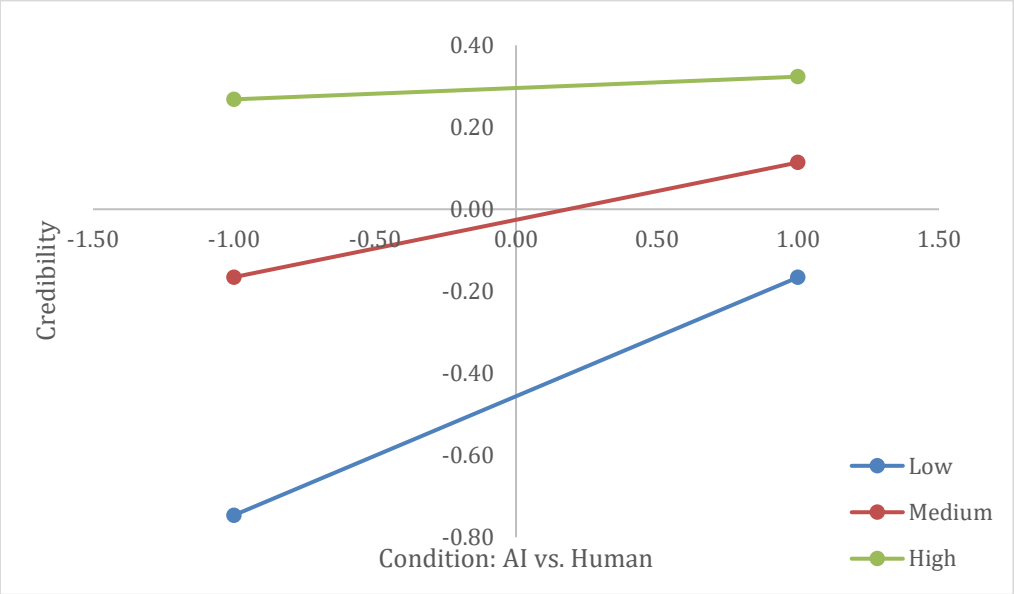


Figure 8 - Perceived Credibility by Report Source and AI-Attitude (Model 4)

The inspection of results from the model predicting accuracy (DV) with AI Attitude as moderator (Model 5) showed that the overall model was significant ($R^2 = 0.41$, $F_{11,48} = 3.01$, $p < 0.001$). The experimental manipulation (AI- vs. Human-generated report) did not have a significant direct effect on accuracy ($\beta = -0.06$, $p = 0.61$), whereas Credibility was a strong positive predictor ($\beta = 0.59$, $p < 0.001$). None of the covariates reached significance. Thus, H1e is rejected. The index of moderated mediation for Credibility

(H2e) was not significant (Index = 0.13, LLCI = –0.11, ULCI = 0.35), leading to rejection of H2e (see table 9 for detailed results).

Table 9 - Results of Model 5. Dependent variable: Perceived Accuracy

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	-0.06	0.11	-0.52	0.61	-0.29	0.17
Credibility	0.59	0.13	4.62	0.00	0.33	0.85
Gender	-0.06	0.20	-0.28	0.78	-0.46	0.34
Age	0.01	0.01	0.95	0.35	-0.01	0.04
Marital status	-0.05	0.16	-0.29	0.77	-0.37	0.28
Parental status	-0.13	0.35	-0.38	0.70	-0.83	0.56
Education	0.07	0.21	0.33	0.75	-0.36	0.49
Income	0.04	0.09	0.43	0.67	-0.15	0.23
Job Position	0.16	0.13	1.27	0.21	-0.10	0.42
Sector	0.04	0.14	0.29	0.78	-0.25	0.33
Size	0.10	0.11	0.91	0.37	-0.12	0.32
Credibility (indirect effect)	0.13	0.12	-	-	-0.11	0.35

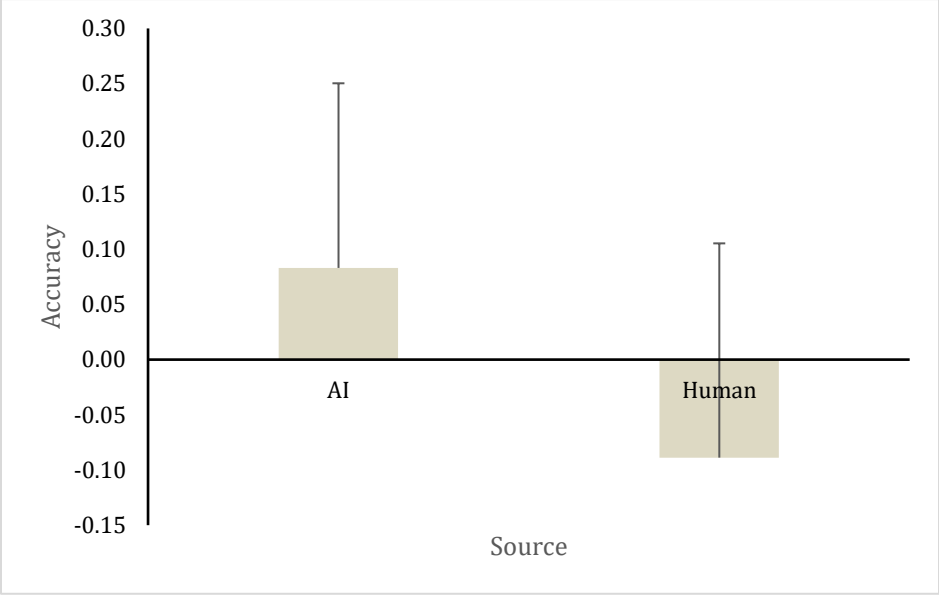


Figure 9 - Accuracy Judgments by Report Source (AI vs. Human)

For the Credibility model with AI Attitude as moderator, the model was overall significant ($R^2 = 0.35$, $F_{12,47} = 2.11$, $p = 0.03$). Condition again showed no effect on Credibility ($\beta = 0.07$, $p = 0.58$), AI Attitude had a significant positive effect ($\beta = 0.42$, $p = 0.01$), but the interaction term was non-significant ($\beta = 0.22$, $p = 0.14$), so H3e is rejected. Company size approached significance ($\beta = –0.22$, $p = 0.06$); all other covariates were non-significant (see table 10 for detailed results).

Table 10 - Results of Model 5. Moderator: Attitude toward AI

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.07	0.12	0.56	0.58	-0.18	0.31
Attitude toward AI	0.42	0.16	2.68	0.01	0.11	0.74
Experimental Manipulation by Attitude toward AI	0.22	0.15	1.48	0.14	-0.08	0.52
Gender	-0.10	0.22	-0.46	0.65	-0.55	0.34
Age	0.01	0.01	0.75	0.46	-0.02	0.04
Marital status	0.07	0.17	0.41	0.69	-0.28	0.42
Parental status	0.26	0.42	0.61	0.55	-0.59	1.10
Education	-0.23	0.23	-1.00	0.32	-0.70	0.24
Income	-0.05	0.10	-0.46	0.65	-0.24	0.15
Job Position	0.19	0.14	1.40	0.17	-0.08	0.47
Sector	-0.17	0.15	-1.17	0.25	-0.48	0.13
Size	-0.22	0.11	-1.95	0.06	-0.44	0.01

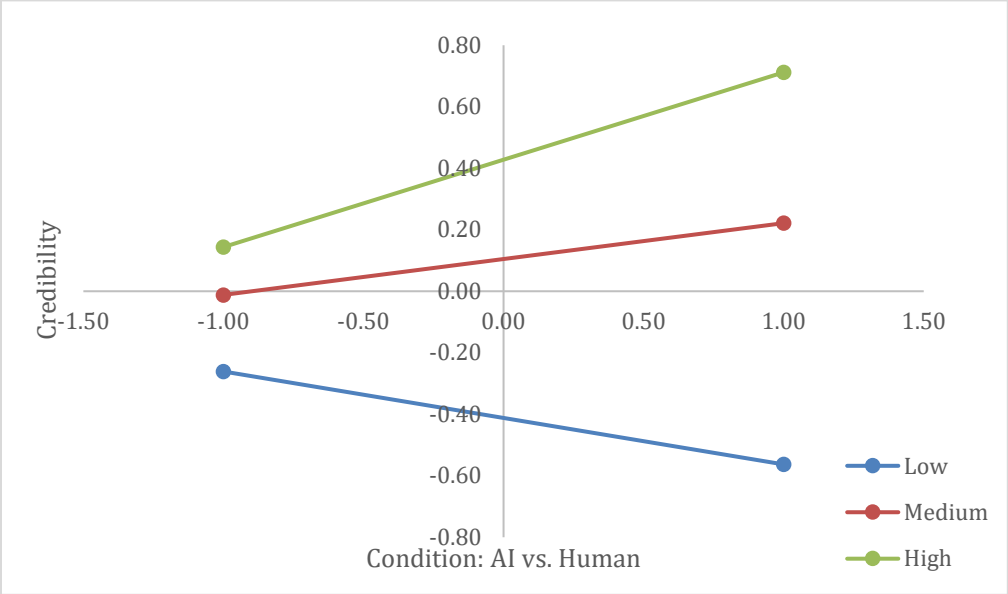


Figure 10 - Perceived Credibility by Report Source and AI-Attitude (Model 5)

For Model 6, the DV model was again significant ($R^2 = 0.41$, $F_{11,48} = 3.01$, $p < 0.001$). The manipulation had no direct effect ($\beta = –0.06$, $p = 0.61$), Credibility remained a strong predictor of accuracy ($\beta = 0.59$, $p < 0.001$), and no covariates were significant. Hence H1f is rejected. The moderated-mediation index for Credibility (H2f) was non-significant (Index = –0.08, LLCI = –0.33, ULCI = 0.12), leading to rejection of H2f (see table 11 for detailed results).

Table 11 - Results of Model 6. Dependent variable: Perceived accuracy

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	-0.06	0.11	-0.52	0.61	-0.29	0.17
Credibility	0.59	0.13	4.62	0.00	0.33	0.85
Gender	-0.06	0.20	-0.28	0.78	-0.46	0.34
Age	0.01	0.01	0.95	0.35	-0.01	0.04
Marital status	-0.05	0.16	-0.29	0.77	-0.37	0.28
Parental status	-0.13	0.35	-0.38	0.70	-0.83	0.56
Education	0.07	0.21	0.33	0.75	-0.36	0.49
Income	0.04	0.09	0.43	0.67	-0.15	0.23

Job Position	0.16	0.13	1.27	0.21	-0.10	0.42
Sector	0.04	0.14	0.29	0.78	-0.25	0.33
Size	0.10	0.11	0.91	0.37	-0.12	0.32
Credibility (indirect effect)	-0.08	0.11	-	-	-0.33	0.12

In the corresponding Credibility model (ESG Attitude moderator), the overall fit was significant ($R^2 = 0.38$, $F_{12,47} = 2.39$, $p = 0.02$). Condition did not affect Credibility ($\beta = 0.14$, $p = 0.25$), ESG Attitude had a strong positive effect ($\beta = 0.40$, $p < 0.001$), and the interaction was non-significant ($\beta = –0.14$, $p = 0.29$), so H3f is rejected. Company size again approached significance ($\beta = –0.22$, $p = 0.05$); all other covariates were non-significant (see table 12 for detailed results).

Table 12 - Results of Model 6. Moderator: Attitude toward ESG

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.14	0.12	1.17	0.25	-0.10	0.39
Attitude toward ESG	0.40	0.14	2.95	0.00	0.13	0.67
Experimental Manipulation by Attitude toward ESG	-0.14	0.13	-1.07	0.29	-0.40	0.12
Gender	-0.37	0.20	-1.84	0.07	-0.78	0.04
Age	0.00	0.01	-0.20	0.84	-0.03	0.03
Marital status	0.10	0.17	0.60	0.55	-0.24	0.44
Parental status	0.45	0.36	1.28	0.21	-0.26	1.17
Education	0.03	0.23	0.14	0.89	-0.43	0.50
Income	-0.05	0.10	-0.47	0.64	-0.24	0.15
Job Position	0.08	0.13	0.62	0.54	-0.18	0.35
Sector	-0.11	0.15	-0.73	0.47	-0.41	0.19
Size	-0.22	0.11	-1.98	0.05	-0.44	0.00

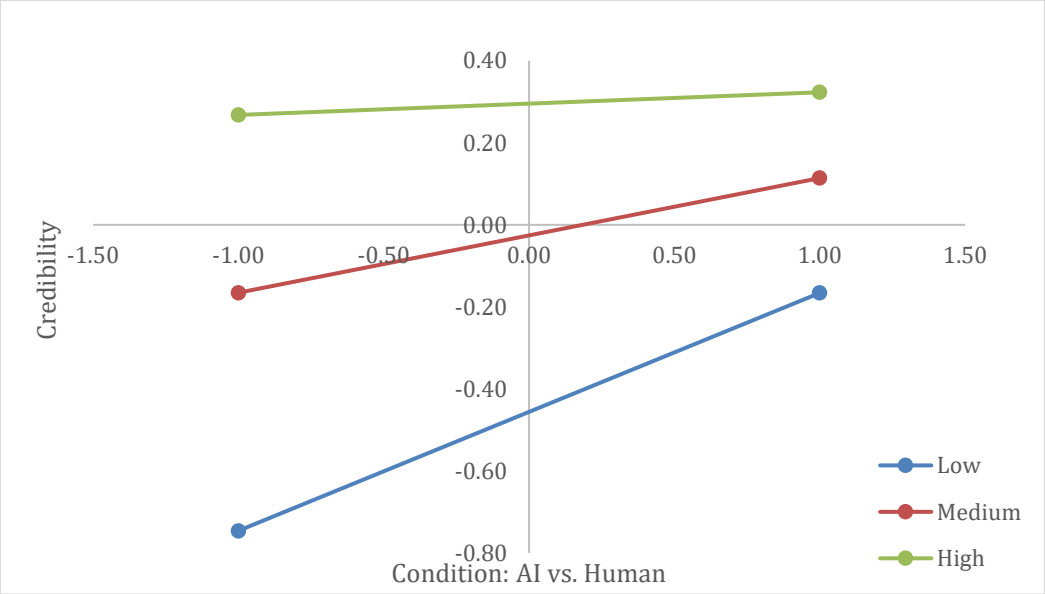


Figure 11 - Perceived Credibility by Report Source and AI-Attitude (Model 6)

Model 7 did not reach significance overall ($R^2 = 0.28$, $F_{11,48} = 1.72$, $p = 0.10$). The experimental manipulation had no effect on influence ($\beta = 0.09$, $p = 0.47$), but Credibility was a significant predictor ($\beta = 0.42$, $p < 0.001$). No covariates were significant, leading to rejection of H1g. The index of moderated mediation for Credibility (H2g) was non-significant (Index = 0.09, LLCI = –0.10, ULCI = 0.24), so H2g is rejected (see table 13 for detailed results).

Table 13 - Results of Model 7. Dependent variable: Perceived influence

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.09	0.13	0.73	0.47	-0.16	0.34
Credibility	0.42	0.14	2.98	0.00	0.14	0.70
Gender	0.26	0.22	1.17	0.25	-0.18	0.70
Age	-0.02	0.01	-1.38	0.17	-0.05	0.01
Marital status	0.15	0.18	0.85	0.40	-0.21	0.51
Parental status	-0.35	0.38	-0.91	0.37	-1.11	0.42
Education	-0.08	0.23	-0.32	0.75	-0.54	0.39
Income	0.12	0.10	1.15	0.26	-0.09	0.32
Job Position	0.03	0.14	0.19	0.85	-0.26	0.31
Sector	0.08	0.16	0.50	0.62	-0.24	0.39
Size	-0.14	0.12	-1.15	0.26	-0.38	0.10
Credibility (indirect effect)	0.09	0.09	-	-	-0.10	0.24

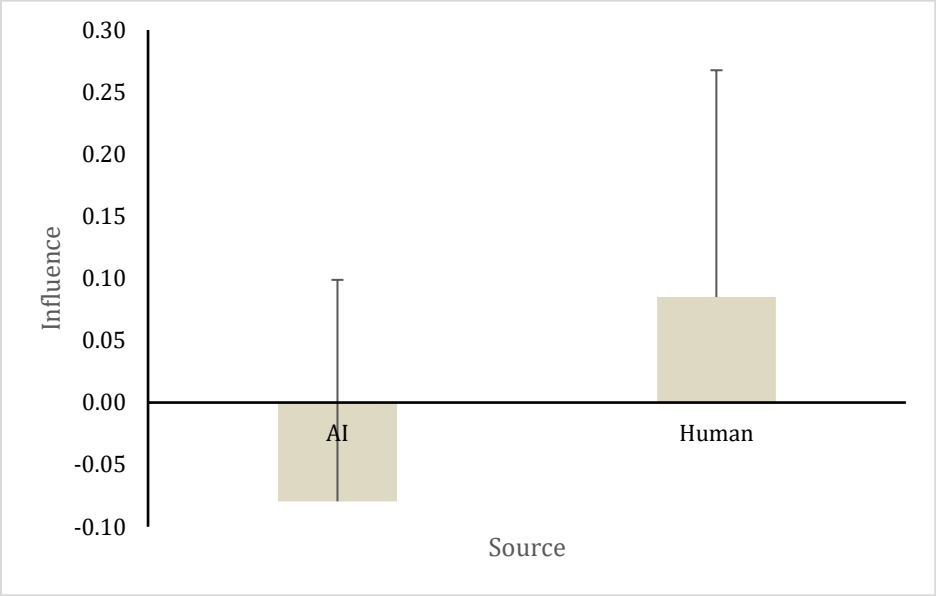


Figure 12 - Perceived Influence by Report Source (AI vs. Human)

For the Credibility model with AI Attitude as moderator, the pattern again mirrored Model 5's Credibility results: overall significant ($R^2 = 0.35$, $F_{12,47} = 2.11$, $p = 0.03$), no main effect of condition ($\beta = 0.07$, $p = 0.58$), a significant effect of AI Attitude ($\beta = 0.42$, $p = 0.01$), and a non-significant interaction ($\beta = 0.22$, $p = 0.14$). H3g is therefore rejected, with company size marginal ($\beta = –0.22$, $p = 0.06$) (see table 14 for detailed results).

Table 14 - Results of Model 7. Moderator: Attitude toward AI

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.07	0.12	0.56	0.58	-0.18	0.31
Attitude toward AI	0.42	0.16	2.68	0.01	0.11	0.74
Experimental Manipulation by Attitude toward AI	0.22	0.15	1.48	0.14	-0.08	0.52
Gender	-0.10	0.22	-0.46	0.65	-0.55	0.34
Age	0.01	0.01	0.75	0.46	-0.02	0.04
Marital status	0.07	0.17	0.41	0.69	-0.28	0.42
Parental status	0.26	0.42	0.61	0.55	-0.59	1.10
Education	-0.23	0.23	-1.00	0.32	-0.70	0.24
Income	-0.05	0.10	-0.46	0.65	-0.24	0.15
Job Position	0.19	0.14	1.40	0.17	-0.08	0.47
Sector	-0.17	0.15	-1.17	0.25	-0.48	0.13
Size	-0.22	0.11	-1.95	0.06	-0.44	0.01

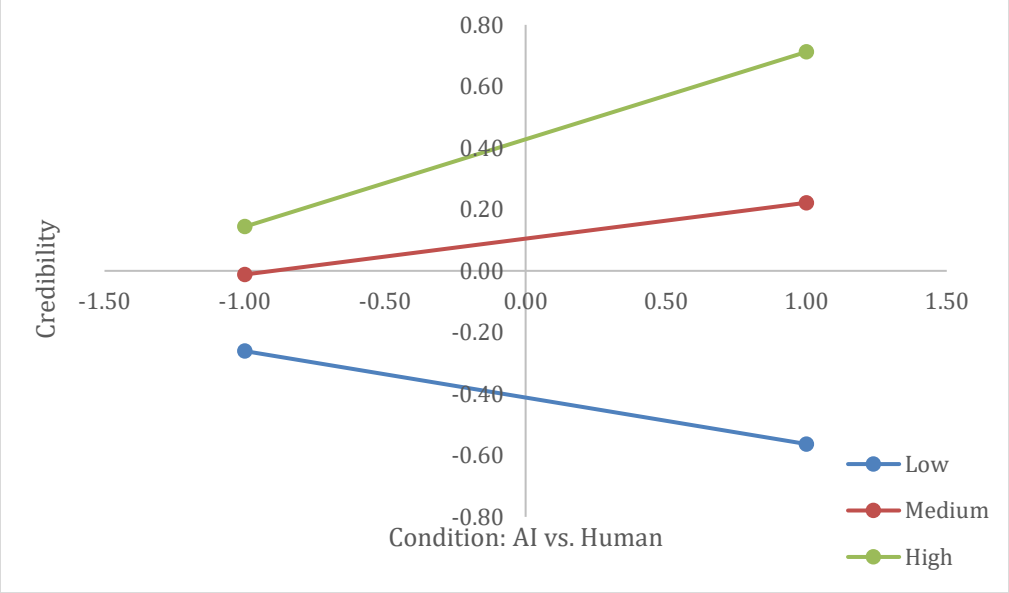


Figure 13 - Perceived Credibility by Report Source and AI-Attitude (Model 7)

Finally, Model 8 was also non-significant overall ($R^2 = 0.28$, $F_{11,48} = 1.72$, $p = 0.10$). Condition had no effect ($\beta = 0.09$, $p = 0.47$), Credibility remained significant ($\beta = 0.42$, $p < 0.001$), and no covariates were significant—H1h is rejected. The moderated mediation index for Credibility (H2h) was non-significant (Index = –0.06, LLCI = –0.29, ULCI = 0.06), leading to rejection of H2h (see table 15 for detailed results).

Table 15 - Results of Model 8. Dependent variable: Perceived influence

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.09	0.13	0.73	0.47	-0.16	0.34
Credibility	0.42	0.14	2.98	0.00	0.14	0.70
Gender	0.26	0.22	1.17	0.25	-0.18	0.70
Age	-0.02	0.01	-1.38	0.17	-0.05	0.01
Marital status	0.15	0.18	0.85	0.40	-0.21	0.51
Parental status	-0.35	0.38	-0.91	0.37	-1.11	0.42
Education	-0.08	0.23	-0.32	0.75	-0.54	0.39
Income	0.12	0.10	1.15	0.26	-0.09	0.32
Job Position	0.03	0.14	0.19	0.85	-0.26	0.31
Sector	0.08	0.16	0.50	0.62	-0.24	0.39
Size	-0.14	0.12	-1.15	0.26	-0.38	0.10
Credibility (indirect effect)	-0.06	0.09	-	-	-0.29	0.06

In the Credibility model with ESG Attitude as moderator, the same pattern held ($R^2 = 0.38$, $F_{12,47} = 2.39$, $p = 0.02$): condition non-significant ($\beta = 0.14$, $p = 0.25$), ESG Attitude significant ($\beta = 0.40$, $p < 0.001$), interaction non-significant ($\beta = –0.14$, $p = 0.29$), and H3h is rejected (company size marginal, $\beta = –0.22$, $p = 0.05$) (see table 16 for detailed results).

Table 16 - Results of Model 8. Moderator: Attitude toward ESG

	β	SE	t	p	LLCI	ULCI
Experimental Manipulation	0.14	0.12	1.17	0.25	-0.10	0.39
Attitude toward ESG	0.40	0.14	2.95	0.00	0.13	0.67
Experimental Manipulation by Attitude toward ESG	-0.14	0.13	-1.07	0.29	-0.40	0.12
Gender	-0.37	0.20	-1.84	0.07	-0.78	0.04
Age	0.00	0.01	-0.20	0.84	-0.03	0.03
Marital status	0.10	0.17	0.60	0.55	-0.24	0.44
Parental status	0.45	0.36	1.28	0.21	-0.26	1.17
Education	0.03	0.23	0.14	0.89	-0.43	0.50
Income	-0.05	0.10	-0.47	0.64	-0.24	0.15
Job Position	0.08	0.13	0.62	0.54	-0.18	0.35
Sector	-0.11	0.15	-0.73	0.47	-0.41	0.19
Size	-0.22	0.11	-1.98	0.05	-0.44	0.00

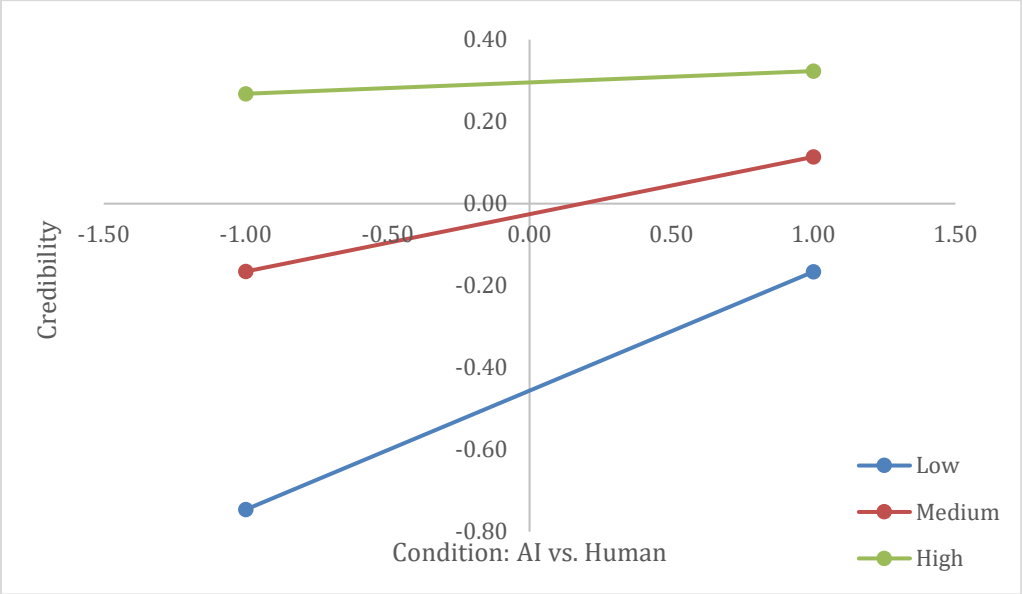


Figure 14 - Perceived Credibility by Report Source and AI-Attitude (Model 8)

4. Discussion

4.1 Main Findings

Across all eight experimental models, the core finding was strikingly consistent: whether managers believed they were evaluating ESG metrics generated by an AI system or by a seasoned human expert had no discernible impact on any of our behavioural or attitudinal outcome measures. Participants in the "AI" and "Expert" conditions allocated their hypothetical budgets in nearly identical patterns, reported comparable levels of confidence in their allocation decisions, and rated the influence of the ESG report on those decisions at the same level, despite being told the source of the data differed. Nor did we observe any differences in how accurate they believed the information to be. In contrast, the single factor that reliably predicted all these outcomes was the credibility that participants attributed to the ESG report itself. Managers who rated the report as more credible consistently exhibited higher decision confidence, stronger perceptions of accuracy, and greater acknowledgment that the report had shaped their choices. This credibility effect emerged regardless of source labelling and remained robust after accounting for demographic and organizational covariates such as age, gender, industry sector, company size, and job position.

Digging deeper into what drove credibility perceptions, we found that individual attitudes, both toward artificial intelligence in general and toward ESG principles specifically, played a pivotal role. Participants who entered the study with a positive orientation toward AI were significantly more inclined to endorse the report's trustworthiness, as were those who held especially strong commitments to environmental, social, and governance values. These predispositions shaped credibility judgments in a direct, linear fashion: the more favourable one's AI- or ESG-related attitudes, the more credible the report appeared, and hence the more confidently and affirmatively one acted upon it. Crucially, we tested whether these attitudes might interact with the source label, hypothesizing that AI enthusiasts might show a source-dependent boost in

credibility when metrics were branded as “AI-generated,” or that ESG devotees might likewise favour the expert attribution, but found no evidence of such moderation. In every model, neither the AI–attitude by source nor the ESG–attitude by source interaction reached significance, indicating that source identity was effectively inert once individual predispositions and overall credibility perceptions were considered.

Taken together, these results reveal that in the context of ESG decision-making, managers do not reflexively defer to or reject information based on whether it comes from an algorithm or a human. Rather, they rely on their own belief systems to judge the inherent trustworthiness of the data, and it is the perceived credibility that ultimately drives their confidence, their sense of accuracy, and the degree to which they allow the ESG information to guide their resource allocations. In sum, the main empirical insight is that source labelling alone has no practical effect on ESG decision outcomes; instead, credibility perceptions rooted in personal attitudes are the decisive force. That said, our ability to detect even subtle source-labelling effects may have been constrained by the study’s modest sample size. We cannot definitively rule out effects that a larger, more powerful design might uncover. Future work with greater statistical power will be needed to confirm whether truly no source effect exists or whether our null findings reflect a limitation of sample size rather than substantive absence of an effect.

4.2 Theoretical and Managerial Implications

Theoretically, our findings demonstrate that perceived credibility, filtered through managers’ pre-existing attitudes toward both AI and sustainability, is the fulcrum upon which trust in ESG metrics pivots. This suggests that classical frameworks of algorithm aversion and algorithm appreciation must be extended to incorporate attitudinal moderators not merely as peripheral covariates but as central, first-order constructs that shape credibility appraisals and downstream behavioural outcomes. In practice, this means that future theoretical work should move beyond dichotomous “source-based” trust models to develop more nuanced, integrative theories that position individual predispositions, such as technology readiness, sustainability commitment, and domain-specific expertise, as key antecedents of credibility perceptions.

Moreover, our evidence challenges the assumption embedded in many trust frameworks that transparency or explainability interventions will uniformly enhance trust; without aligning with managers' underlying value systems, even highly transparent AI outputs may fail to register as credible. Consequently, theoretical advances should explore how credibility emerges from an interaction between system attributes (e.g., explainability, methodological rigor) and user characteristics (e.g., AI attitudes, ESG orientations), and how this interplay governs the actual adoption and utilization of algorithmic advice in high-stakes, value-laden contexts.

From a managerial standpoint, the practical implication is unequivocal: branding sustainability metrics as "AI-driven" or "expert-generated" is unlikely, on its own, to influence managerial behaviour or trust. Instead, organizations should prioritize strategies that actively build and sustain credibility in their ESG reporting processes. First, firms must invest in understanding the AI attitudes and ESG convictions of their decision-makers (through surveys, focus groups, or attitude-assessment tools) and tailor communication and training programs accordingly. Managers who are initially sceptical of AI may benefit from immersive demonstrations that highlight real-world AI successes in ESG analytics, coupled with opportunities for hands-on experimentation, thereby gradually reshaping their attitudinal predispositions and, in turn, their credibility judgments. Conversely, sustainability-oriented managers might be engaged through deep dives into the methodological underpinnings of ESG models, reinforcing how the AI or expert systems align with their core values. Second, reporting platforms, whether AI-based or human-curated, should emphasize credibility-enhancing features: clear documentation of data sources, transparent methodological summaries, and contextual annotations that link metrics to strategic objectives. Embedding interactive "credibility checkpoints," such as confidence intervals, provenance trails, or third-party validations, can further reinforce trust by speaking directly to managers' desire for assurance.

Finally, leadership should foster a hybrid decision-making culture in which AI outputs and expert insights are presented side by side, not as competing authorities but as complementary perspectives. By doing so, companies can leverage the distinct strengths of both human and machine analysis, while signalling to

managers that credibility arises from the robustness of information and alignment with organizational values, rather than from the identity of the information's originator. In sum, success will hinge on shifting the focus from "who" produces ESG metrics to "how" those metrics resonate with the beliefs and expectations of the decision-makers who rely on them.

4.3 Limitations

While we designed the experiment and methods carefully, there are some limitations to consider. These limitations provide context for the results and caution against generalizing beyond the study.

Given the highly specific and difficult-to-reach target population, the final sample size (N = 60) was relatively small and distributed across two experimental conditions (between-subjects design), hence further limiting power. This limited the statistical power, particularly for detecting interaction effects and moderated mediation. In fact, none of the eight models showed statistically significant conditional effects, suggesting that the sample may not have been large enough to capture subtle variations in how the experimental condition interacted with participants' attitudes.

Our between-subjects online experiment was based on a single "Example Company" report and a one-time budget allocation task. While this controls for internal validity it can't capture the iterative, collaborative and high stakes nature of real business decision making or how trust in AI versus expert inputs might change with repeated exposure, peer review or looming deadlines.

We presented only one ESG report (the same one across all conditions except for the source) and underrepresent the richness and complexity of actual sustainability disclosures which often include multiple data sources, narrative and stakeholder feedback. This may have reduced any subtle reactions managers would have had when weighing complex or conflicting information.

All core constructs (trust, credibility, ESG and AI attitudes, decision confidence) were measured via self-reported Likert scales, which are subject to response biases (social desirability, common method variance) and don't give us insight into real world behaviour.

Despite these caveats, the design is still a useful test of the research questions, but readers should keep these in mind.

References

Alon-Barkat, S., & Busuioc, M. (2023). Human–AI interactions in public sector decision-making: "Automation bias" and "selective adherence" to algorithmic advice. Journal of Public Administration Research and Theory, 33(1), 153–169. https://doi.org/10.1093/jopart/muac007

Barnea, A., & Rubin, A. (2010). Corporate social responsibility as a conflict between shareholders. Journal of Business Ethics, 97(1), 71–86. https://doi.org/10.1007/s10551-010-0496-z

Bashkirova, A., & Krpan, C. (2024). Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. Computers in Human Behavior: Artificial Humans, 2(1), Article 100066. https://doi.org/10.1016/j.chbah.2024.100066

Berg, F., Kölbel, J. F., & Rigobon, R. (2020). Aggregate confusion: The divergence of ESG ratings. MIT Sloan Research Paper No. 5822-19. https://doi.org/10.2139/ssrn.3438533

Centre for Sustainability and Excellence. (2025, February 4). Real-time sustainability reporting: How AI is transforming ESG in Canada. Centre for Sustainability and Excellence. https://cse-net.org/real-time-sustainability-reporting-ai/

Correia, A., & Água, P. B. (2024). Harnessing artificial intelligence for enhanced environmental, social, and governance reporting: A new paradigm in corporate transparency. https://doi.org/10.22495/cgrapp15

Deegan, C. (2002). Introduction: The legitimising effect of social and environmental disclosures - A theoretical foundation. Accounting, Auditing & Accountability Journal, 15(3), 282–311. https://doi.org/10.1108/09513570210435852

Dhaliwal, D. S., Li, O. Z., Tsang, A., & Yang, Y. G. (2011). Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting. The Accounting Review, 86(1), 59–100. https://doi.org/10.2308/accr.00000005

Di Maggio, P. J., & Powell, W. W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. American Sociological Review, 48(2), 147–160. https://www.jstor.org/stable/2095101

Dietvorst, B. J. (2017, July 5). When people don't trust algorithms. MIT Sloan Management Review. https://sloanreview.mit.edu/article/when-people-dont-trust-algorithms/

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033

Dirks, K. T., & Ferrin, D. L. (2001). The role of trust in organizational settings. Organization Science, 12(4), 450–467. https://doi.org/10.1287/orsc.12.4.450.10640

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. Academy of Management Review, 20(1), 65–91. https://doi.org/10.5465/amr.1995.9503271992

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human-Computer Studies, 58(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. Management Science, 60(11), 2835–2857. https://doi.org/10.1287/mnsc.2014.1984

Elkington, J. (1997). Cannibals with forks: The triple bottom line of 21st century business. Capstone.

Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99) (pp. 80–87). ACM. https://doi.org/10.1145/302979.303001

Freeman, R. E. (1984). Strategic management: A stakeholder approach. Pitman Publishing.

Friede, G., Busch, T., & Bassen, A. (2015). ESG and financial performance: Aggregated evidence from more than 2,000 empirical studies. Journal of Sustainable Finance & Investment, 5(4), 210–233. https://doi.org/10.1080/20430795.2015.1118917

Gillan, S. L., Koch, A., & Starks, L. T. (2021). Firms and social responsibility: A review of ESG and CSR research in corporate finance. Journal of Corporate Finance, 66, Article 101889. https://doi.org/10.1016/j.jcorpfin.2021.101889

Giudici, P., & Wu, L. (2025). Sustainable artificial intelligence in finance: Impact of ESG factors. Frontiers in Artificial Intelligence, 8, Article 1566197. https://doi.org/10.3389/frai.2025.1566197

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Global Reporting Initiative. (2016). GRI Sustainability Reporting Standards 2016. https://www.globalreporting.org/standards

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, *14*, Article 1191628. https://doi.org/10.3389/fpsyg.2023.1191628

Hodge, K., Subramaniam, N., & Stewart, J. (2009). Assurance of sustainability reports: Impact on report users' confidence and perceptions of information credibility. Australian Accounting Review, 19(3), 178–194. https://doi.org/10.1111/j.1835-2561.2009.00056.x

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). Communication and persuasion. Yale University Press.

Jamali, D., Safieddine, A.M. and Rabbath, M. (2008), Corporate Governance and Corporate Social Responsibility Synergies and Interrelationships. Corporate Governance: An International Review, 16: 443-459. https://doi.org/10.1111/j.1467-8683.2008.00702.x

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360. https://doi.org/10.1016/0304-405X(76)90026-X

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? Communications of the ACM, 59(10), 16–17. https://doi.org/10.1145/2983270

Lewicki, R. J., McAllister, D. J., & Bies, R. J. (1998). Trust and distrust: New relationships and realities. Academy of Management Review, 23(3), 438–458. https://doi.org/10.5465/amr.1998.926620

Lock, I., & Seele, P. (2017). Measuring credibility perceptions in CSR communication: The PERCRED scale. Management Communication Quarterly, 31(4), 584–613. https://doi.org/10.1177/0893318917707592

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Malmendier, U., & Tate, G. (2005). CEO overconfidence and corporate investment. Journal of Finance, 60(6), 2661–2700. https://doi.org/10.1111/j.1540-6261.2005.00813.x

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

Mazzacurati, J. (2021). ESG ratings: Status and key issues ahead. ESMA Report on Trends, Risks and Vulnerabilities No. 1, 2021. European Securities and Markets Authority. https://www.esma.europa.eu/sites/default/files/trv_2021_1-esg_ratings_status_and_key_issues_ahead.pdf

McKinsey & Company. (2021). The state of AI in 2021: AI adoption trends by industry and function. https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2021

Milne, M. J., & Gray, R. (2013). W(h)ither ecology? The triple bottom line, the Global Reporting Initiative, and corporate sustainability reporting. Journal of Business Ethics, 118(1), 13–29. https://doi.org/10.1007/s10551-012-1543-8

Mosier, K. L., & Skitka, L. J. (1997). Automation bias: Decision making and performance in high-tech cockpits. International Journal of Aviation Psychology, 8(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3

Palmucci, D. N., & Ferraris, A. (2023). Climate change inaction: Cognitive bias influencing managers' decision making on environmental sustainability choices. The role of empathy and morality with the need of an integrated and comprehensive perspective. Frontiers in Psychology, 14, 1130059. https://doi.org/10.3389/fpsyg.2023.1130059

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. Human Factors, 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. Academy of Management Review, 20(3), 571–610. https://doi.org/10.5465/amr.1995.9508080331

Sulkowski, A. J. (2024). AI, ESG, and Law: Potential, Limitations, and Strategies Concerning Artificial Intelligence in Sustainability Reporting. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4972787

Sustainability Directory. (2025, April 2). Can we trust AI-driven ESG ratings? Sustainability Directory. https://sustainability-directory.com/question/can-we-trust-ai-driven-esg-ratings/

Sustainability Directory. (2025, January 24). How does data bias affect trust? Sustainability Directory. https://esg.sustainability-directory.com/question/how-does-data-bias-affect-trust/

Söllner, M., Hoffmann, A., & Leimeister, J. M. (2016). Why different trust relationships matter for information systems users. European Journal of Information Systems, 25(3), 274–287. https://doi.org/10.1057/ejis.2015.17

United Nations Global Compact Office. (2004). Who cares wins: Connecting financial markets to a changing world. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/280911488968799581/who-cares-wins-connecting-financial-markets-to-a-changing-world

Wesche, J. S., Hennig, F., Kollhed, C. S., Quade, J., Kluge, S., & Sonderegger, A. (2022). People's reactions to decisions by human vs. algorithmic decision-makers: The role of explanations and type of selection tests. European Journal of Work and Organizational Psychology, 33(1). https://doi.org/10.1080/1359432X.2022.2132940

White, B. (2023, November 30). Potential opportunities and risks AI poses for ESG performance. National Law Review. https://natlawreview.com/article/potential-opportunities-and-risks-ai-poses-esg-performance

Xiao, Y., & Xiao, L. (2025). The impact of artificial intelligence–driven ESG performance on sustainable development of central state-owned enterprises listed companies. Scientific Reports, 15, Article 8548. https://doi.org/10.1038/s41598-025-93694-y