



Kent Academic Repository

Lioliou, Eleni, Seitanidi, M. May and Stadtler, Lea (2026) *Governance under algorithmic opacity: How financial firms construct accountability and control around AI in risk disclosures*. Information Systems Frontiers . ISSN 1387-3326.

Downloaded from

<https://kar.kent.ac.uk/115030/> The University of Kent's Academic Repository KAR

The version of record is available from

<https://doi.org/10.1007/s10796-026-10762-y>

This document version

Publisher pdf

DOI for this version

Licence for this version

CC BY (Attribution)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in **Title of Journal** , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our [Take Down policy](https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies) (available from <https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies>).



Governance under Algorithmic Opacity: How Financial Firms Construct Accountability and Control around AI in Risk Disclosures

Eleni Lioliou¹ · M. May Seitanidi² · Lea Stadtler³

Received: 1 December 2025 / Accepted: 14 May 2026

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

As artificial intelligence (AI) systems reshape financial decision-making, firms face increasing pressure to assure regulators and investors that they remain in control of new technologies that are inherently opaque and difficult to predict. This study examines how financial institutions construct accountability around AI through mandatory risk disclosures. Using a three-year panel of 10-K filings from 73 publicly listed S&P financial firms, we combine keyword-based extraction with supervised machine learning to classify disclosure language along a continuum of AI accountability framing. This continuum ranges from locating accountability within the firm's own governance structures, to situating it within external institutional rules and to dispersing it through technological unpredictability and partial autonomy. Our study advances research on responsible AI governance by showing that AI accountability is not only engineered through formal controls but also constructed discursively, as firms use accountability framings to locate and shape the boundaries of control under conditions of algorithmic opacity. Complementing accounts of AI unpredictability as a sociotechnical condition, our findings show that firms pair claims of control with disclosures of “known unknowns”, presenting AI risks as governance-relevant yet not fully governable. In this manner, firms actively construct what accountability can mean when full explainability is unattainable. We also document how the prevalence of these framings shifts over time, consistent with intensifying regulatory scrutiny and the rise of generative AI.

Keywords AI governance · AI accountability · Responsible AI · Algorithmic opacity · Accountability framings

1 Introduction

The integration of artificial intelligence into financial and operational decision processes represents a major organizational and technological development that is reshaping the financial industry. AI has evolved from an emerging innovation to a core component of organizational infrastructure that shapes how financial institutions analyze markets, manage risk, allocate capital and report performance (Brynjolfsson & McAfee, 2017). Firms increasingly rely on algorithmic models to interpret data and support decisions that directly affect investors, regulators and markets (Gawer, 2021;

Davenport et al., 2020). As algorithms are given a larger role in judgment and prediction, responsibility becomes distributed between human managers and machine agents, blurring the boundaries of control (Agarfalk, 2020; Ananny & Crawford, 2018). This diffusion of agency raises new challenges for demonstrating accountability and transparency around AI use (Agarfalk, 2020).

Traditional accountability frameworks in finance are built on the assumption of traceable human judgement, where actions can be audited, explained and attributed to identifiable decision-makers (Bovens, 2007; Sinclair, 1995; Fama & Jensen, 1983). These frameworks assume a visible chain of logic that links actions to accountable individuals. Nevertheless, the opacity and adaptive complexity of modern AI systems disrupt this logic (Waardenburg & Huysman, 2022; Pachidi et al., 2021; Ananny & Crawford, 2018; Burrell, 2016). Even firm experts often struggle to reconstruct why a particular model produced a given outcome. In highly regulated settings such as financial reporting and risk disclosure, firms face strong institutional expectations to

✉ Eleni Lioliou
e.lioliou@qmul.ac.uk

¹ School of Business and Management, Queen Mary University of London, Mile End, London E14 NS, UK

² Kent Business School, University of Kent, Canterbury, UK

³ Grenoble Ecole de Management, Grenoble 38000, France

provide transparent and accountable accounts of their activities. However, prior research shows that such disclosures are often selective, symbolic and decoupled from underlying practices, reflecting broader tensions between reporting and reality (Cho et al., 2015). These tensions are further intensified in the context of algorithmic systems, where the technical limits of explainability challenge the very possibility of meaningful transparency (Ananny & Crawford, 2018). Under these conditions, firms face a core governance dilemma. Presenting AI as fully controllable risks overstating their capacity to govern inherently opaque systems, while delegating control to institutional actors or technological elements raises questions about diminished managerial oversight. Firms therefore need to navigate competing pressures to demonstrate control while acknowledging limits of what can be explained, attributed or governed.

While AI applications in finance have been extensively studied, existing research has largely concentrated on the technical performance of AI, algorithmic accuracy and data governance (i.e. Shaban et al., 2024; Vijn et al., 2020; Nikou et al., 2019). The fundamental premise of this stream of research is that the behaviour of AI systems can be specified, audited and controlled through improved design, data quality and model transparency. However, the opacity and emergent behaviours of modern AI systems undermine these assumptions (Pachidi et al., 2021; Faraj et al., 2018), blurring the boundaries of agency and complicating established regimes of control (Agarfalk, 2020; Ananny & Crawford, 2018). By treating AI as a purely technical artefact, prior work has overlooked a deeper governance challenge: how accountability can be sustained when its foundational conditions (i.e. transparency, human agency and capacity for oversight) are fundamentally challenged.

Recent scholarship has sought to bridge this gap by developing frameworks of responsible AI governance, emphasizing the interplay between technical, ethical and procedural mechanisms (Spanaki et al., 2022; Papagiannidis et al., 2025; Mikalef et al., 2022; Dennehy et al., 2023; Birkstedt et al., 2023; Niederman & Baker, 2023; Enholm et al., 2022). This work examines how organizations seek to incorporate principles of responsible AI through systems of auditability, ethical oversight and lifecycle documentation that institutionalize responsibility. Nevertheless, accountability is not only engineered through governance mechanisms but also constructed through discourse (Ananny & Crawford, 2018). Building on this perspective, we draw on framing theory to examine how firms construct and locate accountability for AI when full explainability and control cannot be guaranteed. In particular, the primary research question guiding this study is: *How do financial firms discursively construct and locate accountability for AI in their mandatory risk disclosures under conditions of algorithmic opacity?*

In our study, we analyse SEC filings from S&P 500 financial institutions to examine how firms discursively construct accountability for their AI systems. The financial industry offers an ideal setting for studying how organizations frame accountability and control around AI. Financial institutions operate under intense regulatory scrutiny, where mandated disclosures serve as key instruments for generating public trust (Currie et al., 2018; Gozman & Currie, 2014). Because their decisions directly affect markets, investors and systemic stability, these firms must continually demonstrate control, transparency and compliance. This makes financial risk disclosures a rich empirical setting for examining how accountability is constructed and communicated under conditions of algorithmic opacity.

Using supervised machine learning to classify disclosure language, we identify a continuum of accountability framings that position the locus of control inside the organisation, within external regulatory institutions or in technological unpredictability and partial autonomy. Rather than treating these disclosures as purely strategic tools of impression management (Cho et al., 2015; Leung et al., 2015), we conceptualise them as governance instruments through which firms locate and shape the boundaries of accountability for AI. We show that firms pair claims of control with disclosures of “known unknowns”, presenting AI risks as governance-relevant but not fully governable, and that under algorithmic opacity, disclosure itself becomes a governance practice. We also document how these framings shift over time as regulatory scrutiny intensifies and generative AI emerges.

2 Literature Review

2.1 From Traceable Judgement to Algorithmic Opacity: The Evolution of Accountability Frameworks in Finance

Accountability has long been understood as a cornerstone of organizational governance, underpinned by the assumption that actions can be traced to identifiable human agents capable of reasoning and taking responsibility for their decisions. In domains such as risk management, financial reporting and auditing this assumption has provided the foundation for mechanisms of oversight and control. Traditional approaches to accountability have been based on the premise that with sufficient transparency, responsible actors can be located and their decision rationales can be identified (Ananny & Crawford, 2018). This accountability logic presumes not only that control can be implemented, but also that it can be publicly evidenced in ways that allow external audiences to attribute responsibility.

The emergence of artificial intelligence has profoundly disrupted this logic. Financial decision-making and related activities (such as risk assessment, credit scoring, portfolio management and fraud detection) are now increasingly delegated to AI systems that are self-learning, adaptive and opaque in their reasoning. Unlike traditional rule-based automation, contemporary machine-learning models evolve through statistical inference and continuous feedback loops. Their behaviour shifts as new data are introduced, making *ex post* reconstruction of decisions uncertain and accountability difficult to locate (Ananny & Crawford, 2018). Under these conditions, questions of “who is accountable” and “where control resides” become governance problems concerning how accountability is articulated and positioned, rather than merely technical design challenges.

AI systems combine autonomy, learning and opacity in ways that make them extraordinarily powerful yet inherently difficult to predict (Waardenburg & Huysman, 2022). These capabilities transform industries by enabling firms to detect patterns and automate complex decisions at unprecedented speed and scale (Brynjolfsson & McAfee, 2017). However, the same traits render AI a “black box” that is challenging to govern. This is because its decisions stem from complex statistical associations rather than linear reasoning and even developers often struggle to explain specific outcomes (Brasse et al., 2023; Waardenburg & Huysman, 2022; Burrell, 2016).

In response to these challenges, research on Explainable AI (XAI) has sought to restore interpretability and trust in algorithmic decision-making. A growing body of work has proposed methods such as local interpretable model-agnostic explanations (Ribeiro et al., 2016), SHAP values (Lundberg & Lee, 2017) and influence functions (Koh & Liang, 2017) to make the inner workings of complex models more intelligible (Olan et al. 2025a, b; Bauer et al., 2023; Lakkaraju et al., 2019). This research has generated powerful insights into how algorithmic reasoning can be rendered transparent and how interpretability, trust and fairness can be enhanced in AI systems (Spanaki et al., 2025; Brasse et al., 2023). Nevertheless, these approaches remain largely situated within a computer-centred paradigm, focusing on technical interpretability and treating improvements in model transparency as the primary mechanism for addressing broader accountability concerns (Arrieta et al., 2020).

While transparency at the level of code or model logic is often treated as a key requirement for accountability, it does not guarantee that responsibility can be meaningfully attributed. Making algorithms more transparent does not ensure that their outputs can be situated within organisational processes or traced to identifiable human actors. As a result, technical transparency remains partial and often insufficient for governance purposes (Ågerfalk et al., 2022; Mikalef et

al., 2022; Ågerfalk, 2020). As decision-making shifts from human judgement to algorithmic inference, responsibility becomes increasingly diffuse (Novelli et al., 2024), distributed across socio-technical arrangements in which design, deployment and oversight jointly shape outcomes. This diffusion of agency intensifies the challenge of locating accountability across organisational, institutional and technological elements of the AI ecosystem.

The central challenge therefore is not only to make AI technically explainable (Brasse et al., 2023) but to reconsider how accountability is located, institutionalized and sustained within these evolving systems of algorithmic governance. A significant stream of research has recently turned to investigations of responsible AI governance (Papagiannidis et al., 2025; Spanaki et al., 2025; Dennehy et al., 2023; Pappas, et al., 2023; Spanaki et al., 2022; Enholm et al., 2022), aiming to engineer accountability through ethical principles, oversight mechanisms and procedural safeguards. However, comparatively less is known about how firms make accountability arrangements legible in formal communications, or how such communications position the locus of control when opacity constrains what can be demonstrated.

2.2 Responsible AI Governance and the Engineering of Accountability

The maintenance of accountability has become a major concern for organisations as they deploy complex AI systems. As a result, recent research has turned towards responsible AI governance as a means to re-engineer accountability under conditions of algorithmic uncertainty. This work examines how accountability can be embedded through interacting technical, ethical and procedural mechanisms (Enholm et al., 2022; Floridi et al., 2018). Rather than attempting to eliminate opacity, this stream of research concentrates on mechanisms that render AI decision making traceable, auditable and justifiable despite the limits of full comprehension.

Scholars emphasise the deliberate design of mechanisms that operationalise responsibility throughout the AI life-cycle. Enholm et al. (2022) highlight the institutional and moral dimensions of AI adoption, calling for governance practices that align technological development with societal values. Floridi et al. (2018) propose ethical oversight structures that combine human judgement with algorithmic precision. Dennehy et al. (2023) add that accountability depends not only on compliance mechanisms but also on leadership commitment and organisational culture. Li and Goel (2025) extend these ideas through standardised documentation and traceability tools (such as Model Cards and Datasheets) that make algorithmic systems auditable.

This literature advances a view of accountability that is engineered and procedural. In other words, control is maintained through documentation, verification and codified oversight routines that translate ethical principles into auditable artefacts. This “control logic” seeks to make accountability measurable and demonstrable. While this approach has substantial merit, it also rests on the assumption that transparency and compliance alone can secure accountability and it tends to understate the extent to which responsibility must be communicated, interpreted and legitimised by external audiences.

In particular, accountability remains fragile as AI systems continue to yield biased, unpredictable, or opaque outcomes (Gupta et al., 2022). As adaptive models evolve through feedback loops, their behaviour may drift beyond original design intentions. Responsibility becomes distributed across socio-technical assemblages where humans and algorithms jointly shape outcomes (Mikalef et al., 2025; Novelli et al., 2024; Pappas et al., 2023). This diffusion of agency blurs lines of control. In other words, when errors occur, no single actor can easily be held responsible. In financial institutions, this challenge is particularly salient because fiduciary duty requires clear responsibility for decisions (Currie et al. 2018), while algorithmic processes often obscure how those decisions are produced. Consequently, how firms discursively locate where control and accountability reside, shapes how regulators and investors interpret responsibility under conditions of opacity.

These limitations reveal a deeper tension within responsible AI governance. Existing frameworks tend to approach accountability as a design problem, to be addressed through technical and procedural controls. However, when opacity and distributed agency persist, such approaches only partially capture how accountability is established in practice. Organizations must instead articulate control and responsibility in ways that reassure regulators, investors and the public, even when the underlying systems remain partly inscrutable. We therefore argue that accountability under algorithmic opacity must be studied not only as a set of technical controls but also in terms of how responsibility is located and represented in organisational disclosures.

2.3 Algorithmic Decision-Making and the Construction of Accountability

Discourse plays a central role in how organisations construct and sustain legitimacy. Communication is not merely a vehicle for reporting facts but a means through which actors define what is appropriate, credible, and responsible. Prior studies show that legitimacy is actively produced through language and interaction rather than simply resulting from compliance or structural alignment (Christensen et al., 2021; Vaara et al., 2024; Suddaby et al., 2017).

Extant research on corporate communication has largely examined these dynamics through the lens of impression management, explaining how firms strategically use language to shape stakeholder perceptions (e.g., Brennan & Merkl-Davies, 2013; Cho et al., 2015; Leung et al., 2015). For instance, companies may obscure unfavourable performance through complex or vague disclosures, and attribution studies show that managers tend to assign negative outcomes to external factors while claiming credit for positive results (Barton & Mercer, 2005). While this perspective highlights the strategic use of language, it primarily treats disclosure as a tool of persuasion aimed at managing how an underlying organisational reality is perceived.

Our study departs from this perspective in two fundamental ways. First, impression management scholarship assumes a knowable organisational reality that firms present more or less favourably, a stable referent against which disclosure can be evaluated. In mandatory AI risk disclosures, no such stable referent exists. Firms construct accountability around systems whose behaviour they cannot fully trace or predict, meaning there is no underlying governance reality that disclosure merely represents. Second, and consequently, language in this context does not merely signal trustworthiness to external audiences but positions and constitutes where accountability resides when full explainability is unattainable. Mandatory risk disclosures therefore function not as tools of persuasion but as governance artefacts through which accountability is actively located and shaped. In opaque AI contexts, disclosure does not merely influence perceptions of accountability but structures its boundaries.

Research on discursive legitimation demonstrates that organisations frame their actions in socially acceptable terms, using narrative, rhetoric and metaphor to render their activities intelligible and justifiable (Vaara et al., 2024). Through such rhetorical framing, firms express oversight and responsibility even when technical understanding is partial. By translating complexity into narratives of governance, discourse itself becomes a mode of control and accountability.

This dynamic is particularly visible in financial institutions, where accountability is heavily institutionalised (Palermo et al., 2017). Governance mechanisms such as risk frameworks, ethics committees and audit procedures render organisations auditable and structure them for observation and inspection (Power, 2021). Such mechanisms help sustain trust in systems that are too complex to be fully understood, operating not only through their actual effectiveness but also through the reassurance they provide (Power, 1997; Power, 2007; 2021). Communication therefore operates as part of the control apparatus, not supplementary to governance but constitutive of it. This constitutive function is especially consequential in AI contexts, where the object being governed is

itself opaque and no stable referent of accountability exists against which disclosure can be evaluated.

These communicative strategies reveal the growing interdependence between accountability and discourse in algorithmic governance. When AI is framed as a controllable analytical tool, firms reinforce managerial authority. When it is described as embedded within institutional procedures or compliance systems, responsibility is transferred to the organisational level. When it is depicted as autonomous or self-learning, firms implicitly acknowledge new boundaries of control. These discursive moves do not simply describe technical realities but actively shape how oversight, competence and responsibility are located within the wider socio-technical system.

In practice, financial institutions are expected to demonstrate that their AI systems are responsibly governed, even though full intelligibility and accountability remain difficult to achieve in complex systems (Floridi and Cowls, 2022). Accountability, traditionally grounded in human reasoning, is now mediated through adaptive models that evolve beyond direct supervision. Accordingly, disclosures should not be treated only as rhetoric but as institutionalised governance artefacts that shape how external audiences interpret where responsibility and oversight reside. This makes mandatory risk disclosures a particularly consequential site for examining how accountability is constituted and located under algorithmic opacity.

2.4 Framing Theory

Framing theory, originating from Goffman's work on the organization of experience, provides a robust analytical lens for understanding how information is presented and interpreted within organizational contexts, including the domain of financial reporting (Goffman, 1974). This theoretical perspective examines how particular aspects of reality are emphasized to shape audience perceptions and influence subsequent decision-making (Rosenkranz & Pollach, 2016). In financial reporting, framing manifests through rhetorical and narrative strategies that guide stakeholders' interpretations of corporate performance, prospects and governance (Brennan & Merkl-Davies, 2013). Such strategies often involve selectively highlighting favorable information while downplaying or omitting less desirable elements, thereby steering readers toward a preferred understanding (Rosenkranz & Pollach, 2016; Cho et al., 2015).

This process is not merely about transmitting information but about constructing a shared perception of organizational reality, frequently shaped by the desire to manage impressions and maintain legitimacy (Brennan & Merkl-Davies, 2013; Suchman, 1995). Additionally, recent work highlights that framing in corporate narratives is often dynamic

and adaptive, evolving in response to shifting stakeholder expectations and institutional pressures (Arvidsson & Sabelfeld, 2023). However, much of this research has traditionally examined framing in static terms, focusing primarily on impression management rather than on the more fluid processes of ongoing narrative adaptation and change (Arvidsson & Sabelfeld, 2023; Cho et al., 2015).

Building on this literature, we investigate how financial firms frame AI accountability in their risk disclosures. In the empirical sections that follow we examine how 73 S&P financial firms articulate accountability in their risk disclosures. In particular we identify three recurrent framings that locate accountability differently: AI as internally governed within the organisation, AI as institutionally governed through external regulatory forces and AI as unpredictable and partially autonomous, with responsibility shifted toward technological uncertainty and external actors.

3 Methodology

3.1 Data Collection and Corpus Construction

The empirical analysis draws on the 10-K filings submitted to the U.S. Securities and Exchange Commission (SEC) of financial companies listed on the S&P 500 index. In total the filings of 73 firms were analyzed across three consecutive fiscal years (2022–2024). We chose these firms because they represent the leading organizations of the U.S. financial industry and are subject to the market's highest disclosure and governance standards.

We selected 10-K filings as the core corpus of data for several reasons. First, as legally mandated reports submitted to the U.S. SEC, they follow standardized reporting requirements. This format ensures comparability across firms. Second, unlike voluntary ESG or sustainability disclosures, 10-K filings are formal, board-approved and legally binding documents. This makes them particularly suitable for studying how firms frame abstract or emergent risks (such as those associated with artificial intelligence AI) in official narratives that must withstand regulatory and investor scrutiny. Third, these filings provide a longitudinal record of firms' strategic and risk-related priorities, including explicit statements about technological uncertainty and future contingencies. Hence, they offer a unique window into how AI-related risks are framed in corporate discourse.

An initial set of keywords was derived from prior literature on AI governance and accountability (Papagiannidis et al., 2025; Birkstedt et al., 2023; Novelli et al., 2024). We then refined these keywords through exploratory reading of the filings to capture terminology used in practice. The resulting list was iteratively adjusted following manual

review of randomly sampled extracted passages to ensure coverage of terminology used in practice and to remove terms that generated irrelevant extractions. Passages were identified using a dual-filter keyword approach requiring both (1) AI-related terms (approximately 40 keywords; e.g., “artificial intelligence,” “machine learning,” “algorithms,” “generative AI”) and (2) governance-related terms (approximately 90 keywords) covering internal oversight (e.g., “governance,” “control,” “monitoring”), institutional regulation (e.g., “regulation,” “compliance”), and AI unpredictability (e.g., “unpredictable,” “black box,” “explainability”). This approach ensured that the extracted passages discussed AI within governance contexts rather than in neutral technical descriptions. Our full list of keywords is included in the online supplementary material.

To preserve meaning, we employed adaptive context windows, capturing 0–2 surrounding sentences for each AI and governance keyword co-occurrence, depending on sentence length. Sentences of up to 80 characters were expanded with two sentences before and after; sentences between 81 and 200 characters were assigned one sentence before and up to two after; sentences between 201 and 400 characters received one sentence on each side; and sentences exceeding 400 characters were assigned minimal context, with no preceding and at most one following sentence, as these tend to contain substantial standalone disclosure content. This method prioritised contextualised AI discourse over isolated references and enabled us to capture how organisations frame AI within governance narratives, rather than purely operational or capability-focused discussions.

The initial extraction produced 1490 passages that referenced AI in governance contexts. These passages were first reviewed to identify a focused subset for developing and refining the coding scheme. During this stage, we lightly cleaned 216 passages to remove redundant text and formatting issues. This subset formed the foundation for developing and validating the coding categories and it served as the basis that informed our subsequent content analysis.

To ensure that our analyses relied on a consolidated corpus, the full set of 1490 passages was then processed using a Python based deduplication procedure. Passages exceeding 70% similarity within the same filing were collapsed, resulting in a refined dataset of 1203 unique passages. This approach enabled us to remove mechanical repetition common in regulatory filings while preserving passages that contained overlapping but substantively distinct risk narratives.

3.2 Operationalization of the Coding Scheme

Our coding scheme consisted of three accountability framings, ranging from (1) AI framed as internally governed, (2) AI framed as institutionally governed, to (3) AI framed as

unpredictable and partially autonomous. A fourth residual category captured descriptive references to AI that did not foreground governance or accountability.

Internally governed AI passages portrayed AI as a system that remained, or ought to remain, under organisational supervision. Firms described internal structures and processes (such as oversight mechanisms, monitoring routines, and control frameworks) that contained AI-related risks and integrated the technology within existing governance arrangements. Even when external actors misused AI, for example through fraud or cyberattacks, firms typically redirected attention to their own internal countermeasures and resilience measures.

Institutionally governed AI passages framed AI as situated within a broader landscape of formal rules, standards, and oversight bodies. Firms positioned their AI practices in relation to regulatory and compliance structures, highlighting conformity with legal obligations, ethical guidelines, and evolving legislative initiatives such as the EU AI Act. The dominant tone was one of alignment and preparedness for emerging regulatory scrutiny.

Unpredictable or partially autonomous AI passages depicted AI as a technology whose functioning or deployment could exceed human or organisational control. This framing encompassed both internal unpredictability, where AI systems produced opaque or unintended outputs, as well as external misuse where the actions of other actors using AI such as hackers or fraudsters amplified perceptions of limited control.

Descriptive AI passages referred to AI in factual, operational, or competitive terms without explicit links to governance or risk. In these instances, AI was portrayed as part of firms’ standard technological infrastructure, such as process improvement, automation, or innovation initiatives, and as a domain of competition rather than a governance concern. We coded these passages as a separate residual class to distinguish general AI discourse from accountability-oriented framings, though we note that some passages in this category may reflect extraction artefacts rather than substantive governance references. It is therefore retained for quantitative completeness but not qualitatively analysed.

The three accountability framings (internally governed, institutionally governed and unpredictable/partially autonomous AI) can be understood as points along a continuum that reflects shifting loci of control in AI governance. This continuum is informed by governance theory which examines how accountability is distributed across organisational processes and regulatory structures (Bovens, 2007) and by IS scholarship that highlights how technological opacity and unpredictability complicate control (Pappas et al., 2023; Brasse et al., 2023; Faraj et al., 2018). By operationalising these distinctions, our coding scheme captures how firms

narrate responsibility as AI systems become more complex and less easily understood or explained.

To translate the conceptual distinctions into reliable labels, two of the authors jointly reviewed an initial calibration set of 50 passages and compared their classifications in detail. This iterative discussion was used to refine category boundaries and to establish clear decision rules for handling ambiguous cases. For example, when AI failures were attributed to breakdowns in internal oversight, audit mechanisms, or control processes, the passage was assigned to Category 1 (Internally governed AI). In contrast, when failures were linked to the opacity or inherent unpredictability of AI systems, the passage was classified as Category 3 (Unpredictable/Partially autonomous AI). A recurring boundary also arose between internal and institutional governance framings: references to risk management and firm level control mechanisms were coded as Category 1 (Internally governed AI), whereas references to compliance obligations, supervisory authorities, or regulatory enforcement were coded as Category 2 (Institutionally governed AI). Illustrative examples of labelled passages are provided in Appendix 1.

After reaching an agreement on the decision rules, one researcher applied the refined coding scheme to the remaining passages that formed the working subset used for method development. This approach ensured that the labels were applied consistently across passages.

3.3 Model Training Procedure and Evaluation

We adopted SetFit (Tunstall et al., 2022), a few-shot classification framework that fine-tunes a pre-trained sentence transformer through contrastive pair training before attaching a lightweight classification head. The model was initialised from the sentence-transformer backbone paraphrase-mpnet-base-v2, a transformer encoder based on the MPNet architecture that produces contextual sentence embeddings. This architecture is particularly suitable for textual data characterized by nuanced meaning and moderate class imbalance (for example, descriptive references to AI alongside risk and governance language).

We validated the model through 15-fold repeated stratified cross-validation (5 folds $\times$ 3 repeats), ensuring robust evaluation across different data splits while maintaining class balance. The final model was trained on all 216 labelled examples and was then applied to classify the full corpus of 1203 passages.

To mitigate overfitting on our relatively small labelled dataset we trained the SetFit model using 1 epoch with 32 sentence pairs per class. The classifier achieved a mean cross validated accuracy of 77.9% (s.d. 5.6%). Compared to a previous configuration with 2 epochs and 64 pairs per class, this setup reduced the overfitting gap from roughly 22% to 13%,

with only a minor decrease in mean cross-validation accuracy (79.5% to 77.9%). Fold-level scores ranged from 67% to 91%, with most folds clustering around 77% to 79%, indicating that the model captured the dominant patterns in the data.

Our model achieved robust F1-scores in three of the four categories, with institutional governance framings achieving the highest performance (F1 = 0.89), followed by descriptive framings (F1 = 0.81) and unpredictable/partially autonomous framings (F1 = 0.77). Internal governance framings showed the lowest performance (F1 = 0.57), which is consistent with their more varied and less standardised linguistic expressions compared with the clearer lexical patterns associated with institutional governance and descriptive framings. To assess model reliability, we compared automated classifications against a random sample of 50 manually coded paragraphs. The out-of-sample accuracy reached 76% confirming the validity of SetFit's generalization beyond the labelled subset. The full model performance tables (including per-fold accuracies, the classification report and the confusion matrix) are presented in Appendix 2.

To validate the selection of SetFit as the classification model, we conducted a systematic comparison against three baseline classifiers under identical evaluation conditions: Support Vector Machine with TF-IDF features (SVM-RBF), Logistic Regression with TF-IDF features and RoBERTa-base fine-tuned with class-weighted loss. All models were evaluated using the same 15-fold repeated stratified cross-validation protocol (5 folds $\times$ 3 repeats, seed = 42) on the 216-example labelled dataset.

SetFit achieved the highest overall accuracy (77.9%, s.d. = 5.6%) and macro-F1 (0.761), outperforming RoBERTa (76.7%, s.d. = 4.7%, macro-F1 = 0.721), Logistic Regression (70.4%, s.d. = 5.0%, macro-F1 = 0.620) and SVM (68.2%, s.d. = 5.7%, macro-F1 = 0.587). Critically, SetFit demonstrated superior performance on the Internally Governed AI category (F1 = 0.573), which is the most theoretically significant and class-sparse frame in our scheme, compared to RoBERTa (F1 = 0.460), Logistic Regression (F1 = 0.189) and SVM (F1 = 0.087). Given the importance of reliably capturing minority classes, these results support the use of SetFit as a robust modelling choice for this task. The full results are presented in Appendix 3.

The structured dataset enabled a multi-layered analysis that combined quantitative pattern detection with qualitative interpretation. Quantitatively, we examined how the distribution of the three accountability framings and the descriptive AI references varied across firms and years. Qualitatively, we analysed representative passages within each framing to identify how firms discursively locate and construct accountability under conditions of algorithmic opacity. This interpretive stage focused on how language constitutes governance claims (i.e. how firms position control, acknowledge limits

and narrow the scope of responsibility when AI systems cannot be fully explained). Integrating both layers allowed us to trace how accountability framings evolve as regulatory scrutiny intensifies and AI systems become more complex.

4 Findings

4.1 Quantitative Results: Distribution and Dynamics of AI Framings

According to our results AI governance disclosures have expanded rapidly in the SEC filings of S&P financial companies over the past three years. As shown in Table 1, in 2022, 44 of 73 firms reported AI governance related issues (60.3%). This increased to 61 firms in 2023 (83.6%). By 2024, 69 firms included such disclosures (94.5%). The trend shows a clear move toward more explicit references to AI governance related issues. Overall, our results show that S&P financial companies are embedding AI governance more deeply into their risk and compliance narratives. This growing emphasis reflects the rising importance of AI accountability, especially in a sector where errors, bias, or opacity can have direct consequences for markets and consumers.

Table 2 reports the year-by-year distribution of accountability framings. Figure 1 provides a visual representation of these patterns, showing how the prevalence of each framing shifts over time. According to our findings, descriptive references to AI dominated the discourse in 2022, reflecting early stage mentions of AI as part of firms broader technological and competitive infrastructure rather than as a governance concern. Firms also relied heavily on framings that portrayed AI as internally managed, emphasising oversight, monitoring, and organisational control. References to

external institutions and regulatory constraints were present but less prominent, and only a minority of firms described AI as unpredictable or partially autonomous

After the emergence of generative AI, the pattern shifted sharply. “Descriptive AI” as well as “internally governed” framings declined proportionately compared to 2022. Instead, firms increasingly anchored accountability in external institutions and regulatory structures, while also highlighting AI as an unpredictable or partially autonomous force. This shift indicates that organisations at that point in time began to construct accountability less through claims of internal mastery and more through alignment with institutional safeguards and acknowledgements of limits to govern or control.

Our findings therefore reveal a clear transition in how firms construct accountability for AI. After 2023, firms increasingly framed AI within external regulatory regimes and highlighted its unpredictability, signalling a widening recognition that accountability depends both on institutional oversight and on acknowledging the limits of managerial control.

4.2 Robustness Checks

To assess the robustness of our findings to the choice of extraction window, we replicated the full analysis using two alternative context windows: a narrower window (0–1 sentences) and a wider window (0–3 sentences). These produced corpora of 1347 and 1063 passages respectively. The same trained SetFit classifier was applied to all three corpora so that window size was the only source of variation.

As shown in Appendix 4, the overall distribution of accountability framings and the temporal pattern across FY2022–2024 remain substantively stable across all three specifications. In particular, the shift after the emergence of generative AI from Descriptive AI framing toward Institutional and Unpredictable/Partially autonomous framings appears consistently across all window sizes. Minor variation in the Descriptive AI category in FY2022 reflects the fact that narrower windows capture less surrounding descriptive context in shorter early-period passages.

4.3 Qualitative Analysis: Patterns of AI Accountability Framing

4.3.1 Within the Firm’s Oversight: Framing AI as Internally Governed

This framing positions AI as something the organisation can and should control. Firms describe governance mechanisms such as internal procedures, expertise, monitoring systems and defensive actions as the basis for managing AI responsibly. Disclosures in this framing rely on familiar references to oversight, monitoring and control processes and the tone

Table 1 Firm-Level AI Governance Related Discourse Across Fiscal Years

Fiscal Year	Companies Reporting	Companies Not Reporting	Total Companies	% Companies Reporting
2022	44	29	73	60.3
2023	61	12	73	83.6
2024	69	4	73	94.5

Table 2 Accountability Framings by Year

Year	Total Passages	Internally Governed	Institutionally Governed	Unpredictable/Partially autonomous	Descriptive AI
2022	119	26.10%	21.00%	10.90%	42.00%
2023	397	20.40%	34.80%	24.20%	20.70%
2024	685	21.00%	32.80%	26.90%	19.30%

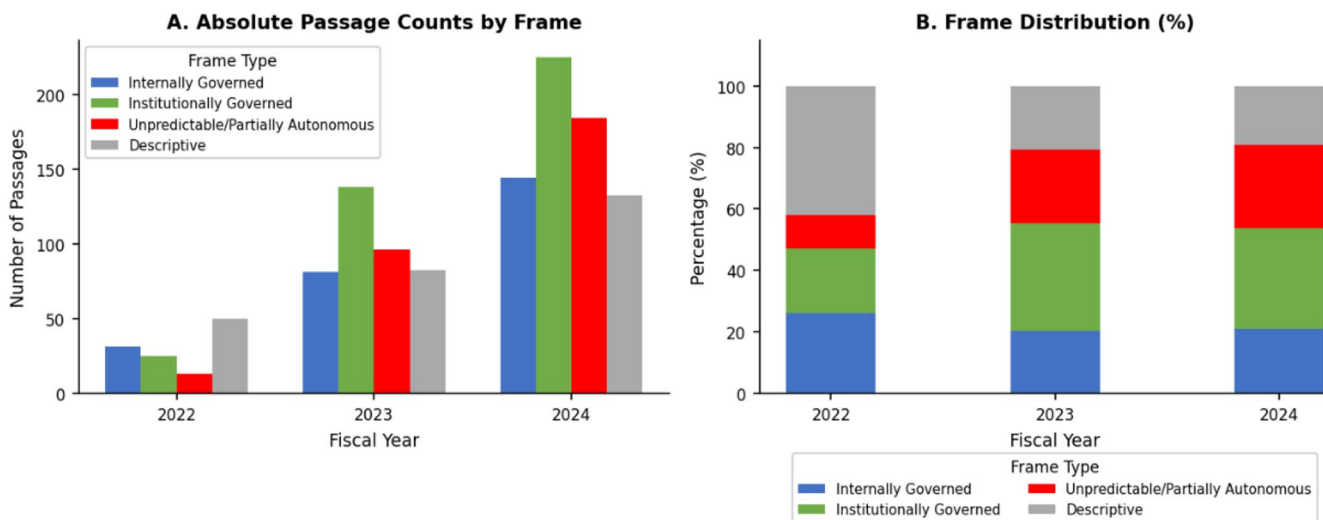


Fig. 1 Distribution of Accountability Framings in AI Governance Related Discourse by Year

is measured and procedural. At the same time, these narratives are often lightly inflected with references to unpredictability or model limitations, which signal that full control may be difficult to achieve without undermining the overall impression of competence. In this way, firms use established governance language to project stability while briefly acknowledging uncertainty.

A recurring pattern is the integration of AI into existing model risk narratives. Rather than treating AI as a distinct category of risk, disclosures embed it in model governance language that emphasises policies, validation, monitoring and revalidation. AI appears as an “emerging risk” layered onto familiar weaknesses such as simplifying assumptions or unstable correlations. For example, one major bank explains that model oversight involves independent validation, ongoing monitoring and periodic revalidation, but notes that “emerging risks from applications that rely on AI” add to inherent model limitations before returning to generic concerns about historical data, market shocks and correlation assumptions. This pattern suggests that AI opacity is absorbed into pre-existing model risk frameworks.

A second pattern appears in cybersecurity disclosures, where AI is positioned within an external threat landscape, with firms warning about “phishing and other forms of social engineering,” “increased cybersecurity risk... related to new and emerging technologies such as AI” and attacks from “foreign governments, organized crime, or terrorist organizations”. This positions AI enabled risk as external and threat-driven. The same passages then pivot to internal responses, stressing that “we seek to continuously monitor for and nimbly react to... malicious cyber activity” and that “we develop our systems to protect our technology infrastructure and data”. Firms thus link AI

related unpredictability to external actors while presenting themselves as vigilant and prepared.

A further pattern concerns the formal institutionalisation of AI governance. Several firms describe AI specific committees, risk forums or control functions that sit within existing risk hierarchies and report into board or group level structures. These bodies are tasked with setting AI policies, overseeing AI model development and monitoring AI related risk, often in coordination with model risk management and technology risk functions. Some firms also reference alignment with external frameworks or standards, reinforcing the impression that AI is being brought under structured, institutional control.

Finally, we observe that AI accountability is sometimes displaced into the extended supply chain. Disclosures highlight exposure arising from “clients, customers and third parties”, as well as from “new and emerging technologies such as AI” deployed in outsourced services or by external actors across the supply chain. In these passages, firms stress their dependence on vendors, cloud providers and other intermediaries and present themselves as vulnerable to the AI practices of those entities. This framing suggests a form of accountability decoupling, in which risk associated with third party AI is narrated as less controllable and more uncertain than the AI and models managed within the firms own governance structures.

4.3.2 Within Institutional Oversight: Framing AI through Regulatory Alignment

This framing situates AI within the domain of external governance structures. Firms describe their compliance with emerging AI regulations, ethical standards and supervisory expectations. This discourse emphasises adherence and

emphasizes that AI use is legitimate because it conforms to institutional frameworks.

Firms repeatedly stress that AI related regulation is rapidly evolving, global and inconsistent across jurisdictions. They talk about “uncertainty regarding the scope of new laws and how existing laws will apply” and say they may not be able to anticipate how to respond to these changing frameworks, especially when state, federal, EU and UK rules diverge. Compliance is framed as costly, complex and resource intensive rather than fully under their control.

Another pattern is the codification of AI governance through specific frameworks and model bulletins. Firms extensively refer to the NAIC AI Model Bulletin, NYDFS guidance, the EU AI Act, NIST AI and other regulations. These documents are described as setting expectations for AI governance frameworks, risk management controls, outcomes testing, bias and unfair discrimination checks and documentation for regulatory exams. In this framing, accountability is increasingly located in formal institutional standards that firms must align with.

An important pattern has also been that regulation of AI is rarely presented in isolation. It is discussed in conjunction to broader regimes on privacy, data protection, cybersecurity, consumer protection and anti-discrimination. Firms highlight that AI rules intersect with GDPR, CCPA, state privacy laws, data localisation, equal opportunity law, algorithmic fairness rules and sector specific insurance rules on unfair discrimination. The overall implication is that using AI triggers a web of overlapping obligations around data governance, transparency, explainability and fairness, which expands both compliance burdens and exposure to enforcement.

4.3.3 Beyond the Firm's Governance Capabilities: Framing AI as Unpredictable and Partially Autonomous

In this framing, firms describe AI as operating beyond their immediate governance capabilities. They stress that AI technology is “highly complex and rapidly developing” and acknowledge that it is “not possible to predict all of the legal, operational or technological risks that may arise” from its use. Models are said to generate outputs that are inaccurate, biased or otherwise harmful in ways that are “not easily detectable” These disclosures convey caution, recognising that even robust internal systems cannot entirely anticipate or handle AI driven risks.

In these passages, AI is described an agent whose behaviour may drift over time, or produce unforeseen legal, operational and reputational risks. Internal governance mechanisms are sometimes briefly mentioned, but they are not the centre of the narrative. What dominates is the concern that even with policies and monitoring in place, firms cannot fully anticipate how AI models will behave or explain

why they produce particular outcomes. Here AI appears less as a fully controllable tool and more as a system whose opacity and evolution outpace established control routines.

Another pattern in this framing links unpredictability to AI-enabled cyber threats. Firms describe an “evolving threat landscape” in which attackers use AI and machine learning to increase the speed, volume and sophistication of phishing, social engineering and other intrusions. Cyber risks are attributed to hostile external actors “including threat actors affiliated with foreign governments, organized crime, or terrorist organizations,” with AI portrayed as a tool that makes their tactics harder to detect and defend against. Disclosures acknowledge defensive measures, such as “defence in depth” strategies and continuous monitoring, but these are framed as attempts to keep up with adversaries rather than as guarantees of control. The emphasis rests on the possibility that AI driven attacks may outstrip existing protections, underscoring AI as an external, adversarial force that introduces uncertainty beyond the firms direct managerial reach.

In a subset of this framing, firms go beyond simply describing AI as a complex technology and instead subtly construct it as an actor in its own right. Outcomes are repeatedly labelled as “unintended”, separating managerial intent from algorithmic results and suggesting that harm arises from the system rather than from deliberate organisational choices. In some disclosures AI systems are also presented as “actors with agency” as when models “take action that is incorrect”, “create forged documents” or “impersonate individuals”, implying an active rather than purely instrumental role. This is reinforced by admissions that such outputs “may not be easily detectable or explainable”, illustrating a picture in which AI behaves in ways that are both autonomous and opaque. These rhetorical moves allow firms to frame AI as a partially independent source of risk, complicating straightforward attributions of accountability.

5 Discussion

Our study demonstrates that AI accountability in financial firms is not only a matter of technical design and formal oversight, but a continuous practice of narrating control under conditions of algorithmic opacity. Against this background our investigation makes several contributions to research on responsible AI governance.

Firstly, our study foregrounds the discursive dimension of AI accountability in financial firms risk disclosures. Research on responsible AI has generated rich insights into the design and governance of AI systems through technical, procedural and organisational mechanisms (Papagiannidis et al., 2025; Akbaringhatar et al., 2025; Pappas et al., 2023; Dennehy et al., 2023; Birkstedt, et al., 2023; Spanaki et al., 2022;

Enholm et al., 2022). Our investigation extends this work by showing how accountability is also constructed through language, as firms narrate control under conditions of algorithmic opacity (Ananny & Crawford, 2018; Agerfalk, 2020). A focus on this complementary discursive dimension matters because it reveals how firms navigate a central governance dilemma. Under algorithmic opacity full control is unattainable. Nevertheless, accountability still needs to be articulated and located, even when full control cannot be demonstrated.

In particular, our study shows that organisations discursively construct accountability through three recurrent framings: AI as internally governed through risk management and internal oversight, AI as institutionally governed through legal and regulatory structures, and AI as unpredictable, partially autonomous and under limited managerial control. Each of these framings positions accountability in a distinct locus of control within the wider sociotechnical system: inside the firm, in external regulatory institutions or dispersed through technological uncertainty. By identifying these discursive accountability framing, we complement prior work on formal and organisational governance mechanisms (Papagiannidis et al., 2025; Akbaringhatar et al., 2025; Pappas et al., 2023) and show that when full control cannot be demonstrated, firms use disclosure language to locate and bound accountability. These disclosures therefore constitute a discursive layer of governance in their own right. Accountability emerges not as a fixed property of organisational control, but as something actively located across organisational, institutional and technological contexts under conditions of algorithmic opacity. Our findings further indicate that discursive accountability does not operate as a substitute to formal governance but as a complementary layer that becomes particularly salient when formal mechanisms reach their limits. Under algorithmic opacity, formal controls cannot fully guarantee explainability or traceability and therefore discourse becomes the means through which accountability is located and bounded.

Secondly, our study advances understanding of how firms narrate AI unpredictability in governance contexts. Scholarly work treats AI unpredictability as a sociotechnical condition that emerges from the interaction of data, models, organisational routines and institutional settings, making outcomes only partially traceable (Bartsch et al., 2025; Brasse et al., 2023; Ananny & Crawford, 2018; Burrell, 2016). This literature also highlights the autonomy and inscrutability of AI systems, whose internal operations often remain opaque even to experts (Bartsch et al., 2025). Complementing these accounts, we show that firms do not simply report this uncertainty but actively construct and organise it within accountability narratives.

In particular, rather than treating opacity as an external constraint, firms often reframe it as a condition that can

be managed, pairing assertions of control with carefully bounded disclosures of “known unknowns”. Prior research on organisational risk disclosure treats the acknowledgment of known unknowns as a transparency-enhancing practice, signalling epistemic honesty about the limits of organisational knowledge (Pinnington et al., 2023). In the context of AI governance however, “known unknowns” take on an additional function. By naming what they do not yet know, firms signal competence and awareness while simultaneously shaping the scope of accountability. Uncertainty is thus presented as governance-relevant but not fully governable, allowing firms to appear prepared and engaged while preserving room to narrow accountability when adverse outcomes are later attributed to opacity, external vendors, or unsettled regulation.

Finally, our study contributes empirically by showing how the use of these accountability framings shifts over time, particularly with the rise of GenAI. Before GenAI, firms relied mainly on descriptive AI and internal accountability framings, with minimal reference to unpredictability. After GenAI, however, disclosures increasingly describe AI as unpredictable and sometimes partially autonomous. This pattern reflects how accountability framings evolve alongside technological change and shifting institutional pressures (Power, 2021; Christensen et al., 2021). By documenting these temporal dynamics, our study shows how the discursive construction of accountability adapts as AI capabilities and regulatory expectations develop.

6 Conclusions

Our study demonstrates that AI accountability in financial firms is not only engineered through formal governance mechanisms but also actively constructed through language. Firms construct AI accountability through three recurrent framings: AI as internally governed through organisational oversight, AI as institutionally governed through legal and regulatory structures and AI as unpredictable and partially autonomous. These framings locate responsibility in different parts of the sociotechnical system, demonstrating that accountability is also constructed through the narratives. Our findings further suggest that rather than describing the sociotechnical complexity of AI directly, firms actively organise uncertainty into accountability narratives that reframe opacity as a governable feature, preserving room to shape the boundaries of what they can reasonably be held accountable for. Finally, our results illustrate the temporal shift following the rise of GenAI. This shift shows how organisations recalibrate these narratives as AI systems become more complex and less traceable. Before GenAI firms relied mainly on descriptive AI or internal governance

framings. After GenAI disclosures increasingly emphasise AI unpredictability and partial autonomy. These dynamics show how accountability is constructed and renegotiated as firms adapt to evolving technologies and shifting institutional pressures.

Our study carries implications for investors, auditors and regulators. By conceptualising financial AI risk disclosures as governance artefacts that emerge where formal control reaches its limits, these stakeholders should attend not only to what firms claim, but to how discursive framings locate accountability, reveal its limits and shape responsibility under conditions of technological opacity.

For investors, this implies reading disclosures as structured risk narratives rather than neutral descriptions of technical reality. Patterns in how accountability is framed may indicate whether control is substantively exercised or primarily constructed through discourse, particularly as firms invoke unpredictability or partial autonomy. For auditors, the findings highlight the need to assess not only the technical accuracy of AI-related disclosures but also their governance coherence. This includes evaluating whether claims about accountability are consistent with the governance structures described elsewhere, and whether invocations of opacity are justified or function to narrow the scope of accountability in practice. For regulators, our findings raise the concern that current disclosure regimes may enable accountability deflection. When unpredictability is normalised as inherent rather than framed as a risk to be mitigated, scrutiny may be pre-empted. Regulators may therefore need to move beyond principles-based disclosure and define mandatory reporting categories (such as data lineage, model drift controls and human-in-the-loop mechanisms), that require demonstrable governance capacity rather than narrative positioning.

Our study is not without limitations. Although machine learning investigation enables the systematic and scalable analysis of a large corpus of disclosures, an inherent limitation of ML models lies in their difficulty in capturing the finer rhetorical nuances that firms intentionally build into their language. In our case, the classifier performed strongly for the institutional, unpredictable and descriptive categories, with high levels of accuracy that indicate relatively clear framing patterns. The main weakness concerned the internal governance frame, where performance was noticeably lower ($F1=0.57$). Our qualitative review of passages in this category suggests that this is less a technical flaw than a property of the discourse itself. When firms describe AI accountability as located inside the organisation, they frequently mix references to internal oversight, regulatory compliance and technological unpredictability within the same passage. While some blending of language was also evident in the other framings, it was more pronounced and systematic in passages associated with internal governance.

These passages often layer internal control language on top of strong institutional and unpredictability signals, which makes them particularly difficult for automated methods that must assign each passage to a single dominant frame. ML is therefore well suited to identifying broad framing patterns across a large corpus, whereas the subtle strategic manoeuvres and mixed signalling in corporate narratives require for further qualitative analysis. Future qualitative work could complement our approach by examining these rhetorical nuances more closely.

Another limitation is that our analysis assigns each passage to a single dominant accountability frame. This design choice supports clear classification and aggregate comparison, but it dilutes cases where other framings appear as strong secondary cues. Future research could extend our findings by examining frame co-occurrence patterns and investigate when and why organizations invoke internal, institutional and autonomous framings together rather than separately. Future research could also explore what strategic or contextual factors drive such layered accountability claims.

Although the keyword extraction procedure was iteratively refined and validated through manual review, some relevant passages may not have been captured.

A further limitation is that our dataset is restricted to publicly listed US financial firms in the S&P index, whose disclosures are shaped by a highly institutionalised reporting environment and may not reflect discursive practices in other industries or regulatory contexts. Future research could extend the inquiry to different sectors or jurisdictions to examine whether similar patterns of blended accountability framing emerge. Additionally, while our analysis shows how firms locate control and accountability using different governance framings, it does not examine how responsibility is discursively distributed across the expanding AI ecosystem. This matters because accountability without responsibility allocation becomes symbolic. Firms may assert strong governance while deflecting liability onto vendors, regulators, data providers or abstract technological forces. As AI systems become increasingly co-developed and externally sourced, understanding how organisations discursively shift responsibility becomes central to identifying oversight, liability and ethical duty. Future work could therefore disentangle how firms pre-emptively allocate responsibility in their disclosures.

Last but not least, future research could extend our analysis by examining how the three accountability framings shape market responses. Different framings may influence investor sentiment, stock volatility or analyst assessments, with unpredictability framings likely to trigger greater caution than internally governed ones. In this way, future work could link discursive accountability framings to tangible financial outcomes.

Appendix 1

Table 3 Illustrative examples per category and justification

Passage	Category	Justification
This committee oversees cybersecurity matters, as well as technology risk management frameworks and methodologies, and monitors their effectiveness. This committee is co-chaired by our CISO and our chief technology officer, who are appointed as chairs by the chairs of the Firmwide Enterprise Risk Committee. To assist the Firmwide Technology Risk Committee in carrying out its mandate, the Firmwide Artificial Intelligence Risk and Controls Committee, which oversees risks associated with the use of AI, reports to the Firmwide Technology Risk Committee.	Internally governed	Formal internal oversight
Our approach to AI may not be successful, which could materially and adversely affect our business, financial condition or results of operations. Given the importance of data to our products and services, AI is becoming an increasingly important part of our business and industry. We have established a Company-wide AI strategy to drive our approach to data protection, licensing and AI integration in our processes, products and services.	Internally governed	Internal policy
On December 9, 2023, the European Parliament and European Council reached a provisional agreement on the European Union Artificial Intelligence Act, which, once formally adopted, will broadly regulate the use of AI within the European Union. European countries, and supranational political organizations like the EU and the Council of Europe, are expected to continue taking an active role in regulating AI in ways that may impact the insurance industry in the future.	Institutionally governed	External regulation
In addition, regulation of AI is rapidly evolving worldwide as legislators and regulators are increasingly focused on these powerful emerging technologies. The technologies underlying AI and its uses are subject to a variety of laws and regulations, including intellectual property, privacy, data protection and information security, consumer protection, competition, and equal opportunity laws, and are expected to be subject to increased regulation and new laws or new applications of existing laws and regulations. AI is the subject of ongoing review by various U.S. governmental and regulatory agencies, and various U.S. states and other foreign jurisdictions are applying, or are considering applying, their platform moderation, privacy, data protection and data security laws and regulations to AI or are considering general legal frameworks for AI. We may not be able to anticipate how to respond to these rapidly evolving frameworks, and we may need to expend resources to adjust our offerings in certain jurisdictions if the legal frameworks are inconsistent across jurisdictions.	Institutionally governed	References to legal frameworks; Focuses on compliance
Because the Generative AI technology is so new, many of the potential risks of Generative AI are currently unknowable; however, specific risks relating to AI and Generative AI could include, among others: "Reputational Damage: Malicious actors could use AI to create deepfakes of the Company's executives or manipulate financial documents, leading to loss of customer trust and significant reputational damage. Moreover, the use of AI trained on inaccurate data sets could result in inaccurate or biased decisions." "Fraudulent Activity: AI could be used to create forged documents or impersonate individuals to commit financial fraud, leading to financial losses and regulatory scrutiny.	Unpredictable/Partially autonomous	AI as a black-box
AI models, particularly generative AI models, may produce output or take action that is incorrect, that result in the release of private, confidential or proprietary information, that reflect biases included in the data on which they are trained, infringe on the intellectual property rights of others, or that is otherwise harmful. In addition, the complexity of many AI models makes it challenging to understand why they are generating particular outputs.	Unpredictable/Partially autonomous	AI as partially autonomous
If we fail to keep pace with technological change, including as a result of artificial intelligence, we could lose clients or have trouble attracting new clients, and our ability to grow may be limited. The markets for our products and services are characterized by constant and rapid technological change, evolving industry standards, frequent introduction of new products and services, and increasing client expectations. Our ability to respond timely to these changes, including by enhancing our current products and services and developing and introducing new products and services, will significantly affect our future success.	Descriptive AI	AI as a tool for competitive positioning
Our future success depends, in part, on our ability to anticipate and respond effectively to the threat and opportunity presented by digital disruption, "big data" and data analytics, and other developments in technology. These may include new applications or insurance-related services based on artificial intelligence, machine learning, robotics, blockchain, the metaverse or new approaches to data mining that impact the nature of how we generate revenue. We may be exposed to competitive risks related to the adoption and application of new technologies by established market participants (for example, through disintermediation or use of the metaverse) or new entrants such as technology companies, "Insurtech" start-up companies and others.	Descriptive AI	AI as a tool for competitive positioning

Appendix 2

Table 4 Per fold Accuracy

Fold	Accuracy
1	0.7727
2	0.7907
3	0.7674
4	0.7907
5	0.7907
6	0.8182
7	0.907
8	0.6744
9	0.7209
10	0.7674
11	0.75
12	0.6977
13	0.8605
14	0.7907
15	0.7907
Mean Accuracy	0.7793
Standard Deviation	0.0562

Table 5 Classification report

Class	Precision	Recall	F1-score
Internally Governed AI	0.517	0.641	0.573
Institutionally Governed AI	0.902	0.886	0.894
Unpredictable/Partially autonomous AI	0.761	0.78	0.77
Descriptive AI	0.896	0.732	0.806
Accuracy			0.779

Table 6 Confusion matrix Appendix 3

Actual \ Predicted	1	2	3	4
1. Internally Governed AI	75	10	22	10
2. Institutionally Governed AI	16	194	7	2
3. Unpredictable/Partially autonomous AI	27	7	124	1
4. Descriptive AI	27	4	10	112

Appendix 3

Model	Accuracy	SD	Macro-F1	F1: Internally Governed	F1: Institutionally Governed	F1: Autonomous/Unpredictable	F1: Descriptive AI
SVM (TF-IDF, RBF kernel)	68.2%	5.70%	0.587	0.087	0.745	0.744	0.776
Logistic Regression (TF-IDF)	70.4%	5.00%	0.620	0.189	0.778	0.758	0.768
RoBERTa (roberta-base)	76.7%	6.60%	0.578	0.163	0.798	0.724	0.708
SetFit (paraphrase-mpnet-base-v2)	77.9%	5.60%	0.761	0.573	0.894	0.770	0.806

Appendix 4

Accountability Frame	Fiscal Year	Window 0–1 (1347 passages)	Window 0–2 (1203 passages)	Window 0–3 (1063 passages)	Notes
1. Internally Governed	2022	29.7%	26.1%	26.4%	Stable across conditions
	2023	21.8%	20.4%	22.0%	Stable across conditions
	2024	21.4%	21.0%	21.8%	Stable across conditions
2. Institutionally Governed	2022	24.6%	21.0%	22.7%	Stable across conditions
	2023	36.6%	34.8%	33.4%	Stable across conditions; post-ChatGPT surge replicated
	2024	37.0%	32.8%	33.7%	Stable across conditions; post-ChatGPT surge replicated
3. Autonomous/Unpredictable	2022	11.9%	10.9%	9.1%	Stable across conditions
	2023	25.4%	24.2%	21.7%	Stable across conditions
	2024	26.7%	26.9%	24.1%	Stable across conditions
4. Descriptive AI	2022	33.9%	42.0%	41.8%	0–1 lower: tighter windows capture less descriptive context
	2023	16.2%	20.7%	22.9%	Some variation but directionally consistent
	2024	15.0%	19.3%	20.4%	Some variation but directionally consistent

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10796-026-10762-y>.

Acknowledgements We are particularly grateful to the Associate Editor and the anonymous reviewers for their thoughtful and constructive feedback.

Authors' Contributions Eleni Lioliou: Conceptualization, Methodology, Data curation, Formal analysis, Writing - original draft, Writing - review & editing. May Seitanidi: Conceptualization, Writing - review & editing. Lea Stadler: Conceptualization, Writing - review & editing. All authors read and approved the final manuscript.

Funding Not applicable.

Data Availability All data analysed in this study originate from publicly available 10-K filings accessible through the U.S. Securities and Exchange Commission's EDGAR database.

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable. No identifiable personal data are included in the manuscript.

Consent to participate Not applicable.

Competing interests The authors declare that they have no competing interests.

References

- Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems*, 29(1), 1–8.
- Ågerfalk, P. J., Conboy, K., Crowston, K., Eriksson Lundström, J., Jarvenpaa, S. L., Ram, S., & Mikalef, P. (2022). *Artificial intelligence in information systems: State of the art and research roadmap*. Association for Information Systems.
- Akbarighatar, P., Pappas, I. O., Purao, S., & Vassilakopoulou, P. (2025). Enacting responsible AI: A configurational analysis of AI principles in practice. *Information Systems Frontiers*, (1), 27. <https://doi.org/10.1007/s10796-025-10618-x>
- Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New media & Society*, 20(3), 973–989.
- Arrieta, A. B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Barredo Arrieta, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.
- Arvidsson, S., & Sabelfeld, S. (2023). Adaptive framing of sustainability in CEO letters. *Accounting Auditing & Accountability Journal*, 36(9), 161–199.
- Barton, J., & Mercer, M. (2005). To blame or not to blame: Analysts' reactions to external explanations for poor financial performance. *Journal of accounting and economics*, 39(3), 509–533.
- Bartsch, S. C., Nguyen, L. H., Schmidt, J. H., Du, G., Adam, M., Benlian, A., & Sunyaev, A. (2025). The present and future of accountability for AI systems: A bibliometric analysis. *Information Systems Frontiers*. <https://doi.org/10.1007/s10796-025-10636-9>
- Bauer, K., Von Zahn, M., & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. *Information systems research*, 34(4), 1582–1602.
- Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133–167.
- Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework. *European law journal*, 13(4), 447–468.
- Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33(1), 26.
- Brennan, N. M., & Merkl-Davies, D. M. (2013). Accounting narratives and impression management. *The Routledge companion to accounting communication* (pp. 109–132). Routledge.
- Brynjolfsson, E., & McAfee, A. (2017). The business of artificial intelligence. *Harvard business review*, 7(1), 1–2.
- Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), Article 2053951715622512.
- Cho, C. H., Laine, M., Roberts, R. W., & Rodrigue, M. (2015). Organized hypocrisy, organizational façades, and sustainability reporting. *Accounting, Organizations and Society*, 40, 78–94.
- Christensen, L. T., Morsing, M., & Thyssen, O. (2021). Talk-action dynamics: Modalities of aspirational talk. *Organization Studies*, 42(3), 407–427.
- Currie, W. L., Gozman, D. P., & Seddon, J. J. (2018). Dialectic tensions in the financial markets: A longitudinal study of pre-and post-crisis regulatory technology. *Journal of Information Technology*, 33(4), 304–325.
- Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, 48(1), 24–42.
- Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., & Pappas, I. O. (2023). Artificial intelligence (AI) and information systems: Perspectives to responsible AI. *Information Systems Frontiers*, 25(1), 1–7.
- Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial intelligence and business value: A literature review. *Information systems frontiers*, 24(5), 1709–1734.
- Fama, E. F., & Jensen, M. C. (1983). Separation of ownership and control. *The Journal of Law and Economics*, 26(2), 301–325.
- Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1), 62–70.
- Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535–545.
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schaffer, B., Valcke, P., & Vayena, E. (2018). AI4People - An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, 28(4), 689–707.
- Gawer, A. (2021). Digital platforms' boundaries: The interplay of firm scope, platform sides, and digital interfaces. *Long range planning*, 54(5), Article 102045.
- Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Harper & Row.
- Gozman, D. and Currie, W. (2014). The role of investment management systems in regulatory compliance: A post-financial crisis study of displacement mechanisms. *Journal of Information Technology*, 29(1), 44–58.
- Gupta, M., Parra, C. M., & Dennehy, D. (2022). Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter? *Information Systems Frontiers*, 24(5), 1465–1481.

- Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885–1894). PMLR.
- Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2019, January). Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 131–138).
- Leung, S., Parker, L., & Curtis, J. (2015). Impression management through minimal narrative disclosure in annual reports. *The British accounting review*, 47(3), 275–289.
- Li, Y., & Goel, S. (2025). Making it possible for the auditing of AI: A systematic review of AI audits and AI auditability. *Information Systems Frontiers*, 27(3), 1121–1151.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Mikalef, P., Benlian, A., Conboy, K., & Tarafdar, M. (2025). Responsible AI starts with the artifact: Challenging the concept of responsible AI in IS research. *European Journal of Information Systems*, 34(3), 407–414.
- Mikalef, P., Conboy, K., Lundström, J. E., & Popović, A. (2022). Thinking responsibly about responsible AI and ‘the dark side’ of AI. *European Journal of Information Systems*, 31(3), 257–268.
- Niederman, F., & Baker, E. W. (2023). Ethics and AI issues: Old container with new wine? *Information Systems Frontiers*, 25(1), 9–28.
- Nikou, M., Mansourfar, G., & Bagherzadeh, J. (2019). Stock price prediction using deep learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management*, 26(4), 164–174.
- Novelli, C., Taddeo, M., & Floridi, L. (2024). Accountability in artificial intelligence: What it is and how it works. *AI & Society*, 39(4), 1871–1882.
- Olan, F., Papadopoulos, T., Spanaki, K., & Jayawickrama, U. (2025a). Social entrepreneurial marketing and innovation in B2B services: Building resilience with explainable artificial intelligence. *Information Systems Frontiers*, (1), 14. <https://doi.org/10.1007/s10796-025-10583-5>
- Olan, F., Spanaki, K., Ahmed, W., & Zhao, G. (2025b). Enabling explainable artificial intelligence capabilities in supply chain decision support making. *Production Planning & Control*, 36(6), 808–819.
- Pachidi, S., Berends, H., Faraj, S., & Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, 32(1), 18–41.
- Palermo, T., Power, M., & Ashby, S. (2017). Navigating institutional complexity: The production of risk culture in the financial sector. *Journal of Management Studies*, 54(2), 154–181.
- Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems*, 34(2), Article 101885.
- Pappas, I. O., Mikalef, P., Dwivedi, Y. K., Jaccheri, L., & Krogstie, J. (2023). Responsible digital transformation for a sustainable society. *Information Systems Frontiers*, 25(3), 945–953.
- Pinnington, B., Benstead, A., & Meehan, J. (2023). Transparency in supply chains (TISC): Assessing and improving the quality of modern slavery statements. *Journal of Business Ethics*, 182(3), 619–636.
- Power, M. (1997). *The audit society: Rituals of verification*. OUP Oxford.
- Power, M. (2007). *Organized uncertainty: Designing a world of risk management*. Oxford University Press.
- Power, M. (2021). Modelling the micro-foundations of the audit society: Organizations and the logic of the audit trail. *Academy of management review*, 46(1), 6–32.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144).
- Rosenkranz, J., & Pollach, I. (2016). The framing and reframing of corporate financial results: How corporate earnings releases become news. *Corporate Communications: An International Journal*, 21(1), 103–119.
- Shaban, W. M., Ashraf, E., & Slama, A. E. (2024). SMP-DL: A novel stock market prediction approach based on deep learning for effective trend forecasting. *Neural Computing and Applications*, 36(4), 1849–1873.
- Sinclair, A. (1995). The chameleon of accountability: Forms and discourses. *Accounting organizations and Society*, 20(2–3), 219–237.
- Spanaki, K., Dennehy, D., Papadopoulos, T., & Dubey, R. (2025). Data-driven digital transformation in operations and supply chain management. *International Journal of Production Economics*, 284, Article 109599.
- Spanaki, K., Karafili, E., Sivarajah, U., Despoudi, S., & Irani, Z. (2022). Artificial intelligence and food security: Swarm intelligence of AgriTech drones for smart AgriFood operations. *Production Planning & Control*, 33(16), 1498–1516.
- Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of management review*, 20(3), 571–610.
- Suddaby, R., Bitektine, A., & Haack, P. (2017). Legitimacy. *Academy of management annals*, 11(1), 451–478.
- Tunstall, L., Reimers, N., Jo, U. E. S., Bates, L., Korat, D., Wasserblat, M., & Pereg, O. (2022). Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.
- Vaara, E., Aranda, A. M., & Etchanchu, H. (2024). Discursive legitimacy: An integrative theoretical framework and agenda for future research. *Journal of Management*, 50(6), 2343–2373.
- Vijh, M., Chandola, D., Tikkiwal, V. A., & Kumar, A. (2020). Stock closing price prediction using machine learning techniques. *Procedia computer science*, 167, 599–606.
- Waardenburg, L., & Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization*, 32(4), Article 100432.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Eleni Lioliou is Reader in International Business and Digital Strategy at Queen Mary University of London and Associate Editor for the Journal of Information Technology. She received her PhD from the London School of Economics. Her research examines the governance of digital sourcing, platforms and AI systems, with a focus on how emerging technologies reshape accountability and control.

M. May Seitanidi is Reader in Strategy and Corporate Social Responsibility, Founder and Lead of the Collaborative Solutions for Social Problems Lab at Kent Business School, University of Kent. Her research focuses on cross-sector collaboration, collaborative value creation and governance, examining how public, private and civil society actors design and sustain collaborative systems for societal outcomes.

Lea Stadler is Associate Professor of Strategy and Sustainability at Grenoble Ecole de Management, France and Deputy Editor at Business & Society. Her research focuses on sustainability and collaboration-related complexities, paradoxes, and strategic responses. She has published more than 20 articles in academic journals and has received international research awards for her work.