MAB MAANDBLAD VOOR ACCOUNTANCY EN BEDRIJFSECONOMIE

Maandblad voor Accountancy en Bedrijfseconomie 96(7/8) (2022): 253–265
DOI 10.5117/mab.96.90108

Research Article

Time to audit your AI algorithms

Iuliana Sandu, Menno Wiersma, Daphne Manichand

Received 8 July 2022 | **Accepted** 12 September 2022 | **Published** 16 September 2022

Abstract

Undoubtedly, the use of algorithms, and Artificial Intelligence (AI) algorithms in particular, has numerous benefits. Fields such as finance, healthcare, automotive, education, and recruitment, to name a few, have demonstrated successful application of AI algorithms. Conversely, cases of bad algorithms abound and lead to lost revenue, discrimination, disinformation, or even bodily harm. Currently, we have surpassed the stage of just observing bad algorithms. New European regulations governing AI force organizations to manage the risks introduced by algorithms and convince the public about the proper functioning of algorithms. In this context, can algorithms be rigorously audited to build public trust and if yes, how? This article aims to answer these questions by building on an auditing framework for model risk management that controls for the novelty introduced by AI algorithms while connecting AI algorithm audit with internal audit terminology.

Relevance for practice

The article aims to guide internal auditors in the task of auditing Artificial Intelligence algorithms.

Keywords

Artificial Intelligence, audit, algorithms, internal audit, model risk

1. Introduction

The urgency to audit AI algorithms is intensified by the regulatory actions happening in the European Union. Upcoming new AI regulations such as the EU Artificial Intelligence (AI) Act (European Commission, 2021), the Digital Markets Act (DMA) and Digital Services Act (DSA) (European Commission, 2022) may drastically change the need to monitor compliance with risk management procedures around the development, implementation and use of AI. The EU AI Act, specifically, is seen as laying the foundations for a more global approach on regulating AI, for which initiatives so far have been rather scattered and limited in scope. Such regulations bring forward the need for the audit of AI algorithms.

Besides current regulatory pressures, the need to audit AI algorithms is fundamentally driven by the fact that individuals and organizations increasingly base their decisions on data and algorithms. Makridakis (2017) describes the present societal impact of the AI revolution as significant, as it affects individuals' lives, work, shopping, entertainment, and employment. AI also influences the operation of organizations. Indeed, AI algorithms can increase the access of poor individuals to financing (Strusani and Houngbonon 2019), support successful breast cancer screening (Barnett et al. 2021), power autonomous cars (Yaqoob et al. 2019), guide students towards success (Page and Gehlbach 2017), and speed up the hiring process (Upadhyay and Khandelwal 2018). At the same time, algorithms can discriminate against students with poorer backgrounds (The Verge 2020), promote sexist recruitment (Dastin 2018), force borrowers who attended historically black schools to pay a penalty when refinancing student loans (Student Borrower Pro-

tection Center 2020), reduce access to care for black individuals (Obermeyer et al. 2019) and even kill pedestrians (BBC 2020). It is not just individuals who suffer when algorithms are insufficiently under control, organizations themselves also suffer. If a company uses a flawed algorithm, it can result in for example loss of revenue. In such an example, financial services firm Transamerica was fined USD 97 million by the US Securities and Exchange Commission, because it relied on an error-prone algorithm that managed investors' portfolios, yet it did not disclose these errors to investors (Securities and Exchange Commission 2018).

There are several dimensions that make AI algorithms impactful but also unpredictable and difficult to control. One particularly salient dimension is the powerful but opaque (or "black-box") nature of some AI algorithms (Burrell, 2016), where it is possible to model complex relationships between data, but it is often difficult to understand fully why the algorithm produced a certain output. Another dimension is the data-intensive nature of some AI algorithms, which is somewhat of a double-edged sword: on one hand this allows for much more fine-grained and precise modeling (Martens et al. 2016), but on the other hand, it means that subtle biases in the data used to train the model get codified in the resulting algorithm and thus perpetuated. Finally, the evolving nature of many AI algorithms, that is, the updates to the algorithm as more data comes in, to further train the algorithm, means that the performance of the algorithm can change over time (Lazer et al. 2014). This means that, in practice, it is increasingly difficult for organizations to control AI algorithms.

In light of the recent regulatory changes, the established societal impact of AI algorithms and the dimensions that make AI algorithms powerful but unpredictable, various attempts have been made to audit AI algorithms (Bandy 2021). However, auditing AI algorithms is understood differently by different users. Some authors refer to expert testing where algorithms can go wrong (Sandvig et al. 2014) while others focus on auditing frameworks aimed at ethical considerations only (Koshiyama et al. 2021). In this article we aim to join the field of internal auditing with that of algorithms in a framework inspired from the internal auditing domain, a domain with tradition in audit. We are motivated to focus on internal audit by the view expressed by Power (Power 2000, p.118) that the "expertise, operational independence and proximity to real times cultures of control" of internal auditors are desirable qualities for an audit. An important role for internal audit is also argued by Raji et al. (2020 p.35): "Internal auditors' direct access to systems can thus help extend traditional external auditing paradigms by incorporating additional information typically unavailable for external evaluations to reveal previously unidentifiable risks."

In this article, we perform a literature review aimed at gaining insights into the elements which are important for the audit of AI algorithms. We use a general audit process definition according to which "auditing is the accumulation and evaluation of evidence about an audit object to determine the degree of correspondence between the characteristic of the audit object and established criteria". By using such a general definition of auditing, we can borrow from other frameworks of auditing who use audit objectives, audit criteria and evidence collected to assess whether the audit object characteristics comply with the audit criteria. More specifically, we build on the Model Risk Management approach (IIA 2018) as the audit object under this framework, the model, has many similarities with an algorithm. Our literature review culminates with a framework aimed to serve as an initial, rigorous, guide towards the audit of algorithms. Essentially, the framework guides the internal auditor towards the elements that should be checked to assess whether we can rely on the output of an algorithm. Although the examples used in this article are primarily related to AI algorithms, we believe that the framework presented can serve the audit of algorithms, in general, due to their similarities to models. It should also be noted that the framework in this article develops audit objectives based on the audit criteria revealed by the literature review but does not expand into the auditing process (e.g., the audit stages, methods and techniques to collect evidence) or forms of reporting on the audit (e.g., levels of assurance).

The article follows with an overview of the European Commission's Artificial Intelligence Act marking the importance of algorithm audit, discusses the challenges to audit that algorithms introduce, investigates different approaches to control the risks of algorithms and ends with a framework aimed at the audit of algorithms.

2. Regulation

A landmark regulation aimed at governing AI algorithms is the Artificial Intelligence Act issued by the European Commission (the AI Act) (European Commission, 2021). The importance of compliance with the AI Act is reflected by the formulated penalties. Non-compliance may lead to fines up to the higher of EUR 30 million and 6% of worldwide annual turnover. From an audit perspective, the most important elements of the AI Act are the definition of AI, the risk-based approach in supervision, and the specific requirements for high and low risk AI systems.

The AI Act uses a broad definition of AI: "artificial intelligence system (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with" (European Commission, 2021, p. 39). The proposed Act distinguishes three types of AI techniques and approaches (European Commission, 2021, Annex I): machine learning approaches (including supervised, unsupervised, reinforcement and deep learning), logic- and knowledge-based approaches

including expert systems, and statistical approaches such as Bayesian estimation. Various stakeholders have proposed amendments to the definition, mostly narrowing the scope (Orgalim 2022). Others, mainly non-profit organizations, argue that the EU AI Act should use the concept of "automated and algorithmic decision-making" (ADM) to truly show the socio-economic impact of AI systems on individuals and society (European Parliament, 2022). Nevertheless, the consensus seems to be that the dynamic nature of the AI algorithms landscape needs to be reflected in the definition.

The risk-based approach proposed by the AI Act recognizes three classes of risk: *unacceptable risk*, *high risk* and *low or minimal risk*. AI applications listed as carrying unacceptable risk are prohibited. Examples of this risk type are practices using techniques beyond a person's consciousness, social scoring techniques likely to cause physical or psychological harm, activities exploiting vulnerabilities of specific groups, or use of real-time remote biometric systems in publicly accessible spaces. High risk AI essentially consists of two lists of industries and activities that to date are recognized as high risk. The first list describes AI systems used as products or safety components of products covered by the sectorial Union law (for example, machinery, personal protective equipment, radio equipment, medical devices, transportation). The second list consists of "other" AI applications with risks that have already materialized or are likely to materialize (for example, biometric identification of natural persons, supply of water, recruitment, access to public benefits and services, access to or assessment in educational and vocational training, creditworthiness, asylum and border control, administration of justice). Low or minimal risk applications are AI systems that are not prohibited or have a high risk.

Before placement on the market, low risk systems have the possibility, but not the obligation, to follow a code of conduct on a voluntary basis. High risk applications described in Annex III of the AI Act, should perform a conformity assessment, and show they have in place:

- An established, implemented, documented and maintained risk management system able to assess, evaluate and mitigate risks
- A data governance approach to ensure the use of high-quality datasets for training and testing learning algorithms
- Technical documentation which tracks the compliance with the requirements of the AI Act and automatic record-keeping to monitor events affecting the AI systems
- Transparency capabilities that enable users to understand details about the functioning of the AI system
- Human oversight such that natural persons can intervene to minimize the risk of AI systems
- An appropriate level of accuracy, robustness and cybersecurity throughout their lifecycle

Overall, the current phrasing of the regulation, from the AI definition to the division of risk categories, is broad such that it accommodates the dynamic nature of AI techniques and risks. Consequently, many stakeholders commented in their position papers that the definition of high-risk was unclear or needed improvement. In such a case, the European Central Bank (ECB) suggested excluding specific creditworthiness applications for natural persons from the high-risk category (European Central Bank 2021). According to the ECB, systems that leverage on the standalone use of linear or logistic regression or decision trees placed under human supervision and which have a minor impact, should be excluded from the high-risk AI category due to ECB's technology neutral approach. The ongoing discussion on the current draft of the legislation has consequences for the timing of the AI Act adoption which was originally planned for mid-2022 but is now expected at the end of 2024.

It is relevant to mention that algorithms and AI are in the focus of US regulators as well. Since approximately 2013, US financial institutions were already regulated on AI under the Supervisory Guidance on Model Risk Management SR 11-7 (Board of Governors of the Federal Reserve System 2011). The main risk driver identified in SR 11-7 was the inherent simplification of reality, intrinsic to any model. According to SR 11-7, awareness about the simplification inherent in any model is important because it spells out the fact that the model has specific capabilities, limitations, and assumptions. Next, in 2022, the National Institute of Standards and Technology (NIST) was directed by the US Congress to collaborate with the private and public sectors to develop an AI Risk Management Framework (RMF). On March 17, 2022, the initial draft was published, to be finalized by the end of 2022 or the beginning of 2023 (NIST 2022). The AI RMF is distinguishable from the EU AI Act in that it is fully voluntary. It addresses risks in the design, development, use, and evaluation of trustworthy and responsible AI products, services, and systems. It recognizes technical (accuracy, reliability, robustness, resilience, or security) and socio technical risks (explainability, interpretability, privacy, safety, bias), and provides guiding principles (fairness, accountability, transparency). NIST shows alignment with other initiatives as it provides a mapping of its applied taxonomy to the OECD Recommendation of the Council on Artificial Intelligence (2019), EU AI Act, and US President Executive Order 13960 (2020). NIST also points out that the EU AI Act (and OECD AI Recommendation) mostly focuses on the ethical principles, and less so on the technical design. The AI RMF describes the importance of a solid governance around AI technologies including the auditing of AI systems. The document recognizes that validity for deployed AI systems is often assessed with ongoing audits or monitoring that confirm that systems behave as intended.

3. Algorithms and the challenges they bring to audit

3.1. Challenges from AI algorithms

In the Cambridge Dictionary, an algorithm is defined as "a set of mathematical instructions or rules that, especially if given to a computer, will help to calculate an answer to a problem" (Cambridge Dictionary 2022). A simple and often used algorithm is sorting. Much more complex algorithms are used in robotics, where a mixture of physicality, linear algebra and statistics are used to map out actions and movement (Alvi 2022). Ismail (2018) describes AI algorithms as "a group of algorithms that can modify its algorithms and create new algorithms in response to learned inputs and data as opposed to relying solely on the inputs it was designed to recognize as triggers. This ability to change, adapt and grow based on new data, is described as intelligence". These definitions give us insight into how difficult it is to control and audit algorithms.

Three main aspects make AI algorithms challenging to control: the input data used to train AI algorithms, the way the algorithm operates, and the autonomous learning performed by an algorithm. It is problematic if the data used to train the algorithm is unrepresentative for the group where the algorithm will be applied. If the algorithm is trained on predominantly European individuals but the intention is to use it on a diverse population, containing other countries, then the algorithm might not function properly as it was trained on the wrong data. Additionally, many AI algorithms are applied without a thorough understanding of how they work and if they can answer the problem for which they are used. It is inaccurate to use algorithms aimed at predicting quantitative variables, such as amount of sales, to predict qualitative variables, such as impairment or no impairment. Lastly, it is often not clear what the algorithm learns from the data, whether it learns to spot the "right" aspect of a problem. For example, a failure of AI was documented when, instead of learning to identify cancerous lesions, AI learned instead to identify images coming from a specific piece of equipment (Barnett et al. 2021). So, all imagines coming from a specific equipment were classified wrongly as "cancerous". All three aspects can culminate in AI algorithms which lead to discrimination against protected groups of citizens, spread disinformation and even cause bodily harm.

Another challenge specific to AI algorithms is that they often use a large volume of structured (for example, tabular data) and unstructured data (for example, video data) which makes it difficult to ensure data integrity and representativeness. Additionally, many AI algorithms are self-learning, continuously improving the algorithm, or adapting to changing circumstances. This introduces difficulties for the validation of the algorithm which changes from a point-in-time validation to a frequent or continuous (automated) monitoring situation. It also requires storing all algorithm changes and historical data used to train the algorithm. AI algorithms are also sensitive to the selection of hyperparameters. Usually simple algorithms, with less or no hyperparameters, can lead to underfitting (missing the underlying patterns in the data) while complex algorithms, with multiple hyperparameters that need tuning, tend to overfit (give a lower training error than the actual test error). In both cases this leads to poor performance of the algorithm in out-of-sample testing or production (Burnham and Anderson 2002). Finally, vendor algorithms tend to be proprietary, and this adds to the lack of transparency inherent in learning algorithms.

3.2. The audit of algorithms and the internal auditor

Considering the challenges introduced by algorithms, and especially AI algorithms, the question arises how an audit on algorithms should be executed and who should audit algorithms. A domain with experience in providing auditing services is the internal audit domain. The main purpose of internal audit is to add value and improve an organization's operations and it does so by performing independent assessments on the effectiveness of governance, risk management, and control processes in an organization (IIA 2022). As organizations increasingly use algorithms (Makridakis 2017), they are confronted with significant challenges as algorithms can have an impact on the governance, risks and processes of an organization and, ultimately, on whether the organization achieves its goals or not. Then, the internal auditor becomes also responsible for auditing algorithms.

If auditing algorithms is expected from internal audit, then the internal auditor should have sufficient skills and experience to perform such an audit. This again makes the role of the internal auditor with respect to algorithms unclear. Various stakeholders still consider that internal auditors should work within the apparent current scope of their main activities and focus on operational and financial risks, on the governance and process rather than the correctness of the algorithms themselves (van Eck and Middelkoop 2020). Differently, in the article "Closing the AI Accountability Gap" Raji et al (2020) emphasize the added value of internal audit compared to external audit, because of the internal audit's direct access to internal systems. This privileged access to internal systems such as the algorithms used in an organization, creates the obligation for the internal auditor to have knowledge about algorithms themselves. Skills and experience in algorithms and in auditing algorithms can be acquired by internal auditors by expanding their current knowledge as suggested by Arnold (2021), similar to how they acquired knowledge in performing other tasks such as controlling the risks of models.

According to Carawan et al. (2018) internal auditors now have an important role in the Model Risk Management domain, as internal auditors can be tasked with assessing the effectiveness of the Model Risk Management Framework used by financial institutions, including the

governance, policies, procedures, and activities conducted to address the risk of model error. Internal auditors are also responsible for understanding when and how the model is used and if it is in line with the model's stated purpose. The currently undocumented practice into model risk management also reveals the fact that internal auditors manage to perform a rigorous audit of the model risks only if they have knowledge on the model, as well. One indication of this can be the trend of including professionals with model knowledge (e.g., statisticians) in audit teams. Similar to the Model Risk Management domain, where the prevalent use of models by financial institutions pushed the internal auditor into assessing the risks of models, the internal auditor working for organizations who use algorithms will have to audit AI algorithms as well. For this, the internal auditor needs a framework as a guide to auditing algorithms.

4. Different approaches to AI control

4.1. Three phases of ethics maturity for algorithms

Kazim and Koshiyama (2020) distinguish between three phases of AI ethics maturity: the principles and guidelines phase, the processes or the ethics-by-design phase, and the AI assurance phase. The first phase involves AI ethics principles and guidelines issued by diverse organizations, ranging from NGOs and governmental bodies to private entities (for example, the Ethics Guidelines for Trustworthy AI issued by EU's High-Level Expert Group on Artificial Intelligence). Despite serving as an important step in the process of AI ethics maturation, merely spelling out AI ethical principles such as transparency, non-maleficence and responsibility mostly fails to converge into implementable instruments (Jobin et al. 2019).

The second phase involves the inclusion of AI ethics principles in professional codes of ethics such that AI has ethics-by-design. For example, the Association for Computing Machinery (ACM) revised in 2018 its 1992 Code of Ethics to include ethics principles for software developers. Such codes of ethics should guide AI developers to the right choices when faced with ethical dilemmas (for example, how much and what user data to collect). But placing the responsibility of developing ethical AI on the sole shoulders of AI developers can turn out to be ineffective when codes of ethics do not actually influence the choices of AI developers (McNamara et al. 2018). The irrefutable conclusion of these approaches was that the responsibility to develop ethical AI was an interdisciplinary problem (Vanhée and Borit 2022) that should be shared by engineers, philosophers, psychologists and many more. Nevertheless, putting an interdisciplinary stamp does not alleviate the elusiveness of getting AI in control in a way that fulfills organizational goals without causing societal or business harm.

The third and current phase of AI ethical approaches concerns the standardization and operationalization of AI ethics. In this stage of AI maturity, the focus is placed on risk-assessments of AI systems. In a risk-based approach to auditing AI, teams of people close to the problem, highly familiar to the environment where the AI system operates, would sit together with the auditors, and generate a risk registry with everything that can *go wrong* with the algorithm. But in the case of AI, where data combines with algorithms and user behavior in unknown ways, teams of AI developers and auditors might not be able to anticipate important risks. In this case, the diversity of the team, not only in terms of expertise but also in terms of other aspects such as race, gender, education, cultural background, might be key to implementing a good risk registry. For example, while a team of Dutch developers of a globally deployed image captioning AI system might not identify the poor performance of AI as a risk, when tasked to identify celebrities from Taiwan, an international AI review team might.

4.2. Internal auditing framework - SMACTR

SMACTR (Scoping, Mapping, Artefact Collection, Testing and Reflection), the internal audit framework proposed by Raji et al. (2020), aims to give an opinion on the compliance of AI systems with the ethical values of the organization. Most importantly, every stage in SMACTR leads to the creation of artefacts, or documentation, useful for the conclusion of the audit. The documentation design is borrowed from other fields such as aerospace (e.g., the Failure Modes and Effects Analysis systematic risk management approach used in safety engineering) but it also mentions specific documentation aimed at AI (e.g., datasheets and model cards).

Guided by commonly recognized ethical principles (Jobin et al. 2019) such as transparency, justice, fairness & non-discrimination, safety & non-maleficence, responsibility & accountability, SMACTR starts off with a Scoping stage where potential risks to ethical principles are identified and analyzed for their potential impact. In one hypothetical example in Raji et al. (2020), a smile detection algorithm used to automatically trigger cameras in physical photo booths can lead to threats to the principle of fairness & non-discrimination if the algorithm disproportionately impacts people with disabilities or different cultural norms on the formality of smiling. SMACTR continues with a Mapping of stakeholders stage and an Artefact Collection stage where the documentation needed to understand the algorithm is collected. Once risks are identified and ranked, testing can be done to check the compliance of the AI systems with the ethical guidelines of the company that are most at risk at the time of the audit. Testing includes mostly technical solutions such as adversarial testing where auditors simulate what an enemy might do to confuse a system, or review of metrics for specific user profiles. The testing stage of SMACTR should confirm the existence of the anticipated

risks. SMACTR ends with a Reflection stage that includes a risk appetite analysis and a decision on the action plans recommended to mitigate the most important risks. At this point, depending on the risk thresholds of the company, the risks that are considered acceptable by the company are proposed to be mitigated with specific action plans. In the case of the smile detection algorithm, a risk mitigation action could be to add a button for users to opt-in to the algorithmic solution.

4.3. Legal, safe and ethical algorithms

A more encompassing framework which tries to operationalize the audit of algorithms, is the framework published in 2021 by an extensive team of researchers gathered from different universities and institutes, including the University College London and the London Stock Exchange (Koshiyama et al. 2021). In the context of this recent publication, the audit of an algorithm is defined as "the research and practice of assessing, mitigating, and assuring an algorithm's safety, legality, and ethics" (Koshiyama et al. 2021, p. 2). According to this framework, there are four dimensions in the audit of an algorithm: Development, Assessment, Mitigation, and Assurance (DAMA, henceforth). While the Development dimension refers to the stages of an algorithm including data setup, feature engineering, model selection, post-processing, and deployment, the Assessment dimension comprises different pressures that act upon algorithms such as fairness, robustness, explainability and privacy concerns (or verticals, as they are called in the paper of Koshiyama et al. (2021)). These two dimensions interact to give rise to specific auditing activities. For example, the data setup stage interacts with the fairness vertical to inspire the focus of the audit on how balanced the data is. But the situation gets even more complicated as there is an additional interaction. The Development and Assessment dimension interact with a third element, the level of access of the algorithm. Consequently, the audit or inspection activities are much more limited when the level of access to the algorithm is restricted to only an indirect observation of the system. Mitigation strategies aim to address the issues discovered in auditing activities. For example, if explainability issues were identified, mitigation strategies like the application of model agnostic local explanations (for example, LIME approach) can be suggested. The Assurance stage contains possible certification opportunities that narrow the scope of the assurance service, such as certification of a system or sector specific certifications.

The DAMA framework presents an interesting view on how a low number of dimensions interacting can lead to numerous auditing activities and that these activities and the subsequent certification of the algorithmic system are impacted by the level of access to the algorithm. It is also valuable to see the focus of the Assessment dimension on four main verticals inspired by the most recent development in Fair AI: fairness, robustness, explainability and privacy. Still, the operationalization of the DAMA framework is not included in the paper and so the concrete steps to take or items to check in an audit are unclear.

4.4. Notable local developments

On September 22, 2020, the Netherlands Court of Audit (NCA) organized a thinking session with more than 30 experts on 5 themes related to AI algorithm use: data driven working, data quality, AI and algorithms, AI at the government, and transparency. One important conclusion of the thinking session was that due to the way algorithms are developed (for example, using historical data which might be biased or inserting bias through development choices), we can never have certainty that algorithms do not discriminate. As such, necessary controls need to be placed on algorithms. In its 2021 publication 'Aandacht voor Algoritmes' (Algemene Rekenkamer 2021), the Netherlands Court of Audit proposes a risk assessment framework aimed to control algorithms. The framework is based on a combination of existing frameworks and is targeted at five areas: governance, model and data, privacy, IT general controls, and ethics. To showcase the usefulness of the framework, the Netherlands Court of Audit tested it on three algorithms and concluded that the current use of algorithms within Dutch municipalities is limited and the algorithms used can be controlled. Although not limited to, the main intended users of the framework developed by the Netherlands Court of Audit are governmental organizations.

In 2021 NOREA, the Dutch Association of chartered IT-auditors, released the Guiding Principles for Trustworthy AI Investigations (De Boer and Van Geijn 2021). The NOREA principles are aimed at guiding Dutch chartered IT-auditors in investigations into AI algorithms. The guidelines are developed using leading practices for trustworthy AI (for example, General Data Protection Regulation, Information Commissioner's Office Guidance on the AI Auditing Framework). Nevertheless, the NOREA principles do not intend to present a comprehensive framework for algorithmic system scrutiny and the advice presented in the guidelines is that, for algorithm control frameworks, existing risk management methodologies such as COSO and COBIT are encouraged to be followed. It is not clear though if such methodologies are a good match with the challenges introduced by AI algorithms.

The DAMA, SMACTR, the NCA and NOREA conceptual frameworks highlighted in this section are aimed at the ethical risks of AI algorithms. In this article we express an opinion that there is a need for a more comprehensive framework, which goes beyond ethical risks to consider other business risks of AI algorithms and of algorithms, in general. We further attempt to step beyond purely conceptual frameworks to design a framework which can be more readily implemented by an auditor in an audit of an algorithm. As such, we develop normative statements (or positions) against which an algorithm can be checked. These normative statements, which describe how things should work and are audit objectives, are

aimed to resonate with auditors who use audit objectives in other audit tasks. For example, in an environmental audit, an audit objective (or normative statement) is that "Environmental policies exist during the reporting period as described in the notes to the environmental data" (Kamp 2002). With a set of normative statements aimed at AI spelled out, we believe that an auditor can perform a rigorous and comprehensive audit of an AI algorithm.

5. A rigorous framework for auditing algorithms

5.1. Inspiration from Model Risk Management

All previous work on the audit of algorithms highlighted in the previous section focuses mainly on the ethical risks introduced by AI algorithms, for example, the risk of discriminating against protected groups of people. The upcoming AI Act also addresses ethical issues around AI. Although they are very important risks, it is salient to regard ethical risks in the bigger picture of the overall risk that the use of AI algorithms carries: the risk that the algorithm produces erroneous results or is used in the wrong way. One area which can serve as a starting point for getting a more comprehensive picture of risks of algorithms, beyond ethical risks, is the area of model risk management. Conceptually, AI algorithms and models are related in that AI "algorithms operate by learning models from existing data and generalizing them to unseen data" (Suresh and Guttag 2021, p. 1). As algorithms are fundamentally the engines of a model, we can borrow from Model Risk Management best practices to identify other risks in using algorithms, besides ethical risks.

The risk management around algorithms can be approached similarly to the Model Risk Management approach, with a "three lines" system (IIA 2020). The first line (for example, the management of the organization) develops and uses an algorithm, being primarily responsible for all relevant risks. The second line (for example, a special function within an organization responsible for risk management, compliance, or internal control) assists in risk management and often also advises management on accepting or rejecting a specific algorithm and under what conditions. As a third line, internal audit gives an independent and objective assessment to a governing body (for example, the management of the organization) regarding the sufficiency and effectiveness of governance and risk management over AI algorithms.

5.2. The Life cycle audit framework for algorithms

Inspired from the Model Risk Management best practice (Garla and Dhillon 2016), the Life cycle framework for auditing AI algorithms (figure 1) recognizes seven phases during which the first line must manage the risks introduced by AI algorithms: Initiation, Development, Implementation, Use, Monitoring, Review and Retirement. Three of those phases take place before the algorithm is put to use (the Initiation, Development and Implementation phases) while three phases are distinguished during the use of the algorithm (Use, Monitoring and Review). During the last phase, the model is retired (Retirement). The second line in the accountability structure over AI algorithms (for example, a special function within an organization responsible for risk management, compliance, or internal control) is responsible for the Validation of the AI algorithm(s) and plays an important role during most of the phases of the algorithm's lifecycle. Importantly, the use of the algorithm needs to be approved by a governing body, advised by the second line's validation function. While the life cycle stages are common to the Model Risk Management approach, the contribution of this article refers to specific aspects in each stage, aimed particularly at AI algorithms.

Figure 1. The Life cycle framework for the AI algorithm audit.

To offer more granularity to the Life cycle framework for AI algorithm audit, we present in Appendix 1: Table A1 an overview of risks per Life cycle phase (including the Validation phase) and normative statements (or positions) against which an algorithm can be checked. While some normative statements in Table A1 are common to the Model Risk Management framework, the highlighted aspects (in blue italic font) are aimed specifically at AI algorithms (Dil et al. 2019). One current limitation of the framework is that it does not spell out the evidence gathering techniques applicable to AI algorithms for every normative statement. It does not do so because the choice of technique for evidence gathering is a creative one and evidence gathering techniques for AI algorithms can range from highly technical approaches (for example, technical toolkits aimed at understanding the workings of an algorithm such as the LIME approach (Ribeiro et. al. 2016)) to less technical

or more traditional approaches (for example, interviews, checking documentation). Nevertheless, the NOREA Guiding Principles for Trustworthy AI Investigations (2021) offers valuable guidance on types of evidence to look for when engaged in investigating the trustworthiness of AI.

Returning to Table A1, the Initiation phase of the Life cycle refers mainly to stakeholders (for example, involving the right stakeholders, making roles and responsibilities clear) and the scope of the algorithm (for example, delineating an area of investigation depending on the purpose and the regulations covering the algorithm). The Development phase is focused on how sound the process of developing the algorithm is (for example, "are the chosen algorithms applied correctly?") and how sound the data used is (for example, "is the data valid?"). The Implementation phase aims to check if the implementation of the algorithms is in line with its design (for example, "when implemented, are deviations from the developed algorithms documented?"). The audit activities in the Use phase concentrate on making sure that the algorithm is used in line with its intended purpose and that the users know how to use the algorithm and can provide feedback on its use. The Monitoring phase delineates what to measure with respect to the algorithm (for example, what performance metrics to monitor) during the life of an algorithm and when to signal an issue (for example, setting up acceptable thresholds for performance metrics). The Review phase follows up on the discoveries made in the Monitoring phase, looks at the continued functioning of the entire algorithm, and aims to assess if necessary improvements are made to the algorithms (for example, re-parameterization is performed). In the Retirement phase, the organization's inventory of algorithms is updated such that algorithms no longer in use are not still available for use or misuse. The Validation of the algorithm aims to timely challenge the development, implementation, use, monitoring, and review of the algorithm (for example, check if all relevant risks are considered). Although it is difficult to claim that a framework is comprehensive enough to support an AI algorithm audit, it is possible that, by focusing on the processes around an AI algorithm, on its life cycle within the organization, such an audit can be performed. Thus, we envisage that this framework can be used for the audit of AI algorithms.

5.3. Hypothetical example

To illustrate the application of the Life cycle audit framework, the hypothetical example of an algorithm supporting an online hotel rating system is used. Such a rating system can be attached to the website of an online travel company who lists hotels, and their ratings, based on the search terms of the customers (for example, location, period, amenities). The rating of hotels can, and does, play a role in the decision of customers to reserve a hotel room (Eslami et al. 2017). The hypothetical example considers that most of the normative positions in the framework are met, with three exceptions where risks are identified

One risk of the rating algorithm is that its purpose is not clearly described and shared with the users (normative statement 2.1 is not met), leading to users being confused by the rating system and eventually not using the rating provided. As an example, users of the rating system might be under the impression that only user reviews matter for the rating, while the rating system has the overall purpose of "rating hotels based on user input and our own domain knowledge". The introduction of domain knowledge in the algorithm can lead to a different rating (for example, a higher rating) than the one obtained from using only user input (for example, averaging user ratings). The auditor of such an algorithm can check the website of the online travel company to verify if the purpose of the rating algorithm is clearly described.

Another risk of the rating algorithm is that its performance is not checked per sub-group (normative statement 4.6 is not met). If the algorithm has a lower performance for the low-to-medium quality hotels sub-group, then users can lose trust in the rating and stop using it. In this situation, the algorithm developer can compare the rating of hotels which are rated on different rating platforms (benchmarking). If the results show a significant difference between the ratings for different hotel categories (for example, low-to-medium quality) on different platforms (for example, Booking versus Expedia), then this is an indication that the algorithm might not perform as expected per sub-groups. The auditor might check if the developer has done the benchmarking and if the validator has challenged the benchmarking.

Rating algorithms which use external user input suffer from a risk of invalid data (normative statement 5.1 is not met). Fake user reviews used in the rating algorithm can materialize into incorrect ratings which lead to a public mistrust in the rating system. To get protection against fake user reviews, the algorithm developer can prompt users to provide a valid email address when making an account on the rating platform or can use other algorithms to check for the existence of fake user accounts. Here too, the auditor would verify the work done by the developer and whether the validation function sufficiently challenged the developer.

6. Conclusion

The wide use of algorithms and the breadth of their impact on society have created pressure for governmental institutions to set regulatory boundaries. The EU AI Act is one regulation spurred from this pressure. The EU AI Act prompts organizations to systematic review the algorithms they use and the ethical risks these algorithms pose to society. Such a systematic look can take the form of an algorithm audit. Moreover, an algorithm audit can play a role in managing the risks of algorithms even when ethical risks are not at the forefront. This is important as algorithms come with serious business risks ranging from providing unreliable information for internal decision-making to fueling public loss of reputation.

The task of auditing algorithms is not an easy one. There is uncertainty as to who can audit algorithms and how a rigorous audit can be done. The internal audit function can play a role in providing insights into the risks that come with the use of algorithms even when ethical risks are not at a forefront. In this article we argue that internal auditors have in their arsenal processes and skills that can be adapted to the new challenge of auditing algorithms. We specifically pinpoint the Model Risk Management approach as something familiar to the internal audit profession and a good starting point for designing a response to the need for algorithm audit. In this way, by keeping connected to the established terminology of the auditing process (e.g., audit objectives, audit criteria) and knowledge (i.e., the Model Risk Management), we aim to show that internal auditors play an important role in assessing the governance and risk management around the use of algorithms, but also the algorithms themselves.

The article produces a framework aimed to serve as an initial guide for the audit of algorithms. The framework develops audit objectives based on the audit criteria revealed by the literature review but offers only an anecdotal example as to how the framework can be applied for an audit. Further investigation is required into the comprehensiveness of the framework, the potential challenges it might reveal when applied in the field (e.g., challenges in collecting evidence) and the recommended composition of the audit team charged with the audit of an algorithm (e.g., in terms of expertise or sociocultural background). An important extension of this article would aim at the development of practical guidance to be used by the auditor when it comes to auditing the algorithm itself (through the system) as there is currently a gap between what the role of the internal auditor should be (i.e., focused at looking both around and through the system) and their current role (i.e., of looking around the system).

- **Dr. Iuliana Sandu** is Academic Director Trustworthy and Accountable AI expert practice at the Rotterdam School of Management and Erasmus Center for Data Analytics.
- **Drs. Menno Wiersma CFA, FRM, ERP** is Senior Manager at Protiviti The Netherlands and responsible for Model Risk Management.
- **Ing. drs Daphne Manichand RA** is Associate Director at Protiviti The Netherlands and responsible for Internal Audit.

References

- Adkins D, Alsallakh B, Cheema A, Kokhlikyan N, McReynolds E, Mishra P, Procope C, Sawruk J, Wang E, Zvyagina P (2022) Prescriptive and descriptive approaches to machine-learning transparency. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 1-9. https://doi.org/10.1145/3491101.3519724
- Algemene Rekenkamer (2021) Aandacht voor Algoritmes. https://www.rekenkamer.nl/binaries/rekenkamer/documenten/rapporten/2021/01/26/aandacht-voor-algoritmes/Aandacht+voor+algoritmes.pdf
- Alvi M (2022) Algorithms Used in Robotics – Explained, HashDork.com. https://hashdork.com/algorithms-used-in-robotics-explained/
- Arnold C (2021) Ethics, technology, and the professional accountant in the digital age. International Federation of Accountants.
- Barnett AJ, Schwartz FR, Tao C, Chen C, Ren Y, Lo JY, Rudin C (2021) A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence 3(12): 1061-1070. https://doi.org/10.1038/s42256-021-00423-x
- BBC (2020) Uber's self-driving operator charged over fatal crash. https://www.bbc.com/news/technology-54175359
- Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency (2011) SR Letter 11-7, Attachment, Supervisory Guidance on Model Risk Management. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.pdf
- Bandy J (2021) Problematic machine behavior: A systematic literature review of algorithm audits. Proceedings of the acm on human-computer interaction 5(CSCW1), 1–34. https://doi.org/10.1145/3449148
- Burnham KP, Anderson DR (2002) Model selection and multimodel inference (2nd Edn.), 488 pp.
- Burrell J (2016) How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data & Society 3(1): 1–12. https://doi.org/10.1177/2053951715622512
- Cambridge Dictionary (2022) Algorithm. https://dictionary.cambridge.org/dictionary/english/algorithm
- Carawan M, Schabel S (2018) Auditing Model Risk Management. The Institute of Internal Auditors, 4-5. https://www.iia.nl/SiteFiles/IIA%20Congres/2018/Presentaties/Colin%20Farquhar%20-%20Auditing%20Model%20Risk%20Management.pdf
- DAMA UK Working Group on 'Data Quality Dimensions' (2013) The six primary dimensions for data quality assessment. http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper-version-1.2-d.d.-3-Sept-2020.pdf
- Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- De Boer M, Van Geijn H (2021) Norea guiding principles trustworthy ai investigations, NOREA De beroepsorganisatie van IT-auditors, 5: 1–41. https://www.norea.nl/download/?id=9720
- Dil S, Baral S, Lau L (2019) Validation of machine learning models: challenges and alternatives, protiviti. https://www.protiviti.com/sites/default/files/united_states/insights/validating-machine-learning-models-whitepaper-protiviti.pdf

- Eslami M, Vaccaro K, Karahalios K, Hamilton K (2017) 'Be careful; things can be worse than they appear': Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In Proceedings of the international AAAI conference on web and social media, Vol. 11, No. 1: 62-71. https://ojs.aaai.org/index.php/ICWSM/article/view/14898/14748
- European Central Bank (2021) Opinion of the European Central Bank of 29 December 2021 on a proposal for a regulation laying down harmonised rules on artificial intelligence. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ%3AJOC_2022_115_R_0005&home=ecb
- European Commission (2021) Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence
- European Commission (2021) Proposal for a Regulation of the European Parliament and of the Council, Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206
- European Parliament (2022) Briefing - Artificial intelligence act. https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf
- European Commission (2022) The Digital Services Act package. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package
- Garla S, Dhillon S (2016) Best Practices for Effective Model Risk Management. SAS, 1–15. https://support.sas.com/resources/papers/proceedings16/SAS6485-2016.pdf
- Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Iii HD, Crawford K (2021) Datasheets for datasets. Communications of the ACM 64(12): 86-92. https://doi.org/10.1145/3458723
- The Institute of Internal Auditors [IIA] (2018). Auditing Model Risk Management. https://iia.no/wp-content/uploads/2019/10/2018-Auditing-Model-Risk-Management.pdf
- The Institute of Internal Auditors [IIA] (2020) The IIA's three lines model. An update on the three lines of defense. https://iia.org.au/sf_docs/default-source/technical-resources/the-iias-three-lines-model---an-update-of-the-three-lines-of-defence.pdf?sfvrsn=4
- The Institute of Internal Auditors (IIA) (2022) The definition of Internal Auditing. https://www.theiia.org/en/standards/what-are-the-standards/definition-of-internal-audit/
- Ismail K (2018) AI vs. Algorithms: What's the Difference? CMSWire. https://www.cmswire.com/information-management/ai-vs-algorithms-whats-the-difference/
- Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nature Machine Intelligence 1(9): 389-399. https://doi.org/10.1038/s42256-019-0088-2
- Kamp AEM (2002) Towards a framework for auditing environmental reports.
- Kazim E, Koshiyama A (2020) AI assurance processes. https://doi.org/10.2139/ssrn.3685087
- Koshiyama A, Kazim E, Treleaven P, Rai P, Szpruch L, Pavey G, Ahamat G, Leutner F, Goebel R, Knight A, Adams J, Hitrova C, Barnett J, Nachev P, Barber D, Chamorro-Premuzic T, Klemmer K, Gregorovic M, Khan S, Lomas E (2021) Towards algorithm auditing: A survey on managing legal, ethical and technological risks of AI, ML and associated algorithms. SSRN Electronic Journal, 35 pp. https://doi.org/10.2139/ssrn.3778998
- Lazer D, Kennedy R, King G, Vespignani A (2014) The parable of google flu: traps in big data analysis. Science 343(6176): 1203–1205. https://doi.org/10.1126/science.1248506
- Makridakis S (2017) The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures 90: 46–60. https://doi.org/10.1016/j.futures.2017.03.006
- Martens D, Provost F, Clark J, Junque de Fortuny E (2016) Mining Massive Fine-Grained Behavior Data to Improve Predictive Analytics, MIS Quarterly 40(4): 869–888. https://doi.org/10.25300/MISQ/2016/40.4.04
- McNamara A, Smith J, Murphy-Hill E (2018) Does ACM's code of ethics change ethical decision making in software development? In Proceedings of the 2018 26th ACM joint meeting on european software engineering conference and symposium on the foundations of software engineering, 729-733. https://doi.org/10.1145/3236024.3264833
- Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency, 220-229. https://doi.org/10.1145/3287560.3287596
- NIST (2022) AI Risk Management Framework: Initial Draft. https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1st-draft.pdf
- Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464): 447-453. https://doi.org/10.1126/science.aax2342
- OECD (2019) OECD Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449) (2019). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
- Orgalim (2022) Position paper - Orgalim amendments to the AI Act. https://orgalim.eu/sites/default/files/attachment/Orgalim%20amendments%20to%20the%20AI%20Act_090222.pdf
- Page LC, Gehlbach H (2017) How an artificially intelligent virtual assistant helps students navigate the road to college. AERA Open 3(4): 2332858417749220. https://doi.org/10.1177/2332858417749220
- Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P (2020) Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, 33–44. https://doi.org/10.1145/3351095.3372873
- Ribeiro MT, Singh S, Guestrin C (2016) 'Why should i trust you?' Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 1135-1144. https://doi.org/10.18653/v1/N16-3020
- Sandvig C, Hamilton K, Karahalios K, Langbort C (2014) An algorithm audit. Data and discrimination: Collected essays, 6–10.
- Securities and Exchange Commission (2018) Transamerica Entities to Pay $97 Million to Investors Relating to Errors in Quantitative Investment Models. Aug. 27th, 2018. https://www.sec.gov/news/press-release/2018-167
- Strusani D, Houngbonon GV (2019) The role of artificial intelligence in supporting development in emerging markets. IFC World Bank Group. https://www.ifc.org/wps/wcm/connect/32e54505-3bfb-419

8-b939-e1e8847715f1/EMCompass-Note-69-Role-of-AI-in-EMs.pdf?MOD=AJPERES&CVID=mNdPiNf

- Suresh H, Guttag J (2021) A framework for understanding sources of harm throughout the machine learning life cycle. In Equity and Access in Algorithms, Mechanisms, and Optimization, 1-9. https://doi.org/10.1145/3465416.3483305
- The Verge (2020) UK ditches exam results generated by biased algorithm after student protests. https://www.theverge.com/2020/8/17/21372045/uk-a-level-results-algorithm-biased-coronavirus-covid-19-pandemic-university-applications
- Thomke S, Manzi J (2014) The discipline of business experimentation. Harvard Business Review 92(12): 17. https://hbr.org/2014/12/the-discipline-of-business-experimentation
- Upadhyay AK, Khandelwal K (2018) Applying artificial intelligence: implications for recruitment. Strategic HR Review 17(5): 255–258. https://doi.org/10.1108/SHR-07-2018-0051
- US President Executive Order 13960, Executive Office of the President (2020) Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government. https://www.presidency.ucsb.edu/documents/executive-order-13960-promoting-the-use-trustworthy-artificial-intelligence-the-federal
- Van Eck M, Middelkoop J (2020) Heeft de auditor een rol bij grip krijgen op algoritmes? Koninklijke Nederlandse Beroepsorganisatie van Accountants, 2 and 5. https://www.nba.nl/globalassets/projecten/visie/kennisgroep-accounttech/2020063_nba_auditor-een-rol-grip-op-algoritmes.pdf
- Vanhée L, Borit M (2022) Ethical By Designer-How to Grow Ethical Designers of Artificial Intelligence. Journal of Artificial Intelligence Research, 73: 619-631. https://doi.org/10.1613/jair.1.13135
- Yaqoob I, Khan LU, Kazmi SA, Imran M, Guizani N, Hong CS (2019) Autonomous driving cars in smart cities: Recent advances, requirements, and challenges. IEEE Network 34(1): 174-181. https://doi.org/10.1109/MNET.2019.1900120
- Zhang CA, Cho S, Vasarhelyi M (2022) Explainable Artificial Intelligence (XAI) in auditing. International Journal of Accounting Information Systems, 100572. https://doi.org/10.1016/j.accinf.2022.100572

Appendix 1

Table A1. The Life cycle framework for the AI algorithm audit.

Lifecycle	Risk	Normative statement (or position) (statements in blue italic font are aimed specifically at AI algorithms)
Initiation	An algorithm is developed lacking adequate governance, without having sufficient support or with too high risk, leading to unnecessarily using resources or creating unnecessary costs.	**1. Evidence of adequate governance**
		1.1 Involved stakeholders (owner, validator, user) are clearly described.
		1.2 The roles of stakeholders are described and followed (for example, the owner approves the use of the algorithm, the validator fulfils the 'second pair of eyes' role for the developer function, the user provides feedback on the use of the algorithm).
		1.3 A clear segregation between roles who need to be segregated (for example, developer and validator) is in place.
		1.4 Diversity within AI development teams across race, gender, sexual orientation, age, economic conditions and more, dependent on potential bias
		1.5 An organizational culture of involving 'experts' for independent feedback on the algorithm in all its life cycle phases, rather than taking algorithm outcomes as truth. (for example, domain experts are involved in the Use phase of the algorithm).
		1.6 The algorithm is correctly listed in the inventory of algorithms.
		2. Preliminary purpose, regulatory environment and risk description
		2.1 The purpose of the algorithm is described and shared with the users.
		2.2 The regulatory risk category (high or low) is determined and specified.
		2.3 The regulatory requirements are listed (for example, *the obligation to perform a conformity assessment as per the EU AI Act*) and used in the development of the algorithm (for example, version control is used in the development of the algorithm).
Development	Inadequate skills are used for the development and/or it is done without sufficiently understanding the context, causing the data or algorithm soundness to miss standards, which compromises the algorithm outcomes.	**3. Documentation requirements are followed:**
		3.1 The internal guidelines (for example, codes of ethics used, *corporate ESG values*) are applied and documented.
		3.2 The input data and the output of the algorithm are described (for example, *datasheets for data sets (Gebru et al. 2021) are used*).
		3.3 How the model works is adequately documented (for example, *model cards (Mitchell et al. 2019) or method cards (Adkins et al. 2022) are used*).
		3.4 A risk registry with all the potential harms that can be caused by the algorithms is in place.
		3.5 For all new algorithms, an impact assessment is performed including the documentation of all the possible risks, including *ethical risks*.
		4. The soundness of the algorithm is evidenced in the documentation:
		4.1 The choice of the algorithm and its settings (for example, *hyperparameters*) is sound and based on theoretical foundations (for example, benchmarked against previous algorithm uses), leading the algorithm to correctly identify relationships existing in the data as opposed to capturing noise in data.

Lifecycle	Risk	Normative statement (or position) (statements in blue italic font are aimed specifically at AI algorithms)
		4.2 The choice of the algorithm is in line with the context where it is applied (*for example, the chosen algorithm is in line with the business purpose; the algorithm development choices with respect to hyperparameters or other settings are in line with the data available (for example, more complex algorithms might require larger data samples))*.
		4.3 Approaches to improve the algorithm (for example, regularization, activation functions, optimizers) are used soundly.
		4.4 Any external tools used (for example, a text parser which extracts features from text data) are understood.
		4.5 Mitigation measures are in place for risks *and they are used conservatively (for example, even if there is only a potential risk to privacy, privacy constraints are placed on the algorithm at development)*.
		4.6 Tests (for example, performance accuracy per sub-groups, sensitivity/scenario analysis, *statistical fairness tests, overfitting detection*) are executed to validate the performance of the algorithm.
		4.7 The results of the algorithm are benchmarked with subject matter expert opinions or other benchmarks (for example, results from other platforms or other algorithms).
		4.8 If necessary, an expert's opinion on the algorithm or on the data is used.
		4.9 If applicable, expert opinions which are overridden are listed and justified.
		4.10 Assumptions and limitations of the algorithm are described (for example, *the model card of the algorithm (Mitchell et al. 2019) describes in what circumstances can the algorithm be used*).
		4.11 The outcomes of the algorithm are in line with corporate ESG values (for example, the results of an algorithm do not interfere with the value of 'diversity').
		4.12 For high-risks applications, decisions made by algorithms can be explained (Zhang et al. 2022) and are interpretable by human subjects.
		5. Data quality is satisfactorily described in documentation:
		5.1 Data is of quality: complete (for example, *data is not biased such that it misrepresents protected groups*), consistent, unique, timely, accurate, valid, complete (DAMA UK Working Group on 'Data Quality Dimensions', 2013), *representative for the population where the algorithm will be used*.
		5.2 Data transformations (for example, scaling, missing data imputation, *feature engineering*) are correct.
Implementation	The implementation does not match with the developed algorithm, data feed quality is poor, or allows for the wrong use of the algorithm, which compromises the outcome during use.	**6. Implementation documentation is up to requirements and in line with development:**
		6.1 The implementation process is documented (for example, the implementation might take place through a randomized controlled experiment (Thomke and Manzi 2014)).
		6.2 The algorithm design is specified (*for example, in a method card (Adkins et al. 2022)*).
		6.3 Changes to the algorithm or the data are described and documented
		6.4 There is Functional and User Acceptance Testing documentation, especially for external tooling.
		6.5 Technical roles and permissions are defined.
		7. Implementation results (algorithm and output) are in line with the design:
		7.1 The algorithm prototype (code, data, model, output) is in line with its implementation.
		7.2 Tests are performed to discover vulnerabilities (for example, fuzz testing).
Use	The use of the algorithm is not in line with the design, or vice versa, causing the algorithm to give the wrong results which are misaligned with the purpose.	**8. The use of the algorithm is documented, in line with practice and in line with the purpose of the algorithm:**
		8.1 There is documentation as to the use of the algorithm.
		8.2 The use of the algorithm is aligned with its purpose and with the documentation.
		9. Training of staff:
		9.1 The staff has knowledge about how to use the algorithm.
		10. Evidence of a formal possibility to give feedback and actual feedback from users:
		10.1 There is a user feedback loop implemented.

Lifecycle	Risk	Normative statement (or position) (statements in blue italic font are aimed specifically at AI algorithms)
Monitoring	The algorithm monitoring is not timely or does not track the correct indicators, whereby it cannot be tracked if the model continues to perform according to expectations.	**11. Monitoring documentation is in line with requirements and lists indicators with thresholds that signal model performance:**
		11.1 Performance metrics (for example, performance accuracy) and acceptable thresholds are defined.
		11.2 The frequency of monitoring is adequate and followed up (for example, *monitoring might be continuous for self-learning algorithms*).
		11.3 The assumptions and limitations of the algorithm hold for the stated purpose and use of the algorithm.
		11.4 Conditional approvals (for example, the algorithm is accepted for immediate use with additional screening for bias) are monitored .
Review	The algorithm review is not in line with its intended use or done timely, whereby it cannot be assured that the algorithm applied is still sufficiently sound and in line with the intended use.	**12. Documentation is in line with requirements and contains a still-fit-for-purpose analysis, and a conclusion to apply 'reparameterization', 'improve' or 'redevelop' in line with requirements:**
		12.1 There is a review frequency set that is followed up.
		12.2 The review contains a still-fit-for-purpose analysis based on monitoring criteria (for example, the use of the algorithm is still in line with its purpose, there is still sufficient knowledge and understanding about the algorithm).
		12.3 The review provides a description and timing of improvement and planned changes in line with findings / weaknesses.
		12.4 If concluded by the review, reparameterization is performed (for example, *dynamic calibration of algorithms where hyperparameters are automatically recalibrated might be done for self-learning algorithms*)
		12.5 If concluded by the review, improvements are performed.
		12.6 If concluded by the review, redevelopment is performed.
		12.7 Previous issues are resolved, findings and recommendations have been implemented according to plan (for example, mitigate a risk by a specific date).
Retirement	An algorithm and data which are no longer being used are not retired, clogging up the inventory or allowing wrong use or use without proper maintenance, or an algorithm still being used is retired, causing failure of procedures.	**13. Retirement procedures are followed in line with requirements:**
		13.1 Dependencies of other algorithms are documented.
		13.2 There is no data redundancy.
		13.3 The algorithm is correctly reflected in the algorithm inventory.
		13.4 Algorithm versions and data are stored for audit purposes.
Validation	Validation is not in line with the process, or without sufficient skills, causing insufficient challenge on the development, implementation and use, which compromises the algorithm quality.	**14.There are efficient controls in place to ensure proper model implementation:**
		14.1 The internal guidelines followed are documented (for example, *codes of ethics*).
		14.2 There is an evaluation of the risk analysis (for example, including *ESG risks*) and classification (for example, *high risk algorithms are correctly identified*).
		14.3 The algorithm implementation can be replicated from the documentation.
		15. There is a fair challenge on algorithm soundness and data quality:
		15.1 There is an evaluation as to whether the development process is suitable to the underlying problem the algorithm is used for (for example, *the development team is sufficiently diverse*, there is conceptual soundness in the choice of the algorithms).
		15.2 There is an evaluation of the performance of the algorithm (for example, using statistical tests, *k-fold cross validation, under/overfitting analysis*, sensitivity analysis, backtesting).
		15.3 Assumptions and limitations of the algorithm are challenged.
		15.4 There is an evaluation of the data quality.
		15.5 There is an evaluation of the Extract-Transform-Load process to identify potential problems from how the data is collected (for example, *potential bias introduced in the data collection stage*).
		16. Findings and recommendations are in line with the weaknesses found, and with requirements:
		16.1 The Validation provides findings and recommendations in a timely manner.
		16.2 The Validation provides a severity level for the risk (for example, untriaged, informational, low, medium, high, or critical).
		16.3 Developers and users are consulted with respect to the findings, recommendations and severity found.
		16.4 The conclusions from the Validation stage are followed up.