IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 October 2025, accepted 31 October 2025, date of publication 10 November 2025, date of current version 19 November 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3631309

RESEARCH ARTICLE

AI Impermanence: Achilles' Heel for AI Assessment?

KATHRIN BRECKER[1], SEBASTIAN LINS[2], NICOLA BENA[3], CLAUDIO A. ARDAGNA[3], MARCO ANISETTI[3], AND ALI SUNYAEV[4]

[1]Institute of Applied Informatics and Formal Description Methods (AIFB), Karlsruhe Institute of Technology (KIT), 76133 Karlsruhe, Germany
[2]Faculty of Economics and Management, University of Kassel, 34109 Kassel, Germany
[3]Department of Computer Science, Università degli Studi di Milano, 20133 Milan, Italy
[4]School of Computation, Information and Technology, Technical University of Munich (TUM), 74076 Heilbronn, Germany

Corresponding author: Kathrin Brecker (kathrin.brecker@kit.edu)

ABSTRACT Scandals have shown that extant assessment methods (e.g., certifications) cannot cater to the impermanent nature of Artificial Intelligence (AI) systems because of their inherent learning capabilities and adaptability. Current AI assessment methods are only limitedly trustworthy and cannot fulfill their purpose of demonstrating system safety. Our interviews with AI experts from industry and academia help us understand why and how AI impermanence limits assessment in practice. We reveal eight AI impermanence-related implications that threaten the reliability of AI assessment, including challenges for assessment methods, the validity of assessment results, and AI's self-learning nature that requires ongoing reassessments. Our study contributes to a critical reflection on current AI assessment ideas, illustrating where their validity is at risk owing to AI impermanence. We provide the foundation for the development of assessment methods that consider impermanence-related implications and are suited to fully leveraging AI capabilities for the benefit of society.

INDEX TERMS Artificial intelligence systems, artificial intelligence, artificial intelligence assessment, system changes, impermanence, learning.

I. INTRODUCTION

Internal and external assessment of Artificial Intelligence (AI) systems is becoming increasingly important because it can contribute to detecting potential risks and undesirable AI behavior (e.g., performance reduction). In essence, AI assessment refers to the attestation and verification of specific AI system characteristics, operations, and management principles to prove compliance with requirements (e.g., fairness principles). AI assessment is increasingly demanded by practitioners and researchers [1], [2], particularly because of popular AI scandals such as Google's algorithm that classified people of color as gorillas or Amazon's Alexa that offered adult content to children [3].

In recent years, various initiatives have emerged that examine how to perform internal or external AI assessment and that develop related methods, such as codes of conduct, ethics principles, certifications, assurance seals, and frameworks for assessment [1]. These AI assessment methods uncover flaws and increase trustworthiness by providing evidence that AI systems fulfill their intended purpose, comply with accepted regulatory standards, or adhere to industry best practices [4]. Regulators foresee a key role of AI assessment owing to its great potential. For example, the novel EU AI Act mandates the internal or external assessment of high-risk AI systems [5], the draft of United Kingdoms' AI Bill proposes third-party audits [6], and the New York City council requires yearly bias audits for AI-based employment decision tools [7].

However, AI assessment faces the risks of limited reliability and validity due to the impermanence of AI systems. Recent failures in practice emphasize adverse consequences of AI impermanence. For example, Zillow's iBuying machine learning algorithm overestimated the value of the houses for which Zillow paid, because the algorithm was not adjusted to changing market conditions [8], [9]. This model drift was

The associate editor coordinating the review of this manuscript and approving it for publication was Tyson Brooks.

only detected lately and led to a loss of $304 million and the closure of the AI system iBuyer in November 2021.

The changing nature of AI systems originates from their characteristics of intelligence, learning, and adaptability to various contexts [10], [11]. For example, Netflix's recommender system uses sophisticated AI algorithms to analyze vast amounts of users' data, learn from their preferences, and adapt the system accordingly [12]. Traditional assessment methods are unsuitable for such impermanent systems that change during operation [13]. AI impermanence presents assessors with the challenge that AI systems cannot be considered "unchangeable end product(s)" [14, p. 3]. Instead, systems change their behavior over time, even after tests [2] and long after deployment [14].

Impermanent AI systems raise concerns regarding the validity period of AI assessment [15], assessment methods' ability to predict future system performance [16] and to provide guarantees for the future [16], [17], [18]. For example, the online grocery delivery service Instacart used a machine learning algorithm to suggest replacements for out-of-stock products and was assessed to provide reliable predictions. However, the system experienced a sudden drop in prediction accuracy from 93% to 61% in March 2020 because customers' shopping habits changed suddenly during COVID-19. Instacart's model training practices had to be adjusted quickly, but it seems that the performance drop was identified too late [19]. AI impermanence thus introduces uncertainties about how AI assessment methods can deal with the self-learning of AI systems and changes in input data [20], [21]. Practitioners are faced with an urgent need to adapt traditional assessment methods to achieve viable AI assessment results. Without such adjustments, organizations are prone to developing AI assessment methods that are unreliable. Ineffective assessment will not only render assessment costs for organizations useless but also fail in reducing customers' uncertainty regarding AI systems. To avoid AI assessment that defeats its purpose, we need to understand the challenges and implications for AI assessment stemming from AI impermanence; ultimately motivating us to conduct this study.

Literature on AI assessments reveals that prior research has extensively analyzed assessment methods in AI (e.g., [14], [22]) and related contexts of systems characterized by fast-paced technological advancements such as cloud services (e.g., [23], [24]). Research has already made valuable contributions, for example, by clarifying how to overcome challenges in assessing ethical AI principles [2]. However, AI assessment research has just recently emerged and focused on developing the first proof of concepts and applicable assessment methods. We are still faced with a limited understanding of the origins of AI impermanence and its impact on AI assessment. Confronted with this research gap, a growing number of researchers have called for further research to uncover the impact of AI impermanence on AI assessment to ensure long-term AI system reliability and detect undesirable changes to mitigate risks (e.g., [3], [21], [25]). For example, research on AI certification calls for continuous assessment to cope with AI impermanence (e.g., [4]). These calls substantiate the need for further research and motivate our study.

We believe that the prevalent limited understanding of AI impermanence is in particular unfortunate for regulatory and assessment bodies, AI providers, and researchers. Traditional assessment methods can be adapted to the AI context only when the impact of AI impermanence becomes clear. Research is also responsible for conveying and explaining to practitioners the misfits arising from impermanence when traditional assessment methods for AI systems are applied. It is important to provide explanations of why assessment methods need to be adapted for AI and how they are influenced by AI impermanence. The promising potential of AI assessment is at risk if assessed systems still lead to major AI scandals. Therefore, we require further knowledge that can guide the adaptation of AI assessment methods, which includes understanding the causes and conditions leading to AI impermanence. We can then examine the impact of the changing nature of AI systems on the performance of AI assessment and their ability to ensure long-term system reliability. In response, we seek to answer the following research question (RQ):

How does AI impermanence impede AI assessment?

We conducted 25 expert interviews to identify AI impermanence-related challenges. We gained detailed insights into the perspectives of industry stakeholders that assess AI systems either internally from a technical perspective or externally as a third party or auditor, as well as stakeholders from other academic institutions. On the basis of these interviews, we reveal how AI impermanence is caused by accident, through updates, or by design. We uncover how these changes impact AI assessment and lead to novel challenges that need to be solved to develop reliable AI assessment methods. Among other things, we explain why AI impermanence hampers differentiating between desired and undesired changes, threatens AI assessment reliability and validity, and requires continuous assessments which are, however, only possible on a limited basis.

This study provides important contributions to research and practice. First, we highlight eight implications for AI assessment validity and generate insights that help address AI impermanence in a differentiated manner. Second, we provide a starting point for future research endeavors that can help address the adverse implications of AI impermanence for AI assessment. Third, our research guides practitioners, such as internal and external assessors, in conducting AI assessment by providing insights into aspects that need to be considered. Through our research, we uncover why AI impermanence leads to the failure of AI assessment in providing long-term guarantees. To support further technological advancement via continuously self-learning AI, we need to

find ways to handle such assessment to use self-learning systems safely and benefit from them. Currently, AI impermanence is the Achilles' heel of assessment, and our study provides first knowledge of how learning aspects and the lack of reliability inherent to AI impede the assessment of AI systems.

This study is structured as follows. We first provide background information on AI systems and AI assessment, followed by outlining the challenges that motivated our study. In Section III, we outline our qualitative research approach by explaining how we conducted expert interviews and analyzed the interview data. Section IV explains the implications of AI impermanence for AI assessment in detail. We discuss our principal findings, contributions to research, implications for research, and limitations in Section V, before presenting avenues for future research in Section VI. Finally, we portray related work that informed our study (Section VII) and provide the study's conclusion in Section VIII.

II. BACKGROUND

A. AI SYSTEMS

We consider diverse AI systems relevant to assessment and follow a broad definition of AI "as the ability of a machine to perform cognitive functions that we associate with human minds such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity" [26, p. iii]. We thereby align with AI regulators, proposing a similarly broad understanding of AI systems to cover the market [27]. Applying this comprehensive and abstract AI system definition that covers a wide range of different AI systems is useful for our study. It helps to ensure that the causes and effects of AI impermanence on AI assessment are captured holistically, independent of a specific AI system type or implementation but instead covering most systems "labeled AI technology" [28, p. 40]. It also considers the fact that there is no single agreed-upon definition of AI (e.g., [29], [30]).

B. AI ASSESSMENT

AI assessment refers to the systematic attestation and verification of an AI system and its responsible provider to verify that system characteristics (e.g., performance, fairness), operations, and management principles comply with legal, ethical, or industry requirements [31]. Assessment methods may include pre-deployment testing, ongoing monitoring, risk analysis, audits, and certifications. The goal of AI assessment is to ensure that the AI system functions as intended across its lifecycle and does not lead to adverse or unintended consequences.

The following three types of stakeholders are relevant when performing AI assessment: (1) AI providers mandating internal assessment or engaging third-party assessors, (2) internal/external assessors executing assessment methods, and (3) customers referring to assessment results to overcome their uncertainty. AI providers pursue assessment due to quality assurance processes and regulatory obligations, or to signal to customers and the public that their AI systems can be trusted. Assessors execute assessment methods and issue a formal proof of conformity if the AI system meets assessment requirements (e.g., an internal audit report, self-made assurance claims, a certificate, or quality seal). Internal assessment is usually conducted by internal auditors, related business units, and the technical developers of AI systems. In contrast, external assessment is conducted by independent third parties (e.g., auditors and certification bodies). For example, the ISO/IEC 42001 has recently gained traction as novel AI management system standard that can be tested and certified by independent parties. Such external assessment is considered more reliable than internal assessment because of the objectivity, credibility, and third-party verification it provides to customers [32]. For example, during a certification assessment, independent assessors review system documentation, check onsite conditions, and conduct employee interviews [31].

C. CHALLENGES OF AI ASSESSMENT

Research on AI assessment highlights that existing assessment methods are largely inadequate to address the peculiarities of AI systems (e.g., [10], [13]).

On the one hand, AI systems lack a precise definition of the system behavior and expected requirements to comply with [4], [33], [34]. Furthermore, such requirements have different applicable definitions that depend on the context where the AI system is intended to operate [33], [35]. On the other hand, AI impermanence, the focus of this paper, challenges assessments from different perspectives.

Early approaches to AI assessment required in fact that AI systems *cannot* be impermanent (e.g., [33]), meaning that the training- and inference-time data distribution shall not change, and the AI model is not altered once assessed. This assumption is however hardly applicable, because biases and drifts can appear at any time (e.g., [10]) and cannot be predicted.

Other assessment methods (e.g., [36], [37], [38]) accept that the AI system can change, but require to precisely detail the environment where the assessment took place which should, in turn, be as diverse as possible. Any changes then trigger a new assessment.

Finally, other methods attempt to reduce impermanence or the uncertainty caused by it. They mostly focus on the generation of extensive test cases to stress the AI system functioning over varying conditions (e.g., [39]) possibly simulating risky real-world settings (e.g., [16]). However, testing remains limited, particularly when the AI system is adapted over time (e.g., [2])

While these works point to impermanence as the main root cause for existing challenges in AI assessment, they do not delve into the concept itself, and barely touch on its

real-world implications on AI assessment, thus pointing to a clear, urgent research gap [3], [21], [25]. In this paper, we therefore aim to shed light on how AI impermanence impacts AI assessment. A more detailed comparison of our contributions with the state of the art can be found in Section VII.

As AI scandals have shown, treating impermanent AI systems as ''final'' end products until undesirable behavior that draws major attention occur involves high risk (e.g., reputation loss and risks for affected individuals or organizations) [3], [14]. Hence, AI impermanence calls for adaptive assessment methods that address issues arising from the changing nature of AI systems, such as assessment timing and validity [13], [14]. We need to better understand the implications of the threats that AI impermanence introduces for AI assessment to provide guidance for assessment development (e.g., meaningful certification validity periods) and avoid the risk of unreliable assessment. Therefore, we aim to first uncover the origins of AI impermanence and contribute to understanding why changes can occur. Afterward, we can explain how they impact AI assessment's long-term validity.

III. RESEARCH METHOD

A. METHOD SELECTION

We applied a qualitative research method and chose an inductive approach. We particularly conducted interviews with experts involved in AI assessment to learn from their experiences and examine the causes of AI impermanence and its implications for AI assessment. Broadly speaking, our research approach is characterized by the iterative process of gathering, interpreting, and comparing data from expert interviews, deriving patterns, and refining them. Inductive methods have proven useful in addressing phenomena that are less understood [40] and creating an abstract analytical schema of a phenomenon [41]. We thus did not apply quantitative methods like experiments, nor did we engage in the development of AI assessment methods, but sought to deeply engage with experts' experiences to truly understand AI's changing nature and uncover how it impacts AI assessment. We first aimed to derive rich descriptions on AI impermanence, as common for qualitative research [42], and then engaged in rigorous reasoning based on empirical data to develop sound explanations about how AI impermanence impacts AI assessment.

To analyze the data from our interviews, we employed grounded theory techniques [43], including *open*, *axial*, and *selective* coding, so that the interview data were analyzed with increasing abstraction levels [43], [44], [45]. Although we considered related research on AI assessment (Table 6), we first started by openly and freely coding the interview data to develop explanations instead of deducing components from existing research. We thus took related literature and related theoretical foundations into account, but not to the extent that they constrained our creativity and idea generation [43].

TABLE 1. Expert interviewee information.

	Technical internal perspective on AI assessment (n=14)	Assessors' external perspective (n=4)	Researchers' scientific perspective on AI assessment (n=8)
Professional AI-related experience	2 years: 1; 3-5 years: 4; >5 years: 3	2 years: 1; >5 years: 3	3-5 years: 2; >5 years: 6
Job roles	(Senior) Data Scientists, (Senior) ML Engineers, (Senior) AI Consultants / Architects, AI Project Leads	AI Audit: Department Head, (Senior) Managers, (Senior) Consultant	University Professors, (Post-Doctoral) Researchers & Data Scientists / Accountable AI
Industry	Leading information technology & consulting companies; Start-ups (insurance, recruiting, healthcare)	Leading auditing firms	Universities & Research Institutions
Organization size	>190,000 employees: 7; 25,000 – 50,000 employees: 4; < 50 employees: 3	>200,000 employees: 4	<10,000 students: 3; 15,000 – 30,000 students: 3; >50,000 students: 2

B. INTERVIEW DATA COLLECTION

We purposely sampled relevant stakeholder groups [46] and acquired 26 experts (Table 1). We selected experienced professionals (with at least 2 years of professional experience) from technical, external assessment, and research job roles, as they are expected to be involved in AI assessment. First, 14 technical AI experts from industry (e.g., machine learning engineers, AI architects, AI consultants, data scientists, and data analysts) were selected owing to their experience with causes leading to AI impermanence, internally assessing AI systems, and avoiding (undesirable) changes in AI system behavior over time. Second, four assessors (e.g., internal assessors, external auditors, and AI assessment initiative members) were selected to gain insights into the AI assessment implications that they observe or expect to find when conducting AI assessment. Finally, eight AI researchers from other universities and research institutions were selected to gather not only the views of practitioners but also theoretical views on AI impermanence and its implications for AI assessment. Similar to previous studies and AI regulations, we took a cross-industry perspective in this study. We thus did not focus on specific industries like automotive. Instead, interviewed assessors conduct audits across industries and most technical AI experts serve cross-industry customers, while some of them focus on insurance, recruiting, and healthcare.

Data collection began in August 2022 and ended in November 2022. We conducted qualitative one-to-one interviews (except one interview with two experts), following the method recommendations of Myers [47]. We applied

TABLE 2. Overview of AI impermanence types, their causes, and example implications for assessment.

Impermanence type	Description	Causes of changes	Example assessment implications
Unintended: change by accident	The AI system performs as desirable during assessment but accidentally deviates from its intended behavior during operation over time.	• Hidden flaws in the input dataset for training • Changes in the real world leading to outdated AI behaviors (e.g., regulatory and societal changes) • Changes in the hardware and technology set-up	• Assess whether the deviation in input data occurs over time (and detect the reasons, e.g., change in technology set-up, nonrepresentative dataset, and use case specific changes over time) • Continuously check for updates in the real world outdating AI system behavior (e.g., new laws and relevant changes for the use case such as people's taste)
Occasional: change by update	AI system enhancement through updates yields undesirable AI behavior.	• Changes in the model due to retraining • Changes due to additions in the technological set-up • (Hidden) flaws in the newly added data	• Assessment cannot consider updates for enhancements blindly as improvements • Every update needs to trigger reassessment • Assessment after updates needs to check for any undesirable aspects added to the AI system (e.g., performance reductions and biases)
Certain: change by design	The AI system constantly evolves, learns, adapts, or optimizes itself during operation, thereby deviating from its intended behavior.	• Evolving AI model implementation (data are fed back and used from operation) • Learning, adaption, or optimization based on, e.g., user feedback or other external influencing factors	• Complexity to assess a continuously moving target to ensure that no undesirable changes occur • Traditional assessment methods are not suited for evolving systems • Change is intended for evolving systems, allowing for desirable change, and excluding undesirable change at the same time can be challenging for assessment

a semistructured interview method that provided structure to experts while leaving room for aspects that were not or could not be considered while preparing the interview guide [47]. During the interviews, we asked the experts about their associations and experiences with changing AI behaviors, causes, and definitions of AI system change; impact on AI assessment; and means and solutions to address such changes. We adhered to the best practices of qualitative research (e.g., [47]), such as applying a nonjudgmental form of listening [42], maintaining distance [48], and conducting the conversations in an open and unpersuasive manner to avoid bias [49]. The interviews took 46 minutes on average and 901 minutes in total. Twenty-two interviews were conducted in German, and three were conducted in English. All interviews were recorded, anonymized, and transcribed. Interviewees provided informed consent prior to the start of the interview recording. In total, we gained 257 pages of transcripts from the 25 interviews.

C. INTERVIEW DATA ANALYSIS

Our data analysis took place iteratively and was started in parallel with data collection, directing our choice of questions on the basis of emerging concepts, that is, challenges for AI assessment resulting from AI impermanence [43], [44], [50]. This procedure enabled us to detect relationships between concepts in an iterative process of constant comparison between the initial data collected and the preliminary results of the analysis [43]. We followed the structured and iterative coding approach proposed by Corbin and Strauss [43], including *open*, *axial*, and *selective* coding, enabling us to iteratively increase the abstraction level of our findings [43], [44], [45]. These grounded theory coding techniques have proven their usefulness and have been widely adopted across research [45], [51]. The tool ATLAS.ti and manual annotations were used to code the transcripts. Online Appendix A provides example supportive evidence for the applied coding techniques.

Since in-depth research on the impermanent nature of AI systems is scarce, we first started to understand the causes and manifestations of AI impermanence to better understand how AI impermanence impacts AI assessment. We engaged in *open coding* as the first step in our coding procedure [43]. We fractured the data according to concepts in the data that might describe relevant aspects of AI impermanence [43], [44]. We aimed to be as open to new concepts as possible, despite reading prior literature that might influence our coding [52]. To avoid such influence, we aimed to set notions from the prior literature aside for initial coding [53].

We first focused on describing AI impermanence as a phenomenon, enabling us later to understand its assessment implications. We turned to the interview transcripts and thus openly coded text segments related to AI impermanence as a concept. For example, the interview passage *"during the operation of the system, it is self-learning based on new data and changes automatically"*[Senior Manager AI Audit] was coded as "self-learning", illustrating that AI impermanence results from the self-learning nature of AI systems. During open coding, we assigned codes to 529 textual segments. We further aggregated our open codes by constantly

comparing coded text segments to achieve higher levels of abstraction and identify the core dimensions of AI impermanence. This aggregation resulted in 14 higher-level code categories, such as ''input data composition'', ''input data change'', ''update'', or ''assessment guarantees''. These categories reflect how AI impermanence can be caused by accident without the intention to change the system (e.g., hidden flaws in the input data set, changes in the real world, changes in the hardware or technology set-up), how AI impermanence can result occasionally through system enhancements (e.g., model retraining, additions, newly added data), or through the continuous evolvement of the system (e.g., learning, adaptation). Table 2 summarizes our derived conceptualization of AI impermanence [54].

Having gained an understanding of AI impermanence as a concept, we continued to look for assessment implications. We followed the suggestions of Corbin and Strauss [43] on *axial coding* and coded for conditions, actions/interactions, and consequences. Here, we looked for conditions uncovering the reasons for AI changes and the manifestations and consequences of AI impermanence as implications for assessment. We then coded actions and interactions to determine how an AI system change was detected, resolved, and addressed during assessment. This method of axial coding enabled us to compare codes and classify them into categories that constitute AI changes and consequences for AI assessment [43], [44]. For example, the following text passages were coded as ''*consequence_noguarantess over time*''for AI assessment due to everchanging AI systems: *"We cannot say that we have looked at the system and the system is fine; therefore, it will stay fine the whole time"* [Manager AI Audit] and *"The honest answer would be we hope and pray. From a technical perspective, you can never know it 100 percent. If other input answers are coming in, then we see what happens"* [Senior Consultant Data & AI]. In total, we identified eight categories of implications that AI impermanence has on AI assessment (Table 3).

As a final step, we applied *selective coding* to structure our categories developed during axial coding into a coherent theoretical framework, a so-called ''core category'' [43]. A core category helps to formulate a storyline for coherent conceptualization of the central phenomenon [43]. Through this step, we aimed to achieve a more abstract conceptualization level and arrange our findings [55]. While reviewing our codes on AI impermanence and assessment implications, we noted that AI impermanence impacts specific phases

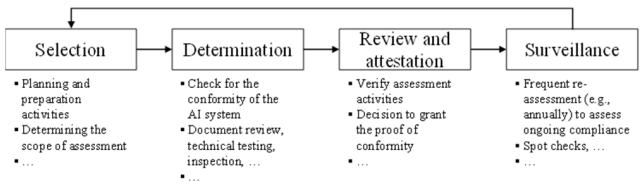

FIGURE 1. Assessment lifecycle according to the ISO/IEC:17000 [56].

across the entire assessment lifecycle. We therefore identified the assessment lifecycle as a potential core category and turned to the literature to identify suitable lifecycles. We discovered that assessment is frequently based on the standard lifecycle proposed by the ISO/IEC:17000, comprising four key assessment functions (Figure 1; [56]).

First, selection functions include planning and preparation activities (e.g., determining the AI system to be assessed and selecting suitable assessment procedures). Second, determination functions comprise key activities undertaken to check for the conformity of the AI system with assessment criteria (e.g., document review concerning developers' decisions to prevent discrimination). Third, review and attestation functions refer to assessors' verification activities in which they examine whether the selection and determination activities and their results are suitable, appropriate, and effective, followed by the decision of whether to grant the proof of conformity (e.g., a certificate). Finally, most assessments involve surveillance functions to execute systematic, repeated conformity assessment to maintain the validity of the conformity statement. We conceptualized each assessment function in detail on the basis of the extant literature and critically compared and mapped our findings to each function (Table 3). We noted that AI impermanence influences primarily the determination, review and attestation, and surveillance functions. For example, we assigned the assessment implication ''*Distinguishing between desired and undesired changes is challenging for assessors*'' to the determination function because the changing nature of AI leads to inabilities in assessing AI behavior.

After comparing and assigning our implication categories to the functions, we again checked our codes and coded text segments to ensure that our core category captures interviewees' perceptions and respective findings. In conclusion, selective coding helped us structure the identified categories of the implications of AI impermanence on three key assessment functions, namely, determination (''*How to Assess Consequences of Change: AI Impermanence Challenges Assessment Methods*''), review and attestation (''*How Reliable are the Assessment Results: AI Impermanence Threatens Validity*''), and surveillance (''*When and How to Reassess: AI Impermanence Impacts the AI Assessment Validity Period*''), discussed in detail in the next section.

In addition to applying coding techniques, our analysis was guided by constant comparison [43], [45] to compare similarities and differences in the thought process with respect to AI impermanence and the causes of (undesirable) changes in AI behavior that impacted AI assessment, especially between the different stakeholder groups of practitioners and researchers, to triangulate our data. Moreover, we used memoing, such as making notes about performing assessments in line with the AI lifecycle model, to capture theoretical ideas (e.g., noting key insights) during the data collection and analysis processes [45], [57]. During the analysis process, we became confident that we reached sufficient theoretical saturation. We noticed that no further AI impermanence causes or conse-

TABLE 3. Overview of AI impermanence-related implications for AI assessment.

Assessment function	Implication categories	AI impermanence-related implication
Determination	How to Assess Consequences of Change: AI Impermanence Challenges Assessment Methods	• No. 1: Distinguishing between desired and undesired changes is challenging for assessors. • No. 2: Assessors cannot anticipate or test (or can anticipate or test only to a limited extent) the external causes of AI impermanence. • No. 3: The limited explainability of self-learning systems hampers reassessment in the case of change.
Review and attestation	How Reliable are Assessment Results: AI Impermanence Threatens Validity	• No. 4: Assessors cannot anticipate and consider (or can anticipate and consider only to a limited extent) rare cases that can have widespread negative impacts. Rare cases may invalidate assessment guarantees and require continuous assessment, which is limited for AI systems. • No. 5: Assessors need to control for reintroduced human bias over time. • No. 6: Assessors can only limitedly declare systems' conformity when the system is extended across regions or organizations.
Surveillance	When and How to Reassess: AI Impermanence Impacts AI Assessment Validity Period	• No. 7: AI assessment validity depends on changes in user input and external conditions, and not only on changes in the system itself. • No. 8: AI systems are intended to change due to their self-learning nature. Surveillance is thus required to allow for and cope with substantial changes in systems.

quences for assessment emerged from additional interviews. We also found strong support for interrelations between elements from ample data and (mostly) densely populated AI impermanence-related implications. Additionally, our findings proved to be robust across interviews. Therefore, after coding all the interviews, we did not believe that further data collection or analysis would generate new findings.

IV. IMPLICATIONS OF AI IMPERMANENCE FOR AI ASSESSMENT

We discuss the implications of AI impermanence on the three main assessment functions: determination (Section IV-A), review and attestation (Section IV-B), and surveillance (Section IV-C). Table 3 summarizes the identified AI impermanence-related implications.

A. HOW TO ASSESS THE CONSEQUENCES OF CHANGE: AI IMPERMANENCE CHALLENGES ASSESSMENT METHODS

During the determination function [56], assessors need to evaluate AI system changes and their implications, which includes assessing external causes that result in system changes, its impact, and the system's conformity after the change. However, it is challenging to assess how much and what change to the system's logic arises from additional training and external conditions due to the opaque learning process of AI.

1) AI IMPERMANENCE-RELATED IMPLICATION NO. 1

Distinguishing between desired and undesired changes is challenging for assessors.

The power that impermanent AI systems yield because of their ability to learn and evolve is a strength that organizations want to leverage. However, organizations and their assessors currently do not have the means to distinguish desirable from undesirable changes during system operations. In particular, because most AI systems are currently not able to learn continuously (i.e., through self-learning), individuals trigger these systems to learn by manually feeding in new data from operation. Despite developers' careful actions (e.g., manual checks for bias contained in novel data), learning from new training data can lead to accidental changes over time that remain unnoticed. More importantly, new data can impact the system in unpredicted ways, for example, by reducing accuracy. Changing the AI system through retraining can then yield the risk of "becoming biased along the journey" [Senior Data Scientist]. In practice, the risk of undesirable changes is therefore one of the main reasons, as stated in the interviews, why organizations typically do not retrain their system and perform any system changes, ultimately hoping to maintain security.

However, even for experts, it can be difficult to assess whether new data yields small or substantial change to the AI system, which even can lead to adverse consequences. Assessors may find it challenging to tell how feeding a certain data set into an AI system during retraining will impact or update the decision logic of that system. The reason for this is the opaque learning nature of AI. An assessor cannot foresee which data points or aspects might cause a system change. Some aspects that an assessor thinks might change the system significantly may not have a major impact. Owing to the opaque nature of especially complex learning systems, assessors often have no other means than to backward reengineer the system to determine the changes caused and their implications.

The interviewees acknowledged that tools to continuously reassess AI systems after updates are starting to enter the market (e.g., tools for machine learning operations). However, the interviewees highlighted that these tools are currently so expensive that they render some AI system use cases economically unviable. Equally severe, these tools are still in their infancy. For example, tools discover deviations in AI system behavior only after their occurrence. Thus, they can help mitigate further damage but do not support fast incident response and provide no guarantees for system behavior. Tools also do not cover all aspects that organizations need to assess during retraining but, instead, focus mostly on specific parts of the system. For example, tools that are specialized in detecting deviating input data and feeding in new data

sources. Some tools can guide AI system updates but need to be set up in advance and executed by individuals. This manual effort threatens the cost-to-benefit ratio in light of AI system use and assessment costs.

On the basis of our interview results, we recommend that assessors seek tool support where possible and affordable (e.g., in light of potential risks from AI system usage) and continue to observe research on technical solutions that support AI system assessment (end2end). We suggest working toward developing service platforms that offer tools for assessors to increase ease of use and reduce integration efforts. Further research can improve assessors' ability to determine the impact of new data input on AI system learning and update its inner functioning without having to backward engineer the system after discovering that the learning has led to undesirable outcomes. Techniques to closed box the opaque learning process in which assessors can hardly tell how much and what change to the system's logic arises from additional training must be developed.

2) AI IMPERMANENCE-RELATED IMPLICATION NO. 2

Assessors cannot anticipate or test (or can anticipate or test only to a limited extent) the external causes of AI impermanence.

The performance and behavior of AI systems depend on the data they receive from the real world as input. The interviewees stressed that they are uncertain about how assessment methods should cater to AI system deviations that result from changes outside the system. Since the future is inherently unpredictable, changes in input data are likely to occur. For example, external changes such as upcoming product substitutions can be difficult to detect in advance by assessors but can later lead to incorrect assumptions about customers' product preferences by the AI system. Similarly, cultural changes can lead to shifts in the ethical understanding of customers that then influence how they interact with the system (e.g., their choice of words or interaction patterns).

One of the challenges in assessing AI systems is the lack of (historical) data for such external changes. Otherwise, developers could have used these data to train and prepare the system, ensuring that the systems remains effective even as the real world evolves. However, during the initial determination, assessors lack access to input data that reflects potential future scenarios because this data has not yet been generated, for instance, by customers interacting with the AI system. Since the external changes that may impact the system have not occurred, assessors cannot anticipate the diverse input data. They thus cannot ensure that the system is prepared for the future real world.

It is also difficult for them to exhaustively simulate the future real world and corresponding input data because they are not entirely known in advance, and aspects with an important impact on the AI system might not be foreseen but are mostly based on assessors' assumptions (e.g., how specifically customers' preferences change). "*Unfortunately, and that is regarding political, economic, and social activities, these data-generating processes rarely or never do us a favor in fulfilling the characteristic of time consistency. [. . .] So, where partially eruptive, partially gradual data-generating processes are deviating sufficiently, all we can do is watch input data between two measuring points. [. . .] I know this sounds very frustrating and disappointing*"[Auditor, Department Head Analytics].

We advise assessors to thoroughly anticipate potential external changes, their sources (whether external or system internal), and the (adverse) consequences of these changes. Like developing security risk concepts, they should document, report, and continuously assess their estimations about system changes. This will not only increase transparency about AI assessment and the root cause for the unintended system change but corresponding documentation can also foster AI accountability. More importantly, assessors need to monitor the real world and simulate potential changes in input data to assess the risk of undesirable behavior due to external changes. This detailed and in-depth consideration of changes in the external environment during assessment is AI-specific in contrast to traditional assessment methods, which focus solely on changes related to the system itself.

3) AI IMPERMANENCE-RELATED IMPLICATION NO. 3

The limited explainability of learning systems hampers reassessment in the case of change.

Limited explainability exacerbates impermanence-related issues for AI assessment. Assessors need to examine whether and how the system has changed after retraining or during continuous learning. Among other things, they need to evaluate whether biases or other undesirable aspects (e.g., performance reduction) result from the change. However, what the system has learned over time to derive its new conclusions remains (partially) uncertain owing to the limited explainability of AI systems. This uncertainty arises from the system's tendency to adapt to changing external factors, such as new input data, weather, or lighting conditions, which can lead to unintended modifications to its logic. For instance, when changes in lighting conditions affect the system's decision-making process, it can be difficult for assessors to identify that the system has incorrectly relied on light as a predictor for a variable that is prone to change. Using such variables can ultimately result in the system producing undesirable outcomes that deviate from its intended purpose.

In particular, self-learning AI systems are very complex. For example, the assessment of neural networks, which are not enabled to evolve but are retrained periodically, is already a very complex endeavor. Among other things, assessors need to check outcomes, labeling, and training data. If this model, which is already very complex to understand (i.e., has limited explainability), is enabled to evolve during operation through self-learning, then it becomes even more difficult for assessors to understand such evolutions and its consequences. This is also one of the reasons that the interviewees stated that organizations prefer probabilistic or rule-based AI systems

with high explainability rather than neural networks. Then, it is easier for them to see what has been learnt and how it changes. Hence, low explainability hinders assessors in trying to manage AI impermanence. As a consequence, interviewees often prioritized safety and reliability over potentially better AI performance through neural networks.

Even if assessors try to understand systems by evaluating their output, there may not be enough time to assess the output between learning cycles and uncover what has changed during learning. *"Of course, for unsupervised learning, it is a lot more difficult to understand how the change is taking place because, in supervised learning, you surveil how and what the model learns so to say. Well, you cannot always explain it fully what the model is doing, but it is most often somewhat understandable. While in unsupervised learning it is mostly very difficult to understand why the behavior has changed when it has changed"*[Data Scientist].

On the basis of our interview results, we suggest that limitations in explainability are inherent to AI systems and need to be accepted, although they negatively impact the detection of AI impermanence. AI systems with high explainability are easier to manage for assessment. However, it is fundamental to AI and the powerful benefits it yields to leverage a variety of complex algorithms that are not understandable by human assessors. In these cases, a degree of uncertainty about AI impermanence that introduces undesirable behavior remains in the assessment results. As explainability is challenging for humans, technical determination tools can support human assessment and reduce (even if not fully) this uncertainty.

B. HOW RELIABLE ARE THE ASSESSMENT RESULTS: AI IMPERMANENCE THREATENS VALIDITY

Review and attestation functions [56] are also impacted by AI impermanence. AI-specific aspects may lead to changes such that the attested criteria may no longer be fulfilled over time. The validity of the proof of conformity issued after assessment is then threatened by changes. There are several aspects particular to AI systems that cannot be observed or checked by assessors, challenging the extent to which the results of AI assessment can be considered reliable with respect to performance guarantees or the appearance of undesirable behavior.

1) AI IMPERMANENCE-RELATED IMPLICATION NO. 4

Assessors cannot anticipate and consider (or can anticipate and consider only to a limited extent) rare cases that can have widespread negative impacts. Rare cases may invalidate assessment guarantees and require continuous assessment, which is limited for AI systems.

There is a risk that the assessment results are only limitedly reliable for rare cases. This uncertainty can have a widespread impact because neglecting rare cases can lead, among other things, to severe negative effects in terms of discrimination and fairness (of those underrepresented groups or cases missing in the training data) or unexpected performance drops in critical industry applications. It is particularly worrying if rare cases cannot be assured at the point of assessment but arise at a later point in time (often as AI scandals). Even if developers make every effort to consider rare cases (e.g., generating additional artificial training data and removing undesirable bias from historical data as much as possible), there remains the risk of especially rare cases not being represented because they are typically not known in advance. Assessment and corresponding proof of conformity (e.g., a certificate) can thus communicate false or at least unreliable signals to AI system customers.

At the beginning of an AI project, it is generally not possible to ensure that all relevant input data are represented because it would require clarity on all relevant conditions that impact desirable outcomes, while outcomes are often not known. For example, a data scientist reported that developers trained an AI system for healthcare to measure the steps taken by users to detect certain illnesses. It performed well during testing. Only later during system operation it became obvious that there was a rare illness that affects the motion profile in different ways. Although the developers first thought that there was an error in the hardware device sending the data to the system, it turned out that the AI system was not able to detect that this person had the observed motion profile because of the rare illness. It is natural for rare cases to not appear in first testing after deployment because it can take time for the AI system to be confronted with rare case data. Thus, the system's performance reduction is observed only during operation at a later point in time. If there is no further assessment during the operation, then performance reduction for rare cases may not be detected.

On the basis of our interview findings, we recommend that assessors specifically consider that, despite assessment, severe rare case issues may arise at later points in time. Limitations in detecting rare cases restrict the validity of the proof of conformity and need to be communicated by assessors clearly to customers and the public. As AI systems are often involved in sensitive aspects relevant to customers' lives (e.g., healthcare and insurance), assessors are thus confronted with the fact that AI systems may work well for cases where (training) data are available, whereas other cases may not be reflected or includable (e.g., because artificial training data cannot be created owing to a lack of knowledge of what would need to be included). Communicating that this limitation is inherent to AI is important to avoid that stakeholders see this limitation purely as a shortcoming of the assessors (or ethical views of developers) who assess the system to the best of their knowledge.

2) AI IMPERMANENCE-RELATED IMPLICATION NO. 5

Assessors need to control for reintroduced human bias over time.

AI assessment is often highlighted as a means to avoid undesirable behavior such as discrimination or unfair decision suggestions that can be brought in during training on the basis of human or societal flaws reflected in training data.

However, uncertainty regarding undesirable human behavior reflected by the AI system remains, even if an assessment is conducted.

Technically, there are limitations to the extent to which AI developers can mitigate risks of undesirable behavior. For example, discriminating attributes can be removed but then reintroduced by other variables. This situation can be challenging for assessors to detect because although mitigatory actions have been taken by developers, they may be ineffective. In contrast, most traditional systems have a clear and rule-based logic, where assessment results are based on system functioning. The potential for deficient functioning can thus be detected. In contrast, AI systems may show a more human-like (faulty) functioning, which they have learned from past data. AI assessment results are thus prone to attesting system safety, despite containing human (unconscious) bias. These deficiencies are not obvious during initial assessment and can cause issues later for the organization and its customers (e.g., reputation loss and legal consequences).

Additionally, despite all efforts during training and initial assessment, self-learning systems can reintroduce biases during their continuous learning process over time when they are confronted with input data that contain biased or discriminatory aspects (e.g., decisions, outcomes, and relationships). ''*It can happen if I say my system should continue learning while I am doing image recognition. For example, it should continue to identify new employees or something such as that. In this case, it can happen that there are more white people because in my society, there are more white people. In addition, it can happen that the system evolves in the direction to recognize white people better than people of color, which is unfair*''[Doctoral Researcher].

To ensure that assessment results remain valid and offer assurance for stakeholders, assessors need to continuously account for the self-learning nature of the AI system. This involves monitoring to ensure that unsuitable data is not being brought in but instead cleansed and corrected, and that system suggestions are not being flawed. A common means to reduce such remaining risks (apart from technical measures such as fairness checks) is to put a human ''in the loop'' while performing reassessment during system operation. However, the interviewees stressed that the quality of this human-based reassessment needs to be doubted for several reasons, and that uncertainty about the assessment results remains in the end. One key reason for this is that human assessors ''in the loop'' can tend to blindly rely on the results of an AI system that in the past provided correct output and that they know was trained by experts [58]. In addition, assessors may not consider systems' decisions suspicious because they match historic decisions and assessors can be unaware of potential biases.

Based on our interview results, we recommend that AI assessors should fulfill high-competency requirements to mitigate the risk of assessors missing biases emerging over time. For example, diversity in assessment groups or external assessors that have an impartial and objective view outside organizational culture should be considered (vs. insiders who may be less critical in terms of past decisions or view them as normal). It is important to note that there is still uncertainty surrounding human (unconscious) bias in AI assessments, and that these aspects may only become apparent at a later point in time. This is consistent with the way that societal norms change over time (e.g., what was seen as ''normal'' decision grounds in the past is now often delineated as being unacceptable).

3) AI IMPERMANENCE-RELATED IMPLICATION NO. 6

Assessors can only limitedly declare systems' conformity when the system is extended across regions or organizations.

Organizations frequently retrain their AI systems to extend the scope of application (e.g., to apply the system to other processes or regions). The interviewees reported that relevant aspects in the data related to specific regions or contexts that can impact AI's decision-making are often overlooked. For instance, in the case of a language translation model, context-specific data such as idiomatic expressions, colloquialisms, or cultural references may be overlooked. Similarly, in a medical diagnosis system, data related to specific diseases prevalent in a particular region or demographic may be ignored. As a consequence, interviewees told us that system extensions (e.g., transfer learning) frequently failed and models need to be rolled back. This is because the AI system's performance is reliant on the data it was trained on, and without adequate representation of the new region or context, the model may not generalize well to the new data.

The extendibility of AI systems also impacts the conclusions of assessors. The proof of conformity issued after system assessment is limited with respect to the extendibility and reuse of AI systems for other regions and contexts. If an AI system is leveraged in a similar context and domain, then the proof of conformity is likely valid. If not, the proof of conformity should communicate the respective limitations in validity. For example, an AI service that is assessed but then runs on different customer data, environments, or domains may bear the risk of adverse AI behavior, despite conducting an AI assessment that checks the general service offering for conformity. *''If someone else wants to use it [AI Service] and wants to be attested, then in such an assessment, it would be checked whether it is a certified tool. If not, then he [the organization] would need to provide evidence that the tool can be used this way.*'' [Manager, AI Audit].

On the basis of our interview results, we conclude that further research is needed on the transferability of AI system assessment results when leveraging these systems in different contexts. We advise that the changing factors surrounding the system when used in different settings, organizations, and use cases need to be carefully assessed. The scope and boundary conditions of assessment results should be clearly specified to increase the truthfulness and reliability of the assessment's assurance signal.

C. WHEN AND HOW TO REASSESS: AI IMPERMANENCE IMPACTS THE AI ASSESSMENT VALIDITY PERIOD

Surveillance functions [56] include reassessment activities to ensure that guarantees remain valid over time. However, reassessing AI systems is challenging because of their impermanent nature and natural vulnerability to (external) influences, such as user interactions.

1) AI IMPERMANENCE-RELATED IMPLICATION NO. 7

AI assessment validity depends on changes in user input and external conditions, and not only on changes in the system itself.

It is challenging to determine how long AI assessment should remain valid before reassessment is needed. AI specifics pose unique challenges because assessment validity is not only tied to system changes but also impacted by changes in the input data from users, the real world, and its operating environment. Real-world data can change abruptly at a certain (unforeseeable) point in time, which can also depend on the context for system operation. The interviewees frequently observed how assessment validity was impacted by context-dependent factors, such as weather, light, or production machines that degrade over time. Thus, the setting on which AI was trained no longer existed. At some point, the system is confronted with new input data that no longer match those on which the system initially trained. For example, if the consumption of rolls is registered at a given point in time, such as three years ago, and an AI technique perfectly predicts the number of rolls to bake, then these predictions are most likely invalid today. Unlike that of traditional systems, AI systems' performance is expected to severely decline either slowly over time or abruptly because of changes in the real world.

Considering fast and evolving changes in the real world, it can thus be risky to grant long assessment validity periods to AI systems, even if the systems themselves have not changed. Foreseeing a general AI assessment validity as long as, for instance, no retraining takes place cannot prevent validity issues that arise from changes in the input data over time. In fact, this situation can even provide a false sense of security to AI customers, who rely on assessment attestations of system performance and intended system behavior. AI assessment cannot provide long-term guarantees because assessors have to be certain that the environment will not change and is also not affected indirectly, although performance decreases over time. The impermanent nature of AI therefore conflicts with traditional assessment considerations that foresee assessment validity as long as the system remains unchanged and require reassessment during surveillance upon substantial system change. For instance, data protection certifications based on the EU General Data Protection Regulation (GDPR) foresee a validity period of three years (including yearly surveillance assessment) if there are no substantial system changes (Art. 42 GDPR). These traditional considerations do not account for external factors, such as changing user input, that are unique to AI systems and severely impact assessment validity. *"I cannot think of a single use case where the data basis would never change"* [Machine Learning Engineer].

On the basis of our interview findings, we suggest that AI assessment cannot have a fixed validity period (e.g., three years) that is set independently from the specific system. We especially advise against traditional assessment premises such as *"no substantial system change appeared"* to justify long assessment validity periods. In the AI context, assessors need to determine the validity period individually for each system by considering the AI impermanence caused by changing system environments and the corresponding data.

The interviewees indicated that defining the right validity period and surveillance intervals is challenging because they depend on the speed of deviation of the input data and the corresponding impact. It is a challenge for assessors to control for changes in the system environment and to set validity periods that are grounded in research and match the AI system. AI assessors need to monitor a broad set of factors that cause input data mismatches, including unexpected user input if the AI system is not applied as intended, hardware changes to generate input data (e.g., cameras to capture images and scanners to digitize data), and changes in the operating environment (e.g., temperature and light). This complexity particularly highlights the need for further research on surveilling and inferencing AI system environments and incoming data. In conclusion, we propose that AI assessment validity should be set based on the ongoing surveillance of changing input data and system environments, in addition to AI system changes.

2) AI IMPERMANENCE-RELATED IMPLICATION NO. 8

AI systems are intended to change due to their self-learning nature. Surveillance is thus required to allow for and cope with substantial changes in systems.

AI systems are designed to learn and adapt during operation based on new input data. This self-learning and impermanent nature of AI systems is unique compared with traditional systems, which evolve through planned and controlled (continuous) changes [11]. A self-learning system is confronted with specific triggers during operation (e.g., deviating real-world data) and then collects its training data for the learning iteration, typically without additional human intervention (e.g., developers engaging in data cleaning to remove those biases reintroduced by the novel data). This self-learning nature of AI systems, which reflects purposeful and intended changes through adaptation, conflicts with traditional assessment practices. They typically require reassessment only in the case of substantial system changes to ensure assessment validity. Instead, the evolution of AI systems requires the continuous reassessment of the AI system to maintain validity. However, AI assessors face the several challenges when implementing the requirement to reassess AI systems at a high frequency (i.e., continuously) during surveillance: (1) assessors' inability to assess triggers for changes, (2) the balancing of the cost-to-benefit ratio of reassessment (i.e.,

continuous reassessment is very cost intensive and requires assessors to perform frequent surveillance), and (3) reassessment must consider the entire AI system.

First, it is challenging for assessors to define when and how often reassessment during surveillance needs to take place because learning triggers are not known to them. Assessors cannot easily classify changing system behavior as adverse. AI system changes caused by continuous learning during operation can hardly be predicted in advance. Assessors need to foresee triggers causing the system to learn and must then cater to flawed data (e.g., distribution) or incoming data in advance to make systems robust. As the system is also prone to manipulated input data (e.g., malicious user interaction or data), assessing AI systems for robustness and preventing undesirable change is an active research topic [59]. Since the self-learning system autonomously decides when to change on the basis of triggers, the system's performance and predictions that it generates as outcomes can differ owing to the updated logic within small or large time intervals. This self-learning concept deviates from the notions of updating and versioning that are assessed in traditional systems. Self-learning systems differ even from systems that are developed in an agile manner (e.g., continuous integration and deployment pipelines) because the AI system changes that stem from self-learning are typically not steered by humans. If, for instance, a cloud system changes, then developers have built and tested the service update prior to deployment. They are also aware of what they intend to change with their intervention. For traditional systems, there is less uncertainty about how the new addition will impact the system, its functioning, and the outcomes that it generates (e.g., performance impacts). However, self-learning AI systems differ in that they are typically not steered by humans but react to external changes autonomously and adapt their system logic based on new data.

Furthermore, the interviewees highlighted that it is not sufficient to reassess only parts of the system after changes but that very often, the entire system needs to be reassessed to understand the potential risks of changes in every round of learning. Otherwise, assessors cannot fully examine the impact of the newly added data. These extensive reassessment procedures are among the main reasons why organizations refrain from using continuously learning systems where they do not have the time (and budget) to perform extensive assessment while the system evolves.

Challenges regarding when and how to reassess show that AI assessment requires adapted surveillance methods. Based on our interview findings, we recommend that AI assessment must embrace AI learning capabilities for technological advancement rather than cutting or restricting learning functionality to better fit with traditional assessment practices. The interviewees indicated that it is challenging to develop new AI solutions for the market that leverage self-learning AI with the lack of means to undergo assessment and gain proof of conformity (e.g., a certificate). *"This is a very special reason, almost the reason why there have to be extra standards for AI because of these changes. Learning is one reason; of course, this goes hand in hand with change"* [Manager, AI Audit]. Restricting AI self-learning due to the lack of suitable assessment methods undermines the potential positive impact of AI technologies on society and the economy. We especially advise researchers and practitioners to investigate how continuously changing AI systems can be assessed adequately rather than to exclude continuous learning from assessment. For example, assessors may employ adequate sampling practices or rely on automated assessment methods. In the end, self-learning AI systems aim to improve their performance through changes and therefore should, in most cases, strengthen the reliability of assessment results.

Assessors need to carefully consider the cost-to-benefit ratio for continuous assessment prior to introducing (unnecessary) market barriers. Allowing AI to learn and then benefiting from its self-learning capabilities may entail accepting certain risks. The development of distinct assessment methods such as certifications for self-learning systems that report risks and ethical considerations in a transparent manner (e.g., self-learning advantages vs. potential negative effects) can be a starting point. Additionally, educating customers about the learning aspects of AI systems to understand the limitations of AI assessment, appropriate methods for surveillance, and dealing with undesirable behavior can help to set the right expectations of customers about assessments.

V. DISCUSSION

In this study, we sought to address prevalent knowledge gaps about the nature of AI impermanence and issues related to assessment methods that cannot address the peculiarities of AI systems because of their impermanence (e.g., [10], [13]). We conducted a qualitative, interview study to reveal how AI impermanence impacts assessment, ultimately demonstrating the Achilles' heel of AI assessment to provide guarantees on systems' compliance with specific requirements. We derived three key categories of assessment implications and detail eight AI impermanence-related implications along the assessment functions of determination, review and attestation, and surveillance. Section VI and Online Appendix B summarizes this study's implications, recommendations, and suggestions for future research to guide the discourse on how to resolve or at least mitigate AI assessment limitations stemming from AI impermanence.

A. REFLECTION AND DISCUSSION OF RESULTS

Our study highlights the need for a paradigm shift in AI assessment, as traditional methods are insufficient to address the implications of AI impermanence. We next critically discuss our key insights and argue that adapted assessment practices are necessary to ensure the safety and compliance of AI systems, while also considering the economic viability of AI adoption.

1) AI IMPERMANENCE-RELATED IMPLICATIONS REQUIRE A SHIFT IN THINKING FROM TRADITIONAL ASSESSMENT METHODS TO ADAPTED PRACTICES FOR AI

While there are frequent calls for AI assessment, the limitations introduced by AI impermanence are less common in the public and scientific discourse. Nevertheless, it is crucial to understand what AI assessment can and cannot provide and where further research and action are needed to adapt traditional assessment methods. Our study shows that disappointment with AI assessment can arise because of the consequences of impermanence-related limitations (e.g., mitigating the risk of undesirable AI behavior, rare cases, and performance reduction). AI scandals have proven that applying traditional assessment methods to AI systems is not suitable for guaranteeing system safety and compliance in general. Traditional assessment premises are unsuitable because of AI specifics, such as its dependence on external system factors, the risk of rare cases, and opaque learning processes. Hence, AI assessment requires adapted methods to leverage self-learning AI advantages fully in practice. Our study highlights that a shift in thinking is required because traditional assessment premises and methods can hardly be used to address the implications of AI impermanence. Future research is needed to develop methods for AI assessment that can attest to compliant systems (Online Appendix B and Section VI).

2) AI IMPERMANENCE-RELATED IMPLICATIONS ARE NEGLECTED OR LACK REASONABLE CONCEPTUAL DIFFERENTIATION IN CURRENT REGULATIONS

The interviewees also pointed out the following prevalent controversy: an AI system that is enabled to continuously adapt itself to external factors is often considered less safe than static systems. One reason for this argumentation is that self-learning systems are claimed to be more difficult to handle for assessment (e.g., because continuous learning causes frequent changes during operation). Yet, this narrow perspective neglects that static AI systems cannot be handled appropriately via traditional assessment. In practice, static systems are often permitted to enter markets, whereas self-learning systems are not. Owing to the lack of adapted assessment practices, static AI systems are attested to be safe, even though they are at risk of becoming incompliant over longer periods because of their impermanent nature.

3) ASSESSMENTS AND AI SYSTEMS RISK CREDIBILITY IF THEY DO NOT ACCOUNT FOR IMPERMANENCE-RELATED IMPLICATIONS

Trust in assessment and AI can decrease when further scandals about assessed systems arise and when the full functionality of AI technologies to react and (continuously) learn from real-world data is limited. For example, such a prevalent limitation includes collecting and checking input data before they are fed into the system in bulk, with a delay in checks. Through this restriction, assessors aim to replicate the traditional assessment premises due to a lack of adapted AI assessment methods because continuously incoming new data flows may cause the system to change more frequently than they can assess for secure operations. The absence of suitable reassessment methods for AI effectively prevents the introduction of self-learning AI technologies in markets (including their benefits and merits, for example, in the medical field). If every learning cycle is faced with the risk of outdating assessment, then it becomes impossible for organizations to operate AI systems that continuously learn in compliant ways and with valid attestations. Moreover, restricting the system from learning to avoid reassessment efforts means that the system decreases in performance over time and cannot adjust to new input data, which also affects assessment validity. On the basis of our study's findings and interviewees' statements, we call for a critical reflection on assessment validity assumptions to self-learning systems compared with static systems and the underlying justifications to develop profound grounds for assessing AI systems that focus on safety for organizations and individuals as a result rather than seeking safety in traditional assessment methods and premises that no longer fit AI systems.

4) AI IMPERMANENCE-RELATED IMPLICATIONS CAN ONLY LIMITEDLY BE ADDRESSED AND AT THE SAME TIME SUCH MEASURES ARE OFTEN COST-INTENSIVE, DIMINISHING THE ECONOMIC VIABILITY OF AI ADOPTION

Finally, the interviewees emphasized that the cost-to-benefit ratio of AI assessment should be considered and that the risks and unique situational demands of AI systems should be prioritized over cost-driven decisions to prevent negative effects and AI scandals. Considering, for example, the detection of rare cases. Interviewees argue that rare cases are challenging to detect, because they tend to have a low economic impact at first sight. While it is ethically questionable or prohibited by law to neglect individual cases regardless of economic relevance, fairness and discrimination issues can arise if an AI system processes cases in which some people are better than others. However, from an economic perspective, the weak economic impact of rare cases can affect the cost-to-benefit ratio of performing constant reassessment. Organizations do not know how long they can benefit from reassessment efforts (e.g., having a valid proof of conformity to show to customers and generating additional revenue for AI providers). The interviewees stated that it can be economically unviable for organizations to allocate (substantial) assessment resources. Rather, organizations frequently want to exploit the profit generated by the AI system if their business case is fulfilled via the current system performance. Such organizations are often willing to tolerate declines in performance over time rather than to plan and calculate via reassessment. As a consequence, changes and errors are detected very late, bearing the risk of novel AI scandals. For some use cases that constantly require high levels of performance, this situation may be more critical than for other use cases. The interviewees argued that reassessment should not introduce additional burdens and

TABLE 4. Overview of practical actions to adress AI impermanence-related implications for assessment.

AI impermanence-related implication	Example Practical Recommendations
No. 1: Distinguishing between desired and undesired changes is challenging for assessors.	• Implement tools to support AI assessment • Use methods to "unblackbox" the AI learning process
No. 2: Assessors cannot anticipate or test (or can anticipate or test only to a limited extent) the external causes of AI impermanence.	• Trace and examine the potential sources of changes • Monitor the AI system and simulate potential changes in the input
No. 3: The limited explainability of self-learning systems hampers reassessment in the case of change.	• Consider the (inherent) limited explainability of AI systems • Rely on explainable AI systems when possible • Accept uncertainty in assessment results when complex AI systems are assessed
No. 4: Assessors cannot anticipate and consider (or can anticipate and consider only to a limited extent) rare cases that can have widespread negative impacts. Rare cases may invalidate the assessment's guarantees and require continuous assessment, which is limited for AI systems.	• Accept that rare cases cannot be ruled out; clearly indicate this assessment limitation • Accept that rare cases cannot be ruled out/cannot be completely present in the training set
No. 5: Assessors need to control for reintroduced human bias over time.	• Accept that assessors may be unconsciously biased, and mitigate this risk though, e.g., skills and diversity
No. 6: Assessors can only limitedly declare systems' conformity when the system is extended across regions or organizations.	• Conduct additional research on the transferability of assessment results across different contexts • Assess the deviating factors surrounding the system when used in different scenarios
No. 7: AI assessment validity depends on changes in user input and external conditions, and not only on changes in the system itself.	• Define the system-specific assessment validity period (opposed to fixed periods) according to potential changes in the system environment, corresponding data, speed of deviation, impact, and mode of surveillance • Do not apply the considerations of a traditional information systems to an AI system to justify longer assessment periods ("no substantial system change") • Consider those factors that cause input data mismatches, such as unexpected user input, sensors, and the environment, as well as errors in the training process/design
No. 8: AI systems are intended to change due to their self-learning nature. Surveillance is thus required to allow for and cope with substantial changes in systems.	• Balance the advantages of self-learning AI and the risk of imperfect assessment

efforts if they are not critically needed. For example, an organization can consciously decide based on its own (economic) risk to not allocate its budget for further reassessment if self-learning-induced changes do not lead to negative effects on humans (e.g., sorting differently colored tomatoes with image recognition). Yet, not considering reassessment for economic reasons is unacceptable for critical AI applications (e.g., smart factory robots in human assembly lines, credit loans, and hiring decisions).

B. CONTRIBUTIONS TO RESEARCH

Our study contributes to research in many ways. First, we complement the literature that has pointed to but not examined AI impermanence-related assessment challenges. Prior research consistently agrees that learning aspects and a lack of reliability are inherent to AI and impede its assessment [10], [14], but how assessment is affected by AI impermanence-related challenges remains unclear in the scientific literature. Our interview study addresses this shortcoming and identifies three categories of AI impermanence-related implications for AI assessment on the basis of the consequences experienced by interviewees and their attempts to mitigate the risk of undesirable AI behavior. Expert interviews with practitioners and researchers provide us with rich insights into AI impermanence (Online Appendix A) so that we can learn about the concrete implications of AI impermanence on the basis of professional experience.

Our study reveals eight AI impermanence-related implications along three important assessment functions that are experienced and discussed among internal and external AI assessors in practice as well as AI researchers from academia. First, we uncover the implications of AI impermanence for the assessment lifecycle and explain why it is challenging for assessors to evaluate the consequences of a change during determination activities. Second, we show how AI impermanence threatens assessment validity and question the reliability of assessment results. Third, we trace why AI assessment validity periods are challenging to determine because AI impermanence introduces doubts about when and how to reassess. We thereby contribute to the research by providing novel insights into the crucial phenomenon of AI impermanence, particularly by highlighting the implications of impermanence along with explaining its underlying causes. We also answer recent calls for more research on AI impermanence [10], [14].

With this study, we not only identify assessment implications arising from AI impermanence but also, more importantly, explain how they emerge and propose corresponding research fields to develop solutions in Section VI and Online Appendix B. Our results have several crucial implications, such as the need for continuous monitoring not only of the AI system itself but also of changes in the real world that impact the ability of AI systems to make accurate predictions. We show that AI impermanence is inherent to AI systems and that its causes cannot be fully mitigated with the help of current AI assessment methods. We also provide the first evidence of and discuss limitations in mitigating AI impermanence-related implications for AI assessment, uncovering novel boundary conditions for AI assessment. We further advance the research by providing diverse starting points by proposing RQs to effectively address the causes of AI impermanence to make

TABLE 5. Overview of future research endeavors to adress AI impermanence-related implications for assessment.

Future Research Endeavor	Related Research Areas	Example RQs for future research
Objectives	AI robustness e.g., [61], [22], [62]	How can assessment ensure AI system corrections that become needed, accidentally at an unforeseeable point in time?
	Fairness e.g., [63], [64], [65], [66]	How can conflicting requirements be assessed? How to select the most relevant requirement to assess?
	Continual learning e.g., [67], [68]	What should be assessed for continuously evolving AI systems, and when?
	Diversity and inclusion e.g., [88][30]	How should assessment deal with existing biases and their reintroduction in AI systems over time?
Methods	XAI e.g., [89]	How can changes in self-learning AI systems be kept understandable so that explainability is assured? Can changes and corresponding impact on AI systems be predicted?
	AI testing e.g., [69], [70]	What techniques and data can be used to reliably assess AI systems, including rare cases?
	Data governance e.g., [71], [72]	How can large-scale datasets be assessed and how they impact AI systems?
	Soft computing e.g., [73], [74]	How to deal with the uncertainty of AI assessments? How to quantify the reliability of an AI assessment? How to deal with advanced AI architectures?
	Continuous certification e.g., [20], [81], [23], [90]	How can the time frame that AI systems remain stable be identified?
Lifecycle	Anomaly and drift detection, forecasting e.g., [75], [76], [77], [78]	How can renewed assessment be managed after improvements? Can changes and corresponding impact on AI systems be predicted?
	Classifier and ensemble selection e.g., [79], [80]	How can the impact of changes be mitigated and AI system compliance be constrained?
	Certification labeling e.g., [37], [84], [85], [38], [86], [87]	How can the assessment results be transferred between different technological set-ups?
	MLOps e.g., [81], [82], [83]	How can the artifacts tracking of MLOps help AI assessment? How can an MLOps-based AI system be evaluated?

this phenomenon more manageable for assessment (Online Appendix B). Future research can build on this study's findings and address open research challenges to adapt existing methods and develop novel assessment methods that mitigate AI impermanence-related implications for AI assessment.

C. IMPLICATIONS FOR PRACTICE

Our results emphasize that AI systems cannot be considered finalized products but that managing AI impermanence requires continuous observation, monitoring, and maintenance. For AI developers and system owners, we help raise their awareness of the fact that they can be confronted with AI system changes, and we provide concrete recommendations for assessment (Table 4).

Recommended actions for practice to address challenges in *distinguishing between desired and undesired changes* (no. 1) are to implement tools to support AI assessment, and use methods to ''closed box'' the AI learning process. To tackle the *hinderance of external causes leading to AI impermanence* (no. 2) we propose tracing and examining the potential sources of changes, and continuously monitoring the AI system and simulating potential changes in the input. We suggest to consider the (inherent) limited explainability of AI systems, relying on explainable AI systems when possible, but also accepting the inevitable uncertainty in assessment results when complex AI systems are assessed (no. 3).

Actionable recommendations for managing *assessment limitations related to rare cases* (no. 4) comprise accepting that rare cases cannot be ruled out and should thus be clearly stated as assessment limitation during communication. A key recommended action for practice *to control for reintroduced human bias over time* (no. 5) is to accept that assessors may be unconsciously biased and hence training is required to improve assessors' skills, besides increasing the diversity of assessment teams and engaging in related mitigation strategies. We argue that assessors can only *limitedly declare systems' conformity when extended across regions or organizations* (no 6.) and therefore need to conduct additional research on the transferability of assessment results across different contexts, and assess the deviating factors surrounding the system when used in different scenarios.

Proposed strategies to *address AI assessment validity* (no. 7) refer to defining the system-specific assessment validity period (opposed to fixed periods) according to potential changes in the system environment, corresponding data, speed of deviation, impact, and mode of surveillance; refraining from applying the considerations of traditional system assessments to an AI system to justify longer assessment validity periods; and considering those factors that cause input data mismatches, such as unexpected user input as well as errors in the training process/design. To deal with *AI's self-learning nature* (no. 8), our suggestions rely on a balancing of the advantages of self-learning AI and the risk of imperfect assessment. Consequences can occur, such as reduced performance or discrimination, despite assessment, which requires practitioners to perform close monitoring throughout the AI lifecycle.

As the interviewees claimed and regulators foresee, whenever an AI system changes substantially, it needs to be reassessed. This study's results show how changes can occur at any point in the lifecycle of AI systems. Therefore, it is impossible to provide guarantees over usual assessment validity periods (e.g., three years), but a continuous or trigger-

TABLE 6. Overview of related work.

Ref.	Focus	AI assessment lifecycle				Raised AI impermanence-related challenges	Solutions
		Selection	Determination	Review and attestation	Surveillance		
Anisetti, Ardagna, Bena, et al. [4]	Preliminary certification scheme for AI-based applications	√	≈	≈	×	- Lack of modeling techniques of AI system - Lack of precise specification of expected system behavior and properties to be assessed - Need to re-define the assessment process	Three-fold assessment of dataset, training process, and resulting model
Arnold et al. [37]	(Self-)declaration of conformity of AI systems	≈	≈	×	√	- Lack of means to communicate countermeasures adopted in the AI system	(Self-)declaration of conformity of the AI system, in terms of basic performance, safety, security, and lineage; including how the claims have been verified
Benedick et al. [39]	Assessment of robustness in univariate time-series classification	×	√	√	×	- Lack of assessment in real-world conditions - Inference-time data can change and impact performance - Unclear relation between dataset and performance for predicting degradation	Injection of real-world perturbations and evaluation of corresponding performance
Dahmen et al. [16]	Real-world conditions testing of AI systems	√	√	×	×	- Lack of assessment in real-world conditions - Lack of precise specification of expected system behavior and properties to be assessed	Five-steps process to generate simulated real-world testing scenarios; generated data are also used for training or validation
De Bruijn et al. [35]	Socio-technical challenges and solutions of XAI	×	×	≈	√	- Lack of alignment between explanations and real-world users' knowledge and expectations - Explanations are context and time-dependent	Roadmap for trusted XAI for creating trust and legitimacy in AI systems, including end-to-end explanations and values embedding
Gebru et al. [84]	(Self-)declaration of dataset characteristics	≈	≈	×	≈	- Lack of focus on training data despite its importance	(Self-)declaration of datasets characteristics
Jenn et al. [91]	Assessment of safety-critical AI systems	≈	≈	≈	×	- Lack of precise specification of expected system behavior and properties to be assessed - Unclear relation between AI system quality and dataset quality - Lack of alignment between explanations and real-world users' knowledge and expectations	Research roadmap on i) non-AI systems showing similar challenges, and ii) AI techniques which already support some desired properties
Mitchell et al. [38]	(Self-)declaration of conformity of AI models	≈	≈	×	≈	- Lack of standards to report the characteristics of AI systems to increase transparency	(Self-)declaration of conformity of AI models, including how the claims have been verified
Mökander and Floridi [2]	Audit process for trustworthy AI	√	≈	×	≈	- Lack of assurance techniques for evolving AI systems - Lack of assessment triggers	Continuous ethics auditing with human oversight
Picard et al [36]	Certification scheme for datasets	√	√	≈	×	- Unclear relation between AI system quality and dataset quality - Lack of agreed-upon definitions of quality	Assessment of dataset quality following a precise set of requirements and corresponding activities
Winter et al [33]	Certification scheme for AI systems	√	√	√	≈	- AI models are perceived as black-box, and their behavior can only be investigated in terms of input-output	Auditing-based certification process based on the identification of the target AI system criticality level

TABLE 6. (*Continued.*) Overview of related work.

						- Inference-time data can change and impact performance - Lack of precise specification of expected system behavior and properties to be assessed	
Yap [34]	Certification scheme for AI systems	√	√	≈	×	- Lack of precise specification of expected system behavior and properties to be assessed - Need to re-define the assessment process	Common Criteria-based certification scheme assessing adherence to best practices and compliance to desired properties
This Study	AI impermanence impact on AI assessment according to interviews	√	√	√	√	- How to assess consequences of change - How reliable are assessment results - When and how to reassess	Roadmap for future research
Note: "√" means that a phase is completely analyzed, "≈" means that it is partially analyzed but it is not the focus of the paper, "×" means that it is not analyzed							

point-based reassessment will likely be required from an AI lifecycle perspective after such changes. We therefore recommend that assessors adapt traditional assessment methods to consider AI impermanence through continuous assessment. Our study also alerts assessors that providing upfront guarantees is a challenge, as at every phase of the AI lifecycle, reassessment can become a requirement. For example, if new or changed underlying technology is used for development or if the statistical population needs to be updated and retrained because the real world has changed. Therefore, practitioners need to consider the types of changes that can arise for the AI system to ensure that such changes are detected.

D. LIMITATIONS

Our study is subject to limitations. First, our study has limitations concerning the number and depth of interviews we conducted to gather information on AI impermanence-related implications for assessment. Our sample of interviewees is relatively diverse in terms of their backgrounds, ranging from data scientists to auditors and researchers. However, at the same time, our sample is homogenous given that most organizations and interviewees operate and live in Europe; thus, conducting further interviews to increase the generalizability of our findings is recommended. In addition, our study may be subject to interpretation and selection biases due to the ambiguity of the language used or the involvement of the interviewers' perspective when constructing knowledge [60]. We find that the implications of AI assessment are repeatedly mentioned across interviewees. However, it is possible that there is further information concerning the implications that has not yet been identified. For example, potential unique industry or sector specific implications that did not become evident across industries. The interviewees may have found it difficult to verbalize some challenges related to AI impermanence because of its conceptual ambiguity and novelty. For example, while interviewees stated that for assessing the risks and consequences of changing AI behavior, the AI type is not essential, we acknowledge that assessment methods and respective metrics depend on the specific AI technologies used. In particular, our explorative interview study does not consider different AI learning techniques (e.g., supervised and unsupervised). Future research may compare AI learning techniques to better explain AI impermanence-induced changes by design. We acknowledge that AI technologies and related research are constantly evolving at a fast pace and therefore may impose additional implications for AI impermanence or resolve identified implications.

Finally, in this initial study on AI impermanence and its impact on AI assessment, we do not examine potential solutions to accelerate AI assessment advancement, thereby mitigating undesirable AI behavior.

VI. FUTURE RESEARCH

We identify three main areas of future research endeavors and corresponding research areas which can contribute to address the implications raised in this paper (Table 5).

A. ASSESSMENT OBJECTIVES

Future research may identify and unambiguously model the (non-functional) requirements to be assessed. Some requirements such as robustness can have multiple definitions with varying relevance depending on the context [22], [61], [62]. Other requirements such as fairness can be conflicting (e.g., with accuracy) and need to be balanced and aligned with stakeholders' expectations [63], [64], [65], [66]. Research on continual learning should lead to the understanding of how requirements definitions and assessment evolve over time [67], [68].

B. ASSESSMENT METHODS

Future research can contribute to address AI impermanence-related implications by building on existing testing and verification techniques while focusing also on the quality of the corresponding results. On the one hand, existing AI

testing techniques (e.g., [69], [70]) are the cornerstone to start with, paired with data governance to analyze and manage the large amount of data involved [71], [72]. Data governance could also contribute to reducing impermanence when data are thoroughly managed. On the other hand, soft computing techniques are commonly used to address uncertainty (e.g., [73], [74]), and should be adapted to measure the uncertainty associated with an AI assessment.

C. ASSESSMENT LIFECYCLE

On the one hand, existing techniques for anomaly detection and forecasting (e.g., [75], [76], [77], [78]) should be adapted to predict possible changes and evaluate their potential impact on the results of past assessments. Similarly, classifier and ensemble selection techniques (e.g., [79], [80]) should be adapted to constrain the evolution of AI systems to not invalidate those results. On the other hand, existing tooling such as Machine Learning Operations (MLOps) (e.g., [81], [82], [83]) should be adapted to track AI systems evolution and possibly the corresponding assessment results, then presented using labeling techniques (e.g., [37], [38], [84], [85], [86], [87]).

In particular, developing and evaluating continuous assessment methods seem promising for mitigating the risks of changing AI behavior, which has also been proposed in the context of everchanging cloud services (e.g., [24]).

VII. RELATED WORK

Prior research already explored AI assessment, from socio-technical analyses to low-level techniques to assess specific properties (Table 6).

Research first focused on increasing the transparency of AI models by disclosing information how they have been developed and, possibly, assessed. Gebru et al. [84] introduced the notion of *Datasheet*, which describes the dataset used for training. Mitchell et al. [38] refined it with *Model Cards*, which include details on datasets, training process, and resulting AI model. Arnold et al. [37] proposed *FactSheets*, which is associated with a specific AI service and extends the aforementioned approaches with additional details on supported non-functional properties such as explainability and fairness. Each change to the service triggers the release of a new, updated FactSheet. Model Cards and FactSheets also require to specify how the disclosed information has been (self-)assessed.

However, self-assessment techniques remain limited, and research then focused on third-party assessments, namely certification and audit. For instance, Picard et al. [36] defined a certification scheme to assess the quality of the datasets used for AI training. It builds on i) the collection of quality requirements ii) refined for the specific system, and iii) a set of assessment activities to be performed, either automatically or manually.

Yap [34] identified the challenges of certifying AI system in contrast with the certification of traditional software-based systems, which are mainly in terms of lack of formal system specifications and means for system assessment. The challenges are addressed by a preliminary certification scheme built on *Common Criteria* which certifies *i*) adherence to common best practices in the AI development process, and *ii)* compliance to desired non-functional properties such as fairness and privacy. Anisetti et al. [4] identified similar challenges and then proposed a certification scheme for AI-based applications. The scheme performs a *multi-dimensional* assessment of the target AI system, evaluating the training and testing data, the training process, and the resulting model. Winter et al. [33] envisioned a certification scheme which organizes the target AI system according to the level of criticality, mostly in terms of the possible harm it can cause to humans. Each level is associated to a set of requirements to be met, then verified by human auditors following an *audit catalog*. The assessment shall be repeated periodically and in response to major changes in the applications. Jenn et al. [91] identified the challenges of certifying AI-based embedded systems, mainly in terms of lack of specification and reliance on data. They then proposed to address these challenges by *i*) studying how non-AI systems, which share common certification issues to AI systems, have been successfully certified, and *ii)*focusing on AI techniques where some desired properties can be easily supported (e.g., decision trees and explainability), possibly with the help of formal methods. Mökander and Floridi [2] focused on audit as the means to assess AI ethics. The envisioned audit process verifies whether the target AI system is compliant against a set of principles and norms, to, possibly, state that such AI system is *trustworthy*. The audit process shall then be *continuous*, and its results used to (re-)improve the AI system.

Other works focused on the assessment under specific circumstances or against specific properties. For instance, Benedick et al. [39] addressed the issue of inference-time data unexpectedly deviating from training-time patterns due to perturbations or shift in the distribution, decreasing the performance of the AI model. The proposed assessment process is built on the injection of realistic perturbations in the training data. Authors also observed that the impact of inference-time data deviation cannot be predicted on the basis of the characteristics of the training dataset only. Dahmen et al. [16] addressed the issue of extensive testing of AI systems in the real world. The proposed five-step process i) identifies the potential risks, ii) describes high-level testing scenarios, iii) translates testing scenarios into low-level machine-readable scenarios where all parameters are defined and digital twins simulating real-world conditions are introduced, and iv) generates concrete testing scenarios by randomly assigning values to the identified parameters. Once the testing scenarios are executed, the generated data are used to i) enrich training and validation datasets and ii) evaluate the depth and breadth of the assessment, in a continuous refinement loop.

Finally, de Bruijn et al. [35] focused on the challenges of XAI from a socio-technical point of view. The challenges refer to the generated explanations that i) may be difficult

to interpret for inexpert users, ii) depend on the context and are used in problems where it may be unfeasible to identify a correct answer, and iii) should change over time. A trusted XAI should then move from explaining the outcome of the AI model to explaining the *decision* took on the basis of the AI model, among the others.

VIII. CONCLUSION

AI impermanence implications can be considered for AI assessments only if they are identified and understood along with their underlying causes rooting in AI's learning nature. This study provides rich descriptions and explanations of eight AI impermanence-related implications and their impact on AI assessment. Thus, improving our understanding of how AI impermanence impedes assessment. We indicate connecting research fields that can inform and contribute to the development of new assessment procedures, criteria, and mechanisms to deal with AI impermanence. This is especially relevant to maintain the credibility of assessments and to manage the expectations to the safety of AI systems as well as the overall reputation of such systems for technological advancement.

REFERENCES

[1] C. d'Angelo, I. Flanagan, I. D. Motsi-Omoijiade, M. Virdee, and S. Gunashekar. (2022). *Labelling Initiatives, Codes of Conduct and Other Self-Regulatory Mechanisms for Artificial Intelligence Applications*. RAND Corporation. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/research_reports/RRA1700/RRA1773-1/RAND_RRA1773-1.pdf

[2] J. Mökander and L. Floridi, "Ethics-based auditing to develop trustworthy AI," *Minds Mach.*, vol. 31, no. 2, pp. 323–327, Jun. 2021, doi: 10.1007/s11023-021-09557-8.

[3] P. Schmidt, F. Biessmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," *J. Decis. Syst.*, vol. 29, no. 4, pp. 260–278, Oct. 2020, doi: 10.1080/12460125.2020.1819094.

[4] M. Anisetti, C. A. Ardagna, N. Bena, and E. Damiani, "Rethinking certification for trustworthy machine-learning-based applications," *IEEE Internet Comput.*, vol. 27, no. 6, pp. 22–28, Nov. 2023, doi: 10.1109/MIC.2023.3322327.

[5] European Parliament, Artificial Intelligence Act. *Corrigendum*. Accessed: Jul. 4, 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf

[6] (2023). *Lord Holmes of Richmond, Artificial Intelligence (Regulation) Bill (HL)*. [Online]. Available: https://bills.parliament.uk/publications/53068/documents/4030

[7] Commitee on Technology. (2021). *A Local Law To Amend the Administrative Code of the City of New York, in Relation To Automated Employment Decision Tools*. [Online]. Available: https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9&Options=ID%7cText%7c&Search=

[8] P. Susarla, D. Purnell, and K. Scott, "Zillow's artificial intelligence failure and its impact on perceived trust in information systems," *J. Inf. Technol. Teaching Cases*, Sep. 2024, Art. no. 20438869241279865, doi: 10.1177/20438869241279865.

[9] A. Datta. (Dec. 13, 2021). *The 500mm+ Debacle At Zillow Offers-What Went Wrong With the AI Models?*. [Online]. Available: https://insideainews.com/2021/12/13/the-500mm-debacle-at-zillow-offers-what-went-wrong-with-the-ai-models/

[10] F. A. Batarseh, L. Freeman, and C.-H. Huang, "A survey on artificial intelligence assurance," *J. Big Data*, vol. 8, no. 1, pp. 1–30, Dec. 2021, doi: 10.1186/s40537-021-00445-7.

[11] N. Berente, B. Gu, J. Recker, and R. Santhanam, "Managing artificial intelligence," *Manag. Inf. Syst. Quart.*, vol. 45, no. 3, pp. 1–64, 2021.

[12] K.-J. Hsiao, Y. Feng, and S. Lamkhede. *Foundation Model for Personalized Recommendation*. Accessed: Jun. 4, 2025. [Online]. Available: https://netflixtechblog.com/foundation-model-for-personalized-recommendation-1a0bd8e02d39

[13] B. J. Taylor, M. A. Darrah, and C. D. Moats, "Verification and validation of neural networks: A sampling of research in progress," *Proc. SPIE*, vol. 5103, pp. 8–16, Aug. 2003.

[14] L. Myllyaho, M. Raatikainen, T. Männistö, T. Mikkonen, and J. K. Nurminen, "Systematic literature review of validation methods for AI systems," *J. Syst. Softw.*, vol. 181, pp. 1–22, Nov. 2021.

[15] D. Applegate and M. Koenig, "Framing AI audits," *Internal Auditor*, vol. 76, no. 6, pp. 29–34, Jan. 2019.

[16] U. Dahmen, T. Osterloh, and J. Roßmann, "Generation of virtual test scenarios for training and validation of AI-based systems," in *Proc. IEEE Int. Conf. Prog. Informat. Comput. (PIC)*, Dec. 2021, pp. 64–71, doi: 10.1109/PIC53636.2021.9687075.

[17] L. Käde and S. Maltzan, "Towards a demystification of the black box: Explainable AI and legal ramifications," *J. Internet Law*, vol. 23, no. 3, pp. 3–13, Jan. 2019.

[18] J. Walmsley, "Artificial intelligence and the value of transpar-ency," *AI Soc.*, vol. 36, no. 2, pp. 585–595, Jan. 2021.

[19] J. Vanian. (Jun. 9, 2020). *How Instacart Fixed Its A.I. and Keeps Up With the Coronavirus Pandemic*. [Online]. Available: https://fortune.com/2020/06/09/instacart-coronavirus-artificial-intelligence/

[20] M. Anisetti, C. A. Ardagna, and N. Bena, "Continuous certification of non-functional properties across system changes," in *Service-Oriented Computing* (Lecture Notes in Computer Science), vol. 14419, F. Monti, S. Rinderle-Ma, A. R. Cortés, Z. Zheng, and M. Mecella, Eds., Cham, Switzerland: Springer, 2023, pp. 3–18, doi: 10.1007/978-3-031-48421-6_1.

[21] P. Cihon, M. J. Kleinaltenkamp, J. Schuett, and S. D. Baum, "AI certification: Advancing ethical practice by reducing information asymmetries," *IEEE Trans. Technol. Soc.*, vol. 2, no. 4, pp. 200–209, Dec. 2021.

[22] T. Gehr, M. Mirman, D. Drachsler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, "AI2: Safety and robustness certification of neural networks with abstract interpretation," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2018, pp. 3–18, doi: 10.1109/SP.2018.00058.

[23] I. Kunz and P. Stephanow, "A process model to support continuous certification of cloud services," in *Proc. IEEE 31st Int. Conf. Adv. Inf. Netw. Appl. (AINA)*, Mar. 2017, pp. 986–993, doi: 10.1109/AINA.2017.106.

[24] P. Stephanow and C. Banse, "Evaluating the performance of continuous test-based cloud service certification," in *Proc. 17th IEEE/ACM Int. Symp. Cluster, Cloud Grid Comput. (CCGRID)*, Madrid, Spain, May 2017, pp. 1117–1126, doi: 10.1109/CCGRID.2017.134.

[25] S. Yanisky-Ravid and S. K. Hallisey, "Equality and privacy by design: A new model of artificial intelligence data trans-parency via auditing, certification and safe harbor regimes," *Fordham Urban Law J.*, vol. 46, no. 2, pp. 428–486, Jan. 2019.

[26] A. Rai, P. Constantinides, and S. Sarker, "Next-generation digital platforms: Toward human-AI hybrids," *Manag. Inf. Syst. Quart.*, vol. 43, no. 1, pp. 3–9, Jan. 2019.

[27] Independent High-Level Expert Group On Arti-ficial Intelligence Set Up By The European Commission. (2019). *A Definition of AI: Main Capabilities and Disciplines*. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[28] Y. K. Dwivedi et al., "Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *Int. J. Inf. Manage.*, vol. 57, pp. 1–47, Apr. 2021, doi: 10.1016/j.ijinfomgt.2019.08.002.

[29] C. Collins, D. Dennehy, K. Conboy, and P. Mikalef, "Artificial intelligence in information systems research: A systematic literature review and research agenda," *Int. J. Inf. Manage.*, vol. 60, Oct. 2021, Art. no. 102383.

[30] D. Monett and C. W. P. Lewis, "Getting clarity by defining artificial intelligence—A survey," in *Philosophy and Theory of Artificial Intelligence* (Studies in Applied Philosophy, Epistemology and Rational Ethics), V. C. Müller, Ed., Cham, Switzerland: Springer, 2018, pp. 212–214, doi: 10.1007/978-3-319-96448-5_21.

[31] J. Lansing, A. Benlian, and A. Sunyaev, "'Unblackboxing' decision Makers' interpretations of IS certifications in the context of cloud service certifications," *J. Assoc. Inf. Syst.*, pp. 1064–1096, Jan. 2018.

[32] J. Luffarelli and A. Awaysheh, "The impact of indirect corporate social performance signals on firm value: Evidence from an event study," *Corporate Social Responsibility Environ. Manage.*, vol. 25, no. 3, pp. 295–310, May 2018, doi: 10.1002/csr.1468.

[33] P. Matthias Winter, S. Eder, J. Weissenböck, C. Schwald, T. Doms, T. Vogt, S. Hochreiter, and B. Nessler, "Trusted artificial intelligence: Towards certification of machine learning applications," 2021, *arXiv:2103.16910*.

[34] R. H. C. Yap, "Towards certifying trustworthy machine learning systems," in *Trustworthy AI- Integrating Learning, Optimization and Reasoning* (Lecture Notes in Computer Science), vol. 12641, F. Heintz, M. Milano, and B. O'Sullivan, Eds., Cham, Switzerland: Springer, 2021, pp. 77–82, doi: 10.1007/978-3-030-73959-1_7.

[35] H. de Bruijn, M. Warnier, and M. Janssen, "The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making," *Government Inf. Quart.*, vol. 39, no. 2, pp. 1–8, Apr. 2022.

[36] S. Picard, C. Chapdelaine, C. Cappi, L. Gardes, E. Jenn, B. Lefevre, and T. Soumarmon, "Ensuring dataset quality for machine learning certification," in *Proc. IEEE Int. Symp. Softw. Rel. Eng. Workshops (ISSREW)*, Oct. 2020, pp. 275–282, doi: 10.1109/ISSREW51248.2020.00085.

[37] M. Arnold, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM J. Res. Develop.*, vol. 63, no. 4/5, pp. 6:1–6:13, Jul. 2019, doi: 10.1147/JRD.2019.2942288.

[38] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[39] P.-L. Benedick, J. Robert, and Y. Le Traon, "A systematic approach for evaluating artificial intelligence models in industrial settings," *Sensors*, vol. 21, no. 18, pp. 1–17, Jan. 2021.

[40] K. Smolander, M. Rossi, and S. Purao, "Software architectures: Blueprint, literature, language or decision?" *Eur. J. Inf. Syst.*, vol. 17, no. 6, pp. 575–588, Dec. 2008, doi: 10.1057/ejis.2008.48.

[41] J. W. Creswell, *Qualitative Inquiry and Research Design: Choosing Among Five Approaches*, 3rd ed., Los Angeles, CA, USA: Sage, 2013.

[42] G. Walsham, "Interpretive case studies in IS research: Nature and method," *Eur. J. Inf. Syst.*, vol. 4, no. 2, pp. 74–81, May 1995.

[43] J. M. Corbin and A. L. Strauss, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 4th ed., Los Angeles, CA, USA: Sage, 2015.

[44] C. Abraham, M.-C. Boudreau, I. Junglas, and R. Watson, "Enriching our theoretical repertoire: The role of evolutionary psychology in technology acceptance," *Eur. J. Inf. Syst.*, vol. 22, no. 1, pp. 56–75, Jan. 2013.

[45] M. Wiesche, M. C. Jurisch, P. W. Yetton, and H. Krcmar, "Grounded theory methodology in information systems research," *Manage. Inf. Syst. Quart.*, vol. 41, no. 3, pp. 685–701, Jan. 2017.

[46] L. A. Palinkas, S. M. Horwitz, C. A. Green, J. P. Wisdom, N. Duan, and K. Hoagwood, "Purposeful sampling for qualitative data collection and analysis in mixed method implementation research," *Admin. Policy Mental Health Mental Health Services Res.*, vol. 42, no. 5, pp. 533–544, Sep. 2015.

[47] M. D. Myers, *Qualitative Research in Business & Management*, 2nd ed., London, U.K.: Sage, 2013.

[48] M. Q. Patton, *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*, 4th ed., Thousand Oaks, CA, USA: Sage, 2015.

[49] H. Heath, "Exploring the influences and use of the literature during a grounded theory study," *J. Res. Nursing*, vol. 11, no. 6, pp. 519–528, Nov. 2006, doi: 10.1177/1744987106069338.

[50] M.-C. Boudreau and D. Robey, "Enacting integrated information technology: A human agency perspective," *Org. Sci.*, vol. 16, no. 1, pp. 3–18, Feb. 2005, doi: 10.1287/orsc.1040.0103.

[51] C. Urquhart, H. Lehmann, and M. D. Myers, "Putting the 'theory' back into grounded theory: Guidelines for grounded theory studies in information systems," *Inf. Syst. J.*, vol. 20, no. 4, pp. 357–381, Jul. 2010, doi: 10.1111/j.1365-2575.2009.00328.x.

[52] H. K. Klein and M. D. Myers, "A set of principles for conducting and evaluating interpretive field studies in information systems," *MIS Quart.*, vol. 23, no. 1, p. 67, Mar. 1999, doi: 10.2307/249410.

[53] C. Urquhart and W. Fernández, "Using grounded theory method in information systems: The researcher as blank slate and other myths," *J. Inf. Technol.*, vol. 28, no. 3, pp. 224–236, Sep. 2013, doi: 10.1057/jit.2012.34.

[54] K. Brecker, S. Lins, and A. Sunyaev, "Artificial intelligence systems' impermanence: A showstopper for assessment?" presented at the Workshop Inf. Technol. Syst., 2023, pp. 1–15.

[55] S. Seidel and C. Urquhart, "On emergence and forcing in information systems grounded theory studies: The case of strauss and Corbin," *J. Inf. Technol.*, vol. 28, no. 3, pp. 237–260, Sep. 2013, doi: 10.1057/jit.2013.17.

[56] *Conformity Assessment-Vocabulary and Gen-eral Principles*, Standard 17000:2020, 2020. [Online]. Available: https://www.iso.org/standard/73029.html

[57] S. Gasson and J. Waters, "Using a grounded theory approach to study online collaboration behaviors," *Eur. J. Inf. Syst.*, vol. 22, no. 1, pp. 95–118, Jan. 2013.

[58] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan, "Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jun. 2023, pp. 1369–1385, doi: 10.1145/3593013.3594087.

[59] A. E. Ciná, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli, "Wild patterns reloaded: A survey of machine learning security against training data poisoning," *ACM Comput. Surveys*, vol. 55, no. 13s, pp. 1–39, Dec. 2023, doi: 10.1145/3585385.

[60] M. D. Myers and M. Newman, "The qualitative interview in IS research: Examining the craft," *Inf. Org.*, vol. 17, no. 1, pp. 2–26, Jan. 2007.

[61] M. Anisetti, C. A. Ardagna, A. Balestrucci, N. Bena, E. Damiani, and C. Y. Yeun, "On the robustness of random forest against untargeted data poisoning: An ensemble-based approach," *IEEE Trans. Sustain. Comput.*, vol. 8, no. 4, pp. 540–554, Oct. 2023, doi: 10.1109/TSUSC.2023.3293269.

[62] S. Sharma, J. Henderson, and J. Ghosh, "CERTIFAI: A common framework to provide explanations and analyse the fairness and robustness of black-box models," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Feb. 2020, pp. 166–172, doi: 10.1145/3375627.3375812.

[63] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–38, Jul. 2024, doi: 10.1145/3616865.

[64] S. Park, S. Kim, and Y.-S. Lim, "Fairness audit of machine learning models with confidential computing," in *Proc. ACM Web Conf.*, Apr. 2022, pp. 3488–3499, doi: 10.1145/3485447.3512244.

[65] S. Segal, Y. Adi, B. Pinkas, C. Baum, C. Ganesh, and J. Keshet, "Fairness in the eyes of the data: Certifying machine-learning models," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, Jul. 2021, pp. 926–935, doi: 10.1145/3461702.3462554.

[66] H. Zhao and G. J. Gordon, "Inherent tradeoffs in learning fair representations," *J. Mach. Learn. Res.*, vol. 32, pp. 15649–15659, Nov. 2022. [Online]. Available: https://www.jmlr.org/papers/v23/21-1427.html

[67] Z. Mai, R. Li, J. Jeong, D. Quispe, H. Kim, and S. Sanner, "Online continual learning in image classification: An empirical survey," *Neurocomputing*, vol. 469, pp. 28–51, Jan. 2022, doi: 10.1016/j.neucom.2021.10.021.

[68] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5362–5383, Aug. 2024, doi: 10.1109/TPAMI.2024.3367329.

[69] H. B. Braiek and F. Khomh, "On testing machine learning programs," *J. Syst. Softw.*, vol. 164, Jun. 2020, Art. no. 110542, doi: 10.1016/j.jss.2020.110542.

[70] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Trans. Softw. Eng.*, vol. 48, no. 1, pp. 1–36, Jan. 2022, doi: 10.1109/TSE.2019.2962027.

[71] M. Janssen, P. Brous, E. Estevez, L. S. Barbosa, and T. Janowski, "Data governance: Organizing data for trustworthy artificial intelligence," *Government Inf. Quart.*, vol. 37, no. 3, Jul. 2020, Art. no. 101493, doi: 10.1016/j.giq.2020.101493.

[72] Q. Lu, L. Zhu, X. Xu, J. Whittle, D. Zowghi, and A. Jacquet, "Responsible AI pattern catalogue: A collection of best practices for AI governance and engineering," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–35, Jul. 2024, doi: 10.1145/3626234.

[73] S. Rizvi, J. Mitchell, A. Razaque, M. R. Rizvi, and I. Williams, "A fuzzy inference system (FIS) to evaluate the security readiness of cloud service providers," *J. Cloud Comput.*, vol. 9, no. 1, p. 42, Dec. 2020, doi: 10.1186/s13677-020-00192-9.

[74] E. Di Nardo and A. Ciaramella, "Advanced fuzzy relational neural network," in *Proc. 13th Int. Workshop Fuzzy Log. Appl.*, 2021, pp. 1–6. [Online]. Available: https://ceur-ws.org/Vol-3074/paper27.pdf

[75] P. R. L. Almeida, L. S. Oliveira, A. S. Britto, and R. Sabourin, "Adapting dynamic classifier selection for concept drift," *Expert Syst. Appl.*, vol. 104, pp. 67–85, Aug. 2018, doi: 10.1016/j.eswa.2018.03.021.

[76] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, pp. 1–37, Apr. 2014, doi: 10.1145/2523813.

[77] A. S. Iwashita and J. P. Papa, "An overview on concept drift learning," *IEEE Access*, vol. 7, pp. 1532–1547, 2019, doi: 10.1109/ACCESS.2018.2886026.

[78] L. Caruccio, S. Cirillo, G. Polese, and R. Stanzione, "An RFD-based approach for concept drift detection in machine learning systems," in *Proc. 28th Int. Conf. Extending Database Technol. (EDBT)*, Mar. 2025, pp. 816–828, doi: 10.48786/edbt.2025.66.

[79] R. M. O. Cruz, R. Sabourin, and G. D. C. Cavalcanti, "Dynamic classifier selection: Recent advances and perspectives," *Inf. Fusion*, vol. 41, pp. 195–216, May 2018, doi: 10.1016/j.inffus.2017.09.010.

[80] I. Khan, X. Zhang, M. Rehman, and R. Ali, "A literature survey and empirical study of meta-learning for classifier selection," *IEEE Access*, vol. 8, pp. 10262–10281, 2020, doi: 10.1109/ACCESS.2020.2964726.

[81] D. Knoblauch and J. Großmann, "Towards a risk-based continuous auditing-based certification for machine learning," *Rev. Socionetwork Strategies*, vol. 17, no. 2, pp. 255–273, Oct. 2023, doi: 10.1007/s12626-023-00148-w.

[82] M. Testi, M. Ballabio, E. Frontoni, G. Iannello, S. Moccia, P. Soda, and G. Vessio, "MLOps: A taxonomy and a methodology," *IEEE Access*, vol. 10, pp. 63606–63618, 2022, doi: 10.1109/ACCESS.2022.3181730.

[83] L. Colombi, A. Gilli, S. Dahdal, I. Boleac, M. Tortonesi, C. Stefanelli, and M. Vignoli, "A machine learning operations platform for streamlined model serving in industry 5.0," in *Proc. IEEE Netw. Oper. Manage. Symp.*, May 2024, pp. 1–6, doi: 10.1109/NOMS59830.2024.10575103.

[84] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, and K. Crawford, "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723.

[85] K. J. M. Matus and M. Veale, "Certification systems for machine learning: Lessons from sustainability," *Regulation Governance*, vol. 16, no. 1, pp. 177–196, Jan. 2022, doi: 10.1111/rego.12417.

[86] K. J. Morik, H. Kotthaus, R. Fischer, S. Mücke, M. Jakobs, N. Piatkowski, A. Pauly, L. Heppe, and D. Heinrich, "Yes we care!-Certification for machine learning methods through the care label framework," *Frontiers Artif. Intell.*, vol. 5, Sep. 2022, Art. no. 975029, doi: 10.3389/frai.2022.975029.

[87] N. Scharowski, M. Benk, S. J. Kühne, L. Wettstein, and F. Brühlmann, "Certification labels for trustworthy AI: Insights from an empirical mixed-method study," in *Proc. ACM Conf. Fairness Accountability Transparency*, Jun. 2023, pp. 248–260, doi: 10.1145/3593013.3593994.

[88] G. Vargas-Solar, N. Bennani, J. A. Espinosa-Oviedo, A. Mauri, J.-L. Zechinelli-Martini, B. Catania, C. Ardagna, and N. Bena, "Decolonizing federated learning: Designing fair and responsible resource allocation," in *Proc. IEEE/ACS 21st Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2024, pp. 1–7, doi: 10.1109/AICCSA63423.2024.10912594.

[89] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan, "Explainable AI (XAI): Core ideas, techniques, and solutions," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–33, Sep. 2023, doi: 10.1145/3561048.

[90] S. Lins, S. Schneider, J. Szefer, S. Ibraheem, and A. Ali, "Designing monitoring systems for continuous certification of cloud services: Deriving meta-requirements and design guidelines," *Commun. Assoc. Inf. Syst.*, vol. 44, pp. 406–510, Jan. 2019.

[91] E. Jenn et al., "Identifying challenges to the certification of machine learning for safety critical systems," in *Proc. Eur. Congr. Embedded Real Time Syst. (ERTS)*, 2020, pp. 1–10.

KATHRIN BRECKER is currently a Researcher with the Institute of Applied Informatics and Formal Description Methods, Karlsruhe Institute of Technology (KIT). Prior to completing her Ph.D. at KIT, she started her career at IBM and was a Senior Consultant with the CIO Advisory Unit of KPMG for several years. She works on research challenges concerned with the adoption, usage, and assessment of artificial intelligence systems. Her main research focus is to study how organizations can purposefully and safely leverage artificial intelligence systems.

SEBASTIAN LINS is currently a Full Professor of information systems with the University of Kassel. He works on research challenges concerned with responsible design and use of trustworthy information systems. One of his main areas of work covers ensuring data protection and information security and performing (continuous) security assessments, thereby fostering trust in emerging technologies.

NICOLA BENA is currently an Assistant Professor with the Department of Computer Science, Università degli Studi di Milano. He has been a Visiting Scholar with Khalifa University and INSA Lyon. His research interests are in the area of security of modern distributed systems with particular reference to certification, assurance, and risk management techniques.

CLAUDIO A. ARDAGNA is currently a Full Professor with the Department of Computer Science, Università degli Studi di Milano, the Director of the CINI National Laboratory on Data Science, and the Co-Founder of Moon Cloud S.r.l. He has been a Visiting Professor with Université Jean Moulin Lyon 3 and a Visiting Researcher with BUPT, Khalifa University, GMU. His research interests are in the area of edge-cloud and AI security and assurance, and data science.

MARCO ANISETTI is currently a Full Professor with the Department of Computer Science, Università degli Studi di Milano. His research interests are in the area of computational intelligence, and its application to the design and evaluation of complex systems. He has been investigating innovative solutions in the area of cloud security assurance evaluation. In this area, he defined a new scheme for continuous and incremental cloud security certification, based on a distributed assurance evaluation architecture.

ALI SUNYAEV is currently a Full Professor of computer science with the Technical University of Munich (TUM). Before joining TUM, he was a Professor with Karlsruhe Institute of Technology, the University of Kassel, and the University of Cologne. His research work accounts for the multifaceted use contexts of digital technologies, with research on human behavior affecting IT applications and vice versa. His research appeared in journals, including *Information Systems Research*, *Journal of Management Information Systems*, *Journal of Information Technology*, *Journal of the Association for Information Systems*, IEEE TRANSACTIONS ON CLOUD COMPUTING, and *Communications of the ACM*.