von Zahn, Moritz; Zacharias, Jan; Lowin, Maximilian; Chen, Johannes; Hinz, Oliver

Navigating AI conformity: A design framework to assess fairness, explainability, and performance

Electronic Markets

Provided in Cooperation with:
Springer Nature

Suggested Citation: von Zahn, Moritz; Zacharias, Jan; Lowin, Maximilian; Chen, Johannes; Hinz, Oliver (2025) : Navigating AI conformity: A design framework to assess fairness, explainability, and performance, Electronic Markets, ISSN 1422-8890, Springer, Berlin, Heidelberg, Vol. 35, Iss. 1, https://doi.org/10.1007/s12525-025-00770-2

Electronic Markets (2025) 35:24
https://doi.org/10.1007/s12525-025-00770-2

RESEARCH PAPER

Navigating AI conformity: A design framework to assess fairness, explainability, and performance

Moritz von Zahn[1] · **Jan Zacharias**[1] · **Maximilian Lowin**[1] · **Johannes Chen**[1] · **Oliver Hinz**[1]

Received: 25 May 2024 / Accepted: 12 February 2025

Abstract
Artificial intelligence (AI) systems create value but can pose substantial risks, particularly due to their black-box nature and potential bias towards certain individuals. In response, recent legal initiatives require organizations to ensure their AI systems conform to overarching principles such as explainability and fairness. However, conducting such conformity assessments poses significant challenges for organizations, including a lack of skilled experts and ambiguous guidelines. In this paper, the authors help organizations by providing a design framework for assessing the conformity of AI systems. Specifically, building upon design science research, the authors conduct expert interviews, derive design requirements and principles, instantiate the framework in an illustrative software artifact, and evaluate it in five focus group sessions. The artifact is designed to both enable a fast, semi-automated assessment of principles such as fairness and explainability and facilitate communication between AI owners and third-party stakeholders (e.g., regulators). The authors provide researchers and practitioners with insights from interviews along with design knowledge for AI conformity assessments, which may prove particularly valuable in light of upcoming regulations such as the European Union AI Act.

Keywords Machine learning · Algorithmic fairness · Explainable AI · Certification · AI auditing · Impact assessment

JEL Classification M15 · L86 · O30

Introduction

Nowadays, organizations adopt artificial intelligence (AI) systems in various application domains, including hiring (Van den Broek et al., 2021), healthcare (Topuz et al., 2018), and credit risk assessment (Moula et al., 2017). Contemporary AI systems can reach prediction performance that exceeds human capabilities by far. As a consequence, AI systems benefit society in various ways, e.g., by advancing environmental initiatives (von Zahn et al., 2024) or by guiding the development of new medications (Fleming, 2018).

Despite the benefits of AI, organizations must be aware of the potential harms associated with its adoption. These include security breaches and data leaks (Michael et al., 2023), privacy-invasive practices (Mökander & Floridi, 2022), declining system performance leading to poorer decision-making (dos Reis et al., 2016), biased outputs that disadvantage certain demographic groups (Barocas & Selbst, 2016), and a lack of accountability for decisions made by the AI (Raji et al., 2020).

In recent years, researchers and journalists have particularly highlighted two major concerns with AI that can cause harm: bias and opacity (Angwin et al., 2016; Bauer et al., 2023; Bringas Colmenarejo et al., 2022; Brown et al., 2021; Jobin et al., 2019). Bias in AI systems can manifest as algorithmic discrimination, as researchers and journalists have revealed in numerous cases where AI systems have yielded disparate outcomes based on individuals' sociodemographic characteristics (see, e.g., Angwin et al., 2016; Cho, 2021; Fu et al., 2021; Lambrecht & Tucker, 2019). Women, for example, are often systematically put at a disadvantage against men when applying for a

Responsible Editor: Ioanna Constantiou.

✉ Moritz von Zahn
vzahn@wiwi.uni-frankfurt.de

[1] Information Systems and Information Management, Goethe University, Frankfurt, Theodor-W.-Adorno Platz 4, Frankfurt Am Main, 60323 Hesse, Germany

Published online: 12 March 2025

bank loan (Fu et al., 2021) or receiving healthcare (Cho, 2021). Biased AI systems may thus harm subgroups of the population, reinforce existing inequalities, and impose legal and reputational risks on organizations. The opacity of AI systems refers to the black-box character of most state-of-the-art AI systems, such as deep neural networks (Kraus et al., 2020). Here, organizations typically face a trade-off, as state-of-the-art AI systems offer high prediction performance but are incomprehensible to humans. The opacity of AI systems implies an inability of organizations to explain critical decisions to external stakeholders, such as regulators. For example, a bank may reject a loan application as a result of a credit default prediction made by a deep neural network. Humans (e.g., auditors of the banking supervision) cannot understand why this prediction has been made, which may cause the bank to face serious legal and reputational risks (Langenbucher, 2020). In summary, while this paper is broadly concerned with the potential harms of AI, we confine our focus to two specific issues: bias and opacity. These challenges are among the most discussed and complex in the field, making them particularly illustrative of the broader risks posed by AI systems.

In order to counteract the potential harms of AI systems, especially with regard to bias and opacity, organizations will increasingly be required to perform conformity assessments (Mökander & Floridi, 2022; Thelisson & Verma, 2024). Conformity assessments determine whether the AI system conforms to particular principles, such as cybersecurity (Junklewitz et al., 2023) and—the focus of this paper—fairness (as opposed to bias) and explainability (as opposed to opacity). In recent years, various regulators and practitioners have called for conformity assessments that encompass the principles of fairness and explainability, as is the case for the Algorithmic Accountability Act in the US (117th Congress, 2022), the National New Generation Artificial Intelligence Governance Expert Committee in China (Roberts et al., 2021), or the AI Act in the European Union (EU) (European Union, 2023). For example, the EU's AI Act demands certain AI systems to undergo "conformity assessment procedures before those systems can be placed on the Union market" highlighting overarching ethical principles such as "non-discrimination and fairness" (European Union, 2023). In the US, the Algorithmic Accountability Act demands organizations to assess whether and how they can improve their AI systems with regard to "fairness, including bias and nondiscrimination" as well as "explainability" and other criteria (117th Congress, 2022). Similarly, the US Institute of Electrical and Electronics Engineers (IEEE) advocates assessing the "compliance in the development or use of artificial intelligence within organizations" (IEEE Standards Association, 2022). These examples illustrate the main reason for organizations to conduct AI conformity assessments, namely, to comply with regulation and industry-wide standards. Another reason for organizations to conduct conformity assessments is to gain a competitive advantage. In this case, organizations may self-commit to conformity assessments and communicate the results to suppliers, customers, and other stakeholders, thus signaling the use of conforming AI systems (Cihon et al., 2021; Roski et al., 2021).

While the necessity seems without question, organizations struggle with the implementation of AI conformity assessments. A major reason relates to the diverse range of definitions proposed by researchers regarding AI conformity. While this holds true for many aspects of conformity (Mökander & Floridi, 2022), it especially applies to fairness and explainability. For example, definitions of AI fairness are numbering in the double digits (c.f. Barocas et al., 2019), and some definitions are mathematically impossible to fulfill at the same time (Kleinberg et al., 2017). Needless to say, the diverse range of definitions poses a pivotal yet challenging task to organizations in selecting an appropriate definition of fairness that suits the specific context (Dolata et al., 2022). This challenge becomes even more pronounced as the chosen definition substantially influences both the impact of AI fairness on prediction performance (Corbett-Davies et al., 2017) and the financial costs incurred by the organization (von Zahn et al., 2022). Similarly, the concept of explainability in AI systems is far from straightforward, as it encompasses a diverse range of ideas and approaches (Dwivedi et al., 2023; Meske et al., 2022) and robust measures to objectively assess transparency still need to be developed (Fresz et al., 2024). The uncertainty surrounding definitions of fairness and explainability is further compounded by broad and often generic regulations. It is mostly unclear how fairness and explainability as overarching principles translate to specific technical requirements within AI systems (Mökander & Floridi, 2022; Veale & Zuiderveen Borgesius, 2021). Adding to the difficulty, AI regulations are typically not static and may be updated periodically. For instance, as stated in Article 42, Sects. 5 and 6 of the AI Act, the European Commission can adapt the requirements for conformity assessments through delegated acts (European Union, 2023). Put differently, AI systems shall conform to overarching principles, but how exactly and to what extent needs still to be determined. Consequently, organizations may be aware of the need for AI conformity assessments to comply with regulations but still lack guidance on the necessary steps to move forward.

Another reason why organizations struggle with conducting AI conformity assessments is the lack of skilled experts (Avin et al., 2021; Benbya et al., 2020). Manually assessing the conformity of AI systems requires expertise in the domain in which the AI system is deployed, as well as knowledge of law, algorithms for artificial intelligence, and arguably even ethics. Considering the existing challenges

organizations face in finding experts for their AI systems (Chui et al., 2022), the distinct combination of skills and knowledge necessary for conducting AI conformity assessments is expected to exacerbate the difficulty in securing qualified professionals.

In this paper, we aim to support organizations in overcoming the aforementioned challenges by providing a design framework and illustrative prototype for conducting AI conformity assessments in a (semi-)automated manner. To the best of our knowledge, we are the first to develop a framework that includes meta-requirements and design principles for systems that develop, share, and assess criteria for AI conformity, and subsequently implement this framework in an illustrative artifact that we test and refine through iterative evaluation. The framework thus lays the foundation for developing, sharing, and assessing industry-wide standards for AI conformity assessments. For this, we follow design science research: we review the literature and legal documents, conduct seven expert interviews extracting concepts and themes, develop the design framework, instantiate our framework in an illustrative running software artifact, and evaluate the artifact in five focus group sessions. Given the high relevance of fairness and explainability in regulatory frameworks and broader discourse, as well as the particular challenges organizations face in conforming to these principles, our focus is centered on these two aspects of AI conformity.

Our work makes important contributions to both theory and practice. First, we propose a design framework that contributes novel design theory on building systems for developing, sharing, and assessing AI conformity. Researchers and practitioners can leverage our framework as a foundation for implementing conformity assessments, including those for self-auditing purposes. Second, our running software artifact provides organizations and regulators with an illustrative practical implementation of our framework. In fact, organizations can directly apply our software artifact to their own datasets and models within the scope of its current implementation, providing an initial assessment that can serve as a foundation for further analysis and refinement. Third, we contribute to the timely topic of AI conformity by providing qualitative evidence on the necessity, challenges, and promises of AI conformity assessments for practitioners building and leveraging AI systems. This qualitative evidence can support practitioners in persuading management of the importance of addressing AI conformity, while also providing a structured overview of anticipated challenges and enabling proactive measures to mitigate them effectively. Last but not least, our design framework represents an initial step toward the collaborative development of industry-wide standards for systematic AI system evaluations, contributing to the literature on (collaborative) governance of AI systems (Birkstedt et al., 2023) and on the standardization of information and communication technology (Hanseth & Bygstad, 2015).

The remainder of this work is structured as follows. Section 2 provides background information, and Sect. 3 presents our design science methodology. Following that, Sect. 4 describes the design framework and our artifact in detail, including the development and results of the evaluation. Section 5 proceeds by discussing our findings and Sect. 6 concludes.

Background

AI conformity assessments

The concept of AI conformity assessments[1] forms a key part of many regulatory frameworks. For example, the European Commission's AI Act proposal defines conformity assessments as "the process of verifying whether the requirements […] of this Regulation relating to an AI system have been fulfilled" (European Union, 2023). Put differently, a conformity assessment presents the process of technically determining to which extent an AI system achieves different criteria. These assessment criteria can be of broad variety, including cybersecurity, technical robustness, environmental sustainability, human oversight and agency, explainability, and fairness (c.f. Thelisson & Verma, 2024).

AI conformity assessments are multifaceted processes that are non-trivial for organizations to implement. As outlined by prior research (Brown et al., 2021; Mökander & Floridi, 2022; Thelisson & Verma, 2024), organizations aiming to conduct AI conformity assessments must first clarify both the purpose and the context of it. The purpose may vary, including regulatory authorities ensuring legal compliance, vendors of AI systems detecting malfunctions and reputational risks, and stakeholders assessing the AI conformity of organizations before engaging with them. The context is crucial, as organizations must grasp the sociotechnical system in which the AI system operates, including intended users and the organizational setting. The AI Act recognizes the importance of context, where conformity assessments' necessity and rigor are determined by the risk the AI system poses to human users

[1] The term "conformity assessment" is closely related to the concept of algorithmic auditing (Brown et al., 2021). For example, the IEEE defines algorithmic audits as "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures" (IEEE, 2008), a definition synonymous with the concept of conformity assessments. In this paper, we adopt the term "conformity assessment" as a broader concept that includes, among other terms, "algorithmic auditing" (Brown et al., 2021) and "impact assessments" (117h Congress, 2022).

(European Union, 2023). Moreover, although organizations may primarily conduct conformity assessments for external stakeholders, there is growing support among practitioners and researchers for the implementation of internal AI assessment frameworks during the development stage (Raji et al., 2020). This would ensure that model developers prioritize not only predictive performance but also a broader range of criteria and ethical principles. However, this approach requires clarifying context-related dependencies earlier in the process, making implementation even more challenging.

Beyond the challenges of context and purpose, the operationalization of assessment criteria presents another difficulty. Prior research highlights that the practical implementation of assessment criteria remains largely unclear (Mökander & Floridi, 2022; Thelisson & Verma, 2024). For instance, while there is consensus on the importance of fairness as a key requirement for AI systems, there is little agreement on how to operationalize fairness in practice (Feuerriegel et al., 2020). These disagreements, along with a lack of coordination, have hindered the development of standardized methods for conformity assessments and left organizations without the means to assess the conformity of their AI systems.

To the best of our knowledge, we are among the first to provide qualitative evidence on the need for and potential of standardized, context-dependent conformity assessments from a practitioner's viewpoint and to develop a design framework for such assessments. We thus lay the foundation for more accessible and effective AI conformity assessments and thereby help organizations communicate the compliance of their AI systems to stakeholders. Given the wide range of contexts and criteria for AI systems, our methodology necessitates a focus on specific cases that reflect current practical challenges. In this paper, we focus on the financial context of fraud detection, a domain where AI has direct implications for human lives, such as the risk of unjustly flagging individuals as fraudulent, and may thus qualify as a "high-risk" application in regulations such as the EU's AI Act. We further focus on two central and highly relevant assessment criteria beyond predictive performance: fairness and explainability (Bauer et al., 2023; Bringas Colmenarejo et al., 2022; Brown et al., 2021; Jobin et al., 2019). We focus on these two criteria because they exemplify the broader challenges in defining and operationalizing assessment criteria for AI conformity. Both fairness and explainability are broad concepts with competing—and sometimes conflicting—definitions and methods of operationalization (Bauer et al., 2023; Feuerriegel et al., 2020). Despite these complexities, they are considered essential by policymakers (see, e.g., Thelisson & Verma, 2024) and have sparked extensive academic discussions, leading to the establishment of dedicated conferences such as the Conference on Fairness, Accountability, and Transparency (FAccT). The following two subsections will detail the criteria in focus.

Fairness in AI systems

The first assessment criterion we focus on is fairness, as researchers have increasingly emphasized the need for it in light of growing evidence of AI systems exhibiting bias against certain individuals. For example, prior research has shown that AI systems in finance may put women at a disadvantage by granting them disproportionately less credit (Fu et al., 2021) and showing them fewer advertisements for high-paying jobs (Lambrecht & Tucker, 2019). Similarly, journalists and researchers have demonstrated how AI in the criminal justice system has falsely classified black defendants as "at risk" more frequently than non-black defendants (Angwin et al., 2016). Such bias can perpetuate existing inequalities, hinder social progress, and expose organizations to legal risks.

As a remedy to bias, researchers have proposed methods to measure and promote fairness in AI systems (see Barocas et al., 2019, for an overview). Over recent years, various definitions of fairness in AI systems emerged and are typically described either at the level of individuals or the level of groups (Dolata et al., 2022; Feuerriegel et al., 2020). The former, individual fairness, relies on a concept of similarity: individuals with similar properties should receive similar outcomes (Dwork et al., 2012). However, in practice, defining a suitable measure of similarity can be challenging. The latter, group-level fairness, stipulates that outcomes associated with the AI system should be equally distributed inside and outside of a group (Dwork et al., 2012; Hardt et al., 2016). Groups are typically identified by an attribute deemed sensitive, such as race, age, or gender (Barocas & Selbst, 2016; Barocas et al., 2019).

Group-level fairness is of particular relevance both in academia (see, e.g., Feuerriegel et al., 2020) and in regulation (see, e.g., Barocas & Selbst, 2016). However, even within group-level fairness exists a broad variety of definitions to choose from (c.f. Barocas et al., 2019). The most prominent examples are statistical parity (Dwork et al., 2012), which represents independence between sensitive attributes and the distribution of predictions, and equalized odds (Hardt et al., 2016), which represents independence between sensitive attributes and the distribution of prediction errors. Crucially, some definitions of group-level fairness are competing and even mathematically impossible to fulfill at the same time (Kleinberg et al., 2017). Moreover, a variety of options arises when considering the sensitive attribute that defines the groups influencing group-level fairness. Here too exists a large variety of different sensitive attributes that depend on the (regulatory) context. For example, under the Equal Credit Opportunity Act in the US, nine different attributes

are deemed sensitive (Smith, 1977). Overall, group-level fairness presents a multifaceted concept with high regulatory relevance that requires careful consideration.

Explainability in AI systems

When adopting AI systems, organizations may nowadays encounter a trade-off between high prediction performance and model interpretability (Meske et al., 2022). State-of-the-art models, such as artificial neural networks, often exhibit the highest accuracy in complex prediction tasks which makes them particularly attractive for organizations (Kraus et al., 2020). However, these models are typically opaque, that is, they are of "black-box" character impeding the ability of human users to understand their outcomes (Meske et al., 2022). The opacity of AI systems can have considerable downsides, such as impaired user trust and restricted contestability (c.f. Rosenfeld & Richardson, 2019).

Researchers have developed methods for explainability as a remedy to opacity in AI systems (Meske et al., 2022). Typically, researchers consider feature-based explanations as state-of-the-art (Bauer et al., 2023; Hsieh et al., 2020), such as SHAP values (Lundberg & Lee, 2017). SHAP is a model-agnostic method that uses additive feature attributions to provide interpretable explanations for black-box model predictions, offering contrastive explanations on an individual level. As a consequence, human users are able to better interpret the outcome of AI predictions (Bauer et al., 2023).

The explainability of AI outcomes promises various societal and business-related benefits (Coussement & Benoit, 2021), such as increased user trust towards AI systems (Rosenfeld & Richardson, 2019) and the ability to communicate the rationale behind AI outcomes to stakeholders (Wang et al., 2022). As a consequence, recent legal initiatives demand organizations to make their data-driven decisions explainable (see, e.g., the proposed Algorithmic Accountability Act in the US, 117th Congress, 2022).

Related tools and frameworks

To the best of our knowledge, there are no well-established AI conformity assessment tools addressing practitioners' needs and promoting industry-wide standardization. However, there are early works on governance frameworks and open-source toolkits aimed at a general assessment of AI systems using questionnaires or common metrics to assess prediction performance, fairness, and other aspects. In the following, we provide an overview of these early works.

CapAI, introduced by Floridi et al. (2022), is a procedure developed to align AI systems with the conformity criteria of the EU's AI Act. It provides organizations with a structured approach to evaluating the ethical, legal, and technical robustness of AI systems in the form of a step-by-step assessment guide. Users can work through the provided checklist, answer key questions, perform the necessary analyses manually, and submit their results to regulatory authorities using information templates. However, to the best of our knowledge, the approach is developed solely from a regulatory perspective, lacks a user-centric design, and is not implemented as a software.

Another related tool is FairX (Sikder et al., 2024). FairX is an open-source benchmarking tool designed to evaluate AI models in terms of fairness, performance, and explainability. The user begins by loading either tabular, image, or custom datasets into the tool. The tool automatically preprocesses the data, trains a model based on the uploaded dataset, and computes different evaluation metrics. Notably, the tool also involves bias mitigation techniques, such as adversarial de-biasing, to counteract fairness problems. Furthermore, there are several other assessment tools that are similar to FairX (see, e.g., AI Fairness 360 by Bellamy et al. (2019), or Fairlearn by Bird et al. (2020)). The focus of these tools, however, lies primarily in implementing existing metrics and methods from the literature, without the aim of guiding organizations in deeply exploring specific aspects, creating and sharing use cases, or documenting the evolution of AI systems over time.

Our framework and software artifact for conformity assessments differs from existing approaches in several ways. Our design framework is based on the academic literature, legal documents, and expert interviews, thus addressing AI conformity assessments more comprehensively. The framework places a strong emphasis on (the development of) industry-wide standards with the ultimate goal of not just implementing metrics and methods but guiding organizations in applying them. For example, we add functionalities that promote the establishment and utilization of standards and best practices for the technical implementation of conformity assessments in different use cases. Moreover, as opposed to related approaches, our framework builds upon distinct user roles throughout the assessment process, allowing different stakeholders, such as AI owners and auditors, to perform specific tasks suited to their expertise. Additionally, the software instantiation of our framework includes a graphical user interface, making the assessment accessible to a broader range of users. The interface presents the results in a clear and interpretable manner so that not only technical experts but also management and regulators can assess the level of AI conformity.

Research methodology

Overall research design

We develop the framework for AI conformity assessments following design science research, which "creates and

evaluates artifacts intended to solve identified organizational problems" (Hevner et al., 2004). In our case, we create a design framework for organizations to conduct AI conformity assessments and, thereby, signal AI conformity to the market and adhere to upcoming regulations. We further instantiate this framework into a running software artifact and evaluate the artifact in focus group sessions.

We adopt the design research cycle proposed by Kuechler and Vaishnavi (2008) as our methodological foundation, thus following an iterative process for continuous evaluation and adaptation of the artifact. The design cycle consists of five phases: problem awareness, suggestion, development, evaluation, and conclusion. In the problem awareness phase, the researchers identify and define the problem by reviewing literature and interviewing experts. In the suggestion phase, they derive meta-requirements and formulate design principles grounded in scientific theories and expertise. The development phase involves instantiating the design framework, in the form of a software artifact, methods, models, or constructs (March & Smith, 1995). The artifact is then evaluated using focus group sessions with experts (Meth et al., 2015), laboratory experiments (Gnewuch et al., 2017), or other established evaluation methods. Finally, the project concludes, and evaluation results inform subsequent iterations if needed.

Design science research is particularly well-suited to address the challenge of how a software tool can support AI conformity assessments, as this challenge is inherently socio-technical in nature. Its unique ability to bridge technological and organizational perspectives enables it to tackle such challenges, with leading scholars even describing design science research as "essential" for addressing socio-technical questions (Abbasi et al., 2024). Furthermore, by integrating theory with practical considerations within its design cycle, design science research enables the development of artifacts that not only advance academic knowledge but also provide actionable solutions for practitioners. This dual focus is particularly valuable in the context of AI conformity assessments, where regulatory, ethical, and operational considerations must be balanced within organizational environments.

Design cycle implementation

In the following, we detail the application of the general design cycle to our specific use case. In Fig. 1, we show the five phases of the general design cycle (left) along with our specific implementation (right).

Problem awareness

In the problem awareness stage, we conduct semi-structured interviews with seven experts in the field of AI conformity (see Table 1). We carefully select experts who employ an organizational, practice-oriented perspective on AI systems. Accordingly, we choose seven experts based in Germany and Switzerland who either consult companies on aspects of AI (Experts 1, 3, 4, 5, 6) or review companies' AI systems from funding (Expert 2) or regulatory perspectives (Expert 7). Importantly, all experts, with the exception of Expert 7, work internationally, engaging with AI systems at least at the European level. We conduct all interviews online and follow a semi-structured protocol. We initially explore the relevance of AI conformity through open-ended questions, avoiding any priming. Subsequently, we delve into hypothetical scenarios to gauge expert insights on AI assessment criteria. Finally, we introduce and search for feedback on the concept of an early-version AI conformity assessment tool, ensuring it does not influence earlier responses. We provide the detailed protocol in Appendix A. For the qualitative analysis of the interview responses, we create recordings and transcripts. We subsequently investigate codes and general themes of the interview via conventional content analysis. Notably, conventional content analysis is particularly suited to study the meaning of text data when existing theory and research are limited (Hsieh & Shannon, 2005). Specifically,

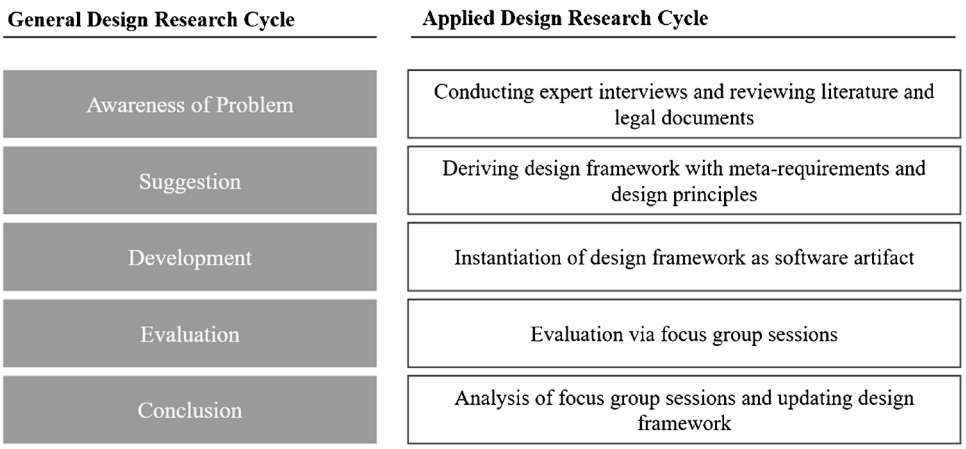


Fig. 1 Our research methodology (based on Kuechler & Vaishnavi, 2008)

Table 1 Overview of interviewed experts

Expert	Title	Organization	Years of relevant experience
1	Managing Director	AI testing center	6
2	Division Manager	Technology research fund	5
3	Head of AI	Technical inspection association	2
4	AI solution architect	Technical inspection association	7
5	Manager	Technical consultancy	9
6	Manager	Technical consultancy	12
7	Senior Risk Manager	National banking supervision	17

we follow Gioia et al. (2013) to capture first-order concepts and construct second-order as well as aggregate themes by clustering and interpretation. To do so, three researchers engage in coding transcripts and subsequently interpreting codes to higher-order concepts. They follow an iterative consensual process, i.e., two researchers conduct the coding and interpretation independently, after which they discuss their results under the guidance of a third, independent researcher to reach a consensus.

Our coding procedure can be illustrated by two related yet disparate quotations. The first quotation says that "most companies, especially small and medium-sized, simply do not deal with [developing conforming AI] at all because they don't have the capacity for it." The second quotation refers to AI auditing, specifically, a "bottleneck, for example, in the medical devices sector, where inspectors are struggling to get the job done at all, especially in the context of the shortage of skilled workers." One coder assigned both quotations to one holistic first-order concept referring to the general need for highly-skilled personnel in the context of AI conformity. By contrast, another coder assigned them to two disparate yet related first-order concepts: "Complexity and multitude of conformity aspects that require specialized experts (from, e.g., computer science and law)," which refers to the need for a variety of experts to successfully develop conforming AI, and "Highly specialized (internal and external) AI auditors are currently lacking", which refers to the scarcity of experts for conducting conformity assessments. After discussing the conflict within the research team, we agreed on these two more specific first-order concepts.

The researchers repeat the process until there is a satisfactory convergence in the interpretations of higher-level codes, as determined by the independent researcher. This approach allows us to systematically derive insights from the interviews to later form the basis for our design framework.

Suggestion

In the suggestion stage, we derive the meta-requirements and design principles of our framework based on our findings of the problem awareness. Design science research literature proposes expert interviews combined with insights from existing literature as an important source for design knowledge (Miah & Genemo, 2016) and, more specifically, the formulation of meta-requirements (Heinz et al., 2024). Thus, we translate each previously identified problem into one meta-requirement according to our understanding which high-level needs a software artifact should fulfill to address these problems. Next, we derive design principles for each meta-requirement, based on our perspective of what specific design components are needed to fulfill all meta-requirements. For the formulation of the design principles, we draw on the framework of action and materiality-oriented design principles according to Chandra et al. (2015).

Development

In the development stage, we instantiate the previously derived meta-requirements and design principles into a running software artifact, acting as a prototype for conducting conformity assessments. We implement the artifact as a web-based application based on Python for computing the metrics, Flask 2.3—a lightweight Python framework—for the web server, MySQL as a database management system, and default Bootstrap themes for the frontend design. This configuration allowed us to add and alter functionalities easily. Designing the artifact as a web application allows easy software sharing with external experts, for instance, during later focus group sessions. The artifact is publicly available and can be accessed at https://github.com/mlowin/conformity_assessment. Of course, it is also possible to run the web server in a private network without access to the Internet. This is especially valuable for companies and use cases involving confidential data.

Evaluation

In the evaluation stage, we aim to assess our design framework and software artifact on AI conformity assessments using focus groups in line with prior work in design science

research (e.g., Hirt et al., 2019; Meth et al., 2015; Zacharias et al., 2022). We follow Venable et al. (2012) to conduct a formative evaluation, that is, we aim to derive potential improvements of our artifact and, thereby, of our design framework (Hevner et al., 2004). With regard to the paradigm of our evaluation, we build upon an artificial setting, that is, we conduct both remote and on-site focus group sessions in which we evaluate our artifact in a hypothetical scenario. The main reason is that the artificial setting allows us to limit the interference of confounding variables and resource constraints. We conduct five separate focus group sessions, each corresponding to a different industry partner, yielding a total of 21 experienced practitioners as participants evaluating our artifact. Table 2 provides an overview and key descriptives of our conducted focus group sessions.

In the focus group sessions, participants evaluate the proposed software artifact by assessing the conformity of an AI system, that is, a machine learning classifier that we implemented using real-world data from fraud detection. The classifier employs gradient boosting (Chen & Guestrin, 2016) and follows standard practices for train-test splitting and hyperparameter tuning (Hastie et al., 2017). The sessions last between 45 and 90 min and consist of three parts:

Part 1. We introduce the hypothetical scenario in which participants need to assess the conformity of an AI system for fraud detection. For this, we briefly recap the problem of bias and opacity in AI as well as upcoming AI regulations to make participants cognizant of the problem. We then present the AI system to be assessed, including the development process, details on the underlying machine learning model and data, and the context of the application in fraud detection. Subsequently, we ask the participants to assess the conformity of the AI system.

Part 2. To introduce our software artifact, we demonstrate all functionalities of the software to the participants on our own screen. Following that, we share the software URL and let each participant conduct the conformity assessment separately on her own laptop. The participant uploads the prediction model and data, answers the set of questions, and ultimately explores the outcome of the assessment on the interactive dashboard. During the conformity assessment, the participant shares her thoughts with us following a think-aloud protocol (Van Someren et al., 1994).

Part 3. We commence the group discussion once all participants are finished with exploring our software. Following the recommendation of Hevner et al. (2010) and Abdel-Karim et al. (2023) to use open-ended questions in focus group sessions, we open the discussion by asking participants open-ended questions on how they experienced the use of our artifact. We jointly discuss the upsides and downsides of our artifact as well as future opportunities and risks. We then conclude the focus group session. In sessions D and E, we send a post-event survey to gather structured feedback on our meta-requirements, design principles, and their implementation within the software.

Conclusion

Finally, in the conclusion stage, we consolidate the findings from the evaluation and refine our artifact based on the newly derived design principles, thus starting the second design cycle. We discuss the strengths of our approach and limitations identified during the focus group sessions. Thereby, we position our work as a foundation for future research and practical application in the development of systematic AI conformity assessment frameworks.

Designing AI conformity assessments

Awareness of the problem

In this phase, we build upon expert interviews complemented by insights from legal documents and the scientific literature to identify both factors that motivate the introduction of conformity assessments in practice and problems that complicate the implementation of such assessments. Figure 2 presents the results of the conventional content analysis. We further provide our

Table 2 Overview of focus group sessions. Note that all participants held positions within the respective organizations that were relevant to the development or utilization of AI systems

Focus group	Nmbr. participants	Industry	Organization size	Location of session
A	3	FinTech	10–49 employees	On-site
B	2	Financial consultancy	> 250 employees	Remote
C	4	Bank	> 250 employees	Remote
D	5	Banking IT-Service	> 250 employees	Remote
E	7	Provider Manufacturer	> 250 employees	Remote

interview protocol in Appendix A and illustrative quotations to each first-order concept in Appendix B. In the following, we detail the results of the phase.

All interviewed experts agree that conformity assessments are necessary for many domains, but name varying motivations for them, with regulation and self-commitment as the prevailing ones (see Fig. 2, upper panel).

Regulation

The most common motivation refers to AI regulation, with most experts stating the EU AI Act as a main driver for conformity assessments (Experts 1, 2, 3, 4, 5, 7). Additionally, other authorities and federal agencies are mentioned as important drivers as well, as Expert 3 puts it: "I also wanted to add that self-declaration also plays a role in other approaches, including the American NIST (National Institute of Standards and Technology), which has also published a risk guide in which the declaration of conformity or self-declaration will also play a major role." Some experts also stress specific regulatory reasons for the need for conformity assessments, such as domain-specific regulation (e.g., in healthcare, Expert 2) and precedents of fines under the EU's General Data Protection Regulation (Expert 3). The stated reasons coincide both with the original content from relevant proposals for and final versions of AI regulation (117th Congress, 2022; European Union, 2023) and with the reasons brought forth by researchers studying AI regulation (e.g., Jobin et al., 2019; Roberts et al., 2021). The European Parliament, for example, approved the AI Act in early 2024, effectively making conformity assessments a "legal obligation" (Thelisson & Verma, 2024).

Self-commitment

In contrast to these reasons, some experts name motivations falling under voluntary self-commitment. Specifically, organizations may voluntarily engage in AI conformity assessments to reap benefits such as yielding deeper insights into the functioning of the scrutinized AI system, potentially informing future development (Expert 3). Furthermore, self-committed conformity assessment reports, effectively communicated to customers and stakeholders, may act as a quality signal to gain a competitive advantage (Experts 1, 3). In this context, Expert 3 explains: "And there could be a competitive advantage. And that would be if there were somehow a kind of seal label where you could say: Okay, this is a safe application. We'll buy it." Notably, previous research largely supports these assertions and consistently highlights the business opportunities in signaling AI conformity to the market (Cihon et al., 2021; Roski et al., 2021).

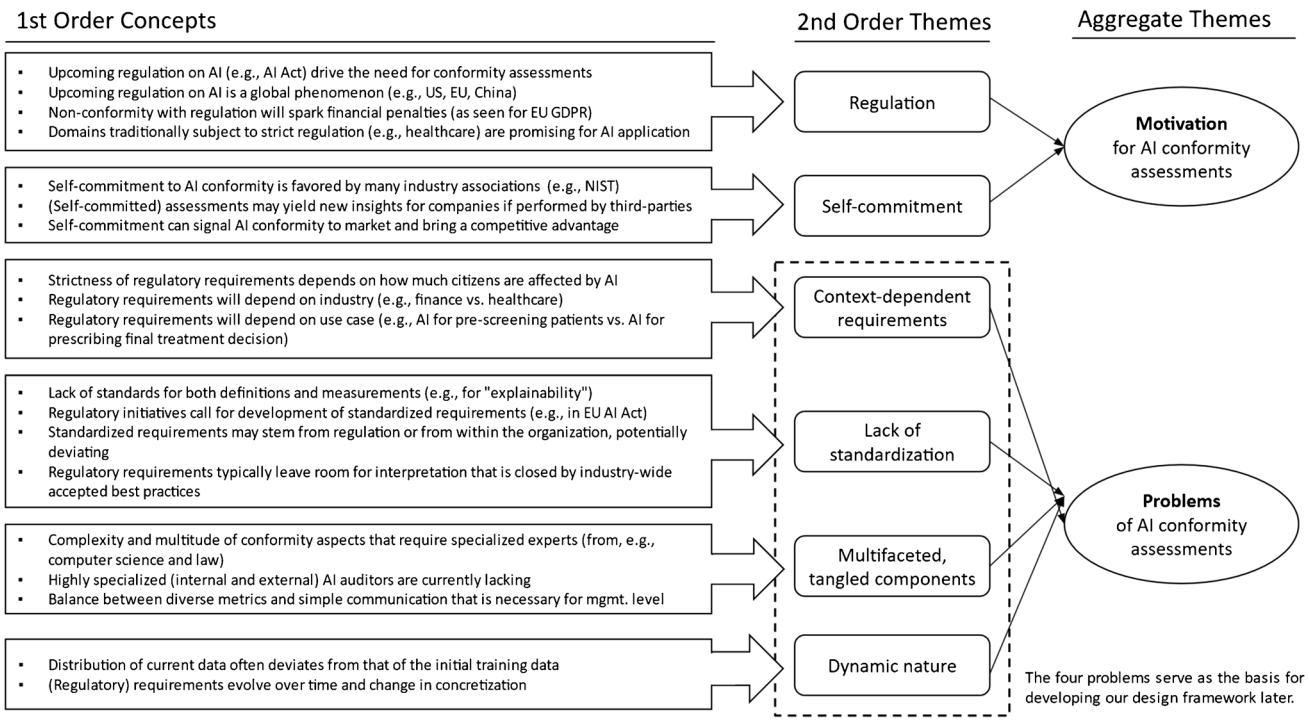


Fig. 2 Data structure of the concepts and themes of the expert interviews (following Gioia et al., 2013)

Overall, the necessity of AI conformity assessments seems undisputed among the interviewed experts.[2] Given this consensus, the interviewed experts name several problems associated with the implementation of AI conformity assessments (see Fig. 2, lower panel).

Context-dependent requirements

First, one hurdle that organizations have to overcome is the context dependency of requirements regarding AI systems (Scantamburlo et al., 2020). The configuration and strictness of conformity assessments depend on various aspects, such as the criticality of the AI system (European Union, 2023) and the affected industry and use case, as noted by Expert 2: "We have already talked about the different sectors that are affected by this. And above all, there are serious differences between individual sectors and these must of course also be considered somehow within the regulatory framework." For example, AI systems in healthcare require more stringent conformity assessments than those in less regulated fields like retail, which can affect the practical operationalization of these assessments, such as the use of relevant metrics and performance thresholds.

Lack of standardization

Second, one problem hindering the widespread implementation of AI conformity assessments is the lack of standardization with regard to current and upcoming requirements. Regulatory requirements tend to be vague and leave leeway for interpretation. For example, as Expert 5 puts it: "Up to now, the regulatory requirements and texts have often said something like: It must be explainable or it must be transparent. And that leaves a lot of room for interpretation. And this […] is often what prevents banks from actually implementing it [..]. Because it is always unclear, even if I have interpreted it for myself, does the auditor on the other side interpret it the same way?." Similarly, the strictness of regulations, such as the AI Act, is often unclear until the proposal is final (Expert 7). This absence of standardization is a crucial hurdle for organizations willing to implement conformity assessments as they lack clear guidance on how to translate vague principles such as algorithmic fairness and explainability into actionable assessment strategies.

As an additional but related insight, many experts emphasized the need for organizations to reach an industry-wide consensus, not only on conformity standards for AI systems but also on how to systematically evaluate these systems against those standards (Experts 1, 3, 4). Indeed, standards play an important role in creating consistent and comparable benchmarks that enable auditors and auditees to align their expectations and ensure fair, accurate, and reproducible evaluations, which both the auditing literature (see, e.g., Burns & Fogarty, 2010) and regulators (see, e.g., the role of standards in the EU AI Act) have repeatedly stressed. Notably, at the conclusion of the interview, when we introduced the idea of a software-supported, semi-automated assessment, the experts largely responded with enthusiasm. One expert, for instance, noted that the use of an agreed-upon software artifact could "foster digitalized communication between organizations and regulators [with regard to their AI systems]" (Expert 7) and thus help establish industry-wide standards.

Multifaceted, tangled components

Third, the complexity and multifaceted nature of conformity assessments that includes the analysis of multiple, partially conflicting ("tangled") metrics is another problem. To successfully navigate the complex landscape of conformity metrics, assessments of AI systems require highly specialized knowledge in different fields (Expert 5, 7). These days, many organizations, especially small and medium-sized ones, lack these competencies, necessitating either extensive investments in skilled experts or hiring external vendors specialized in that field (Expert 1, 3). For example, Expert 1 stresses that "most companies, especially small and medium-sized, simply don't deal with it at all because they don't have the capacity for it." Furthermore, even though technically skilled employees may be aware of the need for conformity assessments, higher-level management often lacks that awareness (Expert 1) while requiring simple communication of the level of AI conformity.

Dynamic nature

Finally, a major concern for our experts is the dynamic nature of conformity assessments, i.e., assessments not being a one-time event but an ongoing process. AI systems have to operate on constantly evolving data distributions and data points that may strongly deviate from the training data (Expert 1). Concrete and actionable requirements formulated in regulations such as the AI Act may also change periodically when new scientific insights and methods emerge, making it necessary for organizations to continuously update existing systems: "Nevertheless, the text is still not entirely

[2] One could argue that the expert sample is biased, as it includes consultants who may soon earn revenue by advising on AI conformity assessments. However, we emphasize that the literature also clearly highlights the necessity of such assessments (see, e.g., Thelisson and Verma, 2024). Additionally, in the final focus group session, we sought confirmation from industry practitioners, who unanimously affirmed the strong relevance of AI conformity assessments.

precise. […] It also changes over time as a result of new findings. And then the text of the regulation would have to be constantly adapted." (Expert 5).

Suggestion

Based on our identified problems, we suggest a design framework consisting of meta-requirements and design principles for AI conformity assessments. We derive additional design principles based on participants' feedback during the initial three focus group sessions in the evaluation phase. Figure 3 presents an overview of our design framework that we detail in the following subsections.

As AI systems are used in different industries and various contexts, decision-makers have to determine appropriate requirements with regard to conformity (Problem 1). Specifically, concepts such as fairness and explainability in AI are by no means clear-cut concepts, as we have learned from the interviewed experts and the literature (e.g., Barocas et al., 2019). As a result, the criteria to which AI systems should conform to heavily depend on the context, which leads to our first meta-requirement:

MR 1: Conformity assessments should be highly adaptable

MR 1 implies two principles. First, the artifact should support varying configurations depending on the industry and use case (DP 1) to accommodate the diverse and context-specific considerations associated with assessment criteria. For instance, the evaluation of fairness of a financial fraud

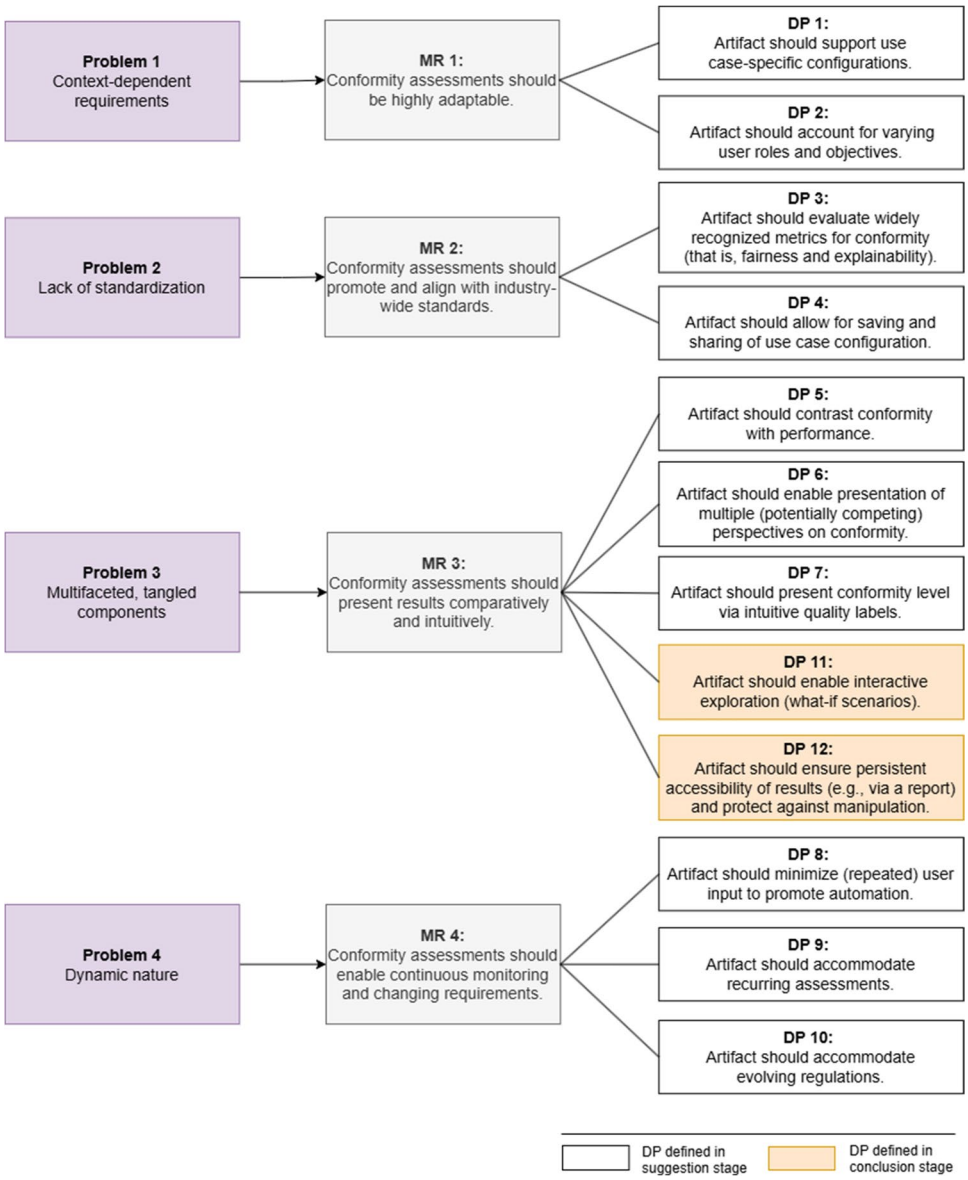


Fig. 3 Design framework consisting of problems identified in the previous stage (left), meta-requirements (middle), and design principles (right) for artifacts conducting AI conformity assessments

detection model might build upon different fairness metrics than a model allocating healthcare resources. Second, given the specialized knowledge needed to identify relevant assessment criteria and apply them in a specific use case, it is important to recognize that the AI owner, who assesses her system, is typically no expert on the requirements regarding fairness and explainability. Accordingly, conformity assessments generally require experts (i.e., users of the artifact) in different roles. The artifact should thus offer support and cater to varying objectives and user roles, i.e., varying expertise, responsibilities, and needs (DP 2).

One of the key problems associated with the widespread adoption of conformity assessments identified in our interviews is the lack of industry-wide standards (Problem 2). For given use cases, neither regulators nor organizations have yet agreed upon the set of relevant metrics and thresholds necessary to assess the conformity of AI systems. This lack of consensus forces organizations to interpret vague guidelines independently, leading to incomparable and inconsistent evaluations within industries. Therefore, with our framework, we aim to promote the establishment of industry-wide standards by providing an artifact that guides organizations in conducting assessments in their specific use cases. This leads to our second meta-requirement:

MR 2: Conformity assessment should promote and align with industry-wide standards

Industry standards for conformity assessments need to rely on well-established metrics for conformity. Our expert interviews, as well as the literature review, highlight that two high-level criteria are crucial in that context: fairness and explainability. Recently, the fields of algorithmic fairness and explainable AI are quickly developing with researchers and practitioners continuously proposing new metrics in an attempt to quantify both criteria (Dolata et al., 2022; Meske et al., 2022). Therefore, we view it as critical that industry standards are grounded in thoroughly researched and widely recognized metrics for these two criteria (DP 3). Naturally, industry standards are yet to be established over time. This requires organizations to discuss possible assessment settings within the industry and regulators, collectively working out best practices. In other words, organizations must find a consensus on which metrics and thresholds to use in certain situations. Once a suitable assessment configuration is agreed upon, all organizations within an industry should have access to it. Therefore, our conformity assessment framework should offer a central use case repository that allows for saving and sharing of successful use case configurations (DP 4).

A further problem associated with conformity assessments is the aforementioned plethora of definitions and, thus, metrics to quantify AI conformity, some of which overlap or even contradict each other (Problem 3). Having computed multiple, potentially contradicting conformity metrics, formulating a clear conclusion on the conformity level of an AI system is a non-trivial task. However, AI owners and other stakeholders need to quickly grasp and report the conformity of their AI systems with regard to the criteria of interest. This leads to our third meta-requirement:

MR 3: Conformity assessments should present results comparatively and intuitively

For example, assessment criteria such as fairness typically involve a trade-off to prediction performance (see, e.g., Corbett-Davies et al., 2017; von Zahn et al., 2022) and, hence, AI owners should be able to compare their systems with regard to conformity vs. performance (DP 5). Similarly, assessment criteria may also be in conflict with one another (Kleinberg et al., 2017) and still be equally relevant to the use case at hand. Considering this dynamic nature, an artifact for conformity assessments should enable the simultaneous presentation of potentially conflicting metrics (DP 6), ensuring a comprehensive evaluation approach. Moreover, MR 3 also implies an intuitive presentation of the output of the assessment to be straightforward to interpret even for non-technical stakeholders such as certain regulators and senior management. Inspired by long-established certification systems in the sustainability domain (Matus & Veale, 2022), we propose that the artifact should convey conformity results via intuitive quality labels (DP 7).

Additionally, conformity assessments are not simply a static prerequisite to deploying AI systems in (high-risk) domains but are of a dynamic nature (Problem 4). The main reasons are that AI systems are frequently retrained (Wilson et al., 2021), data and concepts can shift (as stated by the interviewed experts and, e.g., dos Reis et al. (2016)), and that regulations naturally evolve over time. Accordingly, we formulate a fourth meta-requirement:

MR 4: Conformity assessment should enable continuous monitoring and changing requirements

The dynamic nature of AI necessitates continuous monitoring of both the regulatory requirements and the conformity metrics of the AI systems. To decrease repetitive overhead, it is essential for conformity assessments to minimize manual steps, such as reducing the reliance on user input to promote automation (DP 8). Moreover, artifacts should be tailored to perform recurring assessments (DP 9). For example, over time, fraudsters understand the mechanics of fraud detection models and adapt their behavior accordingly (Abdallah et al., 2016). As a result, metrics reflecting conformity or performance may deteriorate, necessitating the presentation of these metrics

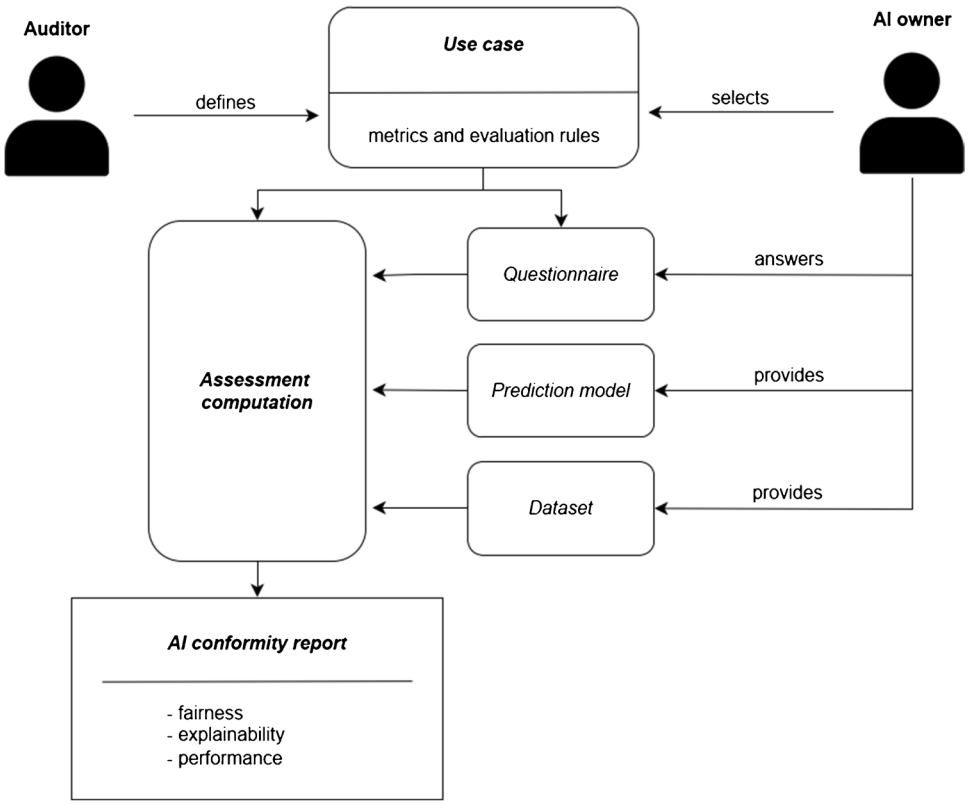


Fig. 4 Workflow of our artifact for AI conformity assessments

in a time-series format to allow exploration for human users. Finally, AI conformity assessments need to accommodate evolving regulations (DP 10). For instance, the modular integration of various metrics into an AI assessment tool can account for such industry-specific and developing requirements.

Development

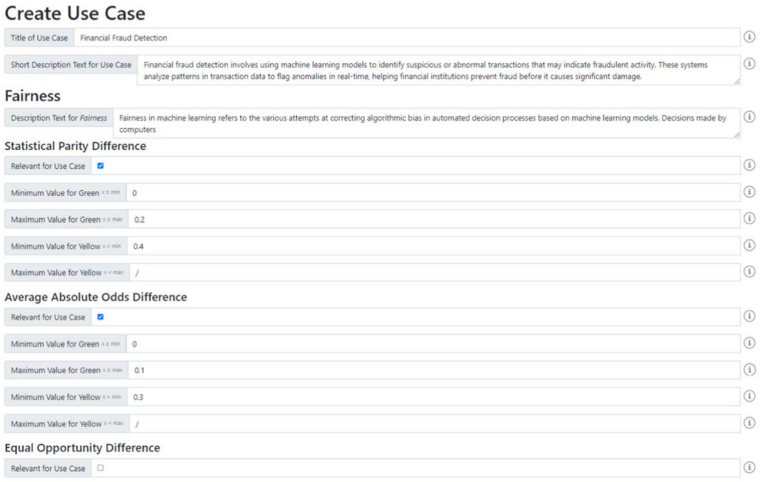
Overview of artifact

To demonstrate how organizations might translate our design framework into a tool for conformity assessments and gain a tangible prototype, we instantiate an illustrative software artifact. Our software semi-automatically computes conformity assessments based on user inputs (see Fig. 4). A company can either use the software artifact as a service or host the software on its own servers to ensure data sovereignty.

We differ between two distinct user roles in our artifact for conformity assessments: the auditor, who specifies the exact criteria for conformity in the given context, and the AI owner, who is responsible for the AI system at hand. The auditor provides use case definitions by describing the context, selecting appropriate metrics and questions from a pre-defined list, possibly defining custom measures, and setting up evaluation rules for mapping the results on quality labels represented by a traffic light system (green, yellow, red), as shown in Fig. 5a. To ensure that our tool allows for saving and sharing of use case configuration, we integrate a use case repository. This feature allows auditors to store configurations of successfully applied use cases for future reuse. The repository is designed to foster industry-wide collaboration, allowing other organizations to select and adapt stored configurations, thereby encouraging the development of best practices across a sector.

The AI owners evaluate a specific AI system by initiating the assessment with key inputs. They begin by selecting the appropriate use case and then providing the required inputs, namely (i) the questionnaire, (ii) the prediction model, and (iii) the data.[3] Once the AI owners

[3] In the initial version of our prototype, AI owners upload the prediction model as a serialized Python object inheriting the sklearn base ClassifierMixin class (specifically, an instance of the XGBClassifier). Additionally, they upload the dataset as a csv file. While this basic implementation facilitates a seamless demonstration of the use case, the subsequent design cycle's artifact can incorporate more sophisticated methods for integrating datasets and models. This may involve leveraging MLOps pipelines or integrating with AI cloud providers and APIs.



a) View on the use case selection screen from the auditor perspective. The auditor creates the use case at hand, specifies relevant metrics and thresholds, and may provide additional information for other users and recipients of the assessment report.

b) High-level view on outcome of conformity assessment. By clicking on one of the three high-level criteria, the user can deep-dive the details of the corresponding outcome.

c) View on outcome of assessment with regard to fairness. By clicking on one of the metrics for assessing fairness, the user can further deep-dive the details.

d) View on outcome with regard to the fairness metric of statistical parity difference. On the left, the user may assess the definition and meaning of the metric. On the right, the user may explore the historical values of the metric at hand for the given AI system.

have provided the inputs, the artifact can automatically generate the AI conformity report by computing all relevant metrics and evaluating the thresholds for quality labels based on the pre-selected use case. On the dashboard that follows, AI owners can explore a broad range of information on the outcomes of the conformity assessment in an interactive manner. Specifically, upon entering the dashboard, the artifact provides a concise summary denoted by a traffic light indicator, effectively conveying the degree of fairness, explainability, and performance (see Fig. 5b). The user may either be content with reviewing the overall assessment or decide to dive deeper by clicking on one of the three criteria (e.g., on fairness, as shown in Fig. 5c). The user can then investigate the details of the assessment with regard to the respective criterion, that is, explore the individual metrics contributing to the overall assessment. For an individual metric, the user can assess a definition and interpretation for the use case at hand as well as explore its values over time (e.g., for statistical parity difference, as shown in Fig. 5d).

Table 3 Overview of metrics implemented in artifact

Fairness	Statistical parity difference	Dwork et al., 2012
	Equalized odds difference	Hardt et al., 2016
Explainability	Questionnaire on explainability	Inspired by Mitchell et al., 2019 and Gebru et al., 2021
	Stability of global explanations	Inspired by Hsieh et al., 2020
Performance	Accuracy	Sokolova et al., 2006
	Balanced accuracy	Sokolova et al., 2006
	Precision	Sokolova et al., 2006
	Recall	Sokolova et al., 2006
	F1-score	Sokolova et al., 2006
	Data drift	Inspired by Reis et al., 2016

For the subsequent evaluation of our artifact, we focus on the use case of fraud detection in finance. AI for fraud detection is well-suited to explore aspects of AI conformity for several reasons. First, AI for fraud detection is well established (Abdallah et al., 2016) and many organizations rely on it (among others, the ones we collaborate with in our evaluation). Second, the domain of fraud detection is legally pertinent, as the outcomes of AI-driven fraud detection—where transactions may be autonomously flagged and blocked—have a direct impact on individuals' ability to execute financial operations. As a consequence, such scenarios are likely to attract regulatory scrutiny. For instance, under certain conditions, the EU AI Act could designate these AI applications as "high-risk," subjecting them to more rigorous conformity standards (European Union, 2023). Finally, fraud detection algorithms frequently incorporate sensitive attributes such as gender (Deepak & Abraham, 2021) and face the challenge of high-class imbalance, with legitimate transactions vastly outnumbering fraudulent ones (Abdallah et al., 2016). This complexity heightens the potential for fairness issues (Barocas & Selbst, 2016), necessitating a thorough evaluation to ensure conformity.

For our use case, we consider a hypothetical company that uses AI for binary classification, that is, classifies transactions as either fraudulent or legitimate using a machine learning model. The model is trained on more than 150,000 historical transactions spanning 12 months[4] and relies on a gradient-boosted forest (Chen & Guestrin, 2016) to predict fraud. Similar to cases from the literature (e.g., Deepak & Abraham, 2021; von Zahn et al., 2022), the company suspects systematically disparate predictions for transactions of men vs. women, potentially resulting in legal and reputational risks as outlined previously. Moreover, the firm leverages feature-based explanations for individual predictions (an increasingly adopted approach in the field, Bhatt et al., 2020) to better understand and potentially mitigate biases. This use case presents us with a tangible context to evaluate our artifact's effectiveness.

While we choose the context of fraud detection for demonstration purposes, we note that our artifact does implement the functionality to add further use cases and metrics via a separate interface for auditors (as proposed by Fig. 4). In fact, we considered several other use cases for the evaluation, such as hiring and credit scoring. In hiring, for example, practitioners often leverage AI (Van den Broek et al., 2021) with several documented violations of AI conformity (see, e.g., Parasurama & Sedoc, 2022). In accordance with employment and anti-discrimination laws (e.g., the General Equal Treatment Act in Germany), auditors may define use cases for hiring. Similarly, in credit scoring, AI frequently creates value (Khandani et al., 2010) but often conflicts with legal frameworks (e.g., with the Equal Credit Opportunity in the US, Smith, 1977). The auditors may define additional use cases for credit scoring to address the unique requirements and laws applicable in this context.

Conformity metrics

For the exemplary use case of fraud detection, we determine the set of suitable metrics to measure fairness, explainability, and performance that is presented in Table 3. Our artifact automatically evaluates these metrics using the provided dataset, model, and questionnaire answers. The auditor estimates the relevance of each metric for a specific use case separately. We briefly describe these metrics in the following.

Fairness: statistical parity difference We include the metric corresponding to statistical parity (Dwork et al., 2012) due to its high relevance within legal frameworks (Barocas & Selbst, 2016). We measure the deviation of statistical

[4] The data is available at https://www.kaggle.com/datasets/dermisfit/fraud-transactions-dataset.

parity (that is, bias) as the difference in the shares of the favorable prediction within a group vs. outside this group. Importantly, a group is typically defined based on an attribute deemed sensitive (cf. Barocas et al., 2019). We compute the "statistical parity difference" in our artifact via

$$\frac{\#\text{ favorable predictions in group}}{\#\text{ all predictions in group}} - \frac{\#\text{ favorable predictions outside group}}{\#\text{ all predictions outside group}}. \quad (1)$$

In our exemplary use case, we consider gender as a sensitive attribute and measure the statistical parity difference between women and the rest of the customers. Notably, interpreting the value resulting from Eq. 1 and determining whether or not it is acceptable depends on the context. If aspiring to perfect fairness as defined by statistical parity, Eq. 1 must equal 0. If following the so-called "80% rule," for example, which is common in many legal frameworks (Barocas & Selbst, 2016; Feldman et al., 2015), fairness would only imply that Eq. 1 yields a value greater than -0.2. In our artifact, auditing experts that define use cases can of course define the range of values considered acceptable for the statistical parity difference and, thereby, tailor it to the given context.

Fairness: equalized odds difference We include the metric corresponding to equalized odds (Hardt et al., 2016) which shifts the emphasis from the distribution of the predictions themselves to the prediction errors. It is one of the most widely used metrics for algorithmic fairness and, in most cases, does not imply fairness when statistical parity difference does (Barocas et al., 2019; Garg et al., 2020), making it a complementary and natural choice for our implementation.

With equalized odds, we measure bias as the average absolute difference between error rates from within vs. outside a group. Similar to statistical parity, groups are typically defined based on sensitive attributes (Hardt et al., 2016). We compute the "equalized odds difference" in our artifact via

$$\frac{1}{2}[|\,\mathrm{FPR}_{\mathrm{ingroup}} - \mathrm{FPR}_{\mathrm{outgroup}}\,| + |\,\mathrm{FNR}_{\mathrm{outgroup}} - \mathrm{FNR}_{\mathrm{ingroup}}\,|]. \quad (2)$$

Analogously, we consider gender as a sensitive attribute in our exemplary use case. Here, $\mathrm{FPR}_{\mathrm{ingroup}}$ refers to the false positive rate for women, $\mathrm{FPR}_{\mathrm{outgroup}}$ refers to the false positive rate for the rest of the customers, and $\mathrm{FNR}_{\mathrm{ingroup}}$ and $\mathrm{FNR}_{\mathrm{outgroup}}$ are the counterparts with regard to the false negative rate. Similar to statistical parity difference, the auditor may manually set the range of values deemed acceptable for the equalized odds difference and tailor it to the context at hand.

Importantly, in this study, we focus on statistical parity difference and equalized odds difference due to their legal relevance, complementarity, and widespread recognition. Future iterations of the artifact will naturally incorporate a broad range of fairness metrics to enhance flexibility and address a variety of contexts and regulatory needs.

Explainability: questionnaire Explainability represents a special criterion as it is hardly feasible to measure the explainability of AI systems solely based on quantitative metrics. Therefore, inspired by existing works aiming to measure the explainability of AI systems (see, e.g., Gebru et al., 2021; Mitchell et al., 2019), we develop a questionnaire that the AI owner fills out as part of the conformity assessment. This questionnaire contains questions pertaining to the prediction model underlying the AI system (e.g., Do you communicate the technical limitations and potential risks of the AI system to users, such as its level of accuracy and/or error rates?) and the used explanations for prediction model outputs (e.g., Does your model provide any explanations for predictions? If yes, what kind of explanations?). The auditor determines for each question whether it is relevant for a use case and which answer of the question represents which color in the traffic light (i.e., green, yellow, and red). Our initial implementation displays the most critical color-coded response. For example, if the answers for a category span green, yellow, and red, the category would ultimately be assigned a red rating to reflect the highest level of concern.

Explainability: stability of feature-based explanations Here, stability refers to the sensitivity of feature-based explanations towards small perturbations in the training data. A low stability score indicates that small perturbations in the training data severely impact the set of features that are reported to be highly important for the AI system, making the global explanations less reliable. Our implementation of this metric is inspired by Hsieh et al. (2020).

Performance We measure the prediction performance based on the most common metrics for binary classification. These include accuracy, balanced accuracy, precision, recall, and the F1-score. For a detailed explanation of each listed performance metric, we refer to the work of Sokolova et al. (2006). Moreover, to account for our interviewed experts' concern about deteriorating performance over time, we implement a metric that measures data drift based on the Kolmogorov–Smirnov test (Darling, 1957; dos Reis et al., 2016). In short, the Kolmogorov–Smirnov test evaluates whether two data series are drawn from the same

probability distribution; under data drift, the statistical properties of data change over time, leading to a change in the underlying probability distribution. Our metric builds upon this evaluation to show the proportion of variables being affected by data drift.

Evaluation

We evaluate our design framework and software artifact following the Framework for Evaluation in Design Science Research (Venable et al., 2012) and build upon focus group sessions to both validate, refine, and add to our design framework. Notably, the evaluation is divided into two distinct iterations. In the first iteration, focus groups A, B, and C assess the initial prototype from the first design cycle, which incorporates DP 1 to DP 10. We use the feedback of these three sessions to formulate two additional design principles, DP 11 and DP 12. In the second iteration, focus groups D and E assess these two newly integrated DP in the updated software artifact. We detail the results of both iterations in the following.

Evaluation of first iteration

In the initial three focus group sessions, our artifact receives overall positive feedback. Participants express that our artifact provides a valuable approach to conducting conformity assessments, highlighting its effectiveness in assessing fairness and explainability in AI systems. They appreciate the user-friendly and interactive interface as well as the comprehensive insights and supporting information provided by the artifact, as expressed by the following comment: "[The tool is] easy, clean, and nicely presented. There is not too much information or cognitive overload" (Manager, focus group B). The think-aloud protocols, capturing participants' usage of the software during the sessions, and the subsequent group discussions further confirm the validity of the meta-requirements and design principles that we have inferred based on our awareness of the problem. For example, many participants particularly appreciate the intuitive nature of the quality labels (see DP 7) displayed through the traffic light system, as it enables them to quickly grasp the level of conformity of the AI system: "The structure is good, clear at the beginning, and the traffic light system immediately shows the performance in each area." (Department head, focus group C). Overall, the positive reception of the artifact and the alignment between its functionality and the intended objectives provides strong evidence of the careful consideration and successful implementation of our design choices.

Participants further make valuable suggestions that substantially contribute to the refinement and enhancement of our design framework. Many participants highly value the abundance of information available on the dashboard but express their desire for the inclusion of what-if scenarios. They emphasize that fairness and performance are influenced by parameters chosen by the AI owner, and having the ability to explore the impact of these choices on specific fairness and performance metrics within the dashboard is highly beneficial. "I would like to see how fairness and performance move when I vary [AI system] parameters like the classification threshold." (Data Scientist, focus group A). For example, consider the scenario where the AI owner viewing the dashboard notices that the AI system disproportionately labels transactions made by women as fraudulent, putting them at a disadvantage. To address this, the AI owner may wish to investigate the impact of adjusting the classification threshold specifically for women. By incorporating, e.g., a slider that allows for parameter adjustments and observing the resulting changes in fairness and performance metrics, the artifact could facilitate the exploration of such hypothetical scenarios. Therefore, we propose an additional design principle: The artifact should enable the exploration of what-if scenarios (DP 11).

Participants in the focus group sessions also share their thoughts on the presentation of results on the dashboard. Our current approach involves a traffic light system to communicate results. According to feedback from participants, this system is highly effective for conveying information to non-technical stakeholders, including upper management, regulators, and customers. However, developers utilizing our software for problem identification and optimization express the need for more detailed quality labels that provide deeper insights. "The traffic lights are suitable for upper management, but they are not optimal for developers." (Senior Data Scientist, focus group C). The participant further argues that for internal purposes, the artifact should communicate the conformity results in a more detailed and actionable way, underlining the importance of a use-case-specific definition of additional user roles (see DP 2).

Finally, participants in the focus group sessions show enthusiasm toward assessing the results on an interactive dashboard and seem to appreciate the ability to explore different levels of granularity. However, they also voice their concern regarding the need for persistent results and proper documentation. In practice, the software would need to generate a documentation report summarizing all results in a comprehensive way. "For validations or internal revisions, [the software] should generate export documents that are tamper-proof […]" (Risk Manager, focus group C). Proper documentation of results ensures the persistence of conformity results and, thus, enables organizations to track their progress over time effectively. This leads to our last design principle: The artifact should ensure persistent accessibility of results and protect against manipulation (DP 12).

Evaluation of second iteration

We integrate the new design principles DP 11 (*Artifact should enable interactive exploration (what-if scenarios)*) and DP 12 (*Artifact should ensure persistent accessibility of results (e.g.,*

via *a report) and protect against manipulation*) into our tool. To address DP 11, we introduce a slider that allows participants to view how performance and fairness metrics change as they adjust the classification threshold, that is, specifying the probability at which an observation is predicted to be of the positive class. To address DP 12, we include a button that, when clicked, generates a PDF report with a structured overview of the results of the conformity assessment, thus ensuring easily accessible and persistent documentation.

In the final two focus group sessions, our artifact once again received overall positive feedback. In particular, the newly integrated slider (implementing DP 11) was found to be useful for data scientists interested in exploring different classification thresholds (Requirements Engineer, focus group E). Another feature that received particularly positive feedback is our separation of the tasks of the AI owner and the auditor. A manager from focus group E, for example, stated: "[The separation of the two user roles] makes total sense… For me, the first impulse here is immediately the four-eyes principle." Intriguingly, we receive ambivalent feedback regarding our central use case repository, as illustrated by the following quote: "Everything that is within a company is certainly possible, everything that is overarching is more difficult" (Manager, focus group E). On the one hand, practitioners seem reluctant to share details on conformity assessments with competitors, even if they do not necessarily include critical information or model specifics. On the other hand, some experts agree that collaborating with other organizations within an industry may be necessary in some contexts: "It depends on the use case. When it comes to regulations that everyone has to comply with [the use case repository] makes sense to be able to compare them" (Data Scientist, focus group E). Still, the reluctance to share relevant information with competitors may become a crucial barrier to developing best practices and, ultimately, establishing assessment standards.

In our post-event survey on the design framework and its implementation in the software artifact, participants strongly agreed on the necessity of the meta-requirements and design principles and agreed, though to a slightly lesser extent, that the prototype effectively implements these principles. We present an overview of the validated design principles and survey results in Appendix C.

Discussion

Interpretation of key findings

Our analysis of literature, legal documents, and expert interviews sheds light on the pressing need for organizations to conduct AI conformity assessments, highlighting both their necessity and inherent complexity. Interestingly, our interviewed experts emphasize that the motivation for organizations to pursue conformity assessments may go beyond regulatory compliance and also stem from the recognition that self-commitment to AI conformity can offer a competitive advantage. This self-imposed commitment can be viewed as one manifestation of corporate digital responsibility (CDR), where companies embrace responsible practices related to digital products, services, and technologies (Mihale-Wilson et al., 2021). Research on CDR also supports the insights provided by the interviewed experts, highlighting the significance of CDR activities in shaping consumer perception. This influence can have a direct impact on consumers' opinions, consumption decisions, and choices of adoption (Carl et al., 2024; Schreck & Raithel, 2018), and ultimately promote a competitive edge in the market. To explore whether conducting AI conformity assessments indeed improves companies' competitive advantage, further research is needed testing our design framework in an organizational setting.

Another remark from the expert interviews that is worth discussing concerns democratization. Specifically, software with a graphical user interface may improve the accessibility of AI conformity assessments for non-technical stakeholders, potentially shifting the control and ownership of the assessments from a limited number of experts to a larger group of people. In line with previous research on democratization in the field of information systems (see, e.g., Awasthi & George, 2020; Zacharias et al., 2022), this shift in control and ownership can substantially benefit organizations by putting more employees in the position of active contributors to data-driven solutions. Moreover, democratization holds the potential to foster an organizational culture that promotes information sharing and embraces diverse perspectives as well as organizational agility (Hyun et al., 2020). Notably, drawing the right conclusions prior research has highlighted that a lack of technical expertise may generally limit broader participation in AI (Birhane et al., 2022), suggesting the need to cautiously assess whether graphical user interfaces can genuinely enable all stakeholders to effectively conduct conformity assessments.

The results of our focus group sessions both confirm the appropriateness of our artifact and design framework for conducting conformity assessments and reveal the potential for further improvement. Participants emphasize that the artifact holds the capability to promote conformity and, more broadly, ethics in AI systems. However, they emphasize the need for versatility in its procedures and presentations. On the one hand, (upper) management seeks intuitive quality labels to facilitate decision-making regarding the deployment of AI systems into production, which is in line with research on managerial decision-making (c.f. Clark Jr et al., 2007). On the other hand, developers need to continuously assess conformity and ethical impact throughout the development process to adhere to "ethics by design" standards

(Iphofen & Kritikos, 2021; Kieslich et al., 2022). In this regard, a software solution emerges as highly suitable, offering comprehensive information and adaptable support tailored to accommodate these diverse needs.

Interestingly, participants expressed curiosity about the availability of information explored within the interactive dashboard in a durable, accessible, and non-manipulatable document. Such a document would hold value in justifying the use of AI systems and effectively communicating conformity to external stakeholders. This observation aligns with the growing demand for AI certification from researchers (Cihon et al., 2021; Matus & Veale, 2022) and policymakers (e.g., European Union, 2023, p. 41), who advocate for its development. In this regard, our findings can serve as a springboard for the advancement of AI certification efforts, providing valuable insights to guide its development and implementation.

Contribution to practice

Our work makes several contributions to practice. First, we propose a design framework and a software artifact for AI conformity assessments, which we provide online.[5] Organizations aiming to prepare for upcoming regulations, such as the AI Act, may adopt and build upon our artifact to conduct conformity assessments and demonstrate compliance with these regulations. Second, our artifact offers a structured approach to assess AI conformity, enabling organizations to streamline the establishment of an AI conformity pipeline with predefined, scientifically derived evaluation metrics. This reduces the effort required during pipeline development and allows organizations to gain an overview of relevant metrics and parameters. Third, by automating much of the technical analysis, our tool reduces reliance on skilled experts and improves accessibility to AI conformity assessments, which in turn may reduce costs. Fourth, the tool facilitates collaboration among various stakeholders within organizations, such as AI developers, product owners, and senior management, and may thus foster a shared understanding of AI conformity challenges and opportunities. Fifth, the qualitative evidence gathered from expert interviews provides organizations with a structured overview of both the necessity and complexity of AI conformity assessments. This overview may help persuade management of the importance of addressing AI conformity and take proactive measures to mitigate anticipated implementation challenges (e.g., preparing a data pipeline for continuous monitoring of conformity metrics). Lastly, the artifact may evolve into a platform for sharing well-defined use cases across organizations in the same industry, potentially incorporating a recommendation system to guide organizations and promote industry-wide standards effectively.

[5] The software artifact and an exemplary use case are available at https://github.com/mlowin/conformity_assessment.

Contribution to theory

Our study makes important contributions to theory. First and foremost, our study contributes prescriptive design knowledge through the proposed design framework (Fig. 3) containing meta-requirements and design principles. This design knowledge serves as a valuable basis for researchers to advance the understanding of software for conformity assessments and AI audits in a broader sense. According to Baskerville et al. (2018), such novel design knowledge constitutes the main theoretical contribution of design science research.

Our work further contributes to the literature on AI governance. Effective governance of AI systems demands robust tools that ensure compliance, promote ethical standards, and enhance accountability across applications (Abraham et al., 2019; Birkstedt et al., 2023; Schneider et al., 2023). Researchers have frequently called for a "collaborative governance" of AI systems (see, e.g., Birkstedt et al., 2023) which extends the focus beyond individual organizations towards collaborative networks with internal and external stakeholders, each with distinct roles and responsibilities. Our design framework can be viewed as a step towards collaborative governance: it assesses AI systems against different societal and business-relevant criteria while considering distinct roles, such as AI owners and auditors, and facilitating collaboration among multiple organizations through sharing use cases and benchmarking AI systems against emerging standards.

Our work also makes a contribution to the literature on standardization of information and communication technology (see, e.g., Costabile et al., 2022; Hanseth & Bygstad, 2015; Lyytinen & King, 2006). A central tenet of this literature is that standardization plays a vital role in managing technologies, including AI systems, by ensuring comparability and interoperability. This aligns with insights from our expert interviews, which underline the necessity of standardizing AI conformity assessments to facilitate consistent and reliable evaluations across organizations within an industry. By contrast, findings from our focus group sessions reveal a practical problem: while practitioners acknowledge the importance of standardization, they are reluctant to share audit-related information with competitors, even when it excludes sensitive customer data or specifics about their AI systems. Yet, sharing insights and experiences with past conformity assessment configurations with other organizations is essential for the emergence of common best practices and, ultimately, industry standards. Our work highlights the importance of addressing this tension and demonstrates that innovative approaches are needed to promote standardization while respecting the concerns of individual organizations.

Transferability and future research opportunities

As outlined previously, focusing on AI for binary classification as well as certain aspects of AI conformity was essential for developing a tangible, functional artifact. Our choice of fairness and explainability stems from both aspects being not only highly relevant in current regulatory discussions but also representing the core challenges that we identified from the expert interviews in the problem awareness phase (e.g., context-dependent requirements and a lack of standardization). However, a key question remains: Can the findings based on these two aspects of AI conformity and the use case of binary classification be transferred to AI conformity in different contexts?

We argue that most findings from our study are indeed transferrable to other contexts. For instance, the feedback from practitioners on the utility of a software-supported assessment, its usability, and features, such as the "drill-down" approach on specific metrics and the traffic light system for simplifying communication, are not tied exclusively to fairness and explainability nor to binary classification. These design features facilitate the sharing of expert standards and interpretations, enabling them to diffuse quickly across industries. This suggests that software tools such as our artifact could support conformity assessments in other domains, independent of the specific aspects under evaluation. Another notable point is that fairness and explainability share similar characteristics and challenges with other aspects of AI conformity, such as privacy, AI safety, and cybersecurity. These aspects, like fairness and explainability, are marked by multiple, often competing definitions and metrics. For instance, privacy can have widely different interpretations depending on the context (Chua et al., 2021), and best practices for AI safety and cybersecurity are continually evolving (Lazar & Nelson, 2023). Therefore, our insights on software-based assessments—designed to facilitate the sharing of standards and enable semi-automated evaluations—may also prove valuable in these areas. Whatever additional criteria are incorporated in the future will likely either be based on quantifiable indicators, such as those reflecting fairness, or survey-based measures, as demonstrated with explainability, both of which were successfully implemented and tested during our focus group sessions.

With regard to extending our findings beyond binary classification to tasks such as regression or content generation (e.g., through generative AI systems, see Feuerriegel et al., 2024), we adopt a more nuanced perspective. Our findings can arguably be transferred to regression tasks, as regression is not fundamentally different from classification. This is evident in the applicability of similar group-level fairness measures (cf. Barocas et al., 2019) and comparable methods for explainability (e.g. Lundberg & Lee, 2017). However, applying our findings to generative AI poses more complex challenges. While the fundamental issues remain similar, such as evidence of bias in large language models and in image generation (Ananya, 2024) and ongoing trust issues related to opacity (Wang et al., 2023) the nature of generative systems introduces additional layers of complexity. While the general approach to AI conformity proposed in this paper could serve as a valuable starting point, our software artifact would require significant extension to accommodate the unique characteristics of generative AI. Further research is particularly crucial given the widespread adoption of systems based on generative AI, such as ChatGPT, DeepSeek, and MidJourney, which amplifies the societal impact and calls for tailored solutions to address the distinct challenges in this domain.

Other contextual factors could also influence the transferability of our findings. For instance, our expert sample, while operating internationally, is entirely based in Germany and Switzerland. This geographic concentration might under certain circumstances affect the applicability of our results to other regions with different regulatory environments or industry practices. While we believe our findings remain valid and make a strong case for "portable principles" (Magnani & Gioia, 2023) that can be transferred to different contexts, future research should include broader testing across diverse geographic and regulatory contexts to further substantiate this claim. Similarly, our artifact was evaluated through focus group sessions using a clickable prototype in a controlled, artificial setting. In real-world applications, integrating the artifact into an organization's IT infrastructure could significantly influence user interaction and adoption. To validate its practical applicability, future research should implement our artifact (or a comparable system) within organizational settings. This would allow for an investigation of its long-term impact on the sociotechnical environment, including its effectiveness, usability, and influence on organizational processes and decision-making.

Conclusion

In this paper, we address the pressing need for organizations to assess the conformity of their AI systems. By leveraging design science research, we have developed a design framework and software artifact that serves as a tool for semi-automated AI conformity assessments, enabling effective communication between AI owners and stakeholders such as regulators. As we look to the future, we envision further advancements in our framework, extending its scope from assessments to certification. These efforts will contribute to adhering to regulations on AI systems, such as the EU AI Act, and empower organizations to navigate the evolving landscape of AI conformity with confidence.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12525-025-00770-2.

Funding Open Access funding enabled and organized by Projekt DEAL. State Hesse,Germany: LOEWE 3, project "TransfAIr".

References

117th Congress. (2022). Algorithmic accountability act. https://www.congress.gov/bill/117th-congress/house-bill/6580/text. Accessed 02 Jan 2024.

Abbasi, A., Parsons, J., Pant, G., Sheng, O. R. L., & Sarker, S. (2024). Pathways for design research on artificial intelligence. *Information Systems Research, 35*(2), 441–459. https://doi.org/10.1287/isre.2024.editorial.v35.n2

Abdallah, A., Maarof, M. A., & Zainal, A. (2016). Fraud detection system: A survey. *Journal of Network and Computer Applications, 68*, 90–113. https://doi.org/10.1016/j.jnca.2016.04.007

Abdel-Karim, B. M., Pfeuffer, N., Carl, K. V., & Hinz, O. (2023). How AI-based systems can induce reflections: The case of AI-augmented diagnostic work. *MIS Quarterly, 47*(4), 1395–1424. https://doi.org/10.25300/MISQ/2022/16773

Abraham, R., Schneider, J., & Vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management, 49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

Ananya,. (2024). AI image generators often give racist and sexist results: Can they be fixed? *Nature News Feature, 627*, 722–725. https://doi.org/10.1038/d41586-024-00674-9

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Avin, S., Belfield, H., Brundage, M., Krueger, G., Wang, J., Weller, A., Anderljung, M., Krawczuk, I., Krueger, D., Lebensold, J., Maharaj, T., & Zilberman, N. (2021). Filling gaps in trustworthy development of AI. *Science, 374*(6573), 1327–1329. https://doi.org/10.1126/science.abi7176

Awasthi, P., & George, J. J. (2020). A case for data democratization. In *AMCIS*. https://aisel.aisnet.org/amcis2020/data_science_analytics_for_decision_support/data_science_analytics_for_decision_support/23

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org. http://www.fairmlbook.org

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review* 671–732. https://doi.org/10.2139/ssrn.2477899

Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems, 19*(5), 3. https://aisel.aisnet.org/jais/vol19/iss5/3

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. https://doi.org/10.1287/isre.2023.1199

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Natesan Ramamurthy, K., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney K. R., & Zhang, Y. (2019). Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4–1. https://doi.org/10.1147/JRD.2019.2942287

Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial intelligence in organizations: Current state and future opportunities. *MIS Quarterly Executive*, 19(4). https://doi.org/10.2139/ssrn.3741983

Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., & Eckersley, P. (2020). Explainable machine learning in deployment. In *Conference on Fairness, Accountability, and Transparency (FAT)* (pp. 648–657). https://doi.org/10.1145/3351095.3375624

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft, Tech. Rep. MSR-TR-2020–32. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022). Power to the people? Opportunities and challenges for participatory AI. *EAAMO '22*. Association for Computing Machinery. https://doi.org/10.1145/3551624.3555290

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: Themes, knowledge gaps and future agendas. *Internet Research, 33*(7), 133–167. https://doi.org/10.1108/INTR-01-2022-0042

Bringas Colmenarejo, A., Nannini, L., Rieger, A., Scott, K. M., Zhao, X., Patro, G. K., Kasneci, G., & Kinder-Kurlanda, K. (2022). Fairness in agreement with European values: An interdisciplinary perspective on AI regulation. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*. https://doi.org/10.1145/3514094.3534158

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society, 8*(1). https://doi.org/10.1177/2053951720983865

Burns, J., & Fogarty, J. (2010). Approaches to auditing standards and their possible impact on auditor behavior. *International Journal of Disclosure and Governance, 7*(4), 310–319. https://doi.org/10.1057/jdg.2010.21

Carl, K. V., Mihale-Wilson, C., Zibuschka, J., & Hinz, O. (2024). A consumer perspective on corporate digital responsibility: An empirical evaluation of consumer preferences. *Journal of Business Economics 94*, 979–1024. https://doi.org/10.1007/s11573-023-01142-y

Cho, M. K. (2021). Rising to the challenge of bias in health care AI. *Nature Medicine, 27*(12), 2079–2081. https://doi.org/10.1038/s41591-021-01577-2

Chandra, L., Seidel, S., & Gregor, S. (2015). Prescriptive knowledge in IS research: Conceptualizing design principles in terms of materiality, action, and boundary conditions. In *Hawaii international conference on system sciences* (pp. 4039–4048). ieee. https://doi.org/10.1109/HICSS.2015.485

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Acm sigkdd international conference on knowledge discovery and data mining*. https://doi.org/10.1145/2939672.2939785

Chua, H. N., Ooi, J. S., & Herbland, A. (2021). The effects of different personal data categories on information privacy concern and disclosure. *Computers & Security, 110*, 102453. https://doi.org/10.1016/j.cose.2021.102453

Chui, M., Hall, B., Mayhew, H., Singla, A., & Sukharevsky, A. (2022). The state of AI in 2022—and a half decade in review. McKinsey & Company, 6. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2021). Ai certification: Advancing ethical practice by reducing information asymmetries. *IEEE Transactions on Technology and Society, 2*(4), 200–209. https://doi.org/10.1109/TTS.2021.3077595

Clark Jr, T. D., Jones, M. C., & Armstrong, C. P. (2007). The dynamic structure of management support systems: Theory development, research focus, and direction. *MIS Quarterly, 31*(3), 579–615. https://doi.org/10.2307/25148808

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Acm sigkdd international conference on knowledge discovery and data mining*. https://doi.org/10.1145/3097983.3098095

Costabile, C., Iden, J., & Bygstad, B. (2022). Building digital platform ecosystems through standardization: An institutional work approach. *Electronic Markets, 32*(4), 1877–1889. https://doi.org/10.1007/s12525-022-00552-0

Coussement, K., & Benoit, D. F. (2021). Interpretable data science for decision making. *Decision Support Systems, 150*, 113664. https://doi.org/10.1016/j.dss.2021.113664

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. *The Annals of Mathematical Statistics, 28*(4), 823–838. https://www.jstor.org/stable/2237048

Deepak, P., & Abraham, S. S. (2021). Fairlof: Fairness in outlier detection. *Data Science and Engineering, 6*(4), 485–499. https://doi.org/10.1007/s41019-021-00169-x

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). A sociotechnical view of algorithmic fairness. *Information Systems Journal, 32*(4), 754–818. https://doi.org/10.1111/isj.12370

dos Reis, D. M., Flach, P., Matwin, S., & Batista, G. (2016). Fast unsupervised online drift detection using incremental Kolmogorov-Dmirnov test. In *Acm sigkdd international conference on knowledge discovery and data mining*. https://doi.org/10.1145/2939672.2939836

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM Computing Surveys, 55*(9), 1–33. https://doi.org/10.1145/3561048

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Innovations in theoretical computer science conference*. https://doi.org/10.1145/2090236.2090255

European Union. (2023). AI act. https://artificialintelligenceact.eu/the-act/. Accessed 08 Oct 2024.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Acm sigkdd international conference on knowledge discovery and data mining*. https://doi.org/10.1145/2783258.2783311

Feuerriegel, S., Dolata, M., & Schwabe, G. (2020). Fair AI: Challenges and opportunities. *Business & Information Systems Engineering, 62*, 379–384. https://doi.org/10.1007/s12599-020-00650-3

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering, 66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Fleming, N. (2018). How artificial intelligence is changing drug discovery. *Nature, 557*, S55–S57. https://doi.org/10.1038/d41586-018-05267-x

Floridi, L., Holweg, M., Taddeo, M., Amaya, J., Mökander, J., & Wen, Y. (2022). CapAI-A procedure for conducting conformity assessment of AI systems in line with the EU Artificial Intelligence Act. *SSRN 4064091*. https://doi.org/10.2139/ssrn.4064091

Fresz, B., Göbels, V. P., Omri, S., Brajovic, D., Aichele, A., Kutz, J., Neuhüttler, J., & Huber, M. F. (2024). The contribution of xai for the safe development and certification of AI: An expert-based analysis. https://doi.org/10.48550/arXiv.2408.02379

Fu, R., Huang, Y., & Singh, P. V. (2021). Crowds, lending, machine, and bias. *Information Systems Research, 32*(1), 72–92. https://doi.org/10.1287/isre.2020.0990

Garg, P., Villasenor, J., & Foggo, V. (2020). Fairness metrics: A comparative analysis. In *IEEE international conference on big data (Big Data)* (pp. 3662–3666). https://doi.org/10.1109/BigData50022.2020.9378025

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Gioia, D. A., Corley, K. G., & Hamilton, A. L. (2013). Seeking qualitative rigor in inductive research: Notes on the Gioia methodology. *Organizational Research Methods, 16*(1), 15–31. https://doi.org/10.1177/1094428112452151

Gnewuch, U., Morana, S., & Maedche, A. (2017). Towards designing cooperative and social conversational agents for customer service. In *ICIS*. https://aisel.aisnet.org/icis2017/HCI/Presentations/1

Hanseth, O., & Bygstad, B. (2015). Flexible generification: ICT standardization strategies and service innovation in health care. *European Journal of Information Systems, 24*(6), 645–663. https://doi.org/10.1057/ejis.2015.1

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In: *Advances in neural information processing systems (NIPS)* (Vol. 29). https://doi.org/10.48550/arXiv.1610.02413

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Heinz, D., Fassnacht, M., Röhrleef, J. H., Sagnier Eckert, L., & Satzger, G. (2024). Designing digital industrial platforms for the circular economy: A requirements catalog. In *ICIS*. https://aisel.aisnet.org/icis2024/gov_strategy/gov_strategy/5

Hevner, A., Chatterjee, S., Tremblay, M. C., & Berndt, D. J. (2010). The use of focus groups in design science research. In *Design research in information systems: Theory and practice* (pp. 121–143). Springer.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105. https://doi.org/10.2307/25148625

Hirt, R., Kühl, N., & Satzger, G. (2019). Cognitive computing for customer profiling: Meta classification for gender prediction. *Electronic Markets, 29*(1), 93–106. https://doi.org/10.1007/s12525-019-00336-z

Hsieh, C.-Y., Yeh, C.-K., Liu, X., Ravikumar, P., Kim, S., Kumar, S., and Hsieh, C.-J. (2020). Evaluations and methods for explanation through robustness analysis. https://doi.org/10.48550/arXiv.2006.00442

Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research, 15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Hyun, Y., Kamioka, T., & Hosoya, R. (2020). Improving agility using big data analytics: The role of democratization culture. In *PACIS*. https://doi.org/10.17705/1pais.12202

IEEE. (2008). IEEE standard for software reviews and audits. IEEE Std 1028–2008. https://standards.ieee.org/ieee/1028/4402/

IEEE Standards Association. (2022). P2863 organizational governance of artificial intelligence working group. https://sagroups.ieee.org/2863

Iphofen, R., & Kritikos, M. (2021). Regulating artificial intelligence and robotics: Ethics by design in a digital society. *Contemporary Social Science, 16*(2), 170–184. https://doi.org/10.1080/21582041.2018.1563803

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Junklewitz, H., Hamon, R., André, A., Evas, T., Soler Garrido, J., & Sanchez Martin, J. (2023). Cybersecurity of artificial intelligence in the AI act. Publications Office Eur. Union, Luxembourg, UK, Tech. Rep. JRC134461. https://doi.org/10.2760/271009

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance, 34*(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Kieslich, K., Keller, B., & Starke, C. (2022). Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. Big Data & Society, 9(1). https://doi.org/10.1177/20539517221092956

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Conference on Innovations in Theoretical Computer Science (ITCS)*. https://doi.org/10.48550/arXiv.1609.05807

Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in business analytics and operations research: Models, applications and managerial implications. *European Journal of Operational Research, 281*(3), 628–641. https://doi.org/10.1016/j.ejor.2019.09.018

Kuechler, B., & Vaishnavi, V. (2008). On theory development in design science research: Anatomy of a research project. *European Journal of Information Systems, 17*(5), 489–504. https://doi.org/10.1057/ejis.2008.40

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science, 65*(7), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

Langenbucher, K. (2020). Responsible AI-based credit scoring–A legal framework. *European Business Law Review, 31*(4). https://doi.org/10.54648/eulr2020022

Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science, 381*(6654), 38. https://doi.org/10.1126/science.adi8982

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems (NIPS)*. https://doi.org/10.48550/arXiv.1705.07874

Lyytinen, K., & King, J. L. (2006). Standard making: A critical research frontier for information systems research. *MIS Quarterly, 30*, 405–411. https://doi.org/10.2307/25148766

Magnani, G., & Gioia, D. (2023). Using the Gioia Methodology in international business and entrepreneurship research. *International Business Review, 32*(2), 102097. https://doi.org/10.1016/j.ibusrev.2022.102097

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems, 15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Matus, K. J., & Veale, M. (2022). Certification systems for machine learning: Lessons from sustainability. *Regulation & Governance, 16*(1), 177–196. https://doi.org/10.1111/rego.12417

Meske, C., Abedin, B., Klier, M., & Rabhi, F. (2022). Explainable and responsible artificial intelligence. *Electronic Markets, 32*, 2103–2106. https://doi.org/10.1007/s12525-022-00607-2

Meth, H., Mueller, B., & Maedche, A. (2015). Designing a requirement mining system. *Journal of the Association for Information Systems, 16*(9), 2. https://doi.org/10.17705/1jais.00408

Miah, S. J., & Genemo, H. (2016). A design science research methodology for expert systems development. *Australasian Journal of Information Systems, 20*. https://doi.org/10.3127/ajis.v20i0.1329

Michael, K., Abbas, R., & Roussos, G. (2023). Ai in cybersecurity: The paradox. *IEEE Transactions on Technology and Society, 4*(2), 104–109. https://doi.org/10.1109/TTS.2023.3280109

Mihale-Wilson, C. A., Zibuschka, J., Carl, K. V., & Hinz, O. (2021). Corporate digital responsibility-extended conceptualization and empirical assessment. In *ECIS*. https://aisel.aisnet.org/ecis2021_rp/80

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Conference on Fairness, Accountability, and Transparency (FAT)*. https://doi.org/10.1145/3287560.3287596

Mökander, J., & Floridi, L. (2022). From algorithmic accountability to digital governance. *Nature Machine Intelligence, 4*(6), 508–509. https://doi.org/10.1038/s42256-022-00504-5

Moula, F. E., Guotai, C., & Abedin, M. Z. (2017). Credit default prediction modeling: An application of support vector machine. *Risk Management, 19*, 158–187. https://doi.org/10.1057/s41283-017-0016-x

Parasurama, P., & Sedoc, J. (2022). Gendered information in resumes and its role in algorithmic and human hiring bias. In *Academy of management proceedings* (Vol. 2022, p. 1713). Academy of Management Briarcliff Manor. https://doi.org/10.5465/AMBPP.2022.285

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Conference on Fairness, Accountability, and Transparency (FAT)*. https://doi.org/10.1145/3351095.3372873

Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. In *Ethics, governance, and policies in artificial intelligence* (pp. 47–79). https://doi.org/10.1007/978-3-030-81907-1_5

Rosenfeld, A., & Richardson, A. (2019). Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems, 33*(6), 673–705. https://doi.org/10.1007/s10458-019-09408-y

Roski, J., Maier, E. J., Vigilante, K., Kane, E. A., & Matheny, M. E. (2021). Enhancing trust in AI through industry self-governance. *Journal of the American Medical Informatics Association, 28*(7), 1582–1590. https://doi.org/10.1093/jamia/ocab065

Scantamburlo, T., Cortés, A., & Schacht, M. (2020). Progressing towards responsible AI. https://doi.org/10.48550/arXiv.2008.07326

Schneider, J., Abraham, R., Meske, C., & Vom Brocke, J. (2023). Artificial intelligence governance for businesses. *Information Systems Management, 40*(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

Schreck, P., & Raithel, S. (2018). Corporate social performance, firm size, and organizational visibility: Distinct and joint effects on voluntary sustainability reporting. *Business & Society, 57*(4), 742–778. https://doi.org/10.1177/0007650315613

Sikder, M. F., Ramachandranpillai, R., de Leng, D., & Heintz, F. (2024). Fairx: A comprehensive benchmarking tool for model analysis using fairness, utility, and explainability. https://doi.org/10.48550/arXiv.2406.14281

Smith, J. F. (1977). The equal credit opportunity act of 1974: A cost/benefit analysis. *The Journal of Finance, 32*(2), 609–622. https://doi.org/10.2307/2326794

Sokolova, M., Japkowicz, N., & Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: A family of discriminant measures for performance evaluation. In *AI 2006: Advances in artificial intelligence* (pp. 1015–1021). Springer. https://doi.org/10.1007/11941439_114

Thelisson, E., & Verma, H. (2024). Conformity assessment under the EU AI act general approach. *AI and Ethics, 4*(1), 113–121. https://doi.org/10.1007/s43681-023-00402-5

Topuz, K., Zengul, F. D., Dag, A., Almehmi, A., & Yildirim, M. B. (2018). Predicting graft survival among kidney transplant recipients: A Bayesian decision support model. *Decision Support Systems, 106*, 97–109. https://doi.org/10.1016/j.dss.2017.12.004

Van den Broek, E., Sergeeva, A., & Huysman, M. (2021). When the machine meets the expert: An ethnography of developing AI for hiring. *MIS Quarterly, 45*(3). https://doi.org/10.25300/MISQ/2021/16559

Van Someren, M., Barnard, Y. F., & Sandberg, J. (1994). *The think aloud method: A practical approach to modelling cognitive* (Vol. 11). Citeseer.

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International, 22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Venable, J., Pries-Heje, J., & Baskerville, R. (2012). A comprehensive framework for evaluation in design science research. In *Design science research in information systems. Advances in theory and practice* (pp. 423–438). Springer. https://doi.org/10.1007/978-3-642-29863-9_31

von Zahn, M., Bauer, K., Mihale-Wilson, C., Jagow, J., Speicher, M., & Hinz, O. (2024). Smart green nudging: Reducing product returns through digital footprints and causal machine learning. *Marketing Science*, forthcoming. https://doi.org/10.1287/mksc.2022.0393

von Zahn, M., 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

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Cheng, Y., Koyejo, S., Song, D., & Li, B. (2023). Decodingtrust: A comprehensive assessment of trustworthiness in GPT models. *Advances in neural information processing systems (NIPS)*. https://doi.org/10.48550/arXiv.2306.11698

Wang, L., Gopal, R., Shankar, R., & Pancras, J. (2022). Forecasting venue popularity on location-based services using interpretable machine learning. *Production and Operations Management, 31*(7), 2773–2788. https://doi.org/10.1111/poms.13727

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., Trindel, K., & Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Conference on Fairness, Accountability, and Transparency (FAccT)*. https://doi.org/10.1145/3442188.3445928

Zacharias, J., von Zahn, M., Chen, J., & Hinz, O. (2022). Designing a feature selection method based on explainable artificial intelligence. *Electronic Markets, 32*(4), 2159–2184. https://doi.org/10.1007/s12525-022-00608-1