Why it Remains Challenging to Assess Artificial Intelligence

Kathrin Brecker
Karlsruhe Institute of Technology
brecker@kit.edu

Sebastian Lins
Karlsruhe Institute of Technology
lins@kit.edu

Ali Sunyaev
Karlsruhe Institute of Technology
sunyaev@kit.edu

Abstract

Artificial Intelligence (AI) assessment to mitigate risks arising from biased, unreliable, or regulatory non-compliant systems remains an open challenge for researchers, policymakers, and organizations across industries. Due to the scattered nature of research on AI across disciplines, there is a lack of overview on the challenges that need to be overcome to move AI assessment forward. In this study, we synthesize existing research on AI assessment applying a descriptive literature review. Our study reveals seven challenges along three main categories: ethical implications, regulatory gaps, and socio-technical limitations. This study contributes to a better understanding of the challenges in AI assessment so that AI researchers and practitioners can resolve these challenges to move AI assessment forward.

Keywords: Artificial Intelligence, Assessment, Challenges.

1. Introduction

Along with the advancing artificial intelligence (AI) adoption across industries, scandals are rising and reveal various issues in contemporary AI-based systems. For example, the risk of discrimination by AI-based systems such as chatbots or image labelling algorithms has caused public attention on potential issues associated with pervasive AI use (Schmidt et al., 2020). Incidents like these have raised calls in academia and practice to develop AI assessments to mitigate risks (e.g., Cihon et al., 2021).

Complementary to internal assessments performed by the AI provider or users, AI assessments conducted by third parties are particularly valuable because they can provide objective and independent evidence on AI's compliance with accepted regulatory standards or industry best practices. AI assessments are not only a concern of AI providers seeking assurance, but also for society because scandals and the AI debate are impacting people in their role as citizens and users of AI solutions.

Despite the growing calls for AI assessment and its auspicious approach, we observe that assessments remain underexplored and mainly at a prototypical stage so far (e.g., test assessments to evaluate specific AI providers). AI providers, users and third parties still struggle with AI assessment conduction (Ebers et al., 2021). For example, it remains unclear what requirements or best practices should be used to assess AI. Today, novel AI-centered regulations are still under development and existing regulations often do not provide sufficient guidance for applying them to the AI context (Yanisky-Ravid & Hallisey, 2019). At the same time, AI assessment remains challenging in terms of technical aspects, such as explainability and AI algorithm reliability.

Prevalent challenges ultimately hamper the achievement of viable AI assessments today and thus require further attention. A broad and detailed understanding of assessment difficulties would increase awareness and enable regulators, AI providers and assessment initiatives to consider potential pitfalls and to develop corresponding solutions. Without understanding prevalent challenges thoroughly, we cannot overcome the lack of viable means for assessment that we observe today. This is needed because AI providers are lacking clear guidance on provisioning trustworthy AI (TAI), hence they cannot seize business opportunities or take risks (Bajarin, 2020).

An ever-increasing number of research articles discusses AI assessment, but they are spread across various disciplines. Predominantly, information systems (IS), computer science, law, and social sciences. Such research examines, among others, how to interpret AI-related laws (e.g., Ebers et al., 2021), methods for explainable AI (e.g., Hanif et al., 2021), or the societal impact and ethical use of AI (e.g., Havrda & Rakova, 2021). Nonetheless, while knowledge on AI assessment challenges is scattered across disciplines, challenges are highly interrelated and interdisciplinary. For example, once regulators increase clarity on what exactly should be assessed, it remains open if and how AI providers can meet the requirements, and how they can perform the corresponding assessments. The existing research discourse has not yet reached a consensus to resolve challenges and move AI assessment forward (Mökander &

URI: https://hdl.handle.net/10125/103275
978-0-9981331-6-4

Floridi, 2021), leading to calls to further examine AI assessment challenges (e.g., Lu, 2021; Maclure, 2021). Identifying the challenges, their interdependencies, and how they are hampering progress can serve as a foundation for future research to examine how the challenges can be sufficiently addressed. Accordingly, we seek to answer the following research question: *What are the challenges hampering AI assessment?*

We conducted a descriptive literature review (Paré et al., 2014) that synthesizes the scattered knowledge on challenges in AI assessment across disciplines to answer this research question. Overall, the review revealed seven challenges grouped into three challenge categories: ethical implications, regulatory gaps, and sociotechnical limitations. The results further suggest that vicious circles can arise if technology design relies on regulations, while policymakers are facing uncertainties due to AI capabilities that limit oversight and control. As normative discussions are ongoing, ethical and legal standards are left undefined and assessment endeavors are lacking a benchmark (Mökander et al., 2021).

Our structured overview of challenges contributes to research on AI assessment, providing a starting point for developing designs for AI assessment. We highlight implications for how AI assessment can be realized and generate insights that help achieving a better understanding of the current situation and why it persists. This is a contribution for both, practitioners, and researchers across disciplines as it provides answers to the question why establishing AI assessments is such a lengthy process and what challenges exactly need to be overcome to move forward.

The remainder of this article proceeds as follows: The next section offers a brief introduction into AI assessment and related research. Then, we explain the research approach applied to identify challenges in AI assessment. Thereafter, we present the derived challenges in AI assessment according to three categories. The subsequent section summarizes our main findings and discusses this study's implications and limitations as well as starting points for future research. We close with our conclusion.

2. Background

2.1 AI Assessment

We align with a broad AI definition and define 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" (Rai et al., 2019, p. iii).

To mitigate potential risks of AI that have become evident in scandals already (e.g., discriminatory AI, privacy issues, or erroneous predictions), practitioners and academia call for internal and third-party AI assessment (d'Angelo et al., 2022; Mökander et al., 2021). Internal assessments are typically performed by internal auditors or related business units. Internal assessments' user influence is limited, as users are aware that self-generated statements may have lower production costs and can be biased or manipulated by AI providers. In contrast, users perceive third parties as more objective and credible, and third-party assessments are verifiable and considered more reliable (Lins & Sunyaev, 2022). Hence, third-party assessments complement providers' internal assessments efforts to reduce user uncertainty.

There are three types of stakeholders involved in AI assessments: AI providers, assessors, and users. AI providers can perform internal assessments or apply for third-party assessments demonstrating to prospective users that their systems are compliant and of defined quality. Users as individuals or organizations scrutinize assessment results to obtain assurance and overcome uncertainty whether the AI-based system meets the requirements. Assessors can be of various types, such as internal auditors or external parties serving as independent intermediaries between AI providers and users. They assess the AI-based system, including aspects such as reviewing system documentation and protection measures, and conducting employee interviews and onsite assessments (Lansing et al., 2018). If the AI-based system meets assessment requirements, a formal confirmation is issued by assessors and the AI provider is entitled to demonstrate system compliance to the public.

There are various AI assessment initiatives run by organizations, developing instruments such as code of conducts, ethics principles, seals and frameworks. (d'Angelo et al., 2022). Across industries, organizations are developing code of conducts and code of ethics, for example, to ensure adherence to ethics and laws when applying AI, to increase employee awareness and to communicate principles for AI use such as non-discrimination (d'Angelo et al., 2022). Besides, certifications are a long established and widely acknowledged third-party assessment type, enabling organizations to demonstrate that their IS comply with standards. They reduce uncertainties for users as they can rely on a trusted third-party to attest that the system is safe for use, especially when system features cannot be observed (e.g., system bias) (Löbbers et al., 2020).

2.2 Related Research

There is a large body of research on AI assessment. Reviewing research reveals four key research streams related to AI assessments (Table 1).

Table 1. Related research streams on AI assessment.

Research stream	Description	Exemplary studies
Leveraging AI capabilities	Evaluating AI value-add, exploring use cases and technical methods to enhance capabilities	- Lebovitz et al., 2021 - Benedick et al., 2021
Trustworthy AI (TAI)	Defining AI trustworthiness, discussing TAI criteria, and realization	- Mökander & Floridi, 2021 - Eschenbach, 2021
Explainable AI (XAI)	Defining AI explainability, examining technical methods to achieve and assess XAI	- Hanif et al., 2021 - Deeks, 2019
Conducting AI assessments	Defining AI assessment criteria and operational realization e.g., by internal auditors, external third parties	- Cihon et al., 2021 - **This study**

First, various research with a focus on leveraging AI capabilities assesses AI value-add for concrete use cases and seeks to further enhance technical methods and possibilities. For example, Lebovitz et al. (2021) investigate the value-add of AI tools for medical diagnosis at a hospital in terms of expert knowledge and performance of the system, examining whether AI meets expectations to enhance medical diagnostics.

Second, both practitioners and researchers are debating on TAI, such as its operationalization. For example, Mökander and Floridi (2021) are addressing ethics-based auditing to develop TAI and Eschenbach (2021) investigates preconditions and challenges for TAI.

Furthermore, related research with a focus on explainable AI deals with the definition of explainability standards and technical methods to achieve and assess AI. For instance, explainable AI techniques such as data-driven and knowledge-driven approaches and corresponding challenges such as understandability (Hanif et al., 2021), or legal regulations on explainable AI and courts role in requirements' definition (Deeks, 2019).

Related research with a focus on AI assessment deals with the development of AI assessment methods and techniques. For instance, the design of AI certifications (Cihon et al., 2021).

While all related research streams provide valuable insights into potential challenges related to AI assessment, they mostly focus on specific challenges in isolation. There is a lack of research that systematically analyzes prevalent assessment challenges and potential interdependencies among them. We, therefore, lack a deeper and broader understanding of potential challenges to understand how they interact and involve different disciplines, and to foster the development of respective solutions to overcome them. Researchers and practitioners may over-see potential pitfalls of their envisioned assessment designs, even if those pitfalls have already been revealed in related research. Hence, in our study we aim to synthesize challenges.

3. Research Approach

We conducted a descriptive literature review (Paré et al., 2014) to synthesize the state of current research on AI assessment and identify challenges, while applying guidelines for literature reviews (i.e. Kitchenham & Charters, 2007; Webster & Watson, 2002).

3.1 Literature Search

The search string was built to reveal articles dealing with AI and assessment performed internally or by third-parties. In addition, we included explainability, transparency and accountability as commonly assessed AI characteristics resulting in the search string:

TI (Artificial Intelligence OR AI) AND TI (Assess OR Evaluat* OR Validat* OR Audit* OR Certif* OR Valuat* OR Explain* OR Transparen* OR Accountab*)*

We applied the search string to six scientific data bases, selected for their access to high quality, peer-reviewed articles in the various disciplines: EBSCOHost, IEEE Xplore, ProQuest, AIS Electronic Library, ScienceDirect, and Emerald Insight. We limited the search to the title due to the broad search string and popularity of the AI term that quickly led to a large amount of search results. Our search yielded a total of 2103 potentially relevant articles, as of January 31, 2022.

We conducted a relevancy check in two stages. First, all 2103 articles were assessed for fit based on their title and abstract, resulting in the exclusion of 2021 articles. Exclusion criteria were applied to articles categorized as off-topic (122), not in English (34), other AI related research (78), AI studies including feasibility assessment or technology application (1366), not research (1), or duplicates (420). Second, the remaining 82 potentially relevant articles were analyzed in their entirety. 50 articles were remaining for analysis after excluding those dealing with other aspects not related to AI assessment challenges such as evaluating market potential.

3.2 Literature Analysis

We applied thematic analysis to our final set of 50 articles (Braun & Clarke, 2006), as a structured approach to identify themes of challenges in AI assessment. Thematic analysis suggests to perform six steps: 'familiarizing yourself with your data', 'generating initial codes', 'searching for themes', 'reviewing themes',

'naming and defining themes' and 'producing the report' (Braun & Clarke, 2006, p. 87).

During *data familiarization*, we noticed that the 50 articles were spread across the disciplines of IS, computer science, law, and social science. Various assessment types were included such as self-assessments, codes of conducts, or certifications. We took notes of, among others, each article's objectives and discipline for later reference during the coding phase.

For *generating initial codes*, we started by reading the full text of the articles and assigned initial codes to relevant text passages providing information on challenges in AI assessment. For instance, the text passage *"since the COMPAS system is owned and developed by a for-profit company, the calculations and proprietary software used to derive a risk score are considered commercially sensitive trade secrets and, therefore, not publicly disclosed for audit."* (Walmsley, 2021, p. 588) was coded as 'dealing with conflicting laws'. We referred back to our notes during data familiarization and particularly looked out for challenges and concepts discussed within and across disciplines.

Often challenges had to be deduced from text passages describing difficulties regarding AI assessment as obstacles were not always described with a concrete name or specific term. We performed multiple rounds of iterative coding to compare codes and segments.

In *searching for themes*, we manually highlighted challenges that we found reoccurring in the same color. As a result, the more papers we included, the more challenges were already part of our color scheme. In the end, after covering all 50 papers of the final set, we arrived at 22 color nuances applied 189 times across all papers. We colored some passages more than once with overlaying nuances, for example, dealing with challenges to assess AI transparency, explainability, trustworthiness and the black box concept as overlaying nuances of one color. We sketched the codes and nuances in their colors in a mind map to see how they relate and differ.

In *reviewing themes*, we considered our set of 22 reoccurring challenge themes, and their relations. We applied Patton's (2002) criteria of internal homogeneity and external heterogeneity for the revision of themes, so that data within themes fits together meaningfully and different themes are clearly distinct from each other. During this process we reduced the initial 22 theme candidates to seven themes as nuances of three main colors.

In *defining and naming themes*, we clustered the seven themes for example, 'applying existing law for AI' or 'impermanence' and identified three main challenge categories, that can make AI assessment a difficult endeavor: ethical implications, regulatory gaps and socio-technical limitations as those where the overarching colors of the nuances. Ethical implications occurred as a preparatory discussion for why certain aspects shall be regulated during assessment, providing input on regulatory gaps that need to be closed and aspects that shall be regulated as recommendations from an ethical perspective. Hence, the legal discourse includes and discusses ethical implications when defining what shall be assessed. Finally, socio-technical limitations occur when examining how assessments can be operationalized.

For *producing the report*, we derived a detailed description of each theme as reported in the following results chapter.

4. Challenges in AI Assessment

4.1. Ethical Implications

The challenge category 'ethical implications' expresses that AI assessment is challenging to conduct because assessment requirements have interlinked ethical aspects. Challenges relate to ethical compliance assessment (C.1) and foreseeing AI's potential consequences for society in the future (C.2).

(C.1) Assessing AI for ethical compliance. Whenever AI is involved in decision-making processes, it needs to be ensured that people are treated fairly and that they are able to understand and contest decisions (Walmsley, 2021). This is especially the case, when the decision refers to sensitive areas of people's lives, such as recruitment and bank lending (Walmsley, 2021). Assessing ethical compliance is a challenge for assessors due to several reasons. In general, AI-related ethical concepts, such as fairness, are complex to consensually operationalize (van Nuenen et al., 2020), and technical limitations, including AI opaqueness, limit insights into the decision making process (Eschenbach, 2021).

More importantly, it is stated that there is a natural "inherent heterogeneity" for assessing ethical questions due to the different ethical perspectives and perceptions (Mantelero & Esposito, 2021, p. 5). Ethical norms and values differ among societies (Bruijn et al., 2021), interdisciplinary decisions differ between contexts (Batarseh et al., 2021), and there is "a lack of consensus around high-level ethics principles" (Mökander & Floridi, 2021, p. 326). Consequently, stakeholders currently struggle to define standards of general validity that can be used for assessments. Therefore, a new research challenge emerging for AI assessment is: *What ethical standards shall AI assessment be based on?*

In the light of the naturally diverse views on ethics, AI developers need to decide on ethical complex trade-offs during model design (Fine Licht & Fine Licht, 2020), such as accuracy vs. fairness (Busuioc, 2021). Developers then draw different conclusions on implementing ethics. The question remains how AI assessments can help to mitigate risks arising from erroneous trade-off decisions and ensure ethically safe systems.

This raises the operational issue of when and under which circumstances an assessor can consider a system as sufficiently safe (Batarseh et al., 2021), and trustworthy (Eschenbach, 2021). An important research question is thus *when shall AI-based systems be considered sufficiently safe and trustworthy?*

What is perceived as ethical shall not be left to developers or assessors, but formal guidance and regulations need to be provided (Mantelero, 2018). However there is a lack of guidance on the ethical implementation of AI, the mechanisms used to oversee human decision making are not applicable (Mökander & Floridi, 2021), and suggestions from the social sciences are too abstract for practical implementation (Schiff et al., 2021). Assessors are thus left uncertain how to assess AI for ethical compliance and deal with normative trade-offs (e.g., deviating fairness concepts or individual vs. societal interests) that arise between ethical principles (Mökander et al., 2021). Hence, it remains challenging to assess the viability of trade-off decisions made by the developers. Therefore, an important research topic is: *How can assessments resolve normative trade-offs to provide guidance for developers?*

Besides, regulators are not providing sufficient legal guidance, leaving it to organizations to deal with ethical concerns (Mantelero & Esposito, 2021). Ethical discussions are a common preceding process for law-making, and thus those regulatory guidance gaps will not be closed until ethical challenges are resolved (Mantelero & Esposito, 2021).

Considering that there is value in applying AI, such as gaining new insights that are not yet known, AI assessments should not undermine or hinder the achievement of AI's promising value contribution (Walmsley, 2021). Yet, assessors struggle to make a decision on the debate whether and which AI use cases shall never be ethically approved (Maclure, 2021) and what is acceptable, for example, in terms of data use and for which potential benefit it needs to be resolved (Batarseh et al., 2021). Similarly, assessors have difficulties to define when the result of an AI-based systems justifies the means, for example, whether there are circumstances under which interests of the public can outweigh rights of the individual (Kriebitz & Lütge, 2020). Defining the right balance to assess ethical requirements in comparison to AI's value contribution remains an open issue.

(C.2) Assessing AI's Future Impact on Society. The diffusion of AI-based systems can have (detrimental) effects on the society. For instance, AI decision making poses the risk to reinforce bias in society (Deeks, 2019), profiling of people based on unacceptable criteria (Bruijn et al., 2021), or influencing human decision making (Busuioc, 2021).

These risks led to calls to incorporate AI's future impact on society for assessments (Havrda & Rakova, 2021). This assessment requires a socio-technical perspective (Schiff et al., 2021), and is a challenging endeavor as it raises fundamental questions (Felzmann et al., 2020), and entails wide-ranging risks that are complex to assess for ethicists as well (Kriebitz & Lütge, 2020). It is predominantly a task for the ethical and social science disciplines to guide such assessments because technical designers of AI-based systems may not be focused on or having the capabilities to assess future technology impact on society (Mantelero & Esposito, 2021). An important research question is thus: *How to measure future impact on society for AI assessment?*

Assessing future impact on society includes foreseeing negative outcomes resulting from bias or system failures, which is challenging for assessors because AI's impact on society is hard to predict and understand (Havrda & Rakova, 2021), especially as it often requires not only to oversee a combination of effects of several AI products and use cases (Schiff et al., 2021), but also requires to draw causal relationships (Havrda & Rakova, 2021).

4.2. Regulatory Gaps

The challenge category 'regulatory gaps' refers to challenges in interpreting and handling regulations or a perceived lack of legal guidance on what requirements AI-based systems have to fulfill. Hence, assessors are lacking a legal fundament and guidelines to refer to for assessment conduction. Relevant challenges result from existing laws' limitations for an application to the AI context and missing AI-specific regulations (C.3), and conflicting laws that introduce hurdles (C.4).

(C.3) Adequacy of existing law for AI assessments. Assessors struggle to apply existing law to assess AI-based systems because they lack guidance and clarity. For instance, there are calls for guidance on the application of product liability regulations to AI (Bertolini & Episcopo, 2021), or assessing AI under the EU GDPR explainability requirement (Hamon et al., 2022).

In addition, we observe a lack of guidance and standards as well as related laws that consider AI specifics. For assessment design and preparation it is important to know what to consider and expect for AI assessment, but "the law is not designed to regulate algorithm-based processes from which potentially incorrect or potentially unjustified or unfair results arise" (Käde & Maltzan, 2019, p. 10). A research challenge is thus: *How shall assessors deal with conflicting laws?*

Legal researchers acknowledge the European Commission's Proposal for an Artificial Intelligence Act (AIA), as an attempt to provide guidance and to close prevalent regulatory gaps. Nonetheless, even before the regulation comes into effect, there have already been

calls for improvement, amendments and further clarification needs (Ebers et al., 2021; Mökander et al., 2021).

Among those needs is a call for providing more guidelines with the AIA to close the gap on assessing legal accountability, for example, resolving questions such as who shall be held responsible for AI-based system failures (Mökander et al., 2021). It is highlighted in law literature that applying existing law for accountability is challenging for the assessment of algorithmic decision-making processes (Käde & Maltzan, 2019), among others, due to algorithmic opacity and black box systems (Yanisky-Ravid & Hallisey, 2019), resulting in a lack of overview and difficulties to question decision outcomes (Busuioc, 2021). Furthermore, accountability assessment is hampered by a lack of standard definitions for explainability requirements it could rely on (Käde & Maltzan, 2019). Therefore, a new research challenge emerging for AI assessment is: *How to define and assess AI accountability?*

Prior research depicts an overall lack of legal standards on legal explainability as a major regulatory gap (Busuioc, 2021; Lu, 2021; Maclure, 2021). The final requirements' definition used to assess explainability will eventually only be derived in litigations based on the views of courts and judges (Deeks, 2019), but not by assessors themselves. The discussion on legal requirements of explainability for certain use cases is hampered by the fact that there are socio-technical limitations in explainability assessment (Lu, 2021). The legal situation for AI explainability has not kept pace with the technological developments of AI and explainability methods (Busuioc, 2021). Socio-technical limitations and research on explainability predominantly found in technical-oriented literature can serve as an input for lawmakers (cf. Section 4.3; C.5).

(C.4) Dealing with Conflicting Laws on Trade Secrets. Assessors need to be able to evaluate how AI developers have implemented the AI-based system to assess, among others, how they solved trade-offs between, for example, accuracy and fairness (Walmsley, 2021). However, trade secrets, for instance as defined in the EU GDPR and the US Freedom of Information Act, are introducing hurdles for AI assessment (Busuioc, 2021). Referring to these trade secret laws, organizations can protect their AI algorithms from public assessment (Lu, 2021). Under such circumstances, assessors are unable to assess the risks AI algorithms hold despite their significant decision power even for sensitive areas of people's lives (Lu, 2021).

A well-known example of a flawed algorithm that became public and significantly impacted people's lives is the COMPAS algorithm, used to decide on prison sentences. The algorithm was not disclosed for public assessment due to trade secrets even in light of a scandal (Walmsley, 2021). Hence, the question is raised how many more of these algorithms exist that go unnoticed until scandals arise (Lu, 2021).

As a result of trade secret laws policymakers are hampering their own efforts towards achieving AI assessments (Busuioc, 2021) and disclosure standards would need to be adjusted to allow for assessment (Lu, 2021). Yet, it remains an open issue to revise current trade secrets and disclosure standards for AI (Lu, 2021), while evaluating options to ensure intellectual property rights, for instance by using other means to mitigate risks and provide transparency and explainability (Havrda & Rakova, 2021). An important research challenge is thus *how can disclosure standards and conflicting laws for AI assessment (e.g., trade secrets) be revised in a way that they do not undermine currently protected rights (e.g., intellectual property rights)?*

4.3. Socio-Technical Limitations

Finally, 'socio-technical limitations' resulting from technical aspects and nature of AI hamper assessment. Relevant challenges result from AI-based system's lack of explainability (C.5) and transparency (C.6.), and AI-based systems' inherent impermanence (C.7).

(C.5) AI-based Systems' Explainability. Socio-technical limitations to provide explainability on AI decisions are causing wide-ranging challenges so that these are extensively discussed in literature. "Explainability can be viewed as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions" (Barredo Arrieta et al., 2020, p. 84). As a general issue, a lack of explainability definition (Hanif et al., 2021) and overview of scattered knowledge on explainability is highlighted (Vilone & Longo, 2021). There is a variety of terms circulating and mixed to describe explainability (Vilone & Longo, 2021), such as interpretability, transparency or comprehensibility (Rawal et al., 2021). Closely related to the explainability concept, the terms opaqueness or opacity are used as umbrella terms as well. Given this conceptual ambiguity, it is not clear what shall be considered for explainability assessment (Vilone & Longo, 2021). Therefore, another critical research question is: *What explainability requirements shall be included for AI assessment?*

There are several technical characteristics of AI that limit assessment possibilities, making explainability a technical problem (Käde & Maltzan, 2019). AI models are complex, opaque and challenging for humans to understand (van Nuenen et al., 2020), especially if the decision making cannot be observed and remains a black box (Käde & Maltzan, 2019). Even for developers it can be challenging to maintain oversight, amid the self-learning capabilities of algorithms when discovering new causal relationships (Walmsley, 2021), and parts of

the system cannot be assessed independently (Dahmen et al., 2021). Hence, even algorithm designers cannot provide full transparency (Walmsley, 2021). This is an issue as it impedes the detection of bias and discrimination or unintended algorithm behavior (Lu, 2021). In addition, human validation of AI decisions is only workable if the algorithm is understood (Hamon et al., 2022). If explainability is not considered during development per se, it is difficult to achieve at later stages, such as during assessments (Batarseh et al., 2021). An important research challenge is thus *how can assessable explainability be established during AI-based systems' design?*

To solve the issue, explainable AI models are built to explain decisions and provide transparency (Vilone & Longo, 2021). However, even with explainable AI models, several challenges remain. On the one hand, assessors cannot demand as much explainability as possible as prioritizing explainability lowers the accuracy of AI models, so that assessors need to consider this trade-off carefully (Hamon et al., 2022; van Nuenen et al., 2020). On the other hand, assessors cannot solely rely on the explanations provided, but need to evaluate and decide whether they can declare them accurate and trustworthy which is a challenging task. Explanation models are only an abstraction that may not reflect the actual decision making process (Bruijn et al., 2021), and themselves are complex and challenging to assess (Eschenbach, 2021). Assessors need to decide whether demanding this additional complexity is appropriate and helpful (Hanif et al., 2021). It needs to be considered that, even if demanded, it is not always possible to provide explanations due to socio-technical limitations (Asatiani et al., 2020), so that explainability requirements can ultimately limit technological advancement (Hanif et al., 2021). Hence, assessing the quality of an explanation introduces new challenges for assessment. An important research question is thus: *How shall explainability models be assessed (e.g., regarding quality, suitability)?*

Currently there is a lack of tools to practically measure and assess the value of an explainability model during an assessment (Rawal et al., 2021). For instance, the extend of human understanding that an explanation needs to provide is not defined (Busuioc, 2021). This is an issue as explainable AI models are often not human understandable (van Nuenen et al., 2020). In addition, different target groups require explanations tailored to their needs and understanding (van Nuenen et al., 2020), so that operationalizing explanations in a meaningful way remains unclear (Vilone & Longo, 2021). Hence, it is challenging for assessors to evaluate the value of explainable AI models (van Nuenen et al., 2020).

At the same time, implementing explainable AI is a challenging task and there is no easy to use or pre-prepared toolbox to choose from (Bruijn et al., 2021). There is a vast amount of explainability methods of varying strength and their inclusion comes with a cost for creation time and computation is required (Hanif et al., 2021). For assessors it can be challenging to detect if an explanation model applied to one context is still valid when the algorithm is applied to another context, and transferability of explanations is an ongoing research field (Rawal et al., 2021).

(C.6) AI-based Systems' Transparency. Explanations on decision making can contribute to providing transparency on AI's inner workings. Nonetheless, even if provided, transparency can introduce severe risks as well (van Nuenen et al., 2020), so that considering undesirable side effects introduces new challenges for assessors. Among the aspects that assessors need to carefully consider are privacy risks (Felzmann et al., 2020), decreasing innovativeness (Vilone & Longo, 2021), and security risks (Fine Licht & Fine Licht, 2020).

First, providing transparency can lead to privacy risks, for example, when explanations include private data (Rawal et al., 2021), acquired during algorithm training (Felzmann et al., 2020). Hence, there is a trade-off between transparency and privacy when providing AI explanations (Rawal et al., 2021). Second, in response to increased transparency, organizations may reduce intellectual property creation, resulting in decreasing innovativeness and advancement on AI technology (Vilone & Longo, 2021). Third, security risks are a concern when introducing transparency as it provides information that can be used to game the system (Fine Licht & Fine Licht, 2020). For example, explanations can reveal potential points for attack, thereby introducing vulnerabilities and requiring concepts for protective measures that need to be developed (Rawal et al., 2021). Therefore, another critical research question is: *How to deal with transparency associated risks when defining explainability requirements for AI assessment?*

Even if transparency and explanations are provided, and risks considered, it is highlighted that assessors shall put special attention on the input data quality serving as the explanation basis, thereby playing a crucial role for any assessment on top of it (Asatiani et al., 2020). If there are flaws in the input data decisions corresponding explanations can be wrong. Yet, assessing input data acquisition and applicability for underlying use cases can be challenging (Yanisky-Ravid & Hallisey, 2019).

(C.7) AI-based Systems' Impermanence. Impermanence is a key feature of AI and at the same time a major impediment for AI assessment. AI adds value by detecting formerly unknown solutions (Applegate & Koenig, 2019). Consequently, expected outcomes are unknown as well. As during learning AI optimizes results, it is challenging to assess whether results of subsequent iterations will be correct without previously being able to know them, just as it is challenging to define

the right point in time for an assessment as it loses validity with the next iteration (Applegate & Koenig, 2019). Assessing what exactly AI learns and whether this will be applicable to new contexts is difficult (Benedick et al., 2021). For example, algorithms may perform well when tested, but fail when applied practically or confronted with perturbations. Hence, it is not sufficient for AI assessment to consider performance during training or a static point in time (Benedick et al., 2021). In contrast, AI would need a continuous audit.

This hampers the provision of guarantees on future algorithm behavior, for instance, regarding discrimination and bias (Yanisky-Ravid & Hallisey, 2019). Assessors struggle to guarantee that an AI-based system is free from bias and discrimination because over time existing stereotypes can be learned and reinforced (Käde & Maltzan, 2019; Walmsley, 2021). Even if excluded for a certain point in time, due to impermanence it cannot be predicted whether new attributes will be learned that result in discrimination and bias (Dahmen et al., 2021), making it challenging to check that a system is not discriminatory at a given point in time to deduce future guarantees. An important research question is thus: *How can AI impermanence be conceptualized and managed in a way that AI assessments can guarantee AI-based systems' long-term safety?*

Finally, impermanence can impact AI algorithm trustworthiness when results change compared to previous iterations which can lead to outdated rationales and adjustment requirements in explanations (Bruijn et al., 2021). Therefore, a new research challenge emerging for AI assessment is: *How to ensure that explanation models remain valid despite AI impermanence?*

5. Discussion

Principal findings. We conducted a descriptive literature review to synthesize existing research on challenges in AI assessment. We found that challenges in AI assessment are discussed across several disciplines. Based on the comparison of existing research we identified three main challenge categories: ethical implications, regulatory gaps and socio-technical limitations. These categories indicate challenge complexity, and it became apparent that it will not be sufficient to discuss assessment hurdles isolated within disciplines. Normative trade-offs are a challenge for AI assessment as there is no common agreement on the basis against which to assess (Mökander et al., 2021). Resolving ethical questions is a task for the ethical and social science disciplines so that the law discipline can include those for regulations to create a baseline for assessment (Mantelero & Esposito, 2021). For the ethical discipline it is challenging to define what the standard should be as "ethics does not provide an answer sheet but a playbook" (Mökander & Floridi, 2021, p. 325). The assessor would still be confronted with the responsibility to check that relevant aspects for ethics assessment were considered and left with the uncertainty to identify what those relevant aspects are (Mökander & Floridi, 2021). The legal discipline faces challenges to make law while there is no clear ethical consensus to be incorporated (Busuioc, 2021). In addition, lawmakers are confronted with a current situation in which AI-based systems are already existent on the market which would need to be revised and technological advancement might be hindered (Yanisky-Ravid & Hallisey, 2019).

Still, socio-technical limitations are complicating the endeavor as there are special challenges within the characteristics of AI, such as opaqueness, impermanence and guaranteeing the exclusion of bias and discrimination, that make it challenging to adhere to those standards during system design and to assess whether those standards once set can be adhered to in the future (Yanisky-Ravid & Hallisey, 2019). Technical-oriented disciplines discuss how opacity can be overcome and risks mitigated while still being able to leverage the full capabilities of AI (Hanif et al., 2021). However, the current socio-technical limitations are subject to ongoing IS and computer science research, and lawmakers and ethicists will have to provide guidance for when the technology can be applied despite the risks (Batarseh et al., 2021). Normative consensus cannot wait for technical-oriented research to resolve reliably bias exclusion and explainability in any case until they suggest regulations. Otherwise, the "vacuum" remains, meaning little guidance for industry, judges and policy makers (Yanisky-Ravid & Hallisey, 2019, p. 473). Scandals have shown that this vacuum is problematic as, despite AI technologies' great potential, they can be a threat for society as well and their adoption is advancing across industries. "Policy makers cannot take a wait-and-see approach" (Maclure, 2021, p. 432).

Implications for Research. Existing literature has discussed challenges in AI assessment according to their respective disciplines of origin. For instance, law literature discourse revolves around the need for new regulations and specific challenges in applying existing law on AI (Yanisky-Ravid & Hallisey, 2019). However, the law discussion falls short in considering socio-technical limitations and possibilities for operationalizing assessments which requires new approaches to cope with (Busuioc, 2021).

Our study reveals seven challenges that hamper advancement on AI assessment, clustered into three main challenge categories. Thereby, we do not only identify challenges, but also provide insights into challenge interdependencies such as hesitations from the legal discipline to suggest too many regulations as those could

hamper technology advancement (Yanisky-Ravid & Hallisey, 2019), or illustrating how in turn socio-technical limitations are hampering the abilities of the legal discipline to find adequate and operationalizable regulations (Busuioc, 2021; Yanisky-Ravid & Hallisey, 2019).

We advance research by providing starting points and corresponding research questions to effectively address challenges such as socio-technical limitations with regard to explainability or algorithm behavior predictability that are discussed within technical-oriented research, and introduce issues for social sciences and ethicists who need to find a perspective on balancing ethics in a way that technology benefits can still be leveraged (Maclure, 2021).

We discuss the disciplines' interrelations, adding to the understanding of why AI assessment remains challenging by illustrating the current state on ethical implications' discussions that are incorporated for the legal assessment basis and need to be operationalized, despite socio-technical limitations.

Implications for Practice. Our results are illustrating to practice that AI assessment needs to be established in a timely manner. This study has shown challenges explaining why this process is so time consuming and complex. At the same time, it clearly points out the interrelations of the challenges and the relevance to start resolving challenges by now. For instance, if AI's legal accountability is not sorted, it practically results in the absence of accountability mechanisms which introduces uncertainty for AI providers, users as well as consulting lawyers and courts.

If those socio-technical limitations that are concerning ethics and legal are not considered during the design phase of AI, it will most likely not be possible for AI providers and users to adjust existing AI-based systems to future regulations. It is important for policymakers to provide guidelines as every day without clearer regulations means uncertainty for market participants and potentially ethically and legally undesirable AI solutions for society (Yanisky-Ravid & Hallisey, 2019).

Finally, AI providers, users and lawmakers that are currently working on the design, development and compliance with AI assessments can take the challenges identified into account and develop solutions to move AI assessments forward. Examining challenges for AI assessments can also inform related AI readiness assessments that examine to what extent organizations are capable of leveraging AI (Holmström, 2022).

Limitations and Future Research. Our study is subject to limitations, paving the way for future research. Limitations of our literature review include an inherent bias for the selection of the search string and the identification of relevant literature. Furthermore, challenges were not always explicitly named with concrete terms, but had to be deduced from the reported issues in literature. Additional empirical work should validate and complement the challenges found as well as examine potential solutions to accelerate AI assessment advancement.

6. Conclusion

The demand for AI assessment realization can only be fulfilled if challenges are identified and addressed. While AI adoption is advancing, assessment has not kept pace so that scandals are rising and measures for risk mitigation are needed. For understanding the current situation and why it persists, challenges in AI assessment need to be first identified and then addressed. This literature review contributes by presenting seven key AI assessment challenges. Our results indicate that the results are linked to three main challenge categories: ethical implications, regulatory gaps, and socio-technical limitations. This categorization reflects the inherent interdisciplinarity and allows to tackle challenges jointly from the disciplines of IS, computer science, law and social sciences. This collaboration and joint approach will be especially relevant for future AI assessments' advancement and progress.

7. References

Applegate, D., & Koenig, M. (2019). Framing AI audits. *Internal Auditor*, *76*(6), 29–34.

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A. (2020). Challenges of explaining the behavior of black-box AI systems. *MIS Quarterly Executive*, *19*(4), 259–278.

Bajarin, T. (2020, June 18). Why it matters that IBM has abandoned its facial recognition technology. *Forbes*. https://www.forbes.com/sites/timbajarin/2020/06/18/why-it-matters-that-ibm-has-abandoned-its-facial-recognition-technology/?sh=5afe3abdafaf

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI). *Information Fusion*, *58*, 82–115.

Batarseh, F. A., Freeman, L., & Huang, C.-H. (2021). A survey on artificial intelligence assurance. *Journal of Big Data*, *8*(1).

Benedick, P.-L., Robert, J., & Le Traon, Y. (2021). A systematic approach for evaluating artificial intelligence models in industrial settings. *Sensors*, *21*(18), 6195.

Bertolini, A., & Episcopo, F. (2021). The expert group's report on liability for artificial intelligence and other emerging digital technologies. *European Journal of Risk Regulation*, *12*(3), 644–659.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Bruijn, H. de, Warnier, M., & Janssen, M. (2021). The perils and pitfalls of explainable AI. *Government Information Quarterly*, 101666.

Busuioc, M. (2021). Accountable artificial intelligence. *Public Administration Review*, *81*(5), 825–836.

Cihon, P., Kleinaltenkamp, M. J., Schuett, J., & Baum, S. D. (2021). AI certification. *IEEE Transactions on Technology and Society*, *2*(4), 200–209.

Dahmen, U., Osterloh, T., & Roßmann, J. (2021). Generation of virtual test scenarios for training and validation of AI-based systems. In *2021 IEEE PIC*.

d'Angelo, C., Flanagan, I., Motsi-Omoijiade, I. D., Virdee, M., & Gunashekar, S. (2022). *Labelling initiatives, codes of conduct and other self-regulatory mechanisms for artificial intelligence applications*. RAND Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RRA1700/RRA1773-1/RAND_RRA1773-1.pdf

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review*, *119*(7), 1829–1850.

Ebers, M., Hoch, V. R. S., Rosenkranz, F., Ruschemeier, H., & Steinrötter, B. (2021). The European commission's proposal for an artificial intelligence act. *J*, *4*(4), 589.

Eschenbach, W. J. von (2021). Transparency and the black box problem. *Philosophy & Technology*, *34*(4), 1607–1622.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science & Engineering Ethics*, *26*(6), 3333–3361.

Fine Licht, K. de, & Fine Licht, J. de (2020). Artificial intelligence, transparency, and public decision-making. *AI & Society*, *35*(4), 917–926.

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G., & De Hert, P. (2022). Bridging the gap between AI and explainability in the GDPR. *IEEE Computational Intelligence Magazine*, *17*(1), 72–85.

Hanif, A., Zhang, X., & Wood, S. (2021). A survey on explainable artificial intelligence techniques and challenges. In *2021 IEEE 25th EDOCW*.

Havrda, M., & Rakova, B. (2021). Enhanced well-being assessment as basis for the practical implementation of ethical and rights-based normative principles for AI. In *2021 IEEE SMC*.

Holmström, J. (2022). From AI to digital transformation. *Business Horizons*, *65*(3), 329–339.

Käde, L., & Maltzan, S. von (2019). Towards a demystification of the black box. *Journal of Internet Law*, *23*(3), 3–13.

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering.*

Kriebitz, A., & Lütge, C. (2020). Artificial intelligence and human rights. *Business and Human Rights Journal*, *5*(1), 84–104.

Lansing, J., Benlian, A., & Sunyaev, A. (2018). "Unblackboxing" decision makers' interpretations of IS certifications in the context of cloud service certifications. *Journal of the Association for Information Systems*, 1064–1096.

Lebovitz, S., Levina, N., & Lifshitz-Assaf, H. (2021). Is AI ground truth really true? *MIS Quarterly*, *45*(3), 1501–1525.

Lins, S., & Sunyaev, A. (2022). Advancing the presentation of IS certifications. *Behaviour & Information Technology*, 1–24.

Löbbers, J., Lins, S., Kromat, T., Benlian, A., & Sunyaev, A. (2020). A multi-perspective lens on web assurance seals. *Electronic Commerce Research*(forthcoming).

Lu, S. (2021). Algorithmic opacity, private accountability, and corporate social disclosure in the age of artificial intelligence. *Vanderbilt Journal of Entertainment & Technology Law*, *23*(1), 99–159.

Maclure, J. (2021). AI, explainability and public reason. *Minds & Machines*, *31*(3), 421–438.

Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Computer Law & Security Review*, *41*, 105561.

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2021). Conformity assessments and post-market monitoring. *Minds & Machines*, 1–28.

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds & Machines*, *31*(2), 323–327.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2014). Synthesizing information systems knowledge. *Information & Management*(52), 183–199.

Rai, A., Constantinides, P., & Sarker, S. (2019). Next-Generation Digital Platforms. *Management Information Systems Quarterly*(43), Article 1, iii–ix.

Rawal, A., Mccoy, J., Rawat, D. B., Sadler B., & Amant, R. (2021). Recent advances in trustworthy explainable artificial intelligence. *IEEE Transactions on Artificial Intelligence*, 1.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., & Lennon, M. (2021). Explaining the principles to practices gap in AI. *IEEE Technology & Society Magazine*, *40*(2), 81–94.

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems*, *29*(4), 260–278.

van Nuenen, T., Ferrer, X., Such, J. M., & Cote, M. (2020). Transparency for whom? *Computer*, *53*(11), 36–44.

Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, *76*, 89–106.

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & Society*, *36*(2), 585–595.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future. *MIS Quarterly*(26), Article 2, xiii–xxiii.

Yanisky-Ravid, S., & Hallisey, S. K. (2019). Equality and privacy by design. *Fordham Urban Law Journal*, *46*(2), 428–486.