AI and Ethics (2023) 3:699–716
https://doi.org/10.1007/s43681-023-00258-9

REVIEW

From ethical AI frameworks to tools: a review of approaches

Erich Prem[1]

Received: 29 July 2022 / Accepted: 5 January 2023 / Published online: 9 February 2023

Abstract
In reaction to concerns about a broad range of potential ethical issues, dozens of proposals for addressing ethical aspects of artificial intelligence (AI) have been published. However, many of them are too abstract for being easily translated into concrete designs for AI systems. The various proposed ethical frameworks can be considered an instance of principlism that is similar to that found in medical ethics. Given their general nature, principles do not say how they should be applied in a particular context. Hence, a broad range of approaches, methods, and tools have been proposed for addressing ethical concerns of AI systems. This paper presents a systematic analysis of more than 100 frameworks, process models, and proposed remedies and tools for helping to make the necessary shift from principles to implementation, expanding on the work of Morley and colleagues. This analysis confirms a strong focus of proposed approaches on only a few ethical issues such as explicability, fairness, privacy, and accountability. These issues are often addressed with proposals for software and algorithms. Other, more general ethical issues are mainly addressed with conceptual frameworks, guidelines, or process models. This paper develops a structured list and definitions of approaches, presents a refined segmentation of the AI development process, and suggests areas that will require more attention from researchers and developers.

Keywords Ethics · Ethical principles · Principlism · AI development · Ethics tools

1 Introduction

Since the inception of Artificial Intelligence, scholars have debated the potential pitfalls, shortcomings, threats, and negative impacts of AI systems [137–139]. Given the experimental and laboratory character of early AI systems, many of these discussions remained mostly theoretical. Although the earlier AI was used, for example, in parameter optimisation, control systems, and later in language and speech processing, the focus of earlier ethical debates were often on aspects that AI systems at that time had not fully realized. This has changed with the advent of AI systems that have become widely used in different applications ranging from robotics to conversational agents and from credit rating to autonomous cars. Although ethical discussions may still have significant hypothetical components, it has also become clear that AI systems raise practical ethical questions regarding their design, use, and longer-term impact.

The intensity of the debate among scholars, policymakers, and, to some extent, among practitioners has reached levels comparable to only a few other technological innovations such as genetics or nuclear power. Although ethical concerns are at least alluded to in the work of Alan Turing, the study of 'computers ethics' intensified in the early 1980s [110, 128]. Since then, the debate has evolved to include a wide range of ethical issues and, in several proposals, how best to address those issues. In addition, researchers in AI and other fields have started to explore diverse directions of research to improve AI systems in response to those ethical concerns, e.g., new technologies for improving the explainability of AI systems. Policymakers have reacted with new rules for designing and operating AI systems. Consequently, there are now hundreds of proposals addressing the ethical aspects of AI systems. So many proposals, frameworks, and ideas have been brought forward that scholars had to systematically analyze them, particularly those relating to "ethical frameworks", e.g. [117, 121].

While the frameworks excel in the identification of ethical issues, they are less convincing in providing practical recommendations for implementation and practice [126, 127, 132]. The main aim of this paper is to review suggestions

✉ Erich Prem
erich.prem@univie.ac.at; prem@eutema.com

[1] Department of Philosophy, The University of Vienna, Vienna, Austria

and approaches in the literature based on the work by Morley and colleagues [127] and to provide a systematic analysis of these ideas from an implementation perspective. The main reason for building on Morley and colleagues is that it was one of the first and most comprehensive analyses of ethical AI principles. The authors study a particularly broad range of ways in which these principles could be implemented in AI systems. Consequently, their work is often referenced, e.g. [126, 140]. The aim is to go from 'what' to 'how', arriving at what Morley et al. called 'the second phase of AI ethics' [127, p. 2147]. I begin with a meta-analysis of the various ethical frameworks to analyze the common structure of principles and to identify the main ethical issues that various approaches are targeting. In a next step, we take a closer look at the design process of AI systems to identify the points at which the approaches can be used. Bootstrapping from the 106 references provided by [127], I propose definitions and a systematic structure for the various approaches. These approaches are then analyzed using the classification scheme and reviewed from an operational and implementation perspective.

For ethicists interested in AI and engineers designing ethical AI systems the results of the analysis exhibit the broad range of proposed approaches while also demonstrating ample room for the development of approaches. The main contributions of this paper are:

- A review and analysis of proposed approaches to creating ethical AI systems confirming that there is a strong focus on algorithmic solutions and a focus on ethical issues for which algorithmic solutions seem possible such as explicability and fairness. The analysis presented here adds privacy and accountability to this list.
- The development of a structured list and definitions of approaches such as software, infrastructure, or methods proposed in the literature for designing ethical AI systems.
- A refined segmentation of the AI development process into nine steps based on the literature and a discussion of the various approaches to be used in the different development steps.
- The identification of possible responses to ethical issues for which only few proposals for technical approaches exist and which will require more work, e.g., labels, good practice, councils, or consent.

The overview and analysis of approaches should lead to a better understanding of the enormously broad design options and their limitations in the building of "ethical" AI systems.

1.1 Ethical AI frameworks

Many frameworks for ethical AI aim to identify potential ethical challenges and propose some remedies to overcome those challenges or mitigate the associated risks. Such frameworks may provide the main relevant *concepts* for discussing the ethical aspects of AI systems and their potential impact, list potential ethical *principles* and *concerns*, and describe rules (in the case of legal frameworks) or *remedies* for addressing them. The latter take the form of recommendations regarding how best to design AI systems or warnings about potential pitfalls. The conceptual dimension usually focuses on explaining concepts underlying ethical principles, e.g., the notions of bias and fairness for AI systems. Principles are often in the form of desirable properties of an AI system, such as transparency of AI systems, privacy of data for the development of AI systems, or human dignity in the application of AI.

In several frameworks, there is little distinction between *concepts* and *principles*. For example, *explainability* can be taken as a requirement (principle) and as a basic concept requiring further conceptual clarification. For *fairness,* several frameworks refer to concepts such as *bias, discrimination, equality* etc. while others may use fairness as both a concept and a principle. Usually, *concepts* are used for the description of concerns, e.g. how fairness may be threatened through unwanted or undetected bias.

Morley and colleagues, for example, describe the *principle* of *explicability* as a requirement for understanding AI systems [127]. Important *conceptual* clarifications include definitions for terms such as understanding, explanation etc. The *concern* is that AI system behaviors should be transparent and that there should be accountability, for which understandability is a prerequisite. There are various possible approaches or *remedies* to make AI systems explicable. They range from documentation to systems that can respond to the "what if" questions and many possible statistical techniques to improve the general understanding of neural network models. Frameworks often frame ethical problems in the form of design challenges that can be addressed by technological means.

As early as 1986, Richard O. Mason proposed four ethical issues of the information age: *privacy, accuracy, property,* and *accessibility*. These principles were not originally conceived for the field of data analytics nor artificial intelligence. Rather, Mason discussed them in the context of the onset of the 'information age', the production of intellectual property, error, and human dignity. Mason's moral imperative was to '*insure that information technology, and the information it handles, are used to enhance the dignity of mankind'* [125, p. 12]. Information systems should not

Table 1 Typical components of ethical frameworks

Concepts	Basic notions relevant for debating the ethical aspects
Principle	Ethical principles (e.g., values)
Concern	How principles are threatened through AI systems use and development
Remedy	Strategies, rules, and guidelines for addressing the concern

unduly invade privacy, must be accurate, protect intellectual property, and be accessible (for all).

Although rather basic, Mason's analysis already contains much of the structure (as defined in Table 1) that later moral recommendations about IT systems would follow. It starts with the explanation and discussion of basic concepts (e.g. responsibility, fidelity etc.) and ethical principles (e.g. accuracy), describes how these values come under threat, and offers some recommendations. Similarly, in its AI white paper [112], the European Commission (EC) discusses relevant ethical concepts (e.g. resilience, human agency and many more) and presents seven key requirements identified in the *Ethics Guidelines for Trustworthy AI* of the High-Level Expert Group [111]:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination, and fairness
- Societal and environmental wellbeing
- Accountability

Besides the general recommendation that trustworthy AI systems adhere to these *principles*, the guideline also includes an *assessment list* for practical use by companies.[1] In the terminology suggested below (cf. definition section and Table 5) such an approach would be characterized as a *checklist*. Given the increasing number of frameworks proposed for developing ethical and responsible AI systems, several scholars published framework analyses and overviews. For example, Floridi and colleagues [115] were among the first to systematically analyze the various ethical guidelines for AI systems. They analyzed 47 principles from six recommendation papers (Table 2); related work [117, 127] reiterated these five principles.

Hagendorff analyzed 21 documents with an average of nine ethical principles [119]. Jobin and colleagues identified *transparency, fairness, non-maleficence, responsibility*, and *privacy* as central to most of the guidelines [121]. Their list contains the previous list of five principles from Table 2 as a subset and the less frequently mentioned *freedom & autonomy, trust, sustainability, dignity*, and *solidarity*.

In summary, scholars who analyzed ethical frameworks arrive at similar or at least compatible results. Their analyses may differ in how they structure various principles, but there is relatively little variation regarding the main identified ethical issues or requirements that the frameworks demand. Although the frameworks can be taken as requirements for an AI systems engineer, they remain very much at the conceptual level without providing clear instructions for *how* to build an ethical system nor what the steps might be to its realization. This lack of concreteness arises to a large extent from the focus on principles as discussed in the next section.

1.2 A focus on principles

Historically, principles have become crucial instruments for achieving ethicality since the advent of modern medical ethics. The onset of this development (also known as *'principlism'*) is dated to the publication of the Belmont report in the late 1970s following ethically dubious medical research [120]. The Belmont report already lists the three basic principles of autonomy, beneficence, and justice that we also find in many ethical AI frameworks. The work of Beauchamp and Childress initiated principlism [107]. Born from real-world application and an urgent demand for practical guidelines, principlism has become a mainstream ethical approach in medical and biomedical practice. Therefore, it is unsurprising that the principles listed in many modern ethical guidelines for research bear so much similarity with those in ethical AI frameworks. There are several caveats, however.

Firstly, policy documents such as the EC guidelines mentioned above strongly focus on principlism and practically adopt or at least implicitly suggest principlism as an approach towards ensuring the ethicality of AI systems. Secondly, philosophical principlism often focuses more on debating their underlying rationales while many framework documents focus on just the set of principles. Thirdly, the principles, although laudable, provide very few concrete constraints on system design. While the frameworks provide a list of ethical objectives, it is far from clear how to realize them and translate them into operationalizable actions [116, 119]. Not only will principles allow for very many different designs, they also do not always lend themselves in a constructive way for designing implementation details. Principles are usually defined at very high levels. They are principles precisely because they are formulated free of any specific context. Therefore, important questions arise within the concrete context of an application as well as the organizational and social context [114]. There is also the additional issue that moral choices are culturally dependent or even

[1] The guidelines were tested by 350 organisations in 2019. The results were fed into the guideline revision process.

Table 2 Overview of five ethical principles and related sub-topics as proposed in [117, 127]

Beneficence	Non-Maleficence	Autonomy	Justice	Explicability
Stakeholder participation Protection of fundamental rights Sustainable and environmentally friendly AI	Resilience to attack and security Fallback plan and general safety Accuracy Privacy and Data Protection Reliability and Reproducibility Quality and integrity of the data Social Impact	Human agency Human oversight	Avoidance of unfair bias Accessibility and universal design Society and democracy Auditability Minimisation and reporting of negative impacts Trade-offs Redress	Traceability Explainability Interpretability

Table 3 Excerpt from [127]: ethical principles and corresponding proposal by the authors for system requirement (excerpt)

Principle	Autonomy	Explicability
Requirement	Human agency: users should be able to make informed autonomous decisions regarding AI systems Human oversight: may be achieved through governance mechanisms such as human-on-the-loop, human-in-the-loop, human-in-command	Traceablilty: the data sets and the processes that yield the AI system's decision should be documented Explainability: the ability to explain both the technical processes of an AI system and the related human decisions Interpretability

team-sensitive, i.e., different teams will interpret principles very differently. This may be acceptable but does not often facilitate an assessment of any concrete design choice. It can also make interoperability between systems difficult.

Also, the ethical frameworks for AI provide lists of principles for AI systems in the form of a collection rather than a hierarchy. This means they are to be fulfilled as much as possible, but sometimes may have to weigh against each other. For example, the principle of *beneficence* may have to be weighed against *privacy* in a medical application. Finally, a few principles typically included in ethical frameworks are rather AI-specific. Examples include system bias and traceability, which are ethical concerns that have become most relevant since the advent of AI or other computer systems. Therefore, these principles lack the existing work and analysis that medical ethics scholars devote, for example, to beneficence.

To put principles to practice, linking them with more tangible system requirements is crucial. However, this phase is not just a mere translation; rather, it will mean research and development of tools, techniques, and technologies in their own right. For example, consider the principles presented in [127] (Table 3).

From a system developer's perspective, it is difficult to understand the listed requirements as the exact system specifications required for building an actual IT system. Most of them are neither concrete nor technical. Some are more actionable than the others, such as the requirement to document the decisions of an AI system. For other requirements, however, the level of abstraction is still too high. For example, it remains unclear what 'interpretability' means and how it should be implemented. Like the higher-level ethical principles (e.g., beneficence, explicability, etc.), weighing requirements against each other persists, e.g., weighing beneficence against explicability in situations where a system's explainability may be opposed to its overall accuracy in a medical application.

In addition, some of the problems are very hard to specify with the necessary algorithmic or mathematical precision. Take, for example, the case of removing bias from a model and ensuring that the AI system treats everybody fair. One of the problems is that fairness has many different interpretations, and it is not straightforwardly clear which mathematical fairness function to use in a selected application context, cf. [124]. On the other hand, although such functions may be hard to design ex-ante, a running system will usually always *implicitly* define a function. The question of mapping the notion of fairness onto a computable function thus is inescapable, albeit very difficult to decide without concrete application context.

Most existing frameworks lack application context and do not consider the practice of AI system development such as the typical trial-and-error approach addressed in the next section nor the practice of software development. There is often an implicit assumption that software is fully controlled by one or only a few designers, which completely ignores the fact that modern software systems, including AI systems

Table 4 A 9-step process segmentation (first column) of the AI application development process based on selected process descriptions in the literature

	AI application development process	Saltz and Dewar [131]	Morley et al. [127]	Prem [130]
1	Business and use-case development	Business understanding	Business and use-case development	
2	System design		Design	
3	Data creation			Data creation
4	Data understanding	Data understanding…		Data/knowledge
5	Pre-processing	… preparation		Pre-processing
6	Model training	Modeling	Building AI application	AI system
7	Test and evaluation	Evaluation	Testing	Test and evaluation
8	Deployment	Deployment	Deployment	
9	Monitoring		Monitoring	

are designed from large software repositories and may be the result of a collaboration of hundreds of programmers [123].

An additional open question is whether ethical principles should be optimized at the level of the individual principle rather than at the whole system level, i.e. should we optimize the AI component for fairness and the whole system for beneficience or is it necessary to consider both at the same time? Similarly, it is unclear whether ethicality can be critically addressed through ethical modules, i.e. through parts of a system that operate in line with an ethical requirement. This effectively addresses the question of compositionality: Can an overall ethical system be achieved from the joint operation of components where each part serves to fulfill one of the principles in an ethical framework? Given the lack of application context, principles also do not consider business contexts and socio-technical system aspects. Thus, they ignore many of the dynamics that occur after an AI system has been deployed. Capturing these aspects means to consider the steps of the whole AI system development process as detailed in the next section.

2 The AI system development process

The ethical frameworks listed above are sets of static principles that, for the most part, are formulated as properties of the targeted AI system, i.e., to the properties of the resulting system. The question then arises which are the various steps of the design process for developing ethical systems as different ethical issues are more relevant than others in the different steps. To include ethical aspects, it is necessary to address a broad perspective that goes beyond just AI modeling. In data analysis and machine learning, however, the focus is usually on model construction and the *machine learning pipeline* (e.g. [133, p. 4]) in an iterative trial-and-error fashion. For example, Saltz and colleagues [131, p. 205] propose five steps from the business case to data understanding, modeling, evaluation, and system deployment. Morley and colleagues [127] add a business case development phase instead of data understanding, split modeling into a training/test data procurement, and an AI application building phase. They also add a monitoring phase at the end. With a focus on the underlying epistemology, Prem [130] starts with the problem domain whose properties lie behind an epistemic boundary. The next steps are (1) the creation of data, (2) understanding of the data (the epistemic domain), (3) preprocessing and formatting it to make it fit for (4) the creation of an AI model. Test and evaluation (validation) of the developed AI model (5) conclude the process. Combining these approaches results in the following process segmentation into nine steps from developing the business case to continuous monitoring (Table 4).

The process could be further detailed, for example, to account for the various test and evaluation regimes or phases of deployment (e.g., for test users, early adopters, and broad roll-out), and decommissioning. But the nine steps in this model should capture most of the relevant ethical decision points during the design of a new AI application. Step 1, development of the business model and the use-case will naturally lend itself to considering the overall system *beneficence* and *non-maleficence*. During the system design phase (step 2), issues such as *stakeholder participation* and *human oversight* will need special attention. An important focus of step 3, data creation, will be *ethical ways of data collection*, data acquisition, and data integrity; this extends to step 4, where *data quality and accuracy* need to be investigated together with potential *biases*. Although steps 5 and 6, pre-processing and training, focus mostly on the technical aspects of data presentation and proper model development, these are potential points of intervention in the AI system for ensuring *explicability* or *interpretability* and improving certain biases. Step 7, test and evaluation, is itself an essential point for checking the *accuracy*, performing tests (e.g., against attacks), and creating data for *auditability*. The actual

Table 5 Overview of approaches in [1–106] to address ethical AI issues

Summaries	Notions	Procedures	Code	Infrastructure	Education	Ex-post assessment and agreement
Overviews and introductions	Frameworks and concepts	Process models	Algorithmic methods	Data sets	Training and tutorial	Audit
Case studies and examples	Criteria and checklists	Guidelines and codes of practice	Design patterns	Online communities		License model
	Declarations	Standards	Software libraries			
	Metrics		Software assistants			
Good practice	*Regulation*	*Consulting*		*Ethics councils and boards*	*Coaching*	*Labels, warnings, consent management*

See text for definitions. Approaches in italics are added from the discussion section.

deployment, step 8, and monitoring, step 9, are crucial for evaluating the longer-term *impact on society, democracy*, and the *environment* and developing further improvements that feed back into any of the steps 1 to 8. This scheme is used below to provide more examples from the analysis of approaches to ethical issues.

3 Approaches, methods, and tools

In the last few years, several proposals attempted to address the various ethical issues, including checklists, standards, computer science or mathematical techniques (e.g., for privacy protection), etc. While some proposals are very technical and concrete, others are more general guidelines that require interpretation and adaptation. The laudable work of Morley and colleagues on which I build analyzed more than 100 proposals for '*tools, methods, and research*' [127, title] to help address various ethical issues. In the following, the tools, methods, and various other proposals are called *approaches.*

3.1 Methodology

The analysis of practical proposals to address ethical issues of AI systems is based on the 106 references in Morley et al., i.e. references [1–106] below. Although the paper mentions various approaches, the focus there is more on the ethical aspects rather than on the analysis of the approaches. The list includes academic papers (app. 60%), documents issued by standardization bodies or associations, publications of consultant firms, and online resources such as software collections and online blogs. In many cases, the approaches are described in papers and are also accompanied by online resources such as software or data in online repositories. Four references had to be excluded from the analysis of approaches as they were no longer available online ([1] and [20]) or where no clear proposal at AI system level could be identified. This includes MIT's work on moral machines [62] and [21] which gives recommendations for regulation. The description of a start-up company [40] and the high-level recommendation to include social scientists in the design team in [98] can be considered outliers but are included except in the overview tables.

The definition of approach categories resulted from a bootstrapping process. First, a preliminary categorization of the approaches was performed. All references were analyzed with respect to the approach and the ethical aspects that it addresses. Many papers address more than one ethical issue (e.g. privacy and explainability) and some propose more than one approach (e.g. an algorithm and software). Based on the resulting list, the categorization was refined to eliminate categories with only a few or no references and to better group the categories as presented below. Finally, I analyzed the resulting classification for approaches, ethical issues addressed, and for cross-dependencies between approaches and ethical issues the most frequent approaches (e.g. for algorithms).

The classification of approaches is not in all cases easy and should be carefully interpreted. Several approaches are so generally described that it is not fully clear how to best categorize them. Others claim to address a range of issues while they may only be clearly formulated for some. A guiding principle in my assessment here was the question of practicality and implementability.

3.2 Approach definition and findings

The following defines the categories developed from the list of references and from the bootstrapping process. I have

grouped the approaches into *summaries, notions, procedures, code, infrastructure, education,* and *ex-post assessments* for clarity and presentation only (Table 5).

3.2.1 Summaries

Many articles can be best described as overviews that summarize various ethical concerns and provide examples of problematic ethical issues arising from AI systems. While many of these papers will only provide high-level recommendations, some include concrete case studies that provide directions for the ethical design of novel AI systems.

- *Overviews and introductions*: Introductory texts and overview articles; explanations of (basic) AI ethics concepts.
- *Case studies and examples*: Analyses of ethical aspects of AI applications and algorithms, examples of ethical issues and how they can be addressed.

3.2.2 Notions

Another large group of papers aims at the clarification of concepts, proposes frameworks and lists criteria to facilitate ethical AI design. This includes checklists and proposals for how to declare characteristics of an AI system or its components. Here I also include papers that aim at measuring ethical aspects and propose metrics for their measurement, in more detail:

- *Frameworks and concepts*: Concepts suggested to support the design of ethical AI systems including high-level abstract concepts; frameworks are structures of concepts that serve as a skeleton for addressing ethical aspects of an AI system often serving as a guide and delineating boundaries between different aspects of ethical systems.
- *Criteria and checklists*: This includes criteria and checklists to support decision making in the design, evaluation, or procurement of systems.
- *Declarations*: Statements describing data, algorithms, and systems to provide insights into aspects of an AI system relevant for assessing ethical aspects, e.g. information about training data or potential system bias. This includes proposals regarding form and content of such statements.
- *Metrics*: a definition, system, or a standard of measuring ethical aspects of a system, e.g. fairness or explainability.

3.2.3 Procedures

Several approaches consist in supporting an ethical AI system design process ranging from less formal guidelines to process models and stricter standards.

- *Process models*: The abstract or visual description of a method or workflow to achieve or improve ethical aspects of systems. The description consists of individual, sequential steps or parts that together provide a model for action such as design, planning, assessment, or improvement.
- *Guidelines and codes of practice*: A set of general rules or an outline of a conduct (policy) (often issued by a professional association) that lays out ethical standards for key aspects of AI design. Codes of practice do not usually carry the same force as standards but are often recommended within a community of practice.
- *Standards*: A generally accepted document that provides requirements, specifications, guidelines, or characteristics that can be used consistently to ensure that AI systems are designed in line with ethical considerations. Standards are often formalized and the result of broad stakeholder consultation.

3.2.4 Code

Many papers directly address coding of ethical AI systems either at the level of improved algorithms or at the software level with libraries or design patterns. A few approaches propose software assistants or running code.

- *Algorithmic methods*: Descriptions of computational techniques for implementing or improving ethical aspects of AI systems. This includes pseudocode, graphical representations, and linguistic descriptions from low-level code to mathematical or computing procedures for implementing computational methods, e.g. privacy techniques.
- *Design patterns*: A general, reusable method or good practice to address an ethical aspect based on an existing solution to an already identified problem. Usually design patterns require adaptation to the problem at hand and can go beyond algorithms in including non-computational aspects, e.g. [91].
- *Software libraries*: a collection of computer code (program code) or modules (executable code)
- *Software assistants*: A computer program (running code) or software agents that operates autonomously or in a dialog with the user.

3.2.5 Infrastructure

Under infrastructure I summarize resources such as the provision of online data sets to help support ethical machine learning and online communities of experts or resources for debate etc.

Table 6 Examples of approaches, type of approach, and main ethical issues addressed

Approach	Ethical issues	Examples
Checklist	Information of stakeholders	[129] describe a data ethics checklist for health care applications
Algorithm	Privacy protection Explaining AI models Fairness robustness	[46] describe the use of generative adversarial networks to generate synthetic datasets [88] use saliency mapping for explainability Optimisation techniques designed to protect against attacks on fairness [71]
Metric	Fairness	The IBM 360 toolkit contains fairness metrics for datasets and ML models [95]
Process model	Value elucidation from stakeholders Responsible AI	IEEE P7000 standard for ethical system engineering [74] describe a process for responsible technology development
Software libraries	Transparency Explainability	Algorithms Tips website [26] Open-source Python library for model inspection[a] XAI machine learning library[b]
Audits	Privacy Fairness	A standardized method for privacy audits [60] Algorithmic audits for detecting discrimination [85]
Software	Privacy	A privacy assistant to elucidate user preferences [70]
Data set	Diversity Bias	Ethics Net[c] Equity Evaluation Corpus[d]
Training and tutorial	Robustness	Online tutorial for improving adversarial robustness [51]
License model	Beneficence	Responsible AI licenses[e] preventing irresponsible use

[a] https://github.com/SeldonIO/alibi

[b] https://github.com/EthicalML/xai

[c] https://www.ethicsnet.com/about

[d] https://saifmohammad.com/WebPages/Biases-SA.html

[e] https://www.licenses.ai/ai-licenses

- *Data sets*: Data bases that can help create ethical systems, in particular data sets for training machine learning algorithms.
- *Online communities*: Groups of people such as experts or users that may help realize ethical systems, often organized as networks or online communities. This often includes online links to resources and services (e.g. for cloud data or computation) and online spaces for debate, exchange, or evaluation.

3.2.6 Education

- *Training and tutorial*: educational material informing about ethical aspects including videos.

3.2.7 Ex-post assessments and agreements

Some approaches do not directly address ethical AI systems design but concern measures to be taken after an AI system is developed such as audits, labels, or licenses.

- *Audit*: a formal examination of the ethical aspects of an AI system including its components, requirements, system behavior, data, or its impact on users.
- *License model*: a pattern (usually a text document) that can be used to create legally binding guidelines governing the use, dissemination, or other aspects of an AI system, e.g. liabilities.

Table 6 provides examples based on [1–106][2] and other references for illustration.

3.3 Classification by approach

The classification of references can be performed according to approaches and by the ethics issue addressed by the approach (Tables 7 and 8). Table 7 presents the classification of all references using the categories defined above without the excluded cases and the outliers [21, 40]. Generally, the classification of approaches is straightforward, for example for most of the algorithmic approaches. However, a few cases are more difficult to decide, especially where there are mixed approaches. In these cases, it was either decided that some papers fall into several categories, e.g. they may propose proposals for algorithms and for measurement (metrics) or to focus on the clearest category. In total, the references were classified as yielding 137 approaches. In addition, several works are very general or are otherwise difficult to translate into designs for AI systems, most notably [10,

[2] Morley et al. list over 100 tools here: (last accessed 2022/06/03) https://docs.google.com/document/d/1h6nK9K7qspG74_HyVlT0Lx97URM0dRoGbJ3ivPxMhaE/edit.

Table 7 Classification of the approaches by category

Approach	Nr	References
Overview, introduction	11	[4, 38, 46, 56, 68, 72, 85, 93, 95, 100, 103]
Case study, examples	*5*	[19*, 24, 26, 75, 93, 100]
Framework/concepts	25	[5, 6, 14, 17, 19, 25, 34, 36, 45–47, 52, 71, 72, 75–77, 80, 84, 90, 94, 96, 98, 104, 105]
Criteria/checklist	5	[5, 25, 41, 66, 76],
Declaration	5	[8, 13, 35, 42, 63]
Metric	5	[15, 22, 46*, 55, 95, 102]
Process model	*6*	[25*, 27, 41*, 46*, 60, 64, 67, 74, 77]
Guideline/code of practice	2	[43, 44]
Standard	1	[46]
Algorithmic method incl. design pattern	36	[2, 7, 9, 11, 12, 15, 18, 22–24, 32, 33, 37, 38, 48–50, 53–59, 73, 79, 81, 82, 85, 87, 88, 91, 97, 99, 102, 103]
Software, SW libraries	17	[16, 29, 39, 61, 65, 69, 70, 79, 83, 86, 89, 92, 95, 99, 101–103]
Online community, collection, data sets	7	[3, 26, 28, 30, 31, 40, 106]
Audit	5	[29, 60, 83, 85, 100]
Training and tutorial	4	[51, 69, 93, 95]
License model	1	[78]

*Borderline cases

Table 8 Classification of the approaches by ethical issue addressed

Ethical issue	Nr	References
General ethical aspects	27	[1, 3, 5, 10, 16, 27, 34, 36, 41, 45–47, 62, 64, 66, 74, 75, 78, 80, 84, 85, 90, 94, 98, 100, 104, 105]
Privacy	22	[6, 7, 12, 15, 19, 40, 43, 44, 49, 60, 67–71, 73, 82, 91, 92, 96, 101, 106]
Fairness, bias	21	[13, 15, 17, 18, 20, 29, 32, 48, 50, 54, 56, 59, 71, 81, 83, 95, 96, 101, 102]
Explainability	18	[15, 22, 37–39, 57, 58, 61, 72, 79, 86–89, 97, 99, 101, 103]
Accountability	12	[8, 9, 14, 24, 25, 35, 42, 52, 53, 63, 77]
Transparency	4	[11, 21, 26, 56]
Correctness, accuracy	4	[23, 33, 55, 99]
Diversity	3	[29–31]
Robustness	2	[51, 65]
Reproducibility	2	[76, 93]

17, 52, 70, 71, 74, 90, 94, 96–98, 104]. On the contrary, the proposals in [64, 84] present rather simple concepts.

A large proportion (26%) of the approaches are classified as algorithmic methods, followed by 18% of approaches that focus on conceptual approaches and frameworks, and 12% software approaches. This means that code (algorithms and software) accounts for 39% of the approaches.

3.4 Classification by ethical objective addressed

The following Table 8 lists all references with respect to the ethical issues that they aim to address. Similar to above, some papers are very general and a few address more than one issue. In total, 115 issues were identified in our set. Also note that the data refers to the claims of the authors rather than an evaluation whether a proposed approach achieves its claims.

A relatively large group of 23% of the approaches address ethical issues at a general level. More than 50% of the approaches address only three concrete issues (privacy 19%, fairness and bias 18%, and explainability 16%). Accountability is addressed by 10% of the approaches. The strong presence of privacy, fairness, and explainability issues in the approaches was already noted by Morley et al. Together, the remaining issues are only addressed with 13% of the approaches.

Table 9 Overview of the number of approaches suggested for the most referenced ethical issues

Ethical issue	Privacy	Fairness	Explainability	Accountability	General
Algorithmic method	7	12	12	3	
Framework/concepts	4	3		4	14
Software, SW collections, tools	4	5	8		
Declaration		3		4	
Metrics		3	2		
Case studies		2			2
Overviews		2	3		3
Process models				2	5

3.5 Relations between ethical issues and approaches

Based on the results presented in Tables 7 and 8, it is now possible to study the relation between ethical issues and proposed approaches. For the issues addressed by more than 10 papers, Table 9 gives the number of approaches greater than one addressing the issue (e.g. 7 papers that address privacy suggest approaching the issue with an algorithm or a computational method).

Fairness, explainability, and privacy are most often addressed with algorithms and software approaches. Other topics and more general aspects are often addressed with conceptual frameworks and process models.

3.6 Practicability

A large proportion of the proposed approaches to ethical issues [1–106] are rather general recommendations oriented at ethical principles and operating on the level of basic ethical concepts. They often aim at clarifying ethical principles or guiding practitioners with very broad suggestions but remain impractical in the sense that they do not clearly suggest details of how to implement an AI system. On the other end there are *algorithms* and *technical methods*, which usually address a specific ethical issue such as explainability or privacy. Although they are precisely specified, algorithms and methods will usually require significant knowledge in computing. *Software tools and libraries,* on the other hand, are already programmed modules that are ready-to-use. While algorithms are often published in combination with online libraries, there are only few approaches that propose running software, e.g. in the form of apps, e.g. [70]. Given that many AI systems are black-box models using data from sensors without any explicit reference to people, purpose, intentions, etc., they may not operate on the right conceptual level for explicit ethical inferences. *Metrics* are usually also clearly implementable and specific. They help create system indicators, e.g., fairness metrics, important for comparing and evaluating systems.

In between those extremes lies a range of proposals that describe stepwise procedures, e.g. [25, 46], provide examples [24], or propose the creation of infrastructure to be used during systems design [3, 28]. For example, *checklists* are a well-known instrument to reduce sources of failure during the design and the operation of technical systems, e.g., when running safety-critical technical systems. It seems only natural to apply them for the case of AI systems to avoid known pitfalls, e.g., bias. Similarly, process models provide elements of an activity in template form to ensure that AI system designers follow the recommended steps of a process.

A very different tool are *communities* and *peer-to-peer* networks supporting the ethical design of a system. Like *data sets* (e.g., for ensuring diversity), they can be considered part of an *infrastructure* for building ethical systems. Perhaps less commonly called infrastructure, *license models* can also be part of libraries that help implement limitations on how an AI system can be rightfully used. Finally, and perhaps insufficiently appreciated to date, are *training material* and *tutorials* that help educate AI system designers.

Several approaches work after the systems was designed and may not directly affect the creation phase. For example, *audits* are structured and independent examinations of a system after its completion. Similarly, *declarations* may be used to provide users or regulatory bodies with information about a system. They are a special form of *user information* that is otherwise missing from 5. The approaches vary greatly in their degree of specificity and operationalizability. Algorithms and software libraries mark one end of the spectrum. They are concrete and usually address a very specific potential ethical shortcoming, e.g. regarding bias or explicability. However, these proposals will typically address only a narrow aspect such as optimisation of a classifier for a given definition of fairness [81] or they are limited to very specific model types and AI architectures, e.g. for Bayes classification [18]. One interesting case describes a design pattern for achieving privacy using Unified Modeling Language [91].

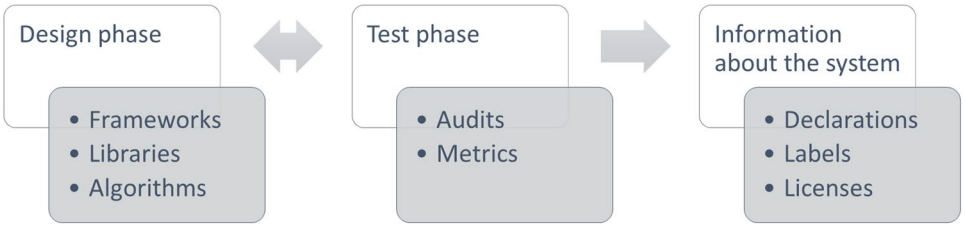

Fig. 1 AI ethics approaches are typically relevant at different points during the development of an AI system. See text for details

Table 10 Overview of approaches and their potential use during the AI development process using the 9-step process model presented in Table 4

	AI development process	Example of proposed approach	Category
1	Business and use-case development	Stimulating public engagement on the ethics of AI [47]	Framework
2	System design	Engineering privacy-by-design [6] IEEE standard for ethical design [46]	Framework Standard, process model
3	Data creation	Datasheets for datasets [35] Data ethics checklist [66]	Declaration Checklist
4	Data understanding	Research method for detecting discrimination [85]	Audit
5	Pre-processing		*n/a*
6	Model training	Confidence-based balancing of fairness and accuracy [33] Optimisation method for fairness in classification [81]	Algorithm Algorithm
7	Test and evaluation	Bias and fairness audit toolkit [83] Evaluation metric for evaluating algorithmic predictions [55] Model cards for trained models [63] Research method for detecting discrimination [85]	Audit Metric Declaration Audit
8	Deployment	Standardized license model to regulate AI system use [78]	License model
9	Monitoring		*n/a*

3.7 Ex ante versus ex-post approaches

The various tools differ substantially regarding the point of intervention in the design process (c.f. Fig. 1). This is an important aspect for understanding how they relate to the AI development process.[3] Frameworks, algorithms, software libraries, etc., usually aim at the ex-ante creation of an ethical system. Audits, checklists, and metrics typically are instruments applied to an AI system once it has been developed, potentially to improve its ethicality in an iterative fashion. Information about a system, declarations and labels can be applied ex-post, including in cases where a comprehensive ethical system is not possible. Communities and data sets can be considered infrastructure supporting the various stages of development.

While the relation of approaches to a high-level perspective of AI system design (Fig. 1) is straightforward, it is more difficult to map the approaches from our data set to the process model described in Table 4. Many approaches can be useful at various steps throughout the development process. For example, the overview, examples, algorithm, and audit procedure described in [85] can be useful for detecting bias of an algorithm or in a data set. Others, especially the frameworks, are broad in addressing many ethical issues. Although they could be relevant for several steps, they are also difficult to operationalize. This is also the reason for not providing a complete categorisation of all approaches to steps here.

Note that some approaches are relevant in a certain development step, but do not necessarily support that step in the development process. For example, [35] proposes the documentation of data sets which needs to be considered during data creation. Its real impact is however later when the system is audited, deployed, or monitored. Similarly, [63] proposes model cards to describe trained model characteristics mostly relevant to users of the system after deployment. Many of the concrete algorithmic methods focus on design features of the AI model (steps 2 and 6). There is, however, a clear lack of approaches in step 5, pre-processing, and in step 9, monitoring in the data. Also, only few authors address the deployment phase (step 8) and the challenge of understanding the data (step 4) used for training AI models. Table 10 provides examples of approaches and how they map on the detailed process model.

[3] [127] also sorted the tools along their list of steps of the AI system development process, i.e., from the business model to monitoring, but with a focus on the ethical issue addressed.

4 Discussion

The analysis of approaches demonstrates a huge interest in improving ethical AI systems design and a broad range of proposals from researchers and practitioners from engineering and other academic fields. Currently, work has intensified on many of the ethical issues. For example, the field of *privacy-preserving* machine learning is now a whole new subdiscipline of machine learning and *explicability* is a major research topic in AI. Similarly, a range of standard process models is being developed with the aim of improving the ethicality of AI systems. For example, IEEE P7000[4] is one of the first standards for ethical system engineering.

A possible reason why fairness and explicability are so often addressed with technical approaches such as algorithms is that these ethical issues appear as properties of the AI system rather than its embedding context. If fairness is mainly understood as a feature of an AI-based classifier, then it is unsurprising that AI engineers aim at improving this function by tuning the learning algorithm. Similarly, if explicability is defined as interpreting an AI system's output in terms of its input, the training data, and the AI system parameters, then engineering work will most likely focus on algorithms that can establish and maintain this relation using concepts that are human-accessible. If, however, fairness is viewed at application level and from a societal perspective, then it is no longer evident that tuning a machine learning algorithm suffices to address the arising ethical issues. Judging an AI system at this level becomes a social and, hence, a political question of what should be considered fair.

Based on the results of the analysis of approaches presented in this paper, such a shift would most likely also entail a shift from the precisely specified algorithms to the other listed tools such as concepts, frameworks, declarations, or process models that more often explicitly consider values or contextual factors. Similarly, some approaches to addressing privacy regard it a mathematical problem about information and data while others may view it as a regulatory issue or one that should be left to an individual's choice. Depending on this stance, an algorithm, code, a regulatory framework, or an information label may be the right answer in terms of which approach to choose for implementing an ethics-oriented AI system. In any case, the various proposed approaches to a single ethical issue, e.g. fairness, are very different from each other. They obviously implement different understandings of what fairness means and have different properties. Designers aiming to improve the ethicality of their AI systems therefore need to carefully consider the different approaches and ensure appropriate design choices.

[4] https://sagroups.ieee.org/7000/.

4.1 Missing ethical issues

As noted by Hagendorff, there are remarkable foci and omissions in the currently developed and published tools [119]. Generally, there is a strong focus on those aspects for which technical solutions *can* be built. The large number of algorithms in our analysis provides further support for this claim. Our categorisation also reconfirms the findings of Morley and colleagues that tools and techniques focus on explainability and on improving fairness. To this list we would add privacy and to some extent also accountability, cf. Table 8.

Democratic control and governance are not central to many AI frameworks and only a few approaches mention these issues at all. Other aspects that are rarely addressed include: existential threats, threats to social cohesion such as echo chambers resulting from algorithmic discourse moderation, abusing AI for political purposes, superiority/inferiority of algorithmic decision making, environmental costs, hidden social costs of AI (e.g., clickworking), or private funding of research, cf. [119]. It is hardly conceivable how such significant ethical issues could be addressed by algorithms or software libraries. However, designing systems with serious consequences is not entirely new in computing. Software engineering for safety-critical systems design developed techniques for managing potentially catastrophic system failure. Today, the design of safety-critical systems follows strict rules and regulations, well-documented methodologies, and certification to ensure acceptable risks, for example, in control systems for nuclear power plants or airplanes. Also, there is work on democratic oversight of systems, albeit much is still in an early phase of concepts and frameworks [135].

4.2 Missing tools

There are several conceivable tools that are not clearly present in the above analysis. Some companies have started to create *ethics councils* and *boards*. The Facebook (Meta) oversight board is perhaps the most prominent example [122].[5] Although *regulation* is mentioned in some approaches, there is very little on how to use it for ethical systems design; an exception is [44] – a guide to GDPR. It is likely that more approaches will address regulation given that there is a clear trend towards more regulation of AI systems, e.g., prohibiting certain use cases such as face recognition[6] [134] or the case of the proposed new EU AI regulatory framework [113].

[5] https://oversightboard.com/.

[6] Biometric Information Privacy Act, IL, US. https://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=3004&ChapterID=57.

An essential tool, especially in industrial practice, is *coaching* and *consulting*. Although the degree to which ethical questions can be delegated to a specialist outside an organization is debatable, consulting plays a crucial role in (ethical) AI systems design, given that many organizations (private and public) lack in-house expertise in the field. Except for the above-mentioned *case studies,* there are very few publications of *good* or *best practices* in ethical design for AI. This situation is very different from, for example, the medical field, where ethical practices are often documented, published, and discussed. This is not the case in software engineering, which generally has a culture very different from the medical field [126].

Finally, *labels* are missing from Table 5 but certainly an important approach used in other technical fields, e.g. labels for consumer white goods. Labels to address accountability are, for example proposed in [108].[7]

As mentioned before, the tools currently support some steps of the AI development process better than others. There is a clear lack of systematic, operationalizable approaches for AI ethics monitoring, only little on deployment, and very little on data creation in the chosen data set.

Many efforts to devise ethics tools assume that ethical problems are solvable in principle, i.e., they are focused on addressing challenges with the intention to completely overcome the ethical issues. A substantially different situation arises when the system cannot be improved towards higher ethical standards. For example, a medical classification system may be developed based on a limited data set that is neither diverse nor unbiased, e.g., it may lack data for female patients. We may still want to deploy such systems as the creation of a new data set may not be feasible in terms of time or costs and using artificial data may not be able to solve the problem at hand. In such situations, the usual approach is to be transparent and *warn* about the identified potential threat or shortcoming.

Note that these information duties tend to shift the burden of ethical decision making to the user. Besides the significant improvement in the transparency (mostly understood as explicability of AI decisions), there are few tools, standards, or guidelines regarding information provision for users or other stakeholders of AI systems. This could concern what information is provided, how it is given, to whom it is addressed, how often the user needs to be informed, and how users effectively *consent*. *Consent* plays a key conceptual role in the practice of digital systems, despite criticism[8] that users often lack understanding of the subject and extent of their consent [136, pp. 125 ff.] There is a need for new approaches addressing consent in a systematic and ethically sound fashion.

4.3 Conclusion

This paper revisited more than 100 articles that aim to contribute to the design of ethical AI systems expanding the work of Morley and colleagues. It developed a structured list and definition of proposed approaches to creating ethical AI systems uncovering an extremely broad spectrum of tools and techniques from algorithms to general frameworks and tools that can become components of an AI ethics infrastructure such as data sets, communities, or license models.

The approaches are spanning the whole range of concreteness from coded software to broad conceptual considerations. The latter are often offered in response to overall ethical concerns at societal level or involving several ethics issues at once. The more concrete and well-specified approaches such as code and algorithms mostly address only few ethical issues and usually only one at a time. This confirms and refines previous analyses in showing that many proposed technical approaches focus on only a few ethical issues, e.g. on explicability and fairness of AI systems. Based on the results presented here, privacy and accountability should be added to this list of most frequently addressed ethics issues and to the list of issues most frequently addressed with algorithmic suggestions.

Several analyses that studied the many ethical frameworks for AI resulted in commensurable sets of principles (see Introduction). However, the analysis here shows no such commensurability ensuing from the more technical work on how to address the various ethics issues. Quite to the contrary, the results demonstrate the enormous breadth and variability of the approaches. The fact that approaches to some ethics issues are developing into whole subfields of machine learning (e.g. explainability, fairness) poses the question of whether simple or succinct technical responses are at all feasible. This does not mean that the efforts to improve the ethicality of AI systems are in vain, but it may imply little reason to expect a universally accepted algorithmic solution to even the clearest ethics issues. The large number of approaches already developed and still under development will create a need to study them in much more detail. It will be important to understand their precise features, the contribution they can make to addressing ethical aspects, their limitations, when to use them and how to further improve them. Topics such as labels, user consent, infrastructure for ethical AI system development, and democratic oversight are areas that require more attention from the side of ethicists and AI engineers.

[7] See also the Swiss Digital Trust label: https://digitaltrust-label.swiss/.

[8] https://blogs.scientificamerican.com/observations/electronic-contracts-and-the-illusion-of-consent/ Accessed 10 November 2022.

Therefore, a tool-based approach to ethical AI systems still raises many questions about the relation of ethics and AI designs, cf. [118]. Following both Floridi and Danks [109], there is a need to study the ethicality of AI systems in concrete application contexts.[9] Learning from ethics in medical practices, we can devise a set of standard situations, i.e., application prototypes and collect, publish, and discuss good ethical practices in these situations. Over time, this should help establish an ethical practice and condemn unethical practices, taking into account specific context, domain ethics, and intended purpose. Such approaches may be combined with audits, labels, declarations, and regulation. Given the enormous breadth of possible approaches to designing AI systems, it is unlikely that principlism alone will achieve their ethicality. Just as medical ethics has evolved to establish best practices, tools such as committees, guidelines, and regulations, AI ethics will require much more research into its practical underpinnings from notions to code, best practices, infrastructure such as described above, education, and communities of practice.

Acknowledgements This research was supported by the "Entrepreneurs as role models" project at the University of Vienna. The author works as an independent strategy and technology consultant at eutema GmbH. He is also a lecturer at TU Vienna and the University of Vienna as well as the Vienna University of Applied Arts.

Funding Open access funding provided by University of Vienna.

Data availability All data generated or analysed during this study are included in this published article.

Declarations

Conflict of interest There are no other employment, financial, or other interests that relate to the submitted work.

References

1. a3i. The Trust-in-AI Framework. (n.d.). http://a3i.ai/trust-in-ai (no longer available online, quoted from [127]).
2. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., Wallach, H. A.: Reductions approach to fair classification. ArXiv: 1803.02453 [Cs]. (2018) Retrieved from http://arxiv.org/abs/1803.02453
3. AI Commons. (n.d.) Retrieved from AI commons website: https://aicommons.com/
4. AI Now Institute Algorithmic Accountability Policy Toolkit. (n.d.). Retrieved from https://ainowinstitute.org/aap-toolkit.pdf
5. AI-RFX Procuement Framework. (n.d.). Retrieved from https://ethical.institute/rfx.html
6. Alshammari, M., Simpson, A.: Towards a principled approach for engineering privacy by design. In: Schweighofer, E., Leitold, H., Mitrakas, A., Rannenberg, K. (eds.) Privacy technologies and policy, pp. 161–177. Springer (2017)
7. Antignac, T., Sands, D., Schneider, G.: Data minimisation: a language-based approach (Long Version). ArXiv: 1611.05642 [Cs]. (2016) Retrieved from http://arxiv.org/abs/1611.05642
8. Arnold, M., Bellamy, R. K. 1E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., … Varshney, K. R.: FactSheets: increasing trust in AI services through supplier's declarations of conformity. ArXiv: 1808.07261 [Cs]. Retrieved from http://arxiv.org/abs/1808.07261 (2018)
9. Arnold, T., Kasenberg, D., Scheutz, M.: Value alignment or misalignment—what will keep systems accountable? AAAI Workshops. (2017)
10. Arnold, T., Scheutz, M.: The "big red button" is too late: an alternative model for the ethical evaluation of AI systems. Ethics Inf. Technol. **20**(1), 59–69 (2018). https://doi.org/10.1007/s10676-018-9447-7
11. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS ONE **10**(7), e0130140 (2015). https://doi.org/10.1371/journal.pone.0130140
12. Bassily, R., Thakkar, O., Thakurta, A.: Model-agnostic private learning via stability. ArXiv: 1803.05101 [Cs]. (2018) Retrieved from http://arxiv.org/abs/1803.05101
13. Bender, E.M., Friedman, B.: Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans. Assoc. Comput. Linguist. **6**, 587–604 (2018). https://doi.org/10.1162/tacl_a_00041
14. Binns, R.: Algorithmic accountability and public reason. Philos. Technol. **31**(4), 543–556 (2018). https://doi.org/10.1007/s13347-017-0263-5
15. Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Presented at the NIPS. (2016) https://arxiv.org/abs/1607.06520
16. Trello board for Agile Ethics for AI (HAI). https://trello.com/b/SarLFYOd/agile-ethics-for-ai-hai Accessed 10 Nov 2022.
17. Butterworth, M.: The ICO and artificial intelligence: the role of fairness in the GDPR framework. Comput. Law Secur. Rev. **34**(2), 257–268 (2018). https://doi.org/10.1016/j.clsr.2018.01.004
18. Calders, T., Verwer, S.: Three naive Bayes approaches for discrimination-free classification. Data Min. Knowl. Disc. **21**(2), 277–292 (2010). https://doi.org/10.1007/s10618-010-0190-x
19. Cavoukian, A., Taylor, S., Abrams, M.E.: Privacy by design: essential for organizational accountability and strong business practices. Identity Inf. Soc. **3**(2), 405–413 (2010). https://doi.org/10.1007/s12394-010-0053-z
20. Chowdhury, R. Tackling the challenges of ethics in AI fairness tool. (n.d.) https://www.accenture.com/gb-en/blogs/blogs-cogx-tackling-challenge-ethics-ai (no longer available online, quoted after [127]
21. Citron, D., Pasquale, F.: The scored society: due process for automated predictions. Wash. Law Rev. **89**(1), 1–33 (2014)

[9] D. Danks suggests a two-tier approach of (i) applied basic research figuring out the application and (ii) basic applied research that investigates foundational questions starting from the application context.

22. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence. In: Cerquitelli, T., Quercia, D., Pasquale, F. (eds.) Transparent data mining for big and small data, pp. 71–94. Springer (2017)
23. Dennis, L.A., Fisher, M., Lincoln, N.K., Lisitsa, A., Veres, S.M.: Practical verification of decision-making in agent-based autonomous systems. Autom. Softw. Eng. **23**(3), 305–359 (2016). https://doi.org/10.1007/s10515-014-0168-9
24. Diakopoulos, N.: Algorithmic accountability: journalistic investigation of computational power structures. Digit. Journal. **3**(3), 398–415 (2015). https://doi.org/10.1080/21670811.2014.976411
25. Diakopoulos, N., Friedler, S., Arenas, M., Barocas, S., Howe, B., Jagadish, H., Zevenbergen, B.: Principles for accountable algorithms and a social impact statement for algorithms (n.d.). Retrieved from FAT ML website: http://www.fatml.org/resources/principles-for-accountable-algorithms
26. Diakopoulos, N., Trielli, D., Yang, A., Gao, A.: Algorithm tips—resources and leads for investigating algorithms in society (n.d.). Retrieved from http://algorithmtips.org/about/ https://www.fatml.org/resources/principles-for-accountable-algorithms
27. DotEveryone. The DotEveryone Consequence Scanning Agile Event. (n.d.) Retrieved from https://doteveryone.org.uk/project/consequence-scanning/, https://doteveryone.org.uk/press-events/responsible-tech-2019/. Accessed 10 Nov 2022
28. Ellpha. (n.d.). Retrieved from https://www.ellpha.com/. Accessed 10 Nov 2022
29. Epstein, Z., Payne, B.H., Shen, J.H., Hong, C.J., Felbo, B., Dubey, A., Rahwan, I.: TuringBox: an experimental platform for the evaluation of AI systems. Proc. Twenty-Seventh Int. Jt. Conf. Artif. Intell. (2018). https://doi.org/10.24963/ijcai.2018/851
30. Equity Evaluation Corpus. (n.d.). Retrieved from https://saifmohammad.com/WebPages/Biases-SA.html
31. Ethics Net. (n.d.). Retrieved from https://www.ethicsnet.com/about
32. Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. ArXiv: 1412.3756 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1412.3756 (2014)
33. Fish, B., Kun, J., Lelkes, Á. D.: A confidence-based approach for balancing fairness and accuracy. ArXiv: 1601.05764 [Cs]. Retrieved from http://arxiv.org/abs/1601.05764 (2016)
34. Friedman, B., Hendry, D.G., Borning, A.: A survey of value sensitive design methods. Found. Trends® Human–Comput. Interact. **11**(2), 63–125 (2017). https://doi.org/10.1561/1100000015
35. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., Crawford, K.: Datasheets for datasets. ArXiv: 1803.09010 [Cs]. Retrieved from http://arxiv.org/abs/1803.09010 (2018)
36. Glenn, J. (n.d.). Futures wheel. Retrieved from ethics kit website: http://ethicskit.org/futures-wheel.html
37. Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation. ArXiv: 1309.6392 [Stat]. Retrieved from http://arxiv.org/abs/1309.6392 (2013)
38. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 1–42 (2018). https://doi.org/10.1145/3236009
39. Hall, P., Gill, N.: H2O.ai machine learning interpretability resources. (n.d.) Retrieved from https://github.com/h2oai/mli-resources/blob/master/notebooks/mono_xgboost.ipynb
40. Hazy. (n.d.). Retrieved from https://hazy.com/
41. Hesketh, P. (n.d.). Ethics cards. Retrieved from Ethics Kit website: http://ethicskit.org/ethics-cards.html
42. Holland, S., Hosny, A., Newman, S., Joseph, J., Chmielinski, K.: The Dataset nutrition label: a framework to drive higher data quality standards. ArXiv: 1805.03677 [Cs]. Retrieved from http://arxiv.org/abs/1805.03677 (2018)
43. ICO. (n.d.-a). Anonymisation: Managing data protection risk-code of practice. https://ico.org.uk/media/1061/anonymisation-code.pdf
44. ICO. (n.d.-b). Guide to the general data protection regulation (GDPR). Retrieved from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/
45. Ideo.org. (n.d.). The field guide to human-centred design. Retrieved from http://www.designkit.org/resources/1
46. IEEE. (n.d.). Artificial intelligence and ethics in design course program. Retrieved from https://innovationatwork.ieee.org/courses/artificial-intelligence-and-ethics-in-design/
47. Involve, DeepMind. (n.d.). How to stimulate effective public engagement on the ethics of artificial intelligence. Retrieved from https://www.involve.org.uk/sites/default/files/field/attachemnt/How%20to%20stimulate%20effective%20public%20debate%20on%20the%20ethics%20of%20artificial%20intelligence%20.pdf
48. Johansson, F. D., Shalit, U., Sontag, D.: Learning representations for counterfactual inference. ArXiv: 1605.03661 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1605.03661 (2016)
49. Joshi, C., Kaloskampis, I., Nolan, L.: Generative adversarial networks (GANs) for synthetic dataset generation with binary classes. Retrieved from https://datasciencecampus.ons.gov.uk/projects/generative-adversarial-networks-gans-for-synthetic-dataset-generation-with-binary-classes/ (2019)
50. Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., Mullainathan, S.: Human decisions and machine predictions*. Q. J. Econ. (2017). https://doi.org/10.1093/qje/qjx032
51. Kolter, Z., Madry, A.: Materials for tutorial adversarial robustness: theory and practice. (n.d.) Retrieved from https://adversarial-ml-tutorial.org/
52. Kroll, J.A.: The fallacy of inscrutability. Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci. **376**(2133), 20180084 (2018). https://doi.org/10.1098/rsta.2018.0084
53. Kroll, J.A., Huey, J., Barocas, S., Felten, E., Reidenberg, J., Robinson, D., Yu, H.: Accountable algorithms, p. 165. University of Pennyslvania Law Review (2017)
54. Kusner, M. J., Loftus, J. R., Russell, C., Silva, R.: Counterfactual fairness. ArXiv: 1703.06856 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1703.06856 (2017)
55. Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., Mullainathan, S.: The selective labels problem: evaluating algorithmic predictions in the presence of unobservables. Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining-KDD '17 (2017). https://doi.org/10.1145/3097983.3098066
56. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., Vinck, P.: Fair, transparent, and accountable algorithmic decision-making processes: the premise, the proposed solutions, and the open challenges. Philos. Technol. **31**(4), 611–627 (2018). https://doi.org/10.1007/s13347-017-0279-x
57. Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions. ArXiv: 1710.04806 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1710.04806 (2017)
58. Lundberg, S., Lee, S.-I.: A unified approach to interpreting model predictions. ArXiv: 1705.07874 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1705.07874 (2017)
59. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Learning adversarially fair and transferable representations. ArXiv: 1802.06309 [Cs, Stat]. Retrieved from http://arxiv.org/abs/180206309 (2018)

60. Makri, E.-L., Lambrinoudakis, C.: Privacy principles: towards a common privacy audit methodology. In: Fischer-Hübner, S., Lambrinoudakis, C., López, J. (eds.) Trust, privacy and security in digital business, pp. 219–234. Cham (2015)
61. Microsoft. (n.d.). InterpretML - alpha release. Retrieved from GitHub website: https://github.com/Microsoft/interpret
62. MIT. (n.d.). Moral machines. Retrieved from http://moralmachine.mit.edu/
63. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Gebru, T.: Model cards for model reporting. Proc. Conf. Fairness, Account. Transpar. – FAT*19 (2019). https://doi.org/10.1145/3287560.3287596
64. New Economy Impact Model. (n.d.). Retrieved from The Federation website: http://ethicskit.org/downloads/economy-impact-model.pdf
65. Nicolae, M.-I., Sinn, M., Tran, M. N., Rawat, A., Wistuba, M., Zantedeschi, V., Edwards, B.: Adversarial robustness toolbox v0.4.0. ArXiv: 1807.01069 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1807.01069 (2018)
66. ODI. (n.d.). Data ethics canvas user guide. Retrieved from https://docs.google.com/document/d/1MkvoAP86CwimbBD0dxySVCO0zeVOput_bu1A6kHV73M/edit
67. Oetzel, M.C., Spiekermann, S.: A systematic methodology for privacy impact assessments: a design science approach. Eur. J. Inf. Syst. **23**(2), 126–150 (2014). https://doi.org/10.1057/ejis.2013.18
68. ONS. (n.d.). The ONS methodology working paper on synthetic data. Retrieved from https://www.ons.gov.uk/methodology/methodologicalpublications/generalmethodology/onsworkingpaperseries/onsmethodologyworkingpaperseriesnumber16syntheticdatapilot
69. OpenMined. (n.d.). Retrieved from https://www.openmined.org/
70. Orcutt, M.: Personal AI privacy watchdog could help you regain control of your data. MIT Technology Review. Retrieved from https://www.technologyreview.com/s/607830/personal-ai-privacy-watchdog-could-help-you-regain-control-of-your-data/ (2017)
71. Overdorf, R., Kulynych, B., Balsa, E., Troncoso, C., Gürses, S.: Questioning the assumptions behind fairness solutions. ArXiv: 1811.11293 [Cs]. Retrieved from http://arxiv.org/abs/1811.11293 (2018)
72. Oxborough, C., Cameron, E., Rao, A., Birchall, A., Townsend, A., Westermann, C. (n.d.). Explainable AI: driving business value through greater understanding. Retrieved from PWC website: https://www.pwc.co.uk/audit-assurance/assets/explainable-ai.pdf
73. Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., Erlingsson, Ú.: Scalable private learning with PATE. ArXiv: 1802.08908 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1802.08908 (2018)
74. Peters, D., Calvo, R. A.: Beyond principles: a process for responsible tech. Retrieved from Medium website: https://medium.com/ethics-of-digital-experience/beyond-principles-a-process-for-responsible-tech-aefc921f7317 (2019)
75. Peters, D., Calvo, R.A., Ryan, R.M.: Designing for motivation, engagement and wellbeing in digital experience. Front. Psychol. **9**, 797 (2018). https://doi.org/10.3389/fpsyg.2018.00797
76. Pineau, J.: The machine learning reproducibility checklist. Retrieved from https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf (2019)
77. Reisman, D., Schultz, J., Crawford, K., Whittaker, M.: Algorithmic impact assessments: a practical framework for public agency accountability. Retrieved from AINow website: https://ainowinstitute.org/aiareport2018.pdf (2018)
78. Responsible AI Licenses. (n.d.). Retrieved from https://www.licenses.ai/about
79. Ribeiro, M. T., Singh, S., Guestrin, C.: Why should I trust you?": Explaining the predictions of any classifier. ArXiv: 1602.04938 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1602.04938 (2016)
80. Royal Society, & British Academy. Data Management and Use: Governance in the 21st Century. (n.d.) Retrieved from https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf
81. Russell, C., Kusner, M. J., Loftus, J., Silva, R.: When worlds collide: integrating different counterfactual assumptions in fairness. In: Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.) Adv. Neural Inf. Process. Syst. 30 (pp. 6414–6423). Retrieved from http://papers.nips.cc/paper/7220-when-worlds-collide-integrating-different-counterfactual-assumptions-in-fairness.pdf (2017)
82. Ryffel, T., Trask, A., Dahl, M., Wagner, B., Mancuso, J., Rueckert, D., Passerat-Palmbach, J.: A generic framework for privacy preserving deep learning. ArXiv: 1811.04017 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1811.04017 (2018)
83. Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., Ghani, R.: Aequitas: a bias and fairness audit toolkit. ArXiv: 1811.05577 [Cs]. Retrieved from http://arxiv.org/abs/1811.05577 (2018)
84. Sampson, O., Chapman, M.: AI needs an ethical compass. This tool can help. Retrieved from Ideo website: https://www.ideo.com/blog/ai-needs-an-ethical-compass-this-tool-can-help (2019)
85. Sandvig, C., Hamilton, K., Karahalios, K., Langbort, C.: Auditing algorithms: research methods for detecting discrimination on internet platforms. Presented at the Data and Discrimination: Converting Critical Concerns into Productive Inquiry" a preconference at the 64th Annual Meeting of the International Communication Association Seattle, WA, USA. (2014)
86. Seldon.io. (n.d.). Alibi. Retrieved from GitHub website: https://github.com/SeldonIO/alibi
87. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. ArXiv: 1704.02685 [Cs]. Retrieved from http://arxiv.org/abs/1704.02685 (2017)
88. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. ArXiv: 1312.6034 [Cs]. Retrieved from http://arxiv.org/abs/1312.6034 (2013)
89. Sokol, K., Flach, P.: Glass-box: explaining AI decisions with counterfactual statements through conversation with a voice-enabled virtual assistant. Proc. Twenty Seventh Int. Jt. Conf. Artif. Intell. (2018). https://doi.org/10.24963/ijcai.2018/865
90. Stahl, B.C., Wright, D.: Ethics and privacy in AI and big data: implementing responsible research and innovation. IEEE Secur. Priv. **16**(3), 26–33 (2018). https://doi.org/10.1109/MSP.2018.2701164
91. Suphakul, T., Senivongse, T.: Development of privacy design patterns based on privacy principles and UML. Int. Conf. Softw. Eng. Artif. Intell. Netw. Parallel/Distrib Comput (SNPD) (2017). https://doi.org/10.1109/SNPD.2017.8022748
92. TensorFlow Privacy. (n.d.). Retrieved from https://github.com/tensorflow/privacy
93. The Turing Way. (n.d.). Retrieved from https://github.com/alan-turing-institute/the-turing-way
94. van de Poel, I.: An ethical framework for evaluating experimental technology. Sci. Eng. Ethics **22**(3), 667–686 (2016). https://doi.org/10.1007/s11948-015-9724-3
95. Varshney, K. R.: Introducing AI fairness 360. Retrieved from IBM website: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/ , https://aif360.mybluemix.net/ (2018)
96. Wachter, S., Mittelstadt, B.: A right to reasonable inferences: re-thinking data protection law in the age of big data and AI.

Columbia Business Law Review, Forthcoming. Retrieved from https://ssrn.com/abstract=3248829 (2018)
97. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. ArXiv: 1711.00399 [Cs]. Retrieved from http://arxiv.org/abs/1711.00399 (2017)
98. Wellcome Data Labs. (n.d.). A new method for ethical data science. Retrieved from https://medium.com/wellcome-data-labs/a-new-method-for-ethical-data-science-edb59e400ae9
99. Wexler, J.: The what-if tool: code-free probing of machine. Retrieved from https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html, https://pair-code.github.io/what-if-tool/ (2018)
100. Wilson, C.: Auditing Algorithms @ Northeastern. Retrieved from http://personalization.ccs.neu.edu/ (2018)
101. XAI Library. (n.d.). Retrieved from https://github.com/EthicalML/awesome-machine-learning-operations
102. Zafar, M. B., Valera, I., Rodriguez, M. G., Gummadi, K. P.: Fairness constraints: mechanisms for fair classification. ArXiv: 1507.05259 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1507.05259 (2015)
103. Zhang, Q., Zhu, S.: Visual interpretability for deep learning: a survey. Front. Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018). https://doi.org/10.1631/FITEE.1700808
104. Zhao, W.-W.: Improving social responsibility of artificial intelligence by using ISO 26000. IOP Conf. Ser. Mater. Sci. Eng. **428**, 012049 (2018). https://doi.org/10.1088/1757-899X/428/1/012049
105. Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., Pasquale, F.: Ten simple rules for responsible big data research. PLOS Comput. Biol. **13**(3), e1005399 (2017). https://doi.org/10.1371/journal.pcbi.1005399
106. Zyskind, G., Nathan, O., Pentland, A. (2015) Enigma: decentralized computation platform with guaranteed privacy. ArXiv: 1506.03471 [Cs]. Retrieved from http://arxiv.org/abs/1506.03471
107. Beauchamp, T., Childress, J.: Principles of biomedical ethics. Oxford University Press, New York (1979)
108. Bertelsmann (n.d.) From principles to practice. An interdisciplinary framework to operationalise AI ethics. Gütersloh, DE. https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf
109. Danks D.: Digital ethics as translational ethics. In: Vasiliu-Feltes, I., Thomason, J. (Eds.) Applied ethics in a digital world (pp. 1–15). IGI Global. https://www.daviddanks.org/s/TranslationalEthics-Final.pdf (2021)
110. Johnson, D.G.: Computer ethics. Prentice-Hall, Englewood Cliffs, NJ (1985)
111. European Commission. Ethics guidelines for trustworthy AI. Directorate-General for Communications Networks, Content and Technology, EC Publications Office. (2019) https://data.europa.eu/doi/https://doi.org/10.2759/177365
112. European Commission. On artificial intelligence – a European approach to excellence and trust. White paper. COM(2020) 65 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020DC0065&from=EN (2020)
113. European Commission. Proposal for a regulation of the European parliament and of the council laying down harmonised ruled on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM/2021/205 final. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN (2021)
114. Fazelpour, S., Lipton, Z.C., Danks, D.: Algorithmic fairness and the situated dynamics of justice. Can. J. Philos. (2021). https://doi.org/10.1017/can.2021.24
115. Floridi, L., Cowls, J., Beltrametti, M., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind. Mach. **28**, 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
116. Floridi, L.: Establishing the rules for trustworthy AI. Nat Mach Intell **1**, 261–262 (2019). https://doi.org/10.1038/s42256-019-0055-y
117. Floridi, L., Cowls, J.: A unified framework of five principles for AI in society. In: Floridi, L. (ed.) Ethics, governance, and policies in artificial intelligence. Philosophical studies series, vol. 144, pp. 5–6. Springer, Cham (2021)
118. Greene D., Hoffmann A.L, Stark L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical Artificial Intelligence and machine learning. Proceedings of the 52nd Hawaii International Conference on System Sciences, pp. 2122–2131. (2019). https://doi.org/10.24251/HICSS.2019.258 https://hdl.handle.net/10125/59651
119. Hagendorff T.: The ethics of AI ethics: an evaluation of guidelines. In: Mind and machines, 30 (1): 99–120 (2019). https://doi.org/10.1007/s11023-020-09517-8 https://arxiv.org/ftp/arxiv/papers/1903/1903.03425.pdf
120. "HEW News" Office of the Secretary, March 5, 1973; Memorandum "USPHS Study of Untreated Syphilis (the Tuskegee Study; Authority to Treat Participants Upon Termination of the Study," from Wilmot R Hastings to the secretary, March 5, 1973.
121. Jobin, A., Ienca, M., Vayena, E.: Artificial intelligence: the global landscape of ethics guidelines. Nat. Mach. Intell. **1**, 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
122. Klonick, K.: The Facebook oversight board: creating an independent institution to adjudicate online free expression. Yale LJ **129**, 2418 (2019)
123. Lee, E.A.: The coevolution: the entwined futures of humans and machines. MIT Press (2020)
124. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics **1**, 529–544 (2021). https://doi.org/10.1007/s43681-021-00067-y
125. Mason, R.O.: Four ethical issues of the information age. MIS Q. **10**(1), 5–12 (1986)
126. Mittelstadt B.: Principles alone cannot guarantee ethical AI. Nat Mach Intell 1(11): 501–507. (2019) Preprint available from: https://arxiv.org/ftp/arxiv/papers/1906/1906.06668.pdf
127. Morley, J., Floridi, L., Kinsey, L., et al.: 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). https://doi.org/10.1007/s11948-019-00165-5
128. Moor, J.H.: What is computer ethics? Metaphilosophy **16**(4), 266–275 (1985). https://doi.org/10.1111/j.1467-9973.1985.tb00173.x
129. Montague, E., Eugene Day, T., Barry, D., et al.: The case for information fiduciaries: the implementation of a data ethics checklist at Seattle children's hospital. J. Am. Med. Inf. Assoc. **28**(3), 650–652 (2021). https://doi.org/10.1093/jamia/ocaa307
130. Prem, E.: A knowledge-based perspective of strategic AI management. In: Tanev, S., Blackbright, H. (eds.) Artifical intelligence and innovation management. World Scientific (2022). https://doi.org/10.1142/9781800611337_0002
131. Saltz, J.S., Dewar, N.: Data science ethical considerations: a systematic literature review and proposed project framework. Ethics Inf. Technol. **21**, 197–208 (2019). https://doi.org/10.1007/s10676-019-09502-5
132. Munn, L.: The uselessness of AI ethics. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00209-w
133. Hapke, H., Nelson, L.: Building machine learning pipelines. O'Reilly Media, Sebastopol, CA (2020)

134. Yew R. J., Xiang A.: Regulating facial processing technologies: tensions between legal and technical considerations in the application of Illinois BIPA. (2022) arXiv preprint arXiv: 2205.07299.
135. Simpson E., Conner A.: How to regulate tech: a technology policy framework for online services. (2021) https://www.americanprogress.org/article/how-to-regulate-tech-a-technology-policy-framework-for-online-services/ Accessed 9 Nov 2022.
136. Schmitt, J.F.: The impact of privacy laws on websites and users. Cuvillier, Göttingen (2022)
137. Forester, T., Morrison, P.: Computer ethics. MIT Press, Cambridge, MA (2001)
138. Dreyfus, H., Dreyfus, S.E.: Mind over machine. Free Press, New York, NY (1986)
139. Dreyfus, H.: What computers still can't do. MIT Press, Cambridge, MA (1979)
140. Hagendorff, T.: Blind spots in AI ethics. AI Ethics **2**, 851–867 (2022)