Using Interpretable Machine Learning for Accounting Fraud Detection – A Multi-User Perspective

Leonhard J. Lösse and Barbara E. Weißenberger

Summary: Machine learning models are increasingly used to identify accounting manipulations based on disclosed information. Still, most approaches focus on accuracy, which at the same time leads to a high number of false-positive predictions that practically hinder their application. The paper analyzes the need for interpretable machine learning techniques from the perspective of primary users such as statutory auditors, enforcement institutions, or investors with respect to their specific legal and organizational frameworks. From this, requirements for additional explanations are derived, which in turn serve as indicators for plausibility checks or as starting points for investigations, thus improving the manageability of predictions and promoting the implementation of the models.

Keywords: Accounting Fraud, Machine Learning, Interpretability, Fraud Detection, Audit, Enforcement

Interpretierbare Vorhersagen durch Machine Learning bei der Aufdeckung von Bilanzbetrug – Analyse aus Nutzerperspektive

Zusammenfassung: Modelle des maschinellen Lernens werden zunehmend genutzt, um Bilanzmanipulationen anhand offengelegter Informationen zu identifizieren. Die meisten Ansätze fokussieren auf Genauigkeit, was gleichzeitig zu einer hohen Zahl falsch-positiver Vorhersagen führt, die eine Anwendung behindern. Der Beitrag analysiert den Bedarf an Verfahren interpretierbaren maschinellen Lernens aus Perspektive der primären Nutzer wie Abschlussprüfer, Bilanzkontrolle und Investoren mit ihren spezifischen rechtlichen und organisatorischen Rahmenbedingungen. Daraus werden Anforderungen an zusätzliche Erklärungen abgeleitet, die wiederum als Indikatoren für Plausibilisierungen oder als Ansatzpunkte für Untersuchungen dienen und so die Handhabbarkeit von Vorhersagen verbessern und die Implementierung der Modelle fördern.

Stichwörter: Bilanzmanipulation, Maschinelles Lernen, Interpretierbarkeit, Betrugsaufdeckung, Abschlussprüfung, Bilanzkontrolle

1. Introduction

After having peaked around the turn of the millennium for the last time, financial accounting fraud has once again become paramount with massive incidents of international interest such as Wirecard or continuing inquiries into potential further case, e.g., Adler

Group (*McCrum*, 2020; *Storbeck*, 2022). The occurrence shows: Accounting fraud, especially if active and meticulously planned, can never be prevented entirely, despite the existence of audits and enforcement structures. Currently, the result is increased sensitivity in the auditing profession, tighter enforcement, and skepticism on the part of investors.

To address this issue and facilitated by an increasing amount of readily available data science applications, current research at the interface of accounting and information systems has been exploring the potential of using machine learning to detect signs of accounting fraud in firms' financial reports as early as possible. The key strength of machine learning[1] is that the underlying model does not have to be explicitly programmed, but rather learns relationships from existing data and observations (*Samuel*, 1959). In this vein, during the last decade research has focused on developing machine learning algorithms that identify accounting fraud as exactly as possible. This has resulted in a most relevant drawback, namely a high number of false-positive predictions, i.e., the erroneous classification of legally compliant cases as fraudulent. As machine learning algorithms typically create black box models that do not allow the user to identify the reason exactly why a financial report is classified as (non-)fraudulent, each false-positive classification results in high costs resulting from, e.g., manual audit efforts and/or losses in reputation or investment opportunities (*Beneish & Vorst*, 2022).

Whether further significant improvements in sensitivity can be achieved while reducing the number of false-positive predictions at the same time is questionable as the amount of training data is restricted due to the limited occurrence of accounting fraud events. Thus, to lever the potential of machine learning algorithms in this field, another approach is considered to be more promising. It consists in trying to make the fraud detection models created by machine learning algorithms more transparent by implementing additional analysis. First approaches even aim to analyze models' interpretability locally at the level of individual prediction, resulting in increased explainability (*Craja et al.*, 2020; *Zhang et al.*, 2022).

We argue that following this avenue of research more closely makes it necessary to also address the diverging perspectives of different user groups of financial statements as well as their legal and operating environments resulting, e.g., in requiring different types of additional explanations. For example, in a highly regulated sector with outstanding professional requirements for expertise, concerns regarding trust, accountability, and efficient implementation of human-machine-interaction-based applications must be considered.

To provide a detailed analysis on this subject, we select audit firms, enforcement authorities, and investors as primary user groups of accounting fraud detection models. For each group, we consider scientific and professional literature as well as legislation on implications arising from legal requirements or organizational and operating circumstances to answer the following research questions:

RQ1: What legal and organizational conditions drive the need for accounting fraud predictions' interpretability?

RQ2: What behavioral interactions must be considered for effective and efficient implementation in a highly regulated setting with high professional requirements?

1 Even though machine learning (ML) is a form of artificial intelligence (AI), it is used synonymously here because the AI systems in this paper are limited to machine learning approaches only.

In a nutshell, our analysis suggests that even though different settings apply for diverging user groups of financial statements, an application of machine learning models for detecting accounting fraud without additional transparency is reasonable only under very narrow assumptions. Enforcement authorities can assess abstract risk for a risk-oriented selection of firms to be audited in the context of sampling examinations and non-professional investors might reduce financial losses through avoiding investments in potentially fraudulent firms. In contrast, user groups in other potential use cases are regularly prevented from applying opaque models by legal and organizational restrictions and further impairing behavioral factors.

Our paper therefore contributes to the literature on machine learning-based accounting fraud detection first by offering a conceptual framework on major users' demands for more detailed explanations on a local level. Requirements are derived from and discussed against the backdrop of legal and organizational environments. Second, our findings also practically contribute by serving as a basis for future applications which focus more intensely on the human-machine interaction by enabling a better understanding of predictions which can subsequently be processed with professional expertise and complemented by human intelligence, and thus, can increase trust as well as necessary professional scepticism in machine learning applications.

The paper is structured as follows. In section 2, an overview on accounting fraud detection as well as on machine learning-based detection approaches and current advances in interpreting algorithmic predictions by so-called interpretable or explainable artificial intelligence (XAI) is given. In section 3, we analyze the legal and operating circumstances of auditors, enforcement institutions, and investors incorporating selected qualitative support and derive individual requirements for interpretable fraud predictions. Finally, the need for explanations is discussed against the background of human-machine interaction in a highly regulated field.

2. Interpretability of Machine Learning-based Accounting Fraud Detection

2.1 Accounting Fraud Detection

According to the *Association of Certified Fraud Examiners* (2022), occupational fraud is a pervasive problem that involves varying costs and damages depending on the type of fraud committed. Within the set of rules of the International Standards on Auditing (ISA), as the most widely used audit standards and applied directly or indirectly in more than 130 jurisdictions (*IAASB*, 2022), fraud is distinguished as intentional manipulation from unintentional errors (ISA 240.2). Among the various types of organizational fraud, e.g. encompassing embezzlement, insider trading, corruption or cover-ups (*Moberg*, 1997), accounting fraud refers to a firm's financial statements and is generally characterized by over- or understatements, first, of assets and liabilities, and second of revenues and expenses. Even though accounting fraud occurs relatively rarely, it regularly results in large financial losses, and in the case of bankruptcy, e.g., with Wirecard (2020), also in job loss as well as the impairment of the business ecosystems a fraudulent firm has been part of. Decreasing the fraud risk by its detection at the earliest possible time is therefore crucial for the efficiency of capital, labor and/or goods markets.

To avoid accounting fraud, firms implement internal control systems by corporate governance exercised via a firm's one- or two-tier board system. Still, these mechanisms might fail due to negligence or top management fraud. It is the role of a firm's statutory auditors to externally monitor the compliance of the accounting systems as well as the resulting financial statements with all existing norms and regulations including professional standards so that they provide a true and fair view on the firm. Thus, undetected fraudulent financial statements are assigned to the auditor's area of responsibility regardless of actual responsibility for the failure, which is commonly known as the audit expectation gap (*Koh & Woo*, 1998; *Ruhnke & Schmidt*, 2014), even if the standards clearly state the limits of a financial statement audit in form of a reasonable assurance (ISA 200.5). Therefore, audit research and practice strive for continuous improvement towards more effective audit procedures and technologies, and the shift from traditional to digital audits including machine-learning based fraud detection which is explicitly seen as an opportunity to reduce this gap (*Fotoh & Lorentzon*, 2023).

As financial statements play a paramount economic role and manipulations can never be fully prevented, national enforcement authorities re-examine the audited financial statements on a sample or ad hoc basis. Within the European Union (EU), competencies are codified in Article 4 of Directive 2004/109/EC ('Transparency Directive') and delegated to national authorities, such as the German Federal Financial Supervisory Authority (BaFin). Well-known counterparts are the Swiss Financial Market Supervisory Authority (FINMA) or SEC.

Besides such regulatory requirements, price mechanisms on capital markets also serve as an indirect external governance mechanism, as investors may act as a kind of market corrective. While the average investor seeks to avoid downside risk by refraining from investing in potentially fraudulent companies, identified risky companies can also be leveraged to profit, if investors sell the shares short and thus speculate on the discovery of possible manipulation. *Massa et al.* (2015), e.g., state "that short selling functions as an external governance mechanism" by a disciplining effect reducing earnings management, with a practical example being Wirecard and the activities of Fraser Perring (*Langenbucher et al.*, 2020).

2.2 Machine Learning-based Detection Approaches

The common core of machine learning-based approaches to detect accounting fraud consists in using disclosed financial data to train models with various algorithms. In contrast to traditional programming techniques which explicitly implement rules to derive a given output from a set of input data, machine learning algorithms use a training-based programming approach by relating input data mathematically to a given output – in our case, e.g., fraudulent vs. non-fraudulent financial statements to develop these rules internally. If the training of the algorithm is successful, its application to a new set of financial statements as input data then allows for a fast and automated classification and thus the detection of accounting fraud. Machine learning models exceed human's capabilities not only because modern hard- and software is able to process vast amounts of data in a short time, but especially because they are able to discover also unknown red flags, i.e., complex relationships between input and output data pointing towards accounting fraud but that are still 'unknown unknowns' to the human mind. This was considered to be a key advantage especially since to date there are but few confirmed theories for identifying financial statement manipulation (*Fanning & Cogger*, 1998). As a result, early machine learning approaches assumed that more flexible algorithms are better able to capture and process more complex changes and relationships between multiple accounts, which humans cannot do due to their limited capacity to absorb and process information in the sense of information overload (*Green & Choi*, 1997). However, research as early as by *Beneish* (1999a) emphasizes the importance of theoretically based variable selection to cover manifestations of manipulations or structures that favor them, which is a first indication that the successful application of machine learning is significantly influenced by human expertise.

During the last decade, the availability of machine learning approaches led to a versatile growth of the research stream in machine learning based empirical accounting research (*Sellhorn*, 2020). Specifically in the field of fraud detection, on the one hand, more diverse input variables were included that went beyond financial and governance variables (*Fanning & Cogger*, 1998) and, e.g., also used textual data from disclosed narratives like management discussions and analysis (MD&A) (*Cecchini et al.*, 2010b; *Goel et al.*, 2010; *Glancy & Yadav*, 2011; *Purda & Skillicorn*, 2015). On the other hand, research addressed the issue of selecting optimal algorithms, which, e.g., included support vector machines in addition to regressions and neural networks (*Cecchini et al.*, 2010a). Further, possible optimizations in the training process, such as undersampling or cross-validation approaches, were analyzed to increase the models' performance (*Perols et al.*, 2017).

As *Dechow et al.* (2011) state, all these approaches are critical as they offer on the one hand the potential to improve the efficiency of the capital markets, but on the other hand are costly in the case of classification errors. In terms of misclassifications accounting fraud detection is inherently faced with highly imbalanced data. If this is not adequately accounted for, models may tend to predict all observations as nonfraudulent and still achieve accuracy measures which might seem to be outstanding (*Perols et al.*, 2017). However, this would neglect varying costs for different prediction outcomes and would severely limit the meaningfulness of traditional performance measures as recall or precision (*Powers*, 2011). In this context, *Zahn et al.* (2022) highlight the potential for overall costs to be reduced if the smaller group of imbalanced data, in this case fraudulent firms, are associated with higher costs in false predictions. *Beneish* (1999b, 1999a) assumes, e.g., a cost ratio for investors of false negatives to false positives of about 20:1 to 30:1. Therefore, an approach consisting only of non-fraudulent predictions cannot be an option due to high costs arising from missed fraud cases. Furthermore, the costs of error differ not only according to the type of error, e.g., whether a case of manipulation was overlooked (false negative) or a company was wrongly classified as fraudulent (false-positive), but different groups of users, i.e. auditors, enforcement institutions, and investors, are also faced with varying misclassification costs. *Beneish and Vorst* (2022) simulate the costs of applying different classification models from the perspective of these user groups. Their results indicate that even though many machine learning-models are highly sensitive with respect to detecting accounting manipulations, this regularly comes at the expense of numerous false-positive predictions. As a bottom line of their research, they conclude that misclassification costs in many cases are too high especially for audit firms.

To address this issue, one approach is to use traditional algorithms, e.g., regression analysis, which is inherently interpretable but lacks the advantages of the machine learning-based approaches. As a compromise, *Gepp et al.* (2021) develop a contemporary ensemble model by training an independent step-wise regression model, thus deriving model coefficients for variables that might drive accounting fraud. Other approaches take a step towards interpretability in advanced decision-tree models, e.g., Random Forests, by using post hoc feature importance to identify variables contributing to the likelihood of fraud (*Bao et al.*, 2020). *Craja et al.* (2020) identify clear textual indicators for accounting fraud in companies' MD&As by using Local Interpretable Model-agnostic Explanations (LIME), eliciting certain phrases related to accounting fraud to provide additional guidance, but it still remains open to what extent concrete starting points for plausibility checks or further investigations arise. Concerning financial variables, *Zhang et al.* (2022) recently succeeded in exemplifying for the first time the use of different approaches for so-called explainable AI (XAI) in the audit context, which also offered initial interpretations of individual predictions and thus enabling the identification of the main drivers for flagging and offering starting points for plausibility checks. Still, a differentiated analysis from the varying point of view of different users in their legal and organizational context is lacking.

2.3 Relevance of Interpretable Machine Learning

The field of interpretable machine learning has received considerable attention, at the latest since 2017/2018 (Figure 1). The central object of research is to develop and apply approaches and procedures to open the so-called black box of AI (*Castelvecchi*, 2016). Although the idea of addressing black box models' transparency is not new, demand has grown because of the actual widespread and increasingly simple application of machine learning-based systems (*Samek & Müller*, 2019). In this context, terms comprising XAI, Explainable AI, or Interpretable Machine Learning refer to a common core (*Adadi & Berrada*, 2018). In spite of the popular and widespread usage of the phrase 'explainable' AI, the term ‚interpretable' (machine learning) is no less common in science. In the following, both terms are used synonymously.

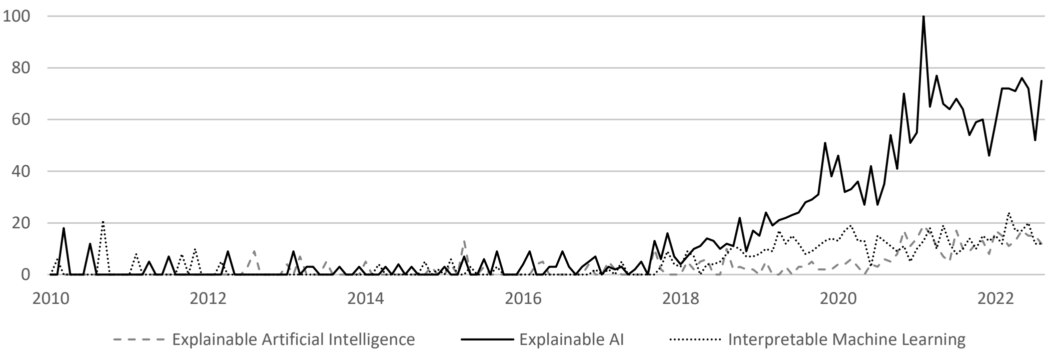

Figure 1: Google trends search related to interpretable machine learning (all categories)

A common understanding of interpretability is crucial to concepts like XAI. We follow *Miller* (2019), who defines interpretability as "the degree to which an observer can understand the cause of a decision" and further equates the terms interpretability and explainability. A widely used approach to substantiate the phenomenon of XAI comprises two essential components, namely "produce more explainable models while maintaining a high level of learning performance" and "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (*Gunning*, 2017).

The first aspect targets the performance interpretability trade-off. In simplified terms, this trade-off means that better algorithmic performance with respect to classification tasks is usually achieved at the expense of interpretability. However, this is not always the case: Sometimes, transparent models with simple underlying relationships can also achieve superior performance, providing so-called model-inherent explanations. But as contemporary machine learning applications are increasingly based on complex relationships between input and output data, e.g., by using algorithms based on support vector machines or neural networks, current approaches to interpretable machine learning provide post hoc explanations based on approximations as, e.g., surrogate models (*Adadi & Berrada*, 2018). This is accomplished by conducting additional analyses to provide insights in how a given classification is achieved, or, in other words, to allow for a peek into to the algorithm's black box. The advantages of such post hoc analyses are twofold. First, they are often model-agnostic, i.e., independent of the underlying algorithm. Second, the trained model itself and its performance remain unaffected because the analyses are applied within the already trained model.

From a technical point of view, interpretable machine learning comprises an ever-growing number of approaches to explain models. Feature relevance explanations, e.g., refer to quantitative scores which variables contribute most to a model's prediction. In contrast, visual explanation aims at illustrating mechanisms, while text explanations go beyond by attempting to provide text-based explanations of models' functionalities (*Arrieta et al.*, 2020). A further aspect of the techniques is scope. A large part of recent empirical research attempts to analyze higher-level relationships between selected variables in experiments, surveys or archival-data based studies. In this context, weights from linear regressions can be described as globally interpretable as they reflect the context at the level of the entire model. The same applies to the previously listed feature relevance explanations such as feature importance. In contrast to global interpretability, some post hoc approaches generate explanations only at the local level, that is within a very much restricted scope. In that case, interpretations are derived on the level of individual predictions. I.e., the approaches are based on slightly manipulating input data around an observation while observing their effects on the prediction, resulting in simplified locally valid models (*Molnar*, 2022).

With post hoc explanations, the second aspect of human user enablement is addressed. These analyses enable an improved information transfer which is more comprehensible and also more targeted, thus enhancing decision making. The diversity of approaches leads to some of them being regarded as more or less human-like and human-friendly (*Adadi & Berrada*, 2018). As the terms indicate, human-like refers to the extent to which the approaches outputs resemble explanations of humans. Human-friendly describes how well explanations can be understood by humans. This is significantly influenced by how contrastive, selective and social the explanations are, which are considered decisive factors for explanations' quality (*Miller*, 2019).

3. Multi-User Demand for Predicting Accounting Fraud

Dechow et al. (2011) highlight that multiple user groups could benefit from effective and efficient machine learning-based fraud detection approaches. We focus on the main groups of users that primarily employ financial statement analysis in practice, which are similarly classified by *Beneish and Vorst* (2022), focusing on the legal as well as organizational environment in which the users operate. On this basis, the individual demand for interpretable models will be derived before discussing their potential in terms of human-machine interaction covering issues of trust, expertise, and accountability.

3.1 Audit

The ISA require auditors to provide reasonable assurance about whether financial statements are free from fraud or error by collecting sufficient evidence to reduce the risk of an erroneous audit opinion to an acceptably low level (ISA 200.5). Risk-oriented planning of audit procedures within a specific audit engagement is essential, where risks have to be identified and determine the audit strategy and program (ISA 300.9 & A8). ISA 315 (revised) therefore requires the identification and assessment of risks at both the financial statement and the assertion level while the risk assessment alone does not constitute audit evidence (ISA 315.4 – 5). This also implies that an isolated prediction would not suffice. Instead, it must be possible to conclude the prediction's driving factors to obtain sufficient appropriate audit evidence (ISA 240.10b). This particularly applies to potential fraud, where specific indicators must be identified on which further audit procedures can be planned (ISA. 240.11). As required by ISA 520.4, analytical procedures involve analyzing plausible relationships. Here, too, driving factors must be identifiable to check relationships and patterns for plausibility.

In each individual audit engagement, the objective is to obtain sufficient audit evidence that "enable[s] the auditor to draw reasonable conclusions" (ISA 200.17). Therefore, according to ISA 200.A32, relevant and reliable evidence is needed. As 'reasonable' implies, both comprehensibility of content and its formal documentation must be ensured. In this context, documentation on matters of risk and related professional judgment is explicitly made essential (ISA 230.A8–A9) which guarantees that the auditor's procedures and decisions can be traced at any time:

> "Documentation of the professional judgments made, where significant, serves to explain the auditor's conclusions and to reinforce the quality of the judgment."

As the Institute of Public Auditors in Germany (IDW) specifies within an examination note, these requirements apply irrespective of the technology:

> "The documentation of the data analysis in the working papers has to be done in such a way that the traceability of the audit results is possible independent of the analysis tool used and the underlying data set." (Translation of German IDW PH 9.330.3.78)

Moreover, quality assurance standards are relevant at the audit firm level, which address risks regarding to the client structure as well as the associated decisions on acceptance, continuation, or resignation of mandates. An assessment of the integrity of management for the acceptance or continuation of a mandate is required, and there must be no information that would cast doubt on this (ISQC 1.26). In Germany, it is specified that liability risks or risks of loss of reputation must be explicitly considered (IDW QS 1.72), including aspects such as aggressive accounting practices, for which explicit individual drivers must be identifiable (IDW QS 1.74).

While audit firms tend to emphasize the performance of their technology in their external communications, professional associations and regulators take a more critical role. The *American Institute of Certified Public Accountants and Chartered Professional Accountants of Canada* (2020) directly question the ability of unexplained approaches being considered as appropriate audit evidence:

> "If the auditor cannot explain or evaluate the results from an AI audit tool, can they conclude that they have obtained sufficient, appropriate audit evidence from the AI audit tool to form an opinion?"

In a similar vein, the *Canadian Public Accountability Board* (2021) refers explicitly to an "explainability risk", which must be considered when implementing advanced technologies:

> "Firms should consider how they will respond to the explainability risk, including how they will test and document (i.e., explain) whether the ATT are achieving their intended purposes."

The professional literature also shows similar challenges across jurisdictions which sound more restrained than the audit firms' communications. E.g., articles from the CPA Journal highlight both, the potential of machine learning based approaches but as well increased documentation requirements, for which it has not yet been conclusively clarified what they must include (*Dickey et al.*, 2019):

> "Finally, although machine learning has great potential, its models are still currently limited by many factors, including […] human understanding and judgment."
> "Rather than simply documenting why certain procedures were performed and explaining why samples were representative of total populations, auditors will need to document their evaluation and application of the data analysis."

Examples from the German WPg underline these challenges and requirements. *Marten and Harder* (2019) share the opinion, that traceability still is a major obstacle for appropriate documentation:

> „One difficulty, however, is the traceability of the evaluations by third parties, since the data and selected parameters are often not documented." (Translated)

Therefore, outputs must be comprehensible and be put into professional context before an actual implementation into decision making processes can be achieved (*Thomas et al.*, 2021):

> "Only technically relevant results on which an auditor can rely in terms of content and whose origin is plausibly comprehensible to him can meaningfully support him in interpreting the data and ultimately in his decision-making." (Translated)

Otherwise, auditors could run the risk of following non-transparent and misleading outputs (*Rapp & Pampel*, 2021):

> "Since the use of AI can lead to audit procedures becoming a kind of black box and results or information ex post no longer being (fully) comprehensible by the auditor in case of doubt, there is a risk – even when used as a decision support tool – that auditors

– consciously or unconsciously – rely on precisely this information that is not (fully) comprehensible.” (Translated)

In addition to the legal perspective, organizational circumstances of audit firms must be taken into account. As the quality standards stipulate, decisions on the acceptance, continuation or resignation of mandates must be carefully considered and incorporated into client portfolio risk management. According to *Johnstone and Bedard* (2004), the risk of accounting fraud is more important than insolvency risks regarding potential liability risks in terms of client portfolio management. Inconsistent results suggest, first, that audit firms generally tend to reject risky engagements rather than actively respond to risk with higher fees (*Johnstone*, 2000), and second, that in case of continuation of a mandate and a higher risk, e.g., for earnings management, higher fees are first enforced as long as the risk is acceptable. Otherwise the mandate is resigned (*Krishnan et al.*, 2013). This risk avoidance can be moderated through expertise. Provided that specialists are available to respond appropriately to the risks, audit firms are more willing to perform audits even under increased risk (*Johnstone & Bedard*, 2003). However, in the extreme case, this development is threatened by market failure due to adverse selection if risk exceeds an acceptable level (*Akerlof*, 1970). That this is not just a theoretical threat is illustrated by the recent case of the Adler Group in Germany, which is no longer able to find an auditor after BaFin announced detected errors (*Bender et al.*, 2022).

Overall, the legal requirements in the context of an individual audit engagement appear to be uniform. The central area of application of the models presented is risk identification and assessment. The various standards require a high degree of traceability here. It must be clearly documented whether and explicitly which risks have been identified, and how these have been addressed by subsequent audit procedures. The same requirements for interpretability arise from quality standards at the more abstract level of the audit firm, primarily to support fundamental decisions on the acceptance or continuation of mandates. From a client portfolio management’s perspective, a general fraud risk prediction can offer a first insight into the overall risk structure. Interpretable predictions on a global level, i.e. identification of driving factors over all firms considered by the model, do not provide sufficient insights for none of the use cases since they only describe which variables contribute to the model’s predictions in general but neglect the clients’ individual risk profile. In order to address risk properly by deploying specialists in a targeted manner, explicit starting points, and thus locally interpretable predictions, are required.

3.2 Enforcement

Despite all regulatory adjustments and advanced techniques, audited statements may still contain errors or even deliberate manipulations. Therefore, unqualified audit opinions should not provide false assurance, as there may be a residual risk of manipulation (*ESMA*, 2020). To analyze the conditions for enforcement, fundamental European regulations are used and examined based on exemplary implementation in Germany. To this end, Art. 24 of the Directive 2004/109/EC delegates the competence to carry out financial reporting enforcement to national authorities, e.g., the German *Federal Financial Supervisory Authority* (*BaFin*).

ESMA provides guidelines for the enforcement design, so that a minimum of generalizability is given. The guideline’s core combines risk-oriented and random selection of firms to be audited (*ESMA*, 2020). The random selection ensures that each company is audited at least once within a certain period of time. In German, the risk-oriented selection distinguishes between concrete and abstract risk. Concrete risk presupposes specific indications of possible misstatements and results in ad hoc examinations. The abstract risk-based selection is more general and is intended to ensure that risky companies, albeit without explicit indications, are audited with a higher probability (*FREP*, 2018). Additionally, *BaFin* only discloses the use of IT-supported market monitoring and largely automated media analysis for risk assessment. Due to confidentiality obligations, the technical design's performance or interpretability remains unclear (*Hanenberg & Kostjutschenkow*, 2021).

Although BaFin does not communicate detailed information about its own systems, it is possible to observe the requirements that BaFin places on risk monitoring systems used by the companies it supervises. It is clear from this, BaFin addresses the lack of explainability of models as published in the BaFin Journal (*Fahrenwaldt & Nohl*, 2022):

> "For BaFin, the Bundesbank and the consultation participants, a key criterion for the successful use of machine learning is the explainability of ML methods." (Translated)

One approach which is explicitly recognized for systems of companies supervised by BaFin is interpretable machine learning (*Federal Financial Supervisory Authority*, 2022):

> "Therefore, XAI should be a part of the validation before the model is put into operation and during operation. XAI could also be used to identify the essential variables and thus construct an almost causal model." (Translated)

Enforcement authorities fundamentally differ from audit firms as they are not profit-oriented, do not bear any potential financial losses, and employees can only be personally prosecuted in the case of at least grossly negligent behavior. But, they are virtual trustees for the reputation, i.e., trust and efficiency of the capital market within a jurisdiction. According to *Ewert and Wagenhofer* (2019), more vigorous enforcement can lead to improved quality of financial reporting, e.g., due to reduced earnings management. But a stricter enforcement can also have a preventive effect in other areas. The areas have in common that stricter enforcement brings a higher probability of detection and is anticipated by the companies to reduce their misconduct (*Shimshack & Ward*, 2005).

From the regulatory perspective, for identifying abstract risk candidates a simple prediction without any explanation might suffice. But a high false-positive rate could also make the use impracticable if too many companies are classified as risky. In contrast, a simple risk prediction is insufficient when assessing the concrete risk for subsequent ad hoc examinations. This applies equally, taking limited resources into account, which should be allocated as targeted as possible to maintain a high reputation of a jurisdiction's capital market. For these purposes, specific indications must be identifiable on the firm level. I.e., locally interpretable predictions are required for plausibility checks by human experts or offering starting points for further investigations.

3.3 Investors

Unlike auditors and enforcement authorities, investors operate in an environment that does not legally expect them to identify accounting manipulations. Therefore, the focus is more on economic aspects than legal framework conditions.

Research on non-professional investors shows that awareness and inclusion of the fraud risk assessment are associated with better overall returns (*Brazel et al.*, 2015). Overall returns can, firstly, be reduced by direct losses due to investments in fraudulent firms, and secondly, by foregone profits due to investments that have not been made because of false-positive fraud predictions (*Beneish & Vorst*, 2022). To evaluate fraud risk, one approach assesses the credibility of management's disclosure by checking the disclosure's inherent plausibility (*Mercer*, 2004). In contrast to non-professional investors, institutional investors have more resources and expertise to assess these risks and take a significantly stronger position towards a company. When institutional investors are distracted, this might, e.g., lead to increased governance risks (*Liu et al.*, 2020). Therefore, especially institutional investors should intensively monitor their current and potential investments.

Instead of avoiding risks, potentially fraudulent firms can be identified as short-selling targets. *Massa et al.* (2015) suggest, "the invisible hand of short selling" can prevent risk due to anticipation and reduced earnings management. In addition to reducing earnings management, short selling can also contribute to detecting fraud (*Fang et al.*, 2016). According to *Karpoff and Lou* (2010), the added value of short selling lies in an earlier detection and price correction closer to the fundamental value.

Overall, demand is more heterogeneous. For some non-professional investors, simple fraud risk predictions might be sufficient. This is in line with *Beneish and Vorst*'s (2022) findings, suggesting that some models might be appropriate for cost-efficient risk assessment under very narrow assumptions. Further requirements can be concluded especially for institutional investors. Locally interpretable predictions can be used for three purposes. First in terms of risk management, more detailed explanations offer additional guidance that helps to differentiate between fraud and non-fraud cases. The statement's plausibility can be better evaluated by guided human expertise and thus improve investors' decisions. Second, short sellers can improve their identification of potential targets, if locally interpretable predictions do not provide a plausible explanation other than for potential fraud. Third, only locally interpretable predictions allow detailed investigations on how models make accurate predictions. It is essential to rule out the possibility, that, e.g., high revenues or profit figures per se lead to a fraud prediction, as fraudulent firms generally tend to overstate these figures.

4. Human-Machine Interaction: Impact of Accountability and Domain Expertise on Trust and Implementability

As machine learning-based accounting fraud detection is affected by technical and legal conditions, the key success factor is an effective human-machine interaction. As behavioral components are crucial, the demand for interpretable predictions therefore interacts with accountability, expertise, and trust in this highly regulated setting.

First, trust in the underlying functionality is a necessary condition for the use of any technical device. As research suggests, auditors tend to rely more on humans than on machine learning models in case of contradictory information (*Commerford et al.*, 2022). This indicates that existing models, taken in isolation, are rather unsuitable for actual deployment. Therefore, ways must be adapted to increase trust. In this vein, *Glikson and Woolley* (2020) argue that transparency and reliability interconnect. Increased transparency can contribute in two ways: in the evaluation of a model and in its actual application.

Alongside traditional prediction performance measures, post hoc analyses of interpretable machine learning can significantly supplement a model's evaluation. Accuracy measures provide information about whether the model correctly predicts test data. However, models can be biased and might not adequately adapt to future changes or unknown cases. Therefore, it is of key interest to examine the models' inherent mechanisms. Evaluating these mechanisms especially enables an examination of possible biases. E.g., a model which correctly identifies fraudulent firms but is characterized by a high false-positive rate requires further investigations. It must be ruled out that false-positives are simply driven, for example, by high sales or high-profit levels but are not able to target manipulated accounts. If investment strategies otherwise falsely exclude these profitable and growth companies on a large scale, i.e., forgone profits, it would make these models inapplicable.

For this detailed examination variables' weights in a global scope, as estimators of a regression, would only offer first insights how a model overall might work. *Bao et al.* (2020), e.g., incorporate feature importance in supplementary analyses to compare the highest scores with most frequently manipulated accounts. However, this approach cannot ensure that the corresponding variable also drove the prediction of a given manipulation. As fraud cases are more complex and could cover opposing effects of different types of manipulations, the expressive power of global explanations is significantly limited. In contrast, approaches offering locally interpretable predictions allow for each individual company to understand why a specific fraud risk was stated. The identification of contributing features for individual predictions can subsequently be compared to actual manipulated accounts. If a model succeeds in indicating manipulated accounts, this will offer supportive indication for a model's quality in detecting fraud and thereby increase trust in a model even before the operative application. Conversely, any significant biases identified could rightly point to a lack of fit of a model which could justifiably entail a lower level of trust or even warn against the use of inappropriate models.

In addition to the previously discussed complementary dimension of evaluation, transparency of systems can further enhance trust during the operative decision making processes (*Mercado et al.*, 2016). Local explanations have a higher degree of transparency, since not only global mechanisms are considered, but explanations at the level of the individual company are made possible. Thus, users can observe, at least in a simplified way, how a model arrived at a certain prediction outcome.

Second, primary users operate in a highly regulated setting when assessing fraud risk. Even if sufficient technical trust is given, it is questionable whether legal certainty exists or whether the lack of it inhibits models' use. According to *Bedué and Fritzsche* (2022) public or legal accountability raises concerns in AI applications. This is likely to be the case in highly regulated environments. As *Beneish and Vorst* (2022) indicate, litigation risks could, e.g., arise in case of positive fraud predictions in previous years if the fraud has not been discovered immediately.

Regarding legal certainty for auditors, audit standards are deliberately formulated in a technology-independent manner to cover a wide range of future developments. However, requirements for traceability and documentation are also imposed on all applications, irrespective of the technology. The extent to which interpretable machine learning can ultimately guarantee legal certainty cannot be conclusively clarified. It is conceivable that the standards could be concretized to explicitly cover these kinds of technologies to ensure legal certainty. Otherwise, it remains to be seen how courts or professional regulators would decide in proceedings. Technically it can be stated, that globally interpretable explanations, e.g. feature importance, are rather unsuitable for legal justifications, especially for further audit planning to address risks, because they try to explain the effects of the model as a whole, but neglect the particularities and possibly diverging effects with respect to individual observations. However, local interpretable predictions explicitly highlight why a certain prediction has been made. Clear documentation on the prediction, its driving factors, and how risks are addressed, e.g., by further audit procedures, might reduce the risk of personal liability (*Krieger et al.*, 2021).

Third, the users' expertise is indispensable. Technical expertise in machine learning, along with transparency, and liability, is a driver of trust in applications (*Bedué & Fritzsche*, 2022). The domain expertise seems to have more complex effects. *Bayer et al.* (2021) show that domain expertise alone has a negative effect on the intention to trust. In contrast, if the level of expertise is high, additional explanations can increase the trusting intention. In this context, *Dikmen and Burns* (2022) point out the danger that additional explanations could convey false certainty and that predictions are followed uncritically. To mitigate the risk, they point out the need for accurate interpretation of the explanations, which requires professional domain knowledge. *Shin* (2021) shows that explanations are helpful for trust building but that a comprehensible possible causability can additionally increase emotional confidence in the application. This assessment is explicitly part of the main task of applying accounting fraud detection models: It is not a matter of blindly following predictions but of questioning them and their drivers, checking their plausibility, and, if necessary, initiating further investigations. This is particularly necessary when developments could not yet be learned from models. Since both, data on financial statements and fraudulent firms, are available with a time lag, time-lagged learning is unavoidable to a certain extent. Therefore, human expertise incorporating recent developments of a company's business environment is essential in addition to the system.

As illustrated and concluded in Table 1, especially in this use case of machine learning models in a highly regulated context, human-machine interaction is crucial. It is essential to properly assess the potential and the limits of a model with technical know-how to be able to place a basic level of trust in it. Appropriate domain expertise and critical consideration of possible liability risks complement the assessment of an application's abilities. As *Liu* (2022) concludes, in an adjacent application area, machine learning applications should be used as complementary guidance, which outperform humans in terms of covering large amounts of hard information to identify conspicuities. The true potential is raised when humans with expertise incorporate additional, partly soft, information to make the most comprehensive assessment, which neither the machine nor the human alone would have been able to do.

		Legal Conditions	Organizational Conditions →	Need for Explanation	Behavioral Implications
Audit	Engagement	Highly regulated ▪ risk assessment for planning of audit program and strategy ▪ detailed documentation of audit procedures and reasonable conclusions	Client acceptance, continuation or rejection decisions ▪ acceptable risk (covered by increased effort/higher fees) vs. unacceptable risk (rejection)	▪ local interpretability is required for risk-oriented efficient and effective audits ▪ documented audit procedures and judgments must be comprehensible	Trust ▪ necessary and can be increased by transparency Accountability ▪ personal liability can prevent deployment ▪ legally compliant use must be ensured Expertise ▪ enables appropriate critical questioning of explanations
	Audit Firm Level	Quality assurance with respect to ▪ clients structure's risk ▪ assessment of integrity ▪ liability and reputational risks	Client portfolio management ▪ adequate audit of all mandates must be ensured ▪ cluster risk, time & personnel resource constraints	▪ local interpretability required ▪ identify potential indicators for cluster risks in the overall client structure ▪ enable checks for plausibility	
Enforcement	Concrete Risk	Must be investigated ▪ specific indication of possible accounting fraud ▪ leads to ad hoc examinations	Effective selection ▪ missed fraud cases lead to reputational damages and ▪ financial losses for various market participants	▪ specific indication requires local interpretability ▪ explicit starting points for further investigations or plausibility checks	Trust ▪ necessary and can be increased by transparency Accountability ▪ justification in case of legal or technical supervision Expertise ▪ domain knowledge might improve the differentiation between fraud and non-fraud firms in the case of abstract conspicuities
	Abstract Risk	Should be investigated ▪ ad hoc or sampling examinations ▪ should be audited with a higher probability	Efficient selection ▪ limited resources must be used accurately and focused ▪ increased risk orientation to improve sample quality	▪ risk prediction might suffice ▪ interpretable predictions allow to increase the proportion of ad hoc-examinations	
Investors	Non-Professional	In general, no legal justification required	Simple investments ▪ diversification ▪ avoid accounting fraud cases vs. risk of foregone profits	▪ no explanation is needed ▪ limited expertise and resources reduce the potential for further analysis ▪ few approaches cost-efficient	Trust ▪ necessary and can be increased by transparency ▪ improved through detailed evaluation that can rule out potential biases which would lead to forgone profits Expertise ▪ enables appropriate critical questioning of explanations
	Professional	In general, no legal justification required	Portfolio risk management ▪ avoid accounting fraud cases vs. risk of foregone profits Short selling ▪ identify potential targets	▪ local interpretability required ▪ check for plausibility (avoid profit forgone) ▪ initial points to investigate (avoid or target fraud firms)	

Table 1: Framework on the Demand for Interpretable Accounting Fraud Predictions

5. Conclusion

This paper identifies factors that inhibit implementing machine learning-based accounting fraud detection models for relevant user groups. Specifically, the perspectives of auditors, enforcement institutions, and investors are considered. Factors are classified as legal or arising from organizational and operating conditions. Particular emphasis has been put on discussing further behavioral implications resulting from the highly regulated setting and outstanding requirements for professional expertise.

Concerning our first research question on legal and organizational conditions that drive users' needs for predictions' interpretability, our analysis provides reasons to assume a significant demand for more transparent models. For auditors and enforcement institutions, transparency of their procedures is a regulatory requirement. This especially intensifies the demand for auditors for interpretable approaches and associated legally compliant documentation. From a business perspective of audit firms, client portfolio management could be improved by more transparent risk assessments. From a national perspective and its enforcement, explainable predictions could enable more efficient allocations of resources. Intensified risk-oriented selections could result in a higher reputation of capital markets, thus increasing trust and overall market efficiency. Investors, on the other hand, are in general not affected by legal obligations. Their need for explanations results from other reasons. Institutional investors could, e.g., benefit from additional and transparent fraud risk assessments to reduce financial losses. However, forgone profits from avoiding investments in false-positive predicted firms might be prevented if conspicuity can be checked for plausibility with professional judgment.

Regarding the second research question, behavioral interactions between accountability, expertise, and trust must be appropriately considered. In general, technical know-how can increase trust in technology. For auditors, it must be ensured that, with appropriate documentation, they can justify conclusions from machine learning-based predictions with legal certainty. Otherwise, even though there might be sufficient trust in the technology's ability, non-compliance would still prevent its use. In contrast, domain know-how can reduce confidence in technology due to stronger questioning of its abilities compared to human expertise. Since these application cases are in a setting of high professional expertise, transparency in the form of additional explanations is required to reverse this effect.

To conclude, this paper offers a conceptual framework on conditions and behavioral requirements of users' needs for an explanation of machine learning models. Our findings can serve as a basis for future applications that focus more intensely on the human-machine interaction. Transparent predictions offer insights into which variables contributed to the prediction. These indicators can subsequently be evaluated by human experts in terms of plausibility. On the one side, this might increase trust in machine learning applications. On the other side, potential weaknesses could be identified, and areas pointed out where a critical basic attitude towards the application seems reasonable.

Findings in this paper are limited to accounting and information systems literature as well as relevant standards and legislation. In addition, we encourage research on an interview basis covering multiple perspectives. Concerning the legal level and here considered ISA, which are directly or indirectly applicable in most jurisdictions, only high-level EU legislation and exemplary German implementation of enforcement have been included. We further encourage broadening research on enforcement, particularly outside the EU.

Finally, the evaluation of models' explanations has been studied far too little up to now (*Vilone & Longo*, 2021). Due to the decisive role of reliability for actual implementation, accounting fraud detection models and their explanations need to be evaluated locally for individual predictions to determine their true potential.

References

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS. 2018.2870 052

Akerlof, G. A. (1970). The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

American Institute of Certified Public Accountants, & Chartered Professional Accountants of Canada. (2020). The Data-Driven Audit: How Automation and AI are Changing the Audit and the Role of the Auditor. Retrieved 30.08.2022, from https://www.aicpa.org/content/dam/aicpa/interestareas/frc/assuranceadvisoryservices/downloadabledocuments/the-data-driven-audit.pdf

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, 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. https://doi.org/10.1016/j.inffus.2019.12.012

Association of Certified Fraud Examiners. (2022). Occupational Fraud 2022: A Report to the Nations. Retrieved 16.08.2022, from https://acfepublic.s3.us-west-2.amazonaws.com/2022+Report+to+the+Nations.pdf

Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting Accounting Fraud in Publicly Traded U.S. Firms Using a Machine Learning Approach. *Journal of Accounting Research*, *58*(1), 199–235. https://doi.org/10.1111/1475-679X.12292

Bayer, S., Gimpel, H., & Markgraf, M. (2021). The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems*, *1-29*. https://doi.org/10.1080/12460125.2021.1958505

Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Bender, R., Fröndhoff, B., & Nagel, L.-M. (2022, August 29). Wirtschaftsprüfer verzweifelt gesucht. *Handelsblatt*, p. 30.

Beneish, M. D. (1999a). The Detection of Earnings Manipulation. *Financial Analysts Journal*, *55*(5), 24–36. https://doi.org/10.2469/faj.v55.n5.2296

Beneish, M. D. (1999b). Incentives and Penalties Related to Earnings Overstatements that Violate GAAP. *The Accounting Review*, *74*(4), 425–457.

Beneish, M. D., & Vorst, P. (2022). The Cost of Fraud Prediction Errors. *The Accounting Review*, *97*(6), 91–121. https://doi.org/10.2308/tar-2020-0068

Brazel, J. F., Jones, K. L., Thayer, J., & Warne, R. C. (2015). Understanding investor perceptions of financial statement fraud and their use of red flags: evidence from the field. *Review of Accounting Studies*, *20*(4), 1373–1406. https://doi.org/10.1007/s11142-015-9326-y

Canadian Public Accountability Board. (2021). Technology in the audit. Retrieved 24.08.2022, from https://cpab-ccrc.ca/docs/default-source/thought-leadership-publications/2021-technology-audit-en.pdf?sfvrsn=f29b51ce_14

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010a). Detecting Management Fraud in Public Companies. *Management Science*, *56*(7), 1146–1160. https://doi.org/10.1287/mnsc.1100.1174

Cecchini, M., Aytug, H., Koehler, G. J., & Pathak, P. (2010b). Making words work: Using financial text as a predictor of financial events. *Decision Support Systems*, *50*(1), 164–175. https://doi.org/10.1016/j.dss.2010.07.012

Commerford, B. P., Dennis, S. A., Joe, J. R., & Ulla, J. W. (2022). Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence. *Journal of Accounting Research*, *60*(1), 171–201. https://doi.org/10.1111/1475-679X.12407

Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, *139*, 113421. https://doi.org/10.1016/j.dss.2020.113421

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting Material Accounting Misstatements *Contemporary Accounting Research*, *28*(1), 17–82. https://doi.org/10.1111/j.1911-3846.2010.01041.x

Dickey, G., Blanke, S., & Seaton, L. (2019). Machine Learning in Auditing. *CPA Journal*, *89*(6), 16–21.

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, *162*, 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

European Securities and Markets Authority. (2020). Guidelines on enforcement of financial information. Retrieved 16.08.2022, from https://www.esma.europa.eu/sites/default/files/library/esma32-50-218_guidelines_on_enforcement_of_financial_information_de.pdf

Ewert, R., & Wagenhofer, A. (2019). Effects of Increasing Enforcement on Financial Reporting Quality and Audit Quality. *Journal of Accounting Research*, *57*(1), 121–168. https://doi.org/10.1111/1475-679X.12251

Fahrenwaldt, M., & Nohl, S. (2022). Maschinelles Lernen in Risikomodellen. *BaFin Journal*, (Februar), 14–16.

Fang, V. W., Huang, A. H., & Karpoff, J. M. (2016). Short Selling and Earnings Management: A Controlled Experiment. *The Journal of Finance*, *71*(3), 1251–1294. https://doi.org/10.1111/jofi.12369

Fanning, K. M., & Cogger, K. O. (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 7(1), 21–41. https://doi.org/10.1002/(SICI)1099-1174(199803)7:1<21::AID-ISAF138>3.0.CO;2-K

Federal Financial Supervisory Authority. (2022). Maschinelles Lernen in Risikomodellen – Charakteristika und aufsichtliche Schwerpunkte: *Antworten auf das Konsultationspapier. Retrieved* 12.08.2022, from https://www.bafin.de/SharedDocs/Downloads/DE/Konsultation/2021/dl_kon_11_21_Ergebnisse_machinelles_Lernen_Risikomodelle.pdf?__blob=publicationFile&v=1

Financial Reporting Enforcement Panel. (2018). Tätigkeitsbericht 2017.

Fotoh, L. E., & Lorentzon, J. I. (2023). Audit digitalization and its consequences on the audit expectation gap: A critical perspective. *Accounting Horizons*, 37(1), 43–69. https://doi.org/10.2308/horizons-2021-027

Gepp, A., Kumar, K., & Bhattacharya, S. (2021). Lifting the numbers game: identifying key input variables and a best-performing model to detect financial statement fraud. *Accounting & Finance*, *61*(3), 4601–4638. https://doi.org/10.1111/acfi.12742

Glancy, F. H., & Yadav, S. B. (2011). A computational model for financial reporting fraud detection. *Decision Support Systems*, *50*(3), 595–601. https://doi.org/10.1016/j.dss.2010.08.010

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

Goel, S., Gangolly, J., Faerman, S. R., & Uzuner, O. (2010). Can Linguistic Predictors Detect Fraudulent Financial Filings? *Journal of Emerging Technologies in Accounting*, *7*(1), 25–46. https://doi.org/10.2308/jeta.2010.7.1.25

Green, B. P., & Choi, J. H. (1997). Assessing the Risk of Management Fraud Through Neural Network Technology. *Auditing: A Journal of Practice & Theory*, *16*(1), 14–28.

Gunning, D. (2017). Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency* (DARPA). Retrieved 21.08.2022, from https://nsarchive.gwu.edu/sites/default/files/documents/5794867/National-Security-Archive-David-Gunning-DARPA.pdf

Hanenberg, L., & Kostjutschenkow, S. (2021). Die neue Bilanzkontrolle. *BaFin Journal*, (Dezember), 14–17.

International Auditing and Assurance Standards Board. (2022). IAASB Public Report 2021. Retrieved 30.08.2022, from https://www.ifac.org/system/files/publications/files/IAASB-Public-Report-2021-Spearheading-Change-Enhance-Confidence.pdf

Johnstone, K. M. (2000). Client-Acceptance Decisions: Simultaneous Effects of Client Business Risk, Audit Risk, Auditor Business Risk, and Risk Adaptation. *Auditing: A Journal of Practice & Theory*, *19*(1), 1–25. https://doi.org/10.2308/aud.2000.19.1.1

Johnstone, K. M., & Bedard, J. C. (2003). Risk Management in Client Acceptance Decisions. *The Accounting Review*, *78*(4), 1003–1025. https://doi.org/10.2308/accr.2003.78.4.1003

Johnstone, K. M., & Bedard, J. C. (2004). Audit Firm Portfolio Management Decisions. *Journal of Accounting Research*, *42*(4), 659–690. https://doi.org/10.1111/j.1475-679X.2004.00153.x

Karpoff, J. M., & Lou, X. (2010). Short Sellers and Financial Misconduct. *The Journal of Finance*, *65*(5), 1879–1913. https://doi.org/10.1111/j.1540-6261.2010.01597.x

Koh, H. C., & Woo, E.-S. (1998). The expectation gap in auditing. *Managerial Auditing Journal*, *13*(3), 147–154. https://doi.org/10.1108/02686909810208038

Krieger, F., Drews, P., & Velte, P. (2021). Explaining the (non-) adoption of advanced data analytics in auditing: A process theory. *International Journal of Accounting Information Systems*, *41*, 100511. https://doi.org/10.1016/j.accinf.2021.100511

Krishnan, G. V., Sun, L., Wang, Q., & Yang, R. (2013). Client Risk Management: A Pecking Order Analysis of Auditor Response to Upward Earnings Management Risk. *Auditing: A Journal of Practice & Theory*, *32*(2), 147–169. https://doi.org/10.2308/ajpt-50372

Langenbucher, K., Leuz, C., Krahnen, J. P., & Pelizzon, L. (2020). What are the wider supervisory implications of the Wirecard case? *SAFE White Paper, No. 74*. Leibniz Institute for Financial Research SAFE. https://doi.org/10.2861/936827

Liu, C., Low, A., Masulis, R. W., & Le Zhang (2020). Monitoring the Monitor: Distracted Institutional Investors and Board Governance. *The Review of Financial Studies*, *33*(10), 4489–4531. https://doi.org/10.1093/rfs/hhaa014

Liu, M. (2022). Assessing Human Information Processing in Lending Decisions: A Machine Learning Approach. *Journal of Accounting Research*, *60*(2), 607–651. https://doi.org/10.1111/1475-679X.12427

Marten, K.-U., & Harder, R. (2019). Digitalisierung in der Abschlussprüfung. *Die Wirtschaftsprüfung*, *72*(14), 761–769.

Massa, M., Zhang, B., & Zhang, H. (2015). The Invisible Hand of Short Selling: Does Short Selling Discipline Earnings Management? *The Review of Financial Studies*, *28*(6), 1701–1736. https://doi.org/10.1093/rfs/hhu147

McCrum, D. (2020). Wirecard: the timeline. Retrieved 30.08.2022, from https://www.ft.com/content/284fb1ad-ddc0-45df-a075-0709b36868db

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent Agent Transparency in Human-Agent Teaming for Multi-UxV Management. *Human Factors*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206

Mercer, M. (2004). How Do Investors Assess the Credibility of Management Disclosures? *Accounting Horizons*, *18*(3), 185–196. https://doi.org/10.2308/acch.2004.18.3.185

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Moberg, D. J. (1997). On Employee Vice. *Business Ethics Quarterly*, *7*(4), 41–60. https://doi.org/10.2307/3857208

Molnar, C. (2022). Interpretable Machine Learning (2nd ed.). Leanpub.

Perols, J. L., Bowen, R. M., Zimmermann, C., & Samba, B. (2017). Finding Needles in a Haystack: Using Data Analytics to Improve Fraud Prediction. *The Accounting Review*, *92*(2), 221–245. https://doi.org/10.2308/accr-51562

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *International Journal of Machine Learning Technology*, *2*(1), 37–63. https://doi.org/10.48550/arXiv.2010.16061

Purda, L., & Skillicorn, D. (2015). Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection. *Contemporary Accounting Research*, *32*(3), 1193–1223. https://doi.org/10.1111/1911-3846.12089

Rapp, D. J., & Pampel, J. (2021). Zur Akzeptanz künstlicher Intelligenz in der Abschlussprüfung. *Die Wirtschaftsprüfung*, *74*(11), 678–689.

Ruhnke, K., & Schmidt, M. (2014). The audit expectation gap: existence, causes, and the impact of changes. *Accounting and Business Research*, *44*(5), 572–601. https://doi.org/10.1080/00014788.2014.929519

Samek, W., & Müller, K.-R. (2019). Towards Explainable Artificial Intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller, R. Goebel, Y. Tanaka, W. Wahlster, & J. Siekmann (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 5–22). Springer.

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*, 210–229. https://doi.org/10.1147/rd.33.0210

Sellhorn, T. (2020). Machine Learning und empirische Rechnungslegungsforschung: Einige Erkenntnisse und offene Fragen. *Schmalenbachs Zeitschrift Für Betriebswirtschaftliche Forschung*, *72*(1), 49–69. https://doi.org/10.1007/s41471-020-00086-1

Shimshack, J. P., & Ward, M. B. (2005). Regulator reputation, enforcement, and environmental compliance. *Journal of Environmental Economics and Management*, *50*(3), 519–540. https://doi.org/10.1016/j.jeem.2005.02.002

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Storbeck, O. (2022). German regulator says Adler overstated 2019 accounts by up to €233mn. Retrieved 30.08.2022, from https://www.ft.com/content/d265bf94-5f84-4df0-883b-c1320d12f65d

Thomas, O., Bruckner, A., Leimkühler, M., Remark, F., & Thomas, K. (2021). Konzeption, Implementierung und Einführung von KI-Systemen in der Wirtschaftsprüfung. *Die Wirtschaftsprüfung*, *70*(9), 551–561.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106. https://doi.org/10.1016/j.inffus.2021.05.009

Zahn, M. von, *Feuerriegel, S., & Kuehl, N.* (2022). The Cost of Fairness in AI: Evidence from E-Commerce. *Business & Information Systems Engineering*, *64*(3), 335–348. https://doi.org/10.1007/s12599-021-00716-w

Zhang, C., Cho, S., & Vasarhelyi, M. (2022). Explainable Artificial Intelligence (XAI) in auditing. *International Journal of Accounting Information Systems*, *46*, 100572. https://doi.org/10.1016/j.accinf.2022.100572

Leonhard J. Lösse, M.Sc., ist wissenschaftlicher Mitarbeiter am Lehrstuhl für Betriebswirtschaftslehre, insbes. Controlling und Accounting an der Heinrich-Heine-Universität Düsseldorf.

Anschrift: Heinrich-Heine-Universität Düsseldorf, Lehrstuhl für Betriebswirtschaftslehre, insbes. Controlling und Accounting, Universitätsstr. 1, 40225 Düsseldorf, Deutschland, Tel.: +49 (0)211/81–15391, E-Mail: Leonhard.Loesse@hhu.de

Barbara E. Weißenberger, Prof. Dr., ist Universitätsprofessorin am Lehrstuhl für Betriebswirtschaftslehre, insbes. Controlling und Accounting an der Heinrich-Heine-Universität Düsseldorf sowie Affiliate Professor of Accounting an der Bucerius Law School, Hamburg.

Anschrift: Heinrich-Heine-Universität Düsseldorf, Lehrstuhl für Betriebswirtschaftslehre, insbes. Controlling und Accounting, Universitätsstr. 1, 40225 Düsseldorf, Deutschland, Tel.: +49 (0)211/81–10190, E-Mail: Barbara.Weissenberger@hhu.de