AI and Ethics
https://doi.org/10.1007/s43681-024-00563-x

REVIEW

The rise of checkbox AI ethics: a review

Sara Kijewski[1] · Elettra Ronchi[2,3] · Effy Vayena[4]

Received: 7 April 2024 / Accepted: 21 August 2024

Abstract

The rapid advancement of artificial intelligence (AI) sparked the development of principles and guidelines for ethical AI by a broad set of actors. Given the high-level nature of these principles, stakeholders seek practical guidance for their implementation in the development, deployment and use of AI, fueling the growth of practical approaches for ethical AI. This paper reviews, synthesizes and assesses current practical approaches for AI in health, examining their scope and potential to aid organizations in adopting ethical standards. We performed a scoping review of existing reviews in accordance with the PRISMA extension for scoping reviews (PRISMA-ScR), systematically searching databases and the web between February and May 2023. A total of 4284 documents were identified, of which 17 were included in the final analysis. Content analysis was performed on the final sample. We identified a highly heterogeneous ecosystem of approaches and a diverse use of terminology, a higher prevalence of approaches for certain stages of the AI lifecycle, reflecting the dominance of specific stakeholder groups in their development, and several barriers to the adoption of approaches. These findings underscore the necessity of a nuanced understanding of the implementation context for these approaches and that no one-size-fits-all approach exists for ethical AI. While common terminology is needed, this should not come at the cost of pluralism in available approaches. As governments signal interest in and develop practical approaches, significant effort remains to guarantee their validity, reliability, and efficacy as tools for governance across the AI lifecycle.

Keywords Artificial intelligence · Health · Tools · Governance · AI ethics · Practical approaches

1 Introduction

The rapid advancements in artificial intelligence (AI) and the ethical challenges involved have resulted in a proliferation of guidelines to aid the development and deployment of ethical, trustworthy and responsible AI. These guidelines, produced by a range of actors such as national governments, private companies, and international organizations, set out broad high-level principles, but have so far paid limited attention to how these principles are to be applied or enforced [1]. Further, while representing a crucial first step on which the development of laws, regulation and standards of AI can build on [2, 3], studies have shown that AI ethics guidance suffers low rates of adoption in practice [3–5]. Ethical, principle-based guidance is commonly described as vague, too general and high-level [2, 6, 7], and as largely lacking mechanisms to facilitate enforcement or translation into practice [1, 8]. In an analysis by AlgorithmWatch of more than 160 documents, only ten include practical enforcement mechanisms [8]. This has prompted a call for a transition "from what to how" in AI ethics [9].

Over the last years, substantial work has thus been dedicated to "lowering the level of abstraction" [10] and to translating ethical principles into actionable and specific practical requirements in AI governance [6, 11]. This has spurred the development of a myriad of practical approaches aiming at providing guidance on ethical AI. Among some of the early, most prominent examples, the Assessment List for Trustworthy AI (ALTAI) is a practical tool developed by the High-Level Expert Group on Artificial Intelligence (AI HLEG), appointed by the European Commission,

✉ Sara Kijewski
sara.kijewski@hest.ethz.ch

[1] Department of Health Sciences and Technology, Chair of Bioethics, ETH Zurich, Zurich, Switzerland

[2] Data and Digital Health Division of Country Health Policies and Systems World Health Organization Regional Office for Europe, Copenhagen, Denmark

[3] Sciences Po, School of Public Affairs, Paris, France

[4] Department of Health Sciences and Technology, Chair of Bioethics, ETH Zurich, Zurich, Switzerland

Published online: 05 September 2024

to translate their Ethics Guidelines for Trustworthy Artificial Intelligence into a self-assessment checklist for developers and deployers [12]. More recently, the United Nations Educational, Scientific and Cultural Organization (UNESCO) developed their own ethical impact assessment tool [13] to ensure that the development of AI aligns with their Recommendation on the Ethics of AI [14]. Additionally, several governments and public institutions have made extensive efforts to develop frameworks or tools aiming to aid the assessment of the possible impacts of the use of AI (see e.g., the Finnish "Assessment framework for non-discriminatory AI systems" [15], the Ada Lovelace Institute Algorithmic Impact Assessment for AI in healthcare [16], and the Dutch "Fundamental Rights and Algorithms Impact Assessment (FRAIA)" [17]). A number of private companies have also developed open-source tools for the assessment and improvement of the trustworthiness of AI systems (see e.g. Holistic AI Open Source Library [18]) or conformity with standards (e.g. Saimple [19]).

The high number and variety of practical guidance tools and approaches being developed by public and private organizations are reflected in the continuously growing catalog of tools and metrics for trustworthy AI by the Organisation of Economic Co-operation and Development (the OECD Catalogue) which, at time of writing, listed 703 technical, educational and procedural tools (stand: March 2024) designed to aid AI actors in the development and deployment of trustworthy AI systems and applications [20]. The OECD Catalogue represents a highly heterogeneous collection of frameworks, codes, toolkits, checklists, software, standards, guidelines, agreements, developed by a broad set of stakeholders, varying by target group, users, sectors, and purpose.

Along with rules, processes, and procedures, such approaches and tools can aid ensuring legal compliance in the development, deployment and use of AI, as well as adherence to social and ethical standards [21]. The original proposed AI Act [22] legislation (21.4.2021), for example, refers specifically to "harmonized standards and supporting guidance and compliance tools" as enablers of compliance in the development, deployment and use of ethically sound AI. These developments signal the growing need for practical guidance. At the same time, the highly heterogeneous landscape and the patchwork of approaches and tools sound warning bells as there appears to be a lack of consensus about what these tools should achieve, how they are validated, and how they should operate.

Focusing on AI for health, we therefore set out to examine and synthesize evidence from literature reviews of the types of practical approaches available, understand how and by whom they have been developed, for what purpose, whether they have been validated and against what criteria, their limitations, gaps, and whether their impact is known. We use the term "practical approach" to capture all tools, toolkits, frameworks, guidance, and methods available for the promotion of ethical, trustworthy and responsible AI in practice. Our research was further guided by the following considerations: first, if any of these practical approaches are to be widely adopted, their promise should be substantiated by their effectiveness; second, if they are developed as a means to assist organizations in complying with ethical standards, information on their provenance and vetting should be easily accessible and verifiable; third, if use of any such approach aims to give users, or more broadly consumers, assurance of an ethical AI product, it should be clear on what basis this assurance is possible.

Given the heterogeneity in the literature, we opted for a scoping review of existing review/survey articles. This type of review is considered helpful for mapping and assessing the breadth and focus of a body of literature on a particular subject [23], to identify gaps and specific research questions in emerging fields [24, 25], and also to gather important insights into the ways, concepts or terms have been used [26]. In the case of complex and diverse literature, scoping reviews are particularly useful [27].

2 Methodology

We performed a scoping review of peer-reviewed scholarly and gray literature on the practical approaches available for the promotion of ethical, responsible, and trustworthy AI in health published between 2019 and 2023. This scoping review follows the PRISMA extension for scoping reviews (PRISMA-ScR) [28]. The data collection process consists of four steps: identification, evidence screening, eligibility and data capture.

2.1 Search and identification

We examine peer-reviewed and gray literature review/survey articles (literature produced without peer-review by government, academics, business and industry in electronic and print formats). Records were searched for between February and May 2023 in the languages English, German, French, Spanish, Italian, Norwegian, and Finnish.

We adopted a multi-step, systematic and comprehensive search strategy that covered both multidisciplinary and more specific databases and search engines of peer-reviewed and non-peer reviewed literature such as pre-prints and conference articles. As a first step, the initial search was conducted in three databases: Scopus (covering among others the database MEDLINE), Web of Science, and Google Scholar. In a second step, we conducted searches of databases serving specific fields such as IEEExplore (engineering and technology), MedRxiv (medicine), and arXiv (natural sciences, engineering, and economics). The search strategy can be found in the Supplementary Files.

Search strings include terms related to reviews (e.g. also survey) of tools or frameworks (incl. standards, checklists, toolkits, assessment, audits, impacts or practical approaches) for ethical (incl. responsible and trustworthy) AI in health domains (e.g. healthcare, medicine, mHealth, digital health). The third step consisted of a Google web search. This search was performed using various terms related to the review of practical approaches for ethical, trustworthy, or responsible AI in health and was conducted using private browsing mode, after logging out from personal accounts and erasing all web cookies and history. For each search, the 100 first search results were followed and screened for relevance. This added one further non-duplicate record to the body of documents. Finally, we exhausted the practice of citation chaining and examined the reference lists of all of the selected documents and identified one additional non-duplicate document. In total 4284 records (5278 before removing duplicates with Rayyan (web version, Rayyan Systems Inc., Cambridge, MA)) were retrieved. A log of the search strategies and results were kept in a Word-document (Microsoft Word for Mac, version 16.74, Microsoft, Seattle, WA). This process is displayed in Fig. 1. With this broad search strategy, we aimed to reduce the risk of missing relevant documents.

2.2 Screening

The articles identified with the help of academic databases were screened for relevance based on title and abstract by SK, EV, and ER. The documents retrieved through web search were screened following a two-step process. SK first screened their title and summary, and second, retained the documents that reviewed tools aimed at promoting ethical/responsible/trustworthy AI, which were not academic articles. Finally, SK identified documents through citation chaining based on document titles and abstracts. All documents not reviewing, surveying or assessing practical approaches to advance ethical AI were excluded, leaving a body of 57 documents for which we retrieved the full-text documents.

2.3 Eligibility and selection

The documents were independently assessed for eligibility by the three authors. Aiming to examine the landscape of practical approaches to AI ethics in health, our inclusion and exclusion criteria were the following. Documents were included that (1) reviewed more than one practical approach, and (2) focused on ethical/responsible/trustworthy AI in general or in health. The first eligibility criteria specifies that the articles should not merely examine a single approach. Further, we do not only include reviews that narrowly focus on practical approaches for AI in health but also on those that focus on approaches for AI in general, as they may be relevant for AI in health. We excluded records reviewing or evaluating approaches focused only on one specific ethical principle (for example fairness) as furthering ethical, responsible or trustworthy AI requires consideration of more than one ethical principle. Although we do not discount the value of these individual approaches, we aimed at understanding the landscape of comprehensive proposals. Finally, records that were published in another language or which were not an article, but a book, full thesis, magazine, or newspaper article, were also excluded.

Any disagreements on selection were resolved through discussion and consensus among the authors. This assessment of eligibility resulted in a body of 19 documents. The full-text articles were then further reviewed jointly by the three researchers, leading to the inclusion of a total of 17 documents in the final analysis.

2.4 Analysis and data capture

The articles were analyzed by all three researchers to identify common topics or descriptors. The analysis identified two main clusters in the corpus of the articles: one on the characteristics of the practical approaches, and one on barriers to adoption. In a second step, following a deductive approach, we analyzed the documents guided by the following four questions: Which types of practical approaches are available for ethical AI? By whom have they been developed, and for what purpose? Where along the AI lifecycle are they meant to be used and by whom? What are the main barriers to their adoption? The analysis consisted of an iterative process performed using Nvivo (1.7.1 (4844), Lumivero, Denver).

3 Results

The final body of documents includes 17 articles, published between 2020 and 2023, with a majority between 2022 and 2023 (see Table 1). Of all of the articles, 15 were published in scientific journals, one was a working paper, and one was an extract of gray literature. Nine of the articles focus on tools for AI in general [9, 30–37] and eight on AI in health [38–45]. The number of practical approaches examined by the articles vary significantly, ranging from 6 to 121.

In the following sections, we aim to answer the research questions and examine the main characteristics of the practical approaches reviewed and the barriers to their adoption.

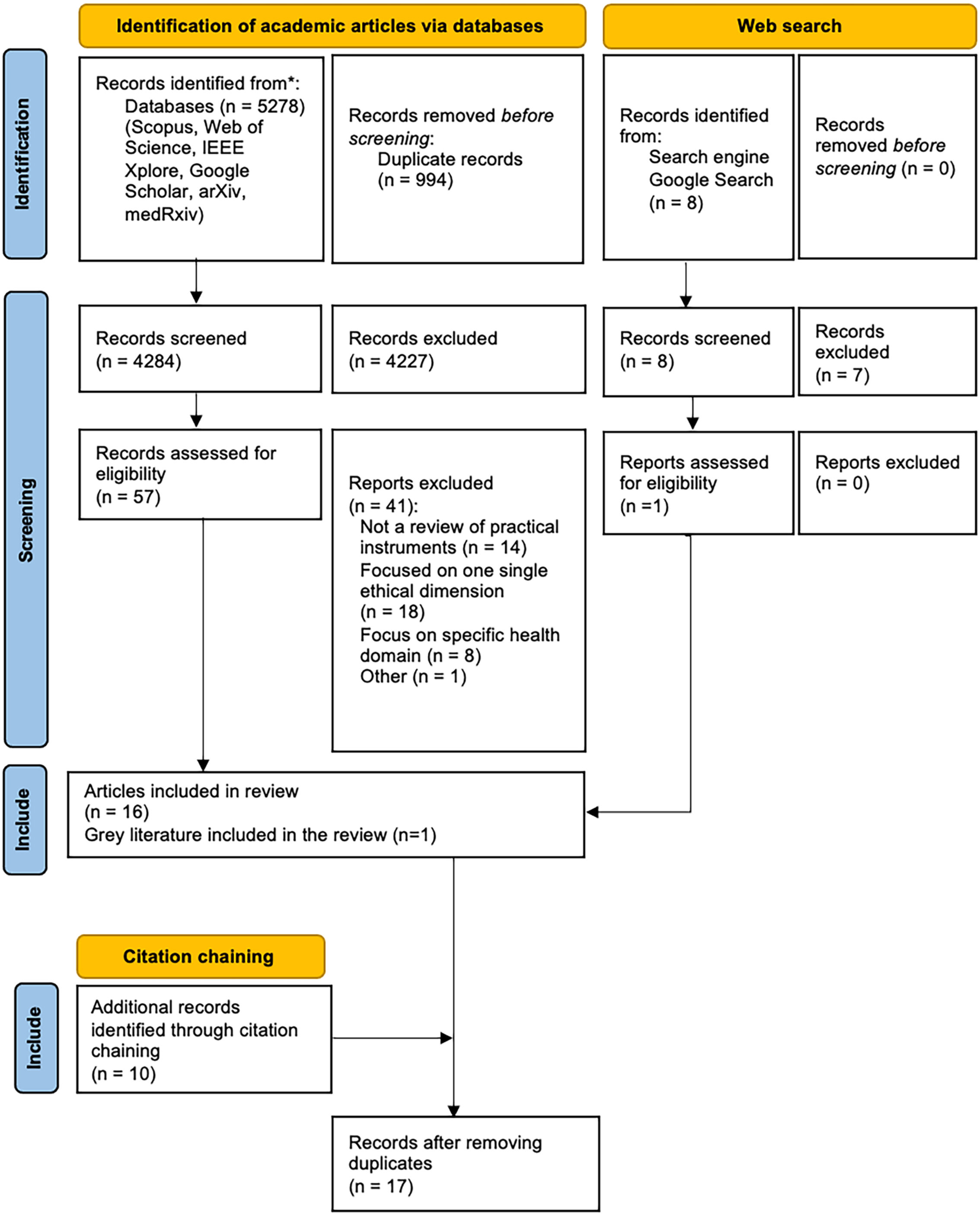
Identification of academic articles via databases
Web search
Identification
Records identified from*: Databases (n = 5278) (Scopus, Web of Science, IEEE Xplore, Google Scholar, arXiv, medRxiv)
Records removed *before screening*: Duplicate records (n = 994)
Records identified from: Search engine Google Search (n = 8)
Records removed *before screening* (n = 0)
Screening
Records screened (n = 4284)
Records excluded (n = 4227)
Records screened (n = 8)
Records excluded (n = 7)
Records assessed for eligibility (n = 57)
Reports excluded (n = 41): Not a review of practical instruments (n = 14) Focused on one single ethical dimension (n = 18) Focus on specific health domain (n = 8) Other (n = 1)
Reports assessed for eligibility (n =1)
Reports excluded (n = 0)
Include
Articles included in review (n = 16) Grey literature included in the review (n=1)
Citation chaining
Include
Additional records identified through citation chaining (n = 10)
Records after removing duplicates (n = 17)

Fig. 1 Adapted PRISMA-ScR (Preferred reporting items for systematic reviews and meta-analyses) 2020 flow diagram for new scoping reviews [29]

3.1 Characteristics of practical approaches for ethical AI

Most articles in our sample set out to examine the practical approaches available, their coverage, their intended users, and the gaps in the current landscape. In total, 15 of the articles classify or organize the reviewed approaches along different dimensions. As an illustration, Ayling and Chapman [33] use the typologies developed from their literature review of sectors, stakeholders, historical practice, and stages in the AI production process as a basis for their classification, while Prem [30] reviews them according

Table 1 Final sample of articles analyzed

Key	Authors	Focus on health	Types of practical approaches examined
[33]	Ayling, J. and Chapman, A. (2022)	No	Audit and assessment tools
[35]	Boza, P. and Evgeniou, T. (2021)	No	Software toolkits and frameworks documentation processes and tools for auditing
[32]	Crockett, K. A. et al. (2021)	No	Toolkits for practical application in SMEs
[43]	Crossnohere, N. L. et al. (2022)	Yes	Frameworks and checklists offering guidance on applying or evaluating AI in medicine
[42]	De Hond, A. A. H. et al. (2022)	Yes	Actionable guidance for AI-based prediction models development, evaluation and implementation
[38]	Garbin, C. and Marques, O. (2022)	Yes	Methods and tools to promote auditing and transparency of datasets and models in ML for healthcare applications
[39]	Goirand, M., Austin, E. and Clay-Williams, R. (2021)	Yes	Technical checklists, organizational and/or evidence-based approaches
[31]	Kaur, D. et al. (2022)	No	Methods, techniques, toolkits, and guidelines
[41]	Lehoux, P. et al. (2023)	Yes	Practice-oriented tools, defined as frameworks and/or sets of principles with clear operational explanations
[45]	Marwood, T. et al. (2022)	Yes	Frameworks and a toolkit for the application of AI in healthcare in Australia
[36]	Minkkinen, M., Laine, J., and Mäntymäki, M. (2022)	No	AI Auditing tools and frameworks
[9]	Morley et al. (2020)	No	Tools and methods to help developers, engineers and designers of ML
[44]	Pradhan, K.B. and Sandhu, N. (2020)	Yes	Frameworks for responsible AI innovation in healthcare
[30]	Prem, E. (2023)	No	Methods and tools (generally defined as Approaches)
[40]	Solanki, P., Grundy, J., and Hussain, W. (2023)	Yes	Guidelines, frameworks, and technical solutions for operationalising ethics in AI for healthcare
[34]	Tidjon, L. and Khomh, F. (2023)	No	Practical guidance of ethical AI principles
[37]	Wong, R.Y., Madaio, M. A, and Merrill, N. (2023)	No	Ethical Toolkits (understood as curated collections of tools and materials)

to the ethical principles addressed, approach category, practicability, and point of intervention in the AI lifecycle.

3.2 Types of practical approaches available for ethical AI

We identified a sizeable and highly heterogeneous body of different practical approaches to help guide ethical implementation. These include not only 'tools, checklists, procedures, methods, and techniques' but also a range of far more general approaches that require interpretation and adaptation such as for research and ethical training/education as well as for designing ex-post auditing and assessment processes. Together, this body of approaches reflects the varying perspectives on what is needed to implement ethics in the different steps across the whole AI system lifecycle from development to deployment. The more than 46 terms used to capture these various practical approaches are rarely defined, and their use is diverse across the examined literature.

The usage of certain terms, e.g. tools and frameworks, is inconsistent across the literature. They are both used as umbrella terms more generally as well as to describe specific approaches. In some of the papers, "tools" is used as a technical term that may specify technical, documentation, implementation or audit and impact assessment tools [33–35], while elsewhere it is used as an umbrella term that also includes toolkits [33, 41]. Similarly, the term "framework" is used to specify both practical tools for application [33] and conceptual models [30, 36]. Occasionally the term is used interchangeably with "guideline" (see e.g. [43],). This ambiguity also applies to the distinction between "tools" and "toolkits". "Toolkits" are by some considered to be collections of resources [32] that include "tools" [37], by others it is understood as a type of "tool" [9, 35, 36].

3.3 Provenance and purpose of practical approaches

Information on the provenance of the practical approaches is reported only by three out of the 17 articles examined [33, 37, 41]. These articles cite technology companies, university centers and academic researchers, non-profit organizations or institutes, open-source communities, design agencies, and government agencies. The private sector has been particularly active in developing the available practical approaches closely followed by academia [33, 37, 41]. A small share of approaches are developed through multisectoral efforts [41].

The diversity of stakeholders developing practical approaches is reflected in the wide range of their intended purposes. While the main objective is to ensure that AI systems are developed, deployed and used in alignment with ethical principles, one study identifies up to 40 distinct purposes. These range from guiding implementation of prominent ethical principles such as nonmaleficence, transparency, privacy and beneficence, to assisting in putting in place ethical data management and addressing feasibility, acceptability and interoperability [41]. Most approaches reviewed by the articles aim to equip stakeholders with the necessary tools or knowledge to address one or few ethical principles [30]. Practical approaches most often seek to advance fairness/bias [9, 30–32, 34, 41], transparency [32, 39, 41, 43, 45], privacy [30, 32, 39, 41], explainability [9, 30, 31], and accountability [30–32].

3.4 Availability of practical approaches throughout the AI lifecycle and target users

The majority of the approaches examined are intentionally designed to aid AI actors at specific stages in the development, deployment and use of AI systems and applications. There is a higher prevalence of practical approaches to guide the design [30, 33, 39] and development of AI systems [38, 43]. Few tools have been developed for the later stages of the AI lifecycle such as monitoring [30, 32, 38, 43], and audit and compliance [32, 41] of AI systems. Whereas frameworks are common for implementation in the design phase, audits, checklists and metrics are more common in the test phase [30]. Finally, it appears that these practical approaches address different ethical principles at various stages of the lifecycle [see e.g., [9, 31, 43]]. To advance explainability, accountability or fairness, for example, most approaches for explainability target the early modeling phases, approaches for fairness are meant to be implemented in the data collection and deployment phase, while approaches promoting accountability often target the planning stage [31].

With regard to target users, most frequently practical approaches target actors involved in the development of AI systems, their delivery, and in quality assurance [33]. Certain categories of practical approaches target distinct groups. Toolkits, for example, are largely aimed at developers, data scientists, designers, technologists, implementation or product teams, analysts and UX teams [37]. Intended users of impact assessment and auditing approaches are mainly decision-makers or actors involved in oversight [33]. Only a few practical approaches target multiple hierarchical levels within organizations [37] or stakeholders outside companies involved in policymaking, governments, civil society organizations, community groups, or users [32, 37]. Few approaches include the broader public in the application of their toolkits [32]. Even where the inclusion of a broad set of stakeholders is a stated aim, guidance is lacking on how to do so in practice [37, 43].

3.5 Barriers to adoption

The articles reviewed here highlight four primary impediments to the adoption and implementation of practical approaches (see Table 2):

1. Skills: Practical approaches are often difficult to use, which can discourage adoption.
2. Absence of Guidance: There is often a lack of clear instructions or support for implementing practical approaches. Without proper guidance, users may struggle to understand how to apply these methods in their specific contexts.
3. Lack of Evaluation Mechanisms and Metrics: Without robust evaluation mechanisms and metrics, it is challenging to assess the effectiveness and impact of practical approaches.
4. Limited Awareness: There is often insufficient awareness of practical approaches among potential users. This can stem from inadequate dissemination of information or a lack of exposure to these methods to relevant target userls.

Limited usability is the most widely cited impediment to the adoption of such approaches. Many practical approaches require a relatively high level of skills, resources and effort to be adopted [9, 30, 32, 33, 37]. The absence of sufficient guidance on how to implement them further impairs their usability [9, 30, 32].

The highly technical nature of toolkits, for example, often necessitates technical skills for their effective utilization. One study finds that, in practice, the actual design and functionality of the majority of toolkits are

Table 2 Obstacles to the implementation of ethical AI practical approaches

High level of skills, resources and effort required for use	[9, 30, 32, 33, 37]
Absence of documentation, specific instructions, examples, case studies or training materials or courses	[32, 34]
Lack of evaluation mechanisms and metrics for success	[31, 39]
Limited awareness of practical approaches	[40]

focused on technical approaches in enacting ethics in AI [37]. This makes it challenging for users without a technical background, e.g. project managers, lawyers or other non-technical stakeholders, to employ such toolkits, or at least presupposes a sufficient level of technical knowledge [37]. Specifically examining small- and medium-sized businesses, another study indicates that the application of toolkits is resource-intensive, demanding substantial time investments from staff resulting in additional workloads [32]. The financial and non-financial costs that may arise for small-sized businesses or organizations in implementing these toolkits can, in practice, act as a disincentive and limit their adoption [9]. Additionally, a large proportion of approaches are general in nature, often aiming at clarifying ethical principles or guiding practitioners with very broad suggestions, lacking specific practical guidance [30].

Many toolkits do not include case studies, use cases, or training material that could facilitate practical application [32]. The most common potential target users such as developers and data scientists, without further guidance and background knowledge in the ethics of AI, would struggle to implement these approaches effectively. Even with standardized approaches and processes [33], applications such as for the assessment of AI models, still would require users to have an understanding of and ability to effectively use the outputs.

More specifically, one study indicates that more than 80 percent of toolkits do not provide educational resources on how to apply them in an organization [32]. Even where courses on AI ethics are made available, another study argues that they most commonly focus on guidelines and standards, rather than raising awareness about the available practical approaches for their effective implementation [34]. This gap in practical guidance, as noted, is particularly challenging for small- and medium-sized businesses, which may require comprehensive training and support materials to successfully implement ethical AI toolkits due to their limited resources [32].

The lack of standards, evaluation mechanisms and measures of successful implementation represents a further obstacle to the adoption of practical approaches [31, 39]. Defining clear indicators for success is essential to assess the efficacy of approaches and determine if they are fit for purpose [39]. Most approaches do not report on whether a formal methodology was used for their development, and none cast light on their real-world applicability and usability and on how their validity, reliability and relevance was ensured [41]. According to one author, less than one-third of practical approaches directly address how to evaluate their successful implementation [39].

While users want concrete, measurable evidence that AI systems meet certain ethical criteria, there is a lack of agreed indicators or systems to test or evaluate these approaches in a way that potential users can understand and trust [31]. Additionally, a lack of awareness also hinders the implementation of such approaches [40]. This lack of awareness may stem from the novelty of practical approaches [31, 39], which may also hinder their adoption.

4 Discussion and conclusion

Our scoping review uncovers a heterogeneous and intricate ecosystem of practical approaches, characterized by diverse and inconsistent terminology as well as lack of consensus on their defining features such as purpose and target audience. At the moment, there appears to be no common understanding of what “tools”, “toolkits” and “frameworks” for ethical AI entail. Clearly defined categories of approaches, and an enhanced understanding of their purpose and capacity, are crucial for policymakers if these approaches are to be implemented for the governance of AI. At the same time, the diversity in terminology, practical approaches, and ethical principles covered by them, implies that there is no single approach to promote AI ethics. The implementation of approaches requires a thorough understanding of the context within which the AI will operate and the potential ethical concerns that must be addressed.

While there is a need to streamline terminology and the understanding of what each specific approach entails, this convergence should not happen at the cost of plurality as some approaches may be more suitable than others depending on context and purpose. There is indeed considerable variation in the way the various practical approaches apply across the AI lifecycle. Few, however, cover all or multiple stages of the AI lifecycle. A substantial share is developed for practical guidance to the earlier phases of the AI lifecycle, i.e., the design and development phases. Practical approaches to guide use and monitoring are largely absent. This may reflect the private sector’s prominent role in both designing and developing AI systems as well as in the creation of practical approaches. Private companies have strong incentives to assess the adequacy of their governance mechanisms in the absence of clear norms or rules and to prevent reputational-related risks. Considering the imperative of aligning AI systems with ethical standards over the entire lifecycle, the accumulation of practical approaches at the early stages highlights the need for a lifecycle-perspective, given the interconnectedness of the different phases. A lifecycle perspective ensures that the potential ethical risks and trade-offs and/or unintended consequences are addressed across the whole process. The existing landscape of practical approaches provides, however, limited assistance for this task.

These considerations bring up three questions:

First, whether the development of practical approaches to AI ethics is emerging as a business opportunity. While this could foster the production of a wide range of available approaches, without proper evaluation and testing of their performance, they are unlikely to sufficiently guide the translation of ethics into practice. The dominance of a few stakeholders such as the private sector and academia in their development has resulted in the "narrowing" down of the ethical requirements addressed by such approaches [46], not reflecting the full breadth of ethical principles. For example, some authors have documented a disproportionate proliferation of approaches addressing specific ethical concerns such as privacy, explainability, fairness [9, 30, 46] and accountability [46].

Second, whether these practical approaches are robust and rigorous enough to be used for monitoring of AI systems and their oversight. While auditing and impact assessment have attracted much attention from researchers, companies, and policymakers, and are considered critical for understanding and minimizing harms from AI systems, there is a paucity of practical approaches for these purposes. This may be symptomatic of the current level of maturity of standards for these processes, largely because of the evolving regulatory environment [47]. Due to the novelty of such practices, professional norms facilitating their implementation are still largely absent [48].

Third, whether effective governance of AI may necessitate context-specific, tailored approaches to ensure adequate oversight. A recently suggested approach involves the development of standards for "ethical disclosure by default" [49]. Rather than imposing uniform ethical norms, this approach would require AI system providers to adhere to minimum standards for procedural consistency in technical testing, documentation, and public reporting. This approach would ensure AI systems are transparent and accountable by design and that users and stakeholders are fully informed about the system's operations, risks, and impacts. This would shift ethical decision-making to stakeholders while limiting the discretion of providers in addressing complex ethical normative issues in the development of AI products and services [49].

Apart from the observed patterns in the current landscape of approaches, the review highlights the presence of significant barriers to their adoption, corresponding to reports of their limited implementation [32, 39, 45]. These include high levels of skills, knowledge and resources required for adoption, the lack of awareness of practical approaches, and the absence of methods or approaches to the measurement of their successful implementation. Measuring successful implementation, as also noted by others [39], is crucial for the assessment of the effectiveness and efficiency of these approaches in advancing ethical AI. There is a clear need for relevant, and practical validation metrics based on standards to assess AI systems' compliance with ethical principles [50]. However, measuring the impact and success of AI ethics in practice remains challenging [9, 51].

Considering that practical approaches are developed largely in the absence of a formal methodology, our understanding of whether they do what they intend and claim to do is limited. Given the diversity of practical approaches and their purposes, the appropriate validation criteria are also likely to vary. Suitable criteria may address their effectiveness and impact, reliability, usability, and stakeholder acceptance.

For the first criteria, the effectiveness and impact of practical approaches on ethical outcomes, quantitative metrics and ethical benchmarks can be useful. Several metrics have recently been proposed, e.g. for fairness [52], which could be used to assess and ensure that such approaches effectively promote ethical principles. Such metrics may also be useful for assessing the reliability and consistency of practical approaches across different contexts. Given the existing barriers to adoption, usability and stakeholder acceptance should be considered as further critical criteria in the validation of approaches. Relevant metrics could include the ease of implementation or the quality of documentation that aids adoption, positive stakeholder feedback or confidence in the practical approach.

While the validation of practical approaches and the measurement of their successful implementation is essential, the low level of their adoption also raises questions concerning incentives: Which incentives are needed to encourage adoption? Will only practical approaches that have "teeth" and aid legal compliance be implemented? Governments are currently taking first steps to enshrine the ethics of AI and data into law. The Danish government adopted a law on disclosure of data ethics, requiring Denmark's largest companies to provide information on compliance with their data ethics policy as part of their annual reporting. Following this, together with business and consumer organizations, the government created a labeling program for IT security and responsible use of data [53]. In Canada, the government has made risk assessment a mandatory step in the design and deployment of systems for automated decision-making. For this purpose, the Algorithmic Impact Assessment Tool [54] was developed. While governments increasingly show interest in and develop these types of practical approaches, much effort still is required to ensure their validity, reliability, and effectiveness if they should be conceived of as governance tools.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43681-024-00563-x.

Funding Open access funding provided by Swiss Federal Institute of Technology Zurich. Fondation Botnar, Swiss National Science Foundation.

Declarations

Conflict of interest EV has consulting fees or other honoraria from Johns Hopkins University and Roche diagnostics; participates on a Digital Ethics Advisory Board for Merck and an Ethics Advisory Panel for IQVIA; and is the co-chair of the WHO expert group on ethics and governance of AI in Health.

References

1. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Minds Mach. **30**, 99–120 (2020)
2. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. New York, NY, USA: Association for Computing Machinery;. pp. 195–200 (2019)
3. Seger, E.: In defence of principlism in AI ethics and governance. Philos Technol. **35**, 45 (2022)
4. Vakkuri, V., Kemell, K.-K., Jantunen, M., Abrahamsson, P.: "This is Just a Prototype": How Ethics Are Ignored in Software Startup-Like Environments. In: Agile processes in software engineering and extreme programming, pp. 195–210. Springer International Publishing, Cham (2020)
5. McNamara, A., Smith, J., Murphy-Hill, E.: Does ACM's code of ethics change ethical decision making in software development? Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. New York, NY, USA: Association for Computing Machinery. pp. 729–733. (2018)
6. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nat Mach Intell. **1**, 501–507 (2019)
7. Khan, A.A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., et al.: Ethics of AI: A systematic literature review of principles and challenges. Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering. New York, NY, USA: Association for Computing Machinery. pp. 383–392. (2022)
8. Haas, L., Giessler, S.,: Thiel V. In the realm of paper tigers – exploring the failings of AI ethics guidelines. In: AlgorithmWatch [Internet]. 28 Apr 2020 [cited 22 Nov 2023]. Available: https://algorithmwatch.org/en/ai-ethics-guidelines-inventory-upgrade-2020/
9. Morley, J., Floridi, L., Kinsey, L., Elhalal, A.: From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci. Eng. Ethics **26**, 2141–2168 (2020)
10. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., Floridi, L.: Ethics as a service: A pragmatic operationalisation of AI ethics. Minds Mach. **31**, 239–256 (2021)
11. Hickok, M.: Lessons learned from AI ethics principles for future actions. AI and Ethics. **1**, 41–47 (2021)
12. High-Level Expert Group on Artificial Intelligence. The Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission; (2020)
13. Ethical impact assessment: A Tool of the Recommendation on the Ethics of Artificial Intelligence. UNESCO;. Available: https://unesdoc.unesco.org/ark:/48223/pf0000386276 (2023)
14. UNESCO. Recommendation on the Ethics of Artificial Intelligence. Paris: United Nations Educational, Scientific and Cultural Organization. Report No.: SHS/BIO/PI/2021/1 (2022)
15. Ojanen, A., Bjork, A., Mikkonen, J.: An assessment framework for non-discriminatory AI. In: Demos Helsinki [Internet]. [cited 13 Nov 2023]. Available: https://demoshelsinki.fi/julkaisut/an-assessment-framework-for-non-discriminatory-ai/ (2022)
16. Algorithmic Impact Assessment: a Case Study in Healthcare. Ada Lovelace Institute. Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/02/Algorithmic-impact-assessment-a-case-study-in-healthcare.pdf (2022)
17. Government of Netherlands. Fundamental Rights and Algorithms Impact Assessment (FRAIA). [cited 7 Nov 2023]. Available: https://www.government.nl/documents/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms (2021)
18. Holistic AI Library. In: Holistic AI [Internet]. [cited 7 Nov 2023]. Available: https://www.holisticai.com/open-source (2023)
19. Saimple: AI Explainability and Robustness Validation Solutions. [cited 13 Nov 2023]. Available: https://saimple.com/ (2023)
20. OECD Catalogue tools for trustworthy AI.: [cited 6 Nov 2023]. Available: https://oecd.ai/en/catalogue/tools (2023)
21. Mäntymäki, M., Minkkinen, M., Birkstedt, T., Viljanen, M.: Putting AI Ethics into practice: The Hourglass Model of Organizational AI Governance. arXiv [cs.AI]. https://doi.org/10.48550/arXiv.2206.00335(2022)
22. European Commission. 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. Commission E, editor. 2021. Report No.: COM/2021/206 final. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206&qid=1658317702094
23. Munn, Z., Peters, M.D.J., Stern, C., Tufanaru, C., McArthur, A., Aromataris, E.: Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. BMC Med. Res. Methodol. **18**, 143 (2018)
24. Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K., Colquhoun, H., Kastner, M., et al.: A scoping review on the conduct and reporting of scoping reviews. BMC Med. Res. Methodol. **16**, 15 (2016)
25. Armstrong, R., Hall, B.J., Doyle, J., Waters, E.: Cochrane update. "Scoping the scope" of a cochrane review. J. Public Health **33**, 147–150 (2011)
26. Pollock, D., Tricco, A.C., Peters, M.D.J., Mclnerney, P.A., Khalil, H., Godfrey, C.M., et al.: Methodological quality, guidance, and tools in scoping reviews: a scoping review protocol. JBI Evid Synth. **20**, 1098–1105 (2022)
27. Peters, M.D.J., Marnie, C., Tricco, A.C., Pollock, D., Munn, Z., Alexander, L., et al.: Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. **18**, 2119–2126 (2020)
28. Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., et al.: PRISMA extension for scoping reviews

(PRISMA-ScR): checklist and explanation. Ann. Intern. Med. **169**, 467–473 (2018)

29. Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow, C.D., et al.: The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372, n71 (2021). https://doi.org/10.1136/bmj.n71
30. Prem, E.: From ethical AI frameworks to tools: a review of approaches. AI and Ethics. **3**, 699–716 (2023)
31. Kaur, D., Uslu, S., Rittichier, K.J., Durresi, A.: Trustworthy artificial intelligence: a review. ACM Comput. Surv. **55**, 1–38 (2022)
32. Crockett, K.A., Gerber, L., Latham, A., Colyer, E.: Building trustworthy AI solutions: a case for practical solutions for small businesses. IEEE Trans. Artif. Intell. **4**(4), 778–791 (2021)
33. Ayling, J., Chapman, A.: Putting AI ethics to work: are the tools fit for purpose? AI and Ethics. **2**, 405–429 (2022)
34. Tidjon, L.N., Khomh, F.: The different faces of AI ethics across the world: a principle-to-practice gap analysis. IEEE Trans. Artif. Intell. **4**, 820–839 (2023)
35. Boza, T., Evgeniou, P.: Implementing AI principles: frameworks, processes, and tools. INSEAD;. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783124 (2021)
36. Minkkinen, M., Laine, J., Mäntymäki, M.: Continuous auditing of artificial intelligence: a conceptualization and assessment of tools and frameworks. Digit Soc. **1**, 21 (2022)
37. Wong, R.Y., Madaio, M.A., Merrill, N.: Seeing like a toolkit: how toolkits envision the work of AI ethics. Proc ACM Hum-Comput Interact. **7**, 1–27 (2023)
38. Garbin, C., Marques, O.: Assessing methods and tools to improve reporting, increase transparency, and reduce failures in machine learning applications in health care. Radiol Artif Intell. **4**, e210127 (2022)
39. Goirand, M., Austin, E., Clay-Williams, R.: Implementing ethics in healthcare AI-based applications: a scoping review. Sci. Eng. Ethics **27**, 61 (2021)
40. Solanki, P., Grundy, J., Hussain, W.: Operationalising ethics in artificial intelligence for healthcare: a framework for AI developers. AI and Ethics. **3**, 223–240 (2023)
41. Lehoux, P., Rivard, L., de Oliveira, R.R., Mörch, C.M., Alami, H.: Tools to foster responsibility in digital solutions that operate with or without artificial intelligence: a scoping review for health and innovation policymakers. Int. J. Med. Inform. **170**, 104933 (2023)
42. de Hond, A.A.H., Leeuwenberg, A.M., Hooft, L., Kant, I.M.J., Nijman, S.W.J., van Os, H.J.A., et al.: Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ. Digit. Med. **5**, 2 (2022)
43. Crossnohere, N.L., Elsaid, M., Paskett, J., Bose-Brill, S., Bridges, J.F.P.: Guidelines for artificial intelligence in medicine: literature review and content analysis of frameworks. J. Med. Internet Res. (2022). https://doi.org/10.2196/36823
44. Pradhan, K.B., Sandhu, N.: Framework to measure responsible innovation compliance in artificial intelligence innovations in healthcare: a review. J. Crit.ical Rev. **7**, 587–590 (2020)
45. Marwood, T., Boyd, J., Khan, U.R., Jade Barclay, S., Jackson. K.: The ethical application of AI in health: a desktop review. digital health crc;. Available: Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3783124 (2022)
46. Palladino, N.: A "biased" emerging governance regime for artificial intelligence? How AI ethics get skewed moving from principles to practices. Telecomm Policy. **47**, 102479 (2023)
47. Costanza-Chock, S., Raji, I.D., Buolamwini, J.: Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery, pp. 1571–1583. (2022)
48. Selinger, E., Leong, B., Cahn, A.F.: AI Audits: Who, When, How... Or Even If? Collaborative intelligence: how humans and AI are transforming our world. MIT Press (forthcoming); (2023)
49. Laux, J., Wachter, S., Mittelstadt, B.: Three pathways for standardisation and ethical disclosure by default under the european union artificial intelligence act. Comput. Law Secur. Rev. (2024). https://doi.org/10.1016/j.clsr.2024.105957
50. Zhou, J., Chen, F.: AI ethics: from principles to practice. AI & Soc. **38**, 2693–2703 (2023)
51. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman L, Hutchinson, B, et al.: Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. New York, NY, USA: Association for Computing Machinery pp. 220–229. (2019)
52. Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I.G., Cosentini, A.C.: A clarification of the nuances in the fairness metrics landscape. Sci. Rep. (2022). https://doi.org/10.1038/s41598-022-07939-1
53. D-seal. Available: https://d-maerket.dk/wp-content/uploads/2023/12/D-seal-pamphlet-english-version_1.0.0.pdf (2024)
54. Treasury Board of Canada Secretariat. Algorithmic Impact Assessment Tool. Available: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html (2024)