AI and Ethics (2022) 2:405–429
https://doi.org/10.1007/s43681-021-00084-x

ORIGINAL RESEARCH

Putting AI ethics to work: are the tools fit for purpose?

Jacqui Ayling[1] · **Adriane Chapman**[1]

Received: 29 May 2021 / Accepted: 10 August 2021 / Published online: 12 September 2021

Abstract
Bias, unfairness and lack of transparency and accountability in Artificial Intelligence (AI) systems, and the potential for the misuse of predictive models for decision-making have raised concerns about the ethical impact and unintended consequences of new technologies for society across every sector where data-driven innovation is taking place. This paper reviews the landscape of suggested ethical frameworks with a focus on those which go beyond high-level statements of principles and offer practical tools for application of these principles in the production and deployment of systems. This work provides an assessment of these practical frameworks with the lens of known best practices for impact assessment and audit of technology. We review other historical uses of risk assessments and audits and create a typology that allows us to compare current AI ethics tools to Best Practices found in previous methodologies from technology, environment, privacy, finance and engineering. We analyse current AI ethics tools and their support for diverse stakeholders and components of the AI development and deployment lifecycle as well as the types of tools used to facilitate use. From this, we identify gaps in current AI ethics tools in auditing and risk assessment that should be considered going forward.

Keywords AI · Ethics · Impact assessment · Audit

1 Introduction

Ethics for AI has been experiencing something of a gold rush in the last few years, with frameworks, guidelines and consultations appearing thick and fast from governments, international bodies, civil society, business and academia. Bias, unfairness and lack of transparency and accountability in AI systems, and the potential for the misuse of predictive models for decision-making have attracted attention across a range of domains from predictive policing to targeted marketing to social welfare [1, 2]. There is disquiet about the ethical impact and unintended consequences of new technologies for society across every sector where data-driven innovation is taking place, and an increasing recognition that even the latest updates to data protection regulation (e.g. GDPR [3]) are not addressing all the ethical issues and societal challenges that arise from these new data pipelines and computational techniques.

This paper sets out to review the landscape of suggested ethical frameworks with a focus on those which go beyond high-level statements of principles (see [4–6] for review of principles), and offer practical tools for application of these principles in the production and deployment of systems. 'Efforts to date have been too focused on the 'what' of ethical AI (i.e. debates about principles and codes of conduct) and not enough on the 'how' of applied ethics' [7, p. 2143]. We can all nod our heads sagely in agreement with principles like fairness and justice, but what does fairness and justice look like in a real-life decision-making context? How are organisations and those within them to reckon with the complex ethical tug-of-war between 'the bottom-line' and upholding ethical principles? To do this, we examine proposed tools to operationalise ethical principles for AI (as opposed to statements of ethical principles), in relation to well-established impact and risk assessment, and audit procedures, that have been used to manage human activities, and new technology.

Societies face a series of complex and difficult problems across multiple domains to which the application of data-driven AI technologies is being eagerly pursued. The ability to collect and store vast troves of data, coupled with increases in computational power, provides the substrate for

✉ Jacqui Ayling
j.a.ayling@soton.ac.uk

[1] Web and Internet Science Group (WAIS), School of Electronics and Computer Science, University of Southampton, Southampton, UK

an explosion of AI applications, particularly machine learning. The kinds of harms that have been of growing concern build on traditional data privacy harms (see for example [8, 9]). Concerns around AI are grouped first around epistemic concerns (the probabilistic nature of insights, the inherent inscrutability of 'black box' algorithms, and the fallibility of the data used for training and input). Then, there are normative concerns about the fairness of decisional outcomes, erosion of informational privacy, and increasing surveillance and profiling of individuals. Algorithmic systems also create problems of accountability and moral responsibility, where it is unclear which moral agent in the process bears (or shares) responsibility for outcomes from a system [10].

Disastrous outcomes like the loss of human life through machine malfunction (think medical applications or autonomous cars), or the hijacking and manipulation of critical systems by bad actors (think military systems, or smart city technologies controlling essential services). These kinds of outcomes pose significant challenges for both government and business and could result in reputational damage, regulatory backlash, criminal proceedings and a loss of public trust [11]. As Daniel Solove presciently noted we risk creating a Kafkaesque world with 'a more thoughtless process of bureaucratic indifference, arbitrary errors, and dehumanization, a world where people feel powerless and vulnerable, without any meaningful form of participation in the collection and use of their information' [12, p. 1398]. It is to meet these challenges that the current interest in ethical frameworks has become so heightened.

In response to increasing public debate and political concern about the negative effects on individuals and wider society of AI, a veritable AI ethics industry has emerged, promoting a variety of different frameworks and tools [13]. Several authors [7, 13–17] have identified different phases in the response to increasing public debate about the impact of AI technologies. In the first phase from 2016 to 2019, many high-level ethical principles for AI were published as evidenced by these catalogues of ethical principles and frameworks for ethical, trustworthy responsible AI [4–6, 18, 19]. This first phase focused on the high-level ethical principles that might best address the impacts of AI and data-driven systems framed as applied ethics and dominated by a philosophical approach as opposed to legal or technical approach.

A second phase saw a more technical approach from the computer science community focusing on fairness, accountability and transparency as an engineering 'ethical-by-design' problem-solving exercise [20–22]. The current phase is seeing a move 'from what to how' [7], with proposals for governance mechanisms, regulation, impact assessment, auditing tools and standards leading to the ability to assure and ultimately, insure AI systems [15]. There is also latterly a shift towards acknowledgement of political, social and justice issues 'beyond the principled and the technical, to practical mechanisms for rectifying power imbalances' [16]. As Crawford [23] argues, AI ethics is not just a 'tech ethics' problem, amenable to 'tech ethics' fixes, but raises deeply political questions about how power is wielded through technology.

Meta-analyses of AI ethics proposals have thus far focused mainly on classifying and comparing the ethical principles suggested, where some convergence can be identified for principles like transparency, fairness, privacy and responsibility [4–6]. What is less clear and needs investigation are other variables for these proposals like scope, applicable context, ownership of or responsibility for the process, method of implementation and representation of stakeholders. There are already established governance methodologies for assessing and mitigating the impact of new technologies, processes, and infrastructure across the domains of environment [24], information privacy [25], data protection [26] and human rights [27]. Impact assessment and audit methodologies take core values and combine them with a process for the public, outside experts, and policymakers to consider complex social and technical questions.

This work provides an assessment of the myriad of frameworks, principles, templates, guidelines and protocols that have arisen around AI through the lens of known best practices for impact assessment and audit of technology. We as a community have taken the first steps in identifying that there is a problem to be addressed and started to identify how to apply this by proposing tools to manage ethical challenges and risks. However, maturity in these thoughts is still to be achieved. Looking at the environmental movement of the mid-twentieth century, in which ethical considerations for many diverse parties, application of technology and societal concerns all converged, there are parallels for best practice in the current AI ethics, impact assessment and audit conversations. There are also robust, long-established audit and assurance practices in other sectors like financial services. In particular, we look back to the impact assessment and audit tools, processes and procedures to identify the gaps in our current approaches. We then lay out the methodology that we will use to identify the holes in current mechanisms by content analysis of a range of pertinent aspects using typological schema. These have been developed by review of previous best practice and current discussions around AI governance. Next, we present the current mechanisms being used to encourage these ethical practices in AI technology. We then provide an overview of the development of impact assessment and audit, and the key components as they related to understanding impact across participants, technology and processes. Using the typologies created from the review of previous practice, we analyse current AI frameworks according to these criteria to identify the gaps in current approaches. This paper contributes to the literature by mapping the current landscape of suggested tools

for ethical assessment of AI systems, placing these tools in a historical tradition of managing the impacts of technology, thereby exposing possible areas for strengthening these tools in practice.

2 Background to impact assessment and audit practices

2.1 Impact assessment

Ethical tools and frameworks for AI do not spring like Dionysus fully formed from Zeus' thigh, they are part of a development of governance tools to tackle health, environmental and privacy impacts of technology that began in the 1960's. Impact and risk assessment is 'a type of fact-finding and evaluation that precedes or accompanies research, or the production of artefacts and systems, according to specified criteria. Assessing the impact of some X upon some Y has been practiced for generations, and has engendered debates over methods, purpose, focus, policy relevance, terminology, and efficacy' [13, pp. 6–7]. These assessments are shaped by notions of relevance (what is important to society and which phenomena are worthy of attention), evidence (identification of causes and effects), and normative claims (what is good, acceptable or tolerable) [28, p. 4].

2.2 Technology assessment

Technology assessment (TA) is a practice that began with the US Office of Technology Assessment (OTA) 1972–1995 [29, 30]. TA was 'foremost an attempt to gain political control over the potential negative effects of technological development by means of early warnings. TA was supposed to predict unintended negative consequences of technical innovations to facilitate more adequate policy-making' [31, p. 544]. In the 1990s, Europe also developed its own TA institutions like the Scientific Technological Options Assessment (STOA) and recent activities include setting up the STOA Centre for AI [32]. Several different varieties of TA have been developed, for example in the Netherlands and Denmark, TA was extended to address issues of participation. Instead of the traditional TA model with panels of experts producing reports for policy-makers, participatory TA (pTA) includes contributions from a much wider group of stakeholders like lay people, journalists, trade unions and civil society groups [33]. pTA uses various forms of public deliberation including focus groups, citizens' assemblies and consensus conferences to gather data for reporting [34]. The lack of an ethical dimension to TA has also led to suggestions for an ethical TA (eTA) [31, 35], which mirror many of the concerns found in AI ethics frameworks [5].

2.3 Environmental impact assessment

Environmental Impact Statements (EIS) were pioneered in the US by the 1969/70 National Environmental Policy Act (NEPA), leading to many other jurisdictions enacting environmental legislation Environmental Impact Assessments (EIA) broadened the process to include the identification of future consequences and included in the process public consultation and review mechanisms[25]. '[T]he role of the stakeholders or parties at interest plays such a critical role in technology assessment, and involvement of citizens in environmental impact statements is mandated by law' [29, p. 374]. These assessments are part of many jurisdictions planning and/or environmental legislation, intended to allow stakeholders, including the public in its widest definition, to contribute to decision-making on infrastructure development like dams and roads [36, 37]. It should be noted though that there is a lack of clear definition in EIA literature and practice as to what 'participation' actually means [38]. There are also specific assessment techniques for products and materials to assess environmental impact which map life cycles [39].

Environmental Risk Assessment (ERA) developed as a separate practice from the broader scope of EIAs, using formal quantitative analysis of probabilities for undesirable outcomes of a process or substance [36, 40]. Fiscal Impact Analysis (FIA) is an economic impact tool commonly used in land use planning decisions [41], and EIAs often includes forms of cost–benefit analysis (CBA) [42].

2.4 Social and human rights impact assessment

EIAs and ERAs were criticized for focusing on only biophysical and economic impacts and not including the social and cultural impacts of proposed developments or technologies, leading to the development in the 1990s of Social Impact Assessments (SIA). SIA is not a widely applied form of assessment 'largely because of the challenge of defining, predicting and measuring social change and impact, in addition to legal and regulatory frameworks that are persistently weak or ineffectual in terms of social impact' [43, p. 91]. They still remain fairly uncommon, but have been used in policy impact assessments, for example, by the IMF to try and understand the impact of macro-economic policy changes [44].

2.5 Privacy and data protection impact assessment

The concept of privacy, which underpins modern data protection legislation, is essentially normative and represents the cultural and historical values of societies. In the Western tradition, there are two core assumptions, the first appealing to a 'natural' divide of the public (the state and politics,

work and business) and the private realm (the realm of the home, family, body and personal property, where the individual is considered the best judge of their privacy interests. The second assumption posits privacy as a prerequisite for the liberal democratic state. There are shifting social norms around the value and definition of privacy, with debates revealing tensions, for example, between the goals of privacy vs security and privacy vs economic growth [45].

The 'fair information practices' (FIP) movement emerged in the US in the work of Westin [46, 47], in response to growing societal concerns over the collection and processing of personal data in both the public and private sector. It was not until the mid-1990s that Privacy Impact Assessments (drawing on the model of EIAs) emerged in various forms across different jurisdictions [48]. By 2007, the UK Information Commissioners Office published a handbook describing a methodology for conducting a PIA, which was further developed in Europe into a Data Protection Impact Assessment (DPIA), a key tool in the latest iteration of data protection regulation, the GDPR [3].

Privacy impact assessments developed to meet the need for public trust in information processing by identifying and managing risks. This is part of a wider move in industrialised societies to manage potential risks of new technologies, processes or products that can also be seen in TA and EIA [13]. DPIA's use checklists and risk assessments to document the data processing and any necessary mitigations if risks are identified in an iterative review process [26].

2.6 Audit

There are long-established techniques for auditing processes and systems, for example in the financial sector where there exist globally agreed standards like Generally Accepted Accounting Principles (GAAP) and International Financial Reporting Standards (IFRS) [49]. These rules lay down the process for transparent 3rd-party auditing which have been adopted into law by the majority of jurisdictions around the world. There are also audit and assurance standards in safety critical engineering for industries like aviation, nuclear power, or more recently, autonomous vehicles [50, 51].

Audit techniques are also used for third-party verification for accreditation to international industry standards e.g. International Organization for Standardization (ISO) [52]. An audit consists of the examination of evidence of a process or activity, like financial transactions or an engineering process, and then evaluation of the evidence against some standards or metrics, which could be a regulation or standards regime [53], or internal management metrics [49, 54], as illustrated in (Fig. 1).

To conduct an audit, there first needs to be a set of auditable artifacts that record decisions, systems and processes. Brundage et al. define as this as problem space for current AI production in that they 'lack traceable logs of steps taken in problem-definition, design, development, and operation, leading to a lack of accountability for subsequent claims about those systems' properties and impacts' [55]. This is where some of the technical tools addressing AI ethics (see "Discussion") can become part of an audit process by providing evidence for evaluation by auditors. Audit also requires non-technical governance processes [15] to ensure consistency with relevant principles or norms [56].

Impact assessments like EIA, and audits such as those conducted in the finance sector have well-established protocols regulated by legal requirements. Independence of assessment and audit is used to ensure transparency and places liabilities on both the parties assessing and the assessed parties. 'Whether the auditor is a government body, a third-party contractor, or a specially designated function within larger organisations, the point is to ensure that the auditing runs independently of the day-to-day management of the auditee' [56, p. 2]. External assessment provides publicly available documents which can also serve a broader range of stakeholders beyond the entity or process in question to include users, customers and wider society.

2.7 Risk assessment and techniques

While a myriad of processes, tools and applications of these tools at various parts of the production cycle exist across the historical impact assessment and audit activities above, one of the key elements is risk assessment.

Modern conceptions of risk (risk = accident × probability) became a fully-fledged part of modern societies with the risk assessment practices developed in response to concerns over the impact on the environment and human health from human activity in the form of development, technologies and industrial processes and materials. In 1969, in an article entitled "What is our society willing to pay for safety?" [57] articulated a systematic and quantitative approach to risk, and introduced the concept of trade-offs between risks and benefits [58]. Debates within the environmental movement, and the associated legal and organisational structures that grew out of this period, came to be famously characterised by Beck in the 1980s as the 'Risk Society' [59]. Beck posited that the project of modernity had become not how to distribute wealth or goods, but how to distribute the risks, or 'bads', of modern industrial society, where technical experts are given pole position to define agendas and impose bounding premises a priori on risk discourses' [59, p. 4].

A risk-based approach has developed throughout the latter part of the 20th and into the twenty-first century, taking the methodology and approaches from environmental management and risk assessment and applying them to areas like occupational health and safety, business risk (financial, operational, reputational), quality and information security.

Risk assessment techniques vary from quantitative to qualitative approaches [40] depending on the sector and application. Risk assessments often rely on scoring or 'traffic light' systems for ranking risks [60], and highlighting those areas that need treatment, either in the form of mitigation (changing the risk score) or in taking measures (like insurance) or documenting decisions to 'trade off' the risk against the potential benefits. Risk assessments are also used for achieving compliance with the existing regulatory frameworks. The latest European iteration of data protection (GDPR) also takes a risk-based approach to privacy protections for data subjects. Many of the ethical frameworks proposed for AI build on these models and approaches to risk assessment.

For the business sector, managing reputational risk is an important consideration and providing evidence of responsible behaviour has direct links to both users/customers and also to investors and boards. Many investors use Environmental Social and Governance (ESG) assessments, where they look for evidence of compliance with international standards and norms, where the risks (especially reputational) could impact across all three areas of ESG assessments for investors. Business-focused AI ethics tools fall into the suite of tools organisations deploy to protect the core value. Managing risk allows institutions to 'adopt procedures and self- presentation to secure or repair credibility' [59, p. 4], a core purpose of contemporary risk management strategies [61].

Risks in AI can manifest as either underusing the technology and missing out on value creation and innovation, or overusing/misusing the technology [62]. Floridi et al. [63] draw attention to risk that results from not using the technology, and how these risks need careful trade-offs to ensure the greatest benefit. As Jobin et al. [5] note in their systematic review of global AI guidelines, conflicts can be identified in the different proposals 'between avoiding harm at all costs and the perspective of accepting some degree of harm as long as risks and benefits are weighed against each other. Moreover, risk–benefit evaluations are likely to lead to different results depending on whose well-being will be optimized for and by which actors. Such divergences and tensions illustrate a gap at the cross-section of principle formulation and their implementation into practice' [5, p. 396].

2.8 Stakeholder theory and participation

The influential European Commission's report on 'Trustworthy AI' proposes that 'management attention at the highest level is essential to achieve change. It also demonstrates that involving all stakeholders in a company, organisation or institution fosters the acceptance and the relevance of the introduction of any new process (whether or not technological). Therefore, we recommend implementing a process that embraces both the involvement of operational level as well as top management level' [64, p. 25]. A wide-ranging network of stakeholders can be plotted in the production and deployment of new technologies that extend far beyond the domain of engineers and developers (see Table 3).

Since the development in the 1980s of corporate stakeholder theory [65], it has become common parlance to refer to 'stakeholders' across a range of organizational domains. Stakeholder theory provides a well-established framework that allows us to:

1. Identify and describe all interested and affected parties in the deployment of a technology
2. Acknowledge stakeholders have legitimate interests in technology
3. Affirm that all stakeholders have intrinsic value, even if their concerns do not align with the concerns of the technology producers
4. Identify the responsibilities of parties with relation to a given process [66].

Table 3 (below) identifies the broad categories of public and private sector stakeholders who either have direct roles in the production and deployment of AI technologies, or who have legitimate interests in the usage and impact of such technologies. Stakeholder theory has long challenged the assumption that a company's exclusive obligation is to their shareholders or investors, with business leaders increasingly recognizing the need for a wider set of obligations beyond the narrow vision of 'shareholder primacy' [67].

2.9 Technical and design tools

Another active space in the AI ethics debate is within the AI/ML community itself where much attention and research has been focused on metrics like fairness, accountability, explainability and transparency.[1] A range of computational approaches have been suggested, offering quantitative metrics for fairness, methods to 'debias' training data sets, test models against protected characteristics and provide explanations of 'black box' algorithms, packaged up into AI fairness toolkits [20, 68–70]. These toolkits have been criticised for offering a 'reductionist understanding of fairness as mathematical conditions' [71, p. 1], and reflect a longer history of attempts to reduce (un)fairness to a metric [72]. Studies with ML developers highlight that considerations of a model's context, and the specificity of the domain in which it is used, are vital to improve features like fairness

[1] E.g. new conferences have been created like ACM FAccT https://facctconference.org/index.html and high profile conferences in the AI/ML space increasingly including work on ethical problems like NeurIPS https://neurips.cc/.

[73]. Many would argue that in fact, developing ethical AI requires not only technical 'fixes' but the deployment of social science disciplines is vital to address negative outcomes [73–75].

Other suggestions focus on design processes, for example, awareness raising for design teams in workshop style events [76, 77], or participatory design processes [78]. The human–computer interaction (HCI) community is also concerned to translate previous work in, for example, Value Centred Design, to address the issues in human–AI interactions [79].

3 Methodology

This study draws from the rich impact assessment and audit literature from other domains to develop a typology for comparative document analysis of proposed AI ethics tools. To understand how proposed AI ethics tools might be applied, it is first necessary to understand what they are offering, how they differ, and to identify any gaps. This understanding can be used to refine and develop these tools for future use. The AI ethics documents themselves provide the data for study, which have been analysed using qualitative content analysis, 'a research technique for making replicable and valid inferences from data to their contexts' [80, p. 403]. Typologies of salient features were developed in response to research questions (see Table 9), using a review of related literature and AI ethics documents, and iteratively refined. Typologies are useful heuristics to enable systematic comparisons [81], and extensive related literature was reviewed to build representative typologies for the tool types under examination which would yield useful comparisons across a diverse range of documents.

The research process is set out in Fig. 2. The process began with a systematic collection of AI ethics documents using the document types and keywords detailed in Table 2. We used a combination of web searches, citation scanning and monitoring of relevant social media and news items to identify suitable candidates between May 2019 and December 2020. Other collections of AI ethics documents were also used both as sources of relevant documents, and for validation [4, 5, 18, 82]. The initial search yielded $n = 169$ documents. Many of these documents are drafted by public, private or not-for-profit organisations and constitute 'grey literature' not typically found in academic databases [71]. Academic sources were also included, particularly as the private sector is active in producing and publishing academic papers on this topic [83].

The lead researcher on this project has a background in environmental management techniques, and has previously been trained to conduct ISO:14000 audits [84]. Reflecting on the processes used to provide assurance for the environmental impact of organisations, we considered there to be parallels in the need to implement processes to assure the ethical design and deployment of AI systems. This background knowledge informed the decision to reflect on impact assessment and audit processes in other domains, many of which have a long lineage. We wanted to understand better the features of the current proposed tools for implementing AI ethics, to assess if these tools are fit for purpose, and where gaps might be illuminated by previous practice.

Table 1 Key terms and background literature

Key terms from initial content analysis of ethics frameworks	Background literature review to build content analysis
Impact assessment	Technology Assessment
Audit	Environmental Impact Assessment
Technical tool	Social and Human Rights Impact
Design tool	Assessment
Application stage	Privacy and Data Protection Impact
Stakeholder	Assessment
Risk assessment	Risk Assessment
Procurement	Audit
Type of author	Technical and Design Tools
	Stakeholder theory

This initial data set was analysed using a qualitative content analysis methodology [85] to elicit frequently applied terminology and approaches. The documents were stored in Zotero reference manager and coded in an MS Excel spreadsheet iteratively to identify recurrent key words and concepts that were used to describe their main purpose, type of document, author and audience. The key terms derived from this process are shown in Table 1.

From this, we devised a set of sub-questions with which to query the data which shaped the categories and codes we developed (see Table 9). These questions were considered the salient features that would allow us to understand and compare the AI ethics tools. Key terms were then used to search for literature that mirrored these terms across different domains as show in Table 1. The deep background literature review of previous practices was used to identify categories which became the codebook (see Table 9). This was a reflective process where we identified principles and categories across domains and used them create typological sets as follows:

The next step was to narrow down the initial large data set of $n = 169$ documents, which contained many documents that were statements of principles or discussions of AI ethics. We were only interested in those documents that would give an organisation or practitioner a concrete tool to apply to AI production or deployment. See [86] for a discussion of why principles are not enough on their own, and how we need to bridge to gap between principles and practice. We

Table 2 Criteria for sample identification

Criteria	Inclusion	Exclusion
Document type	Codes, principles, checklists, risk assessments, reports, white papers, academic research, technical tools, documentation, impact assessments, audits, guidelines, standards, registers, contracts, policy documents, recommendations, webpages, institutional reports, declarations, professional ethics	Opinion articles, speeches, audio/visual materials, images, legislation
Keywords	AI, artificial intelligence data—ethics, stewardship, big data machine learning, deep learning algorithms predictive analytics automated decision-making advanced analytics automated scoring, profiling, aggregating, sorting data science digital technology	Traditional data protection, privacy
Type of content	Practical proposals for implementing ethics for AI, including both model and data	Ethical principles and frameworks without proposals for how to apply these principles
Author	Public, private and not-for-profit sector (including NGO's), academic research, standards bodies	Authors not representing an organization, or not peer-reviewed publication
Language	English	
Availability	Public, online	
Data collection time period	May 2019 to December 2020	
Document publication date	2016–2020	Pre-2016 and post 2020

excluded all that did not contain practical tools to apply ethical principles (see Table 2), leaving a data set $n = 39$ documents that offered practical tools to operationalise ethical principles in the production and deployment of AI systems.

3.1 Typology of stakeholder types

After review of stakeholder theory [65, 66, 87–89], a categorisation of key stakeholder groups relevant across both public and private sector was developed, adapting a typology from [64, 90–92]. Table 3 presents a typology of stakeholders that has been adapted and extended from the identification of possible stakeholders described in [64, p. 25], where it is interesting to note the table did not include users or customers, or shareholders. We have therefore extended the categories to mirror the roles in the public sector, and also widened the stakeholders beyond the confines of the production or deployment of the technologies to include all stakeholders who are affected or have in interest in the process.

Table adapted from [64, p. 25], [90–92].

3.2 Typology of tool types for impact assessment

Table 4 shows the key features of impact assessments derived from our literature review.

3.3 Typology of tool types for audits

Table 5 shows key processes mapped from the review of audit techniques.

3.4 Internal vs external process

Table 6 shows the codes we created to identify if the tool was designed for internal organisational use, or provided for third-party inspection.

3.5 Technical and design tools

A sub-set of tools being suggested for operationalising ethical AI comprise design and engineering tools for use in specific stages of the production pipeline (see Table 7 Typology of technical and design tools.) These are either materials for use in design teams in workshop style events [76, 77], tools for producing documentation of the design, build and test process [21, 68], or technical tools for testing models, protecting privacy and security, testing for bias [69, 93, 94], or tracking provenance of data [95].

Table 3 Typology of stakeholders

Stakeholder	Public sector	Private sector	
Voiceless	Environment Marginalised or excluded groups	Environment Marginalised or excluded groups	Impacts on physical environment, ecosystems and its members, energy and raw material extraction and use. Workers in extractive or digital industries (e.g. mining, content moderation, data annotation). Traditionally marginalised groups with limited voice in society (e.g. the poor, minority ethnic groups, refugees and immigrants, disabled, incarcerated, women, children)
Vested interest	Citizen	Shareholders Investors	The electorate have a right to transparent processes, and should have the ability to contribute to decision-making (participation). Shareholders and investors also have fiduciary duty to consider the ethical behaviour of their investment vehicles
Decision-makers	Elected Official Chief Executive Director	Senior Management (C-suite) Board	Senior management discusses and evaluates the AI systems' development, deployment or procurement and serves as an escalation board for evaluating all AI innovations and uses, when critical concerns are detected. It involves those impacted by the possible introduction of AI systems and their representatives throughout the process via information, consultation and participation procedures
Legal	Compliance/Privacy Legal Department Policy	Compliance/Privacy Legal Department Corporate Responsibility Department	The responsibility department monitors the use of an ethical assessment and its necessary evolution to meet the technological or regulatory changes. It updates the standards or internal policies on AI systems and ensures that the use of such systems complies with the current legal, regulatory and policy frameworks and to the values of the organisation
Delivery	Delivery Managers Service Managers Domain Experts	Product Managers Service Development or equivalent	The Product and Service Development department uses an ethical assessment to evaluate AI-based products and services and logs all the results. These results are discussed at management level, which ultimately approves the new or revised AI-based applications
Quality Assurance	Policy Service delivery staff Quality assurance	Quality Assurance	The Quality Assurance department (or equivalent) ensures and checks the results of an ethical assessment and takes action to escalate an issue higher up if the result is not satisfactory or if unforeseen results are detected
HR	HR	HR	The HR department ensures the right mix of competences and diversity of profiles for developers of AI systems. It ensures that the appropriate level of training is delivered inside the organisation
Procurement	Procurement	Procurement	The procurement department ensures that the process to procure AI-based products or services includes an assessment of ethics
Developer	Data Scientists/Engineers Developers Project Managers	Developers Project managers	Developers and project managers include an ethical assessment in their daily work and document the results and outcomes of the assessment

Table 3 (continued)

Stakeholder	Public sector	Private sector	
Users	Service users	Users Customers	Participation of users in development, and/or publication of assessments for public interrogation. (NB: this layer is missing from the EU categories)
Oversight	Independent Oversight Bodies Expert Committees Freedom of Information Requests Regulators Courts	Independent Review/Oversight Bodies Expert Committees Regulators Courts	Public Sector governance has a variety of structures aimed at accountability and transparency and compliance with the law

Table 4 Typology of impact assessment methods

Impact assessment	
Checklist; questionnaire	Widely deployed tool across impact assessments and audits to describe activity and interrogate aspects of project or process. Can be used for both potential projects and to documentation for audit
Baseline study	Commonly used in EIA and policy assessments to ascertain baseline conditions against which proposed projects or policy can be measured
Participation process	Mandated part of EIA process, public stages of EIA involve scoping and review, and publicly available documentation
Cost–benefit analysis	Assessment tool to compare economic costs with potential benefits
Risk assessment	Can be qualitative or quantitative, frequently translated to a scoring or traffic light output
Life-cycle assessment	Assessment technique for products or materials to calculate environmental or health impacts
Change measurement	Commonly used in policy or human rights impact assessment to determine impacts
Expert committee	Used in assessment process to provide expert evidence or domain knowledge
Governance process	Business and administrative processes to document activity and provide verifiable documentation
Procurement process	Structured process to assess the impact of a purchasing decision

Table 5 Typology of audit methods

Audit	
Checklist; questionnaire	Widely deployed tool across impact assessments and audits to describe activity and interrogate aspects of project or process. Can be used for both potential projects and to documentation for audit
Documentation	Audits require artifacts for inspection and assessment such records of processes, materials, outcomes and decisions
Reporting	Output from audits is commonly in the form of auditors' reports
Governance process	Business and administrative processes to document activity and provide verifiable documentation

Table 6 Typology of internal vs external process

Internal vs external assessment/audit	
Internal/self-assessment	Designed to be used only as internal organisational tool. Outcomes assessed only by internal parties. No process for wider transparency or participation
External/3rd party	Designed to be used by external auditors, standards body. May include provision for publication of results/outcomes for wider transparency

3.6 Production and deployment process for AI Systems

AI systems go through stages of production, from initial definition of a use case, development of a business case, through the design, build, test and deploy process [96]. Assessment and audit tools can be applied at different stages of the process (or attempt to capture cover the whole pipeline), and can be focused on the data flowing through the pipeline, or the attributes of the model, or both. Table 8 defines codes for these stages. The pipeline for deployment of systems often includes selling the AI system to a customer, who will deploy the system, at which point ethical considerations can be included in the procurement process.

Table 7 Typology of technical and design tools

Technical and design tools	
Workshop materials	Materials produced for use by design teams as workshop or discursive events e.g. scenarios, design cards, agile design events
Documentation	Technical documentation like logs and incident reports, technical descriptions
Technical tools	Specific technical applications for addressing issues like privacy, security, bias, transparency, provenance in models and data

Table 8 Typology of when tool used and if applied to data and/or model

Stage in process tool used/applied to data and/or model	
Business/use case	A problem space, or area for improvement is identified, and the use case and business case are developed
Design	Business case is translated into design requirements for engineers
Training data collection	Training and test data are identified, collated, cleaned and prepared for training the model
Building	AI application is built
Testing	The system is tested
Deployment	The system goes live
Monitoring	System performance is monitored as it performs in the wild
Procurement of system	Third party buys system for their own use
Data	Depending on the focus of the tool, either the data pipeline is the main object of assessment, or the model itself
Model	

3.7 Document analysis

A total of $n = 169$ items were identified under the broad category of AI-related ethics frameworks, which after application of the exclusion criteria resulted in a final list of $n = 39$ ethics tools see Appendix 1. The documents were analysed using qualitative content analysis [97], through the development of a codebook of variables to identify key features (see Table 9 below). This was an iterative process where the codes were refined during the process of reading and coding the material.

The terms impact assessment and audit are used in differing ways in the domain of AI ethics tools made coding of these documents complex. As Carrier and Brown [98] note, there is much ambiguity over the use of the term 'audit' in relation to AI ethical assessment being used by what they term as the 'AI ethics industry'. Across the landscape of AI ethics audit and impact assessment tools, terms are often used loosely, or are used interchangeably. In a recent Ada Lovelace Institute report, 'Examining the black box', algorithmic *audit* is into divided into two types, a narrow 'bias audit' or a broader 'regulatory inspection' which addresses 'compliance with regulation or norms, necessitating a number of different tools and methods, typically performed by regulators or auditing professionals' [99, p. 3]. Algorithmic *impact assessment* is divided into an ex ante risk assessment, and what the report terms an 'algorithmic impact evaluation' which assesses the effects of an application after use [99, p. 4]. The codes reflect a decision by the researchers to define 'impact assessment' as an ex ante process which was predicting possible impacts, with audit being an ex-post-process for examining ongoing activities. This is not necessarily reflected in the language of the documents themselves, depending on the author and field or discipline from which they originated.

The coding process was consisted of reading and re-reading the documents and coding them against the typologies to create the results (see Table 10). The research methodology used is a reflexive and adaptive [85], creating a robust process for relating the document data to their context as shown in the diagram in Fig. 3. Despite this, the limited size of the team analysing and coding the documents presents a limitation in that often validity of qualitative analysis is considered to be justified by the process of recurrent iterations with different coders [100]. Despite this limitation, we believe we have made every effort, from the conception and planning of the project, through to development of typologies and coding of results, to consider where bias and omission could occur in the process [101]. We believe we have developed categories for assessing AI ethics tools that reliably surface salient features which can be used to compare across disparate types of tool or procedure.

We also have not set out to provide an exhaustive review of the computational techniques in the AI/ML research to address ethical issues like fairness and explainability, for this, see for example [102, 103].

Table 9 Sub-questions and derived codes

Question posed to documents	Codes
Which sector were the authors/users from?	Public sector Private sector Not-for-profit Academic research
Which stakeholder would either use the tool, or engage with the results? [See Table 3 for detailed category breakdown]	Voiceless Vested interest Decision-makers Legal Delivery Quality assurance Procurement HR Developer Users Oversight
What type of tool was it? Which strategies did it employ?	Impact Assessment Checklist questionnaire Baseline study Participation process Cost–benefit analysis Risk assessment Life-cycle assessment Change measurement Expert committee Business process Procurement process Audit Checklist questionnaire Documentation Reporting Business process Technical Tools Workshop materials Documentation Technical tests
Were these tools for use internally, or have external elements?	Internal/self-assessment External/3rd party
Which stage in AI production and use was the tool used?	Business/use case Design Training data collection Building Testing Deployment Monitoring Procurement of system
Was the tool appropriate for addressing the model, data, or both?	Model Data

4 Results

The data set of 39 proposed tools for ethical AI was coded using the typologies developed from the literature review of sectors, stakeholders, historical practice, and stages in AI production process, and shown in Table 10. The documents are arranged in ascending year of publication, with the majority of documents being produced in in 2019/2020, 2020 comprising half the total. Some judgement was required in coding these documents as to whether they were an impact assessment or audit, as the terms are used with varying meanings across the AI ethics documents.

Key findings:

- The focus has moved from data to models from 2017 to 2020. Earlier documents were often concerned with issues around 'big data', with concerns shifting to models and algorithms. This does not mean that data are not considered in these later iterations (particularly training

Table 10 Overall results for coded document set ($n = 39$) (see Appendix 1 for document details)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39
Title	Risks, Harms and Benefits Assessment	AI and Big Data A blueprint for a human rights, social and ethical impact assessment	ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY	An Ethical Toolkit for Engineering/Design Practice	Ethical Data and Information Management: Concepts, Tools and Methods	Ethical OS	Ethics & Algorithms Toolkit (beta)	AI Fairness 360	AI Procurement in a Box	AI-RFX Procurement Framework	Algorithmic Impact Assessment (AIA)	Codex for Data- Based Value Creation	Consequence Scanning – doteveryone	IBM Watson OpenScale	IEEE SA - The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical	Model Cards for Model Reporting	Model Ethical Data Impact Assessment	ODI Data Ethics Canvas	Understanding artificial intelligence ethics and safety: A guide for the public sector	A Proposed Model AI Governance Framework - Second Edition	AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems	Algorithm Register	Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment	Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic	Co-Designing Checklists to Understand Organizational Challenges and Opportunities	Corporate Digital Responsibility	Data Ethics Framework	Datasheets for Datasets	Empowering AI Leadership	Fairlearn: A toolkit for assessing and improving fairness in AI	IEEE Draft Model Process for Addressing Ethical Concerns During System Design P7000/D3	IEEE 7010 Recommended Practice for Assessing the Impact of Autonomous and Intelligent	Responsible AI	Standard Clauses for Municipalities for Fair Use of Algorithmic Systems	Toward situated interventions for algorithmic equity: lessons from the field	Value-based Engineering for Ethics by Design	Welcome to the Artificial Intelligence Incident Database	White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions
Year	2017	2018	2018	2018	2018	2018	2018	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2019	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020	2020
Sector produced by																																							
Public Sector							•				•										•		•	•		•		•					•		•				
Private Sector					•			•						•	•		•							•	•	•		•		•	•	•	•	•			•		
Not-for-Profit	•					•	•		•	•		•	•					•	•											•								•	•
Academic Research		•	•	•			•			•						•	•			•		•			•	•	•		•					•				
Sector used by																																							
Public Sector	•	•	•				•		•		•		•		•			•	•	•	•		•	•	•	•		•	•		•	•	•		•		•	•
Private Sector	•	•		•	•	•		•		•	•	•	•	•	•	•	•	•	•			•		•	•	•	•		•	•	•	•	•	•	•	•	•	
Not-for-Profit	•																									•											•	
Academic Research				•													•								•	•					•			•			•	
Stakeholder Type: ● Applying Tool ○ User of tool output																																							
Voiceless																							○			•	○												
Vested Interest							○								○					○	○		○		○	•	○						○		○			○	
Decision Makers	○	○	○	○	○		○		•	○	○	○			○		○	○	○	○	○			○	○	○	•	○		•	○	○	○		○	○	○	○	○
Legal	•	•	•				○				•							•					•		•		•		○					•				
Delivery			•	•	•	•	•	•	•	•	•	•	•		•		•		•	•	•	•		•	•	•		•		○	•	•	•	•		•	•	
Quality Assurance			•	•	•	•		•		•					•	•		•		•	•			•	•	•		•		○	•	•	•	•		•	•	
Procurement							•		•	•	•								•	•	•			•				•		○					•				•
HR																																							
Developer			•	•	•	•	•	•		•	•	•	•	•	•	•	•	•			•	•		•	•	•		•	•	○	•	•	•	•		•	•	
Users																			•	•																			
Oversight							○		○						○		○	○	○		○		○	○	○	○	○	○	○		○	○	○		○	○		○	○
Impact Assessment	•	•	•	•		•	•				•	•						•	•	•	•			•		•		•							•			
Checklist questionnaire	•	•	•	•		•	•				•	•						•	•	•	•			•		•		•							•			
Baseline study																																•							
Participation process			•				•																			•											•	•		
Cost-benefit analysis	•																																						
Risk assessment	•				•		•				•			•				•			•			•				•									•		
Life-cycle assessment																																							
Change measurement																																							
Expert committee		•																																					
Governance process	•			•								•						•		•	•															•			
Procurement process							•		•	•										•																			•
Audit					•				•	•				•	•		•						•		•		•		•		•	•	•	•	•		•	•
Checklist questionnaire									•	•																				•									•
Documentation					•									•	•		•			•			•		•		•		•		•	•	•				•	
Reporting					•									•	•		•								•							•	•				•	
Governance process					•									•	•										•					•		•	•					
Technical & Design Tools						•							•			•					•							•		•					•			
Workshop materials						•							•			•					•				•											•			
Documentation														•															•					•			•		
Technical tools								•						•																	•			•					
Internal or external assessment/audit																																							
Internal/ self-assessment	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•			•	•	•	•		•
External/ 3rd party															•								•									•	•					•	
Stage in process tool applied to: ● data and/or ◊ model																																							
Business/use case	•	◊	◊						◊	◊	• ◊	•	• ◊					•	•	• ◊		◊		◊	◊	◊	• ◊	•		◊		◊				◊	◊		◊
Design	•	◊		◊		• ◊	◊				• ◊	•	• ◊			◊	◊		•	• ◊		◊		◊	◊	◊	• ◊	•		◊		◊	◊	◊		◊	◊		
Training data assembly					•			• ◊				•				◊	◊		•	• ◊									•				◊				◊	
Building							◊	• ◊				• ◊														◊					◊								
Testing								• ◊				• ◊					•			◊						•			•		◊		◊	◊				
Deployment											• ◊	•		◊	◊		•	•		◊	◊					◊	• ◊				◊		◊			◊		
Monitoring												• ◊		◊	◊				•	◊																◊		
Procurement	•		◊				◊		◊	◊		• ◊						•			◊			◊	◊					◊				◊	◊				◊

Fig. 1 Audit process

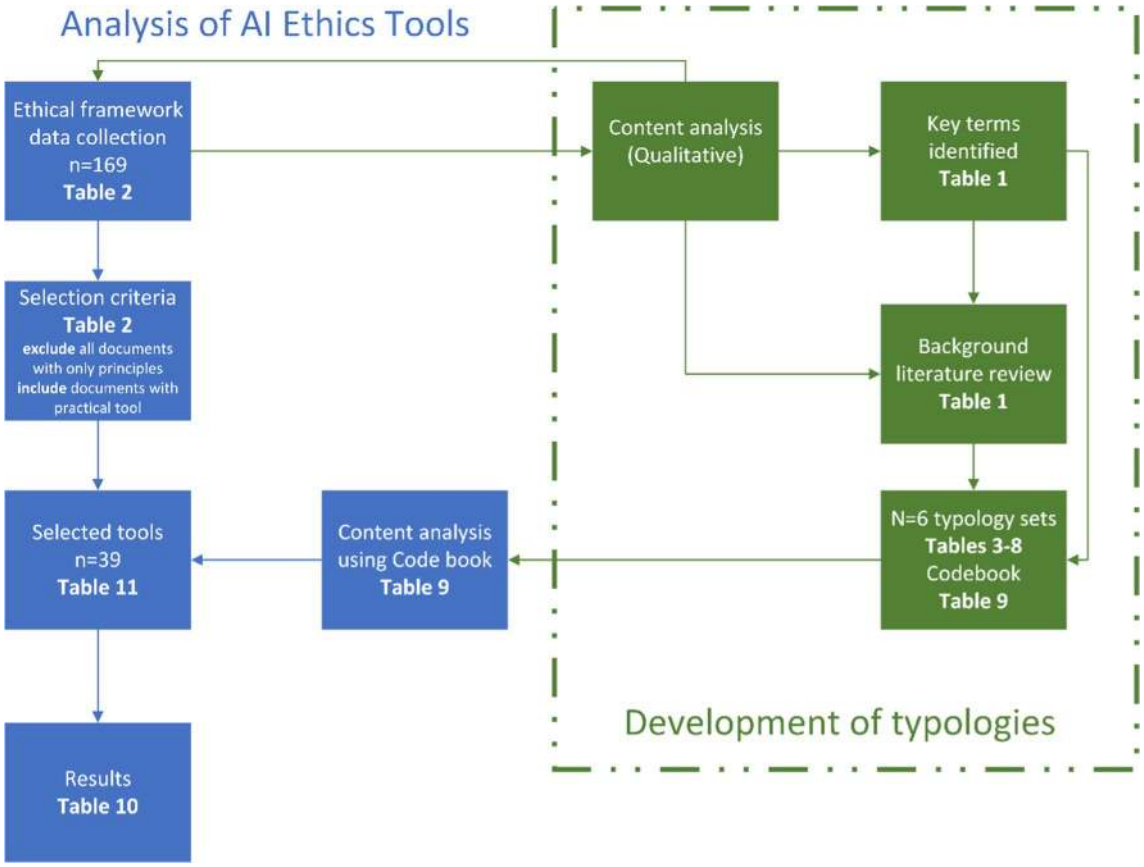

Fig. 2 Flow diagram of methodology

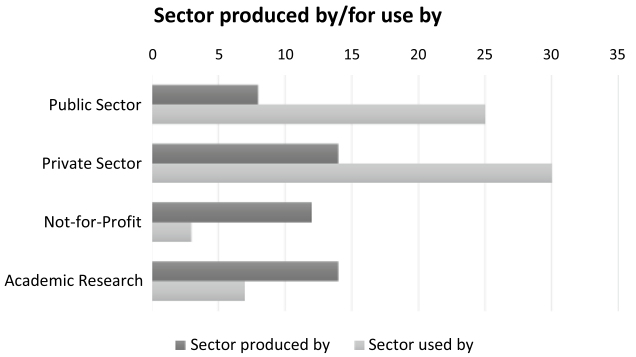

Fig. 3 Sector produced by/for use by

and test data), but the focus shifts from a more traditional data protection approach.

- Stakeholder types directly using the tools are clustered around the product development phase of AI (developers, delivery, quality assurance), with the output from the tools (reporting) being used by management Decision-Makers.
- There is little participation in the assessment or audit process by certain stakeholder groups (Voiceless, Vested Interests and Users) who are not included in the process of applying the tools or interacting with the outputs as tools for transparency or decision-making. Perhaps most surprising is how little inclusion there is of Users/Customers in these tools.
- Nearly all of the tools are for Internal Self-assessment, with only the IEEE standards requiring any kind of external verification, and the two examples of public registers providing explicit transparency.
- Techniques and practices deployed by other forms of Impact Assessment (like EIAs) are not present or rarely suggested in ethical AI impact assessments (Participation process, Baseline study, Life-cycle assessment, Change measurement or Expert committees.)
- Checklists/questionnaires are ubiquitous across Impact Assessment tools. Audit tools less frequently use Checklists but do require Documentation of processes.
- The output from the tools can provide documentation for Oversight from external actors, but as the majority are internal activities, there is generally no process or requirement for the wider publication of the results of these tools.

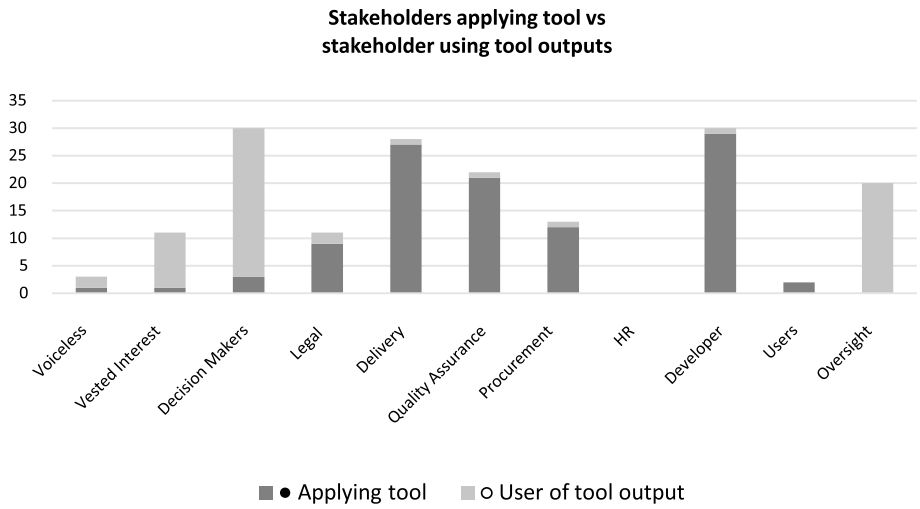

Fig. 4 Stakeholder applying tool vs stakeholder using product from tool

- A third of the Impact Assessment tools focus on Procurement processes for AI systems from 3rd-party vendors, indicating the need for not only producers of AI products to engage with ethical assessment, but also the customers for these products, who will be the ones deploying the products.

Figure 3 illustrates the main sectors who are either producing ethical AI tools, and compares this to those sectors for whom the tool is intended for use. It shows the main sectors targeted by ethics tools are the public and private sector, which reflects the main sectors where AI systems are being designed and produced (private sector), and the concerns around deployment of AI in public sector institutions. There is also interest from the academic community in AI ethics tools and how to address these issues, with the not-for-profit sector (civil society, NGO's and think tanks) also looking to provide solutions to ethical issues in AI production and deployment, although not-for-profit are not producers of AI systems, some sectors of not-for-profit (like development agencies) do deploy these systems. It is interesting to note that it can be difficult to separate academic research from private, corporate research in AI as there is strong cross-fertilization between these, with scientists moving between sectors, and technology companies funding their own research outputs, and funding university research.

Figure 4 shows the number of tools that include which type of stakeholder in their terms of reference either as producers of artifacts, or consumers of the product. For example, a developer team uses an ethics tool to assess a system which produces an output (e.g. report). This output can then be released to other stakeholders who can act on or respond to the findings. As might be expected, the stakeholders who are likely to be applying the tool are mainly in the production side of AI systems (developer, quality assurance and delivery roles), with the results of the tool being used by

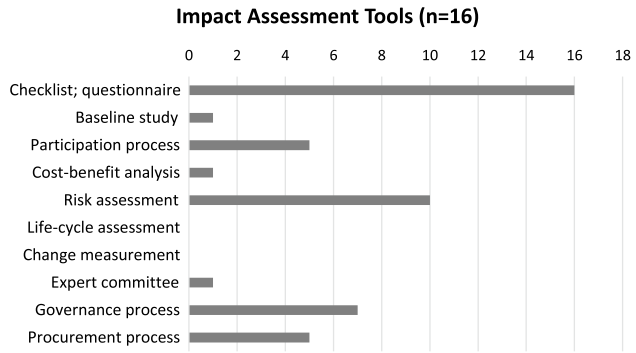

Fig. 5 Impact assessment tools

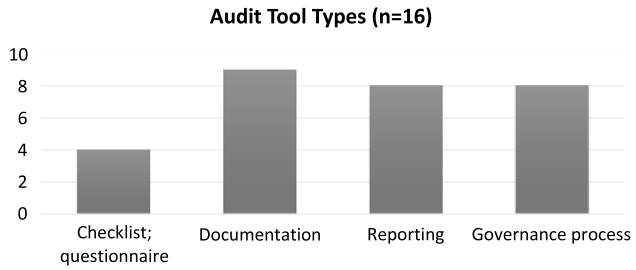

Fig. 6 Audit tools

decision-makers and senior staff. The tools can also comprise evidence for shareholders and citizens, and oversight bodies. Despite participation processes being recommended in some impact assessments (see Fig. 5 Impact Assessment Tools), we can see that the range of stakeholders involved in the proposed tools only really captures those involved in producing AI systems, or procuring them, and wider stakeholders (to whom negative impacts of deployment of an AI system actually accrue, i.e. users and wider stakeholders in society) are not included in these processes.

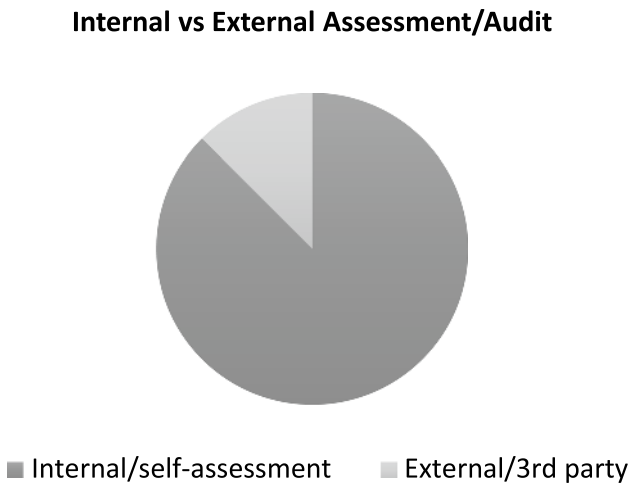

Fig. 7 Internal/self-assessment vs external/3rd-party assessment/audit

Figure 5 represents the number of component tools used within an Impact Assessment. A checklist or questionnaire is used in all 16 Impact Assessments coded in the study (as compared to only 4/16 audit tools Fig. 6 Audit Tools). It is a structured way to record proposals, decisions and actions, and can also be used to embed a governance process for the process of applying an ethical tool. Risk assessments were also commonly included, often embedded as part of the checklist process. Impact assessments were also used as part of a procurement process to assess ethical impacts and risks of purchasing an AI system. Unlike other types of impact assessment like EIA, little attention was paid to measurement of baseline conditions or predicting change. There were also omissions in these proposed tools for AI which did not include the types of impacts that would be measured in a life-cycle assessment for a product or process, leaving out key considerations like resource or energy use and sustainability.

Figure 6 shows the tools types identified in ethics tools that are categorised as audits. The focus of these is on appropriate documentation for verification and assurance in the audit process, the reporting process and on having appropriate governance mechanisms in place.

Figure 7 illustrates whether the ethical tool is an internal assessment or audit, as opposed to a verification process from a 3rd party. External verification only occurs in 5 of the 35 tools analysed, surfacing in either the certified standards from IEEE or in tools like incident databases which are designed for transparency.

Figure 8 breaks down these tools into workshop and design tools, forms of technical documentation, and tools for testing or monitoring data and models. The workshop materials do not fit into an impact assessment or audit framework and are not designed to provide verifiable evidence of process, but more to elicit 'ethical thinking' from design

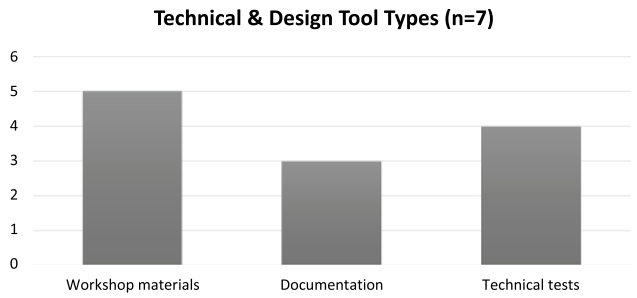

Fig. 8 Technical and design tools

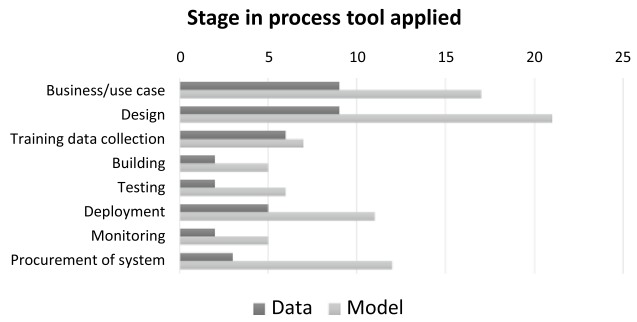

Fig. 9 Stage in process tool applied

teams, unlike the documentation tools which can provide evidence for audits. The technical tests are part of creating robust systems and can also provide an audit trail.

Figure 9 illustrates the stage in the production process pipeline that the proposed tools apply to, and also categorises these tools depending on whether they are focused on the data or the model. Many of the tools are designed for use early in the process—at the use case and design phase, where the main focus on the model is found. The attention to the model is also more marked in the deployment and procurement process, with data also being an important object for assessment early in the process.

5 Discussion

Reviewing the landscape of tools for applying ethical principles to AI, our research reveals some key themes emerging. Emerging from our analysis, there are three key areas where tools are being developed—impact assessment, audit and technical/design tools. As Fig. 10 illustrates, these approaches target different stages of AI system development and provide different outcomes. Ex ante impact assessments are used at the early stages of use case development, and for procurement processes to provide a predictive decision-making tool for whether a proposed AI system should progress to development and/or be deployed or purchased and what are

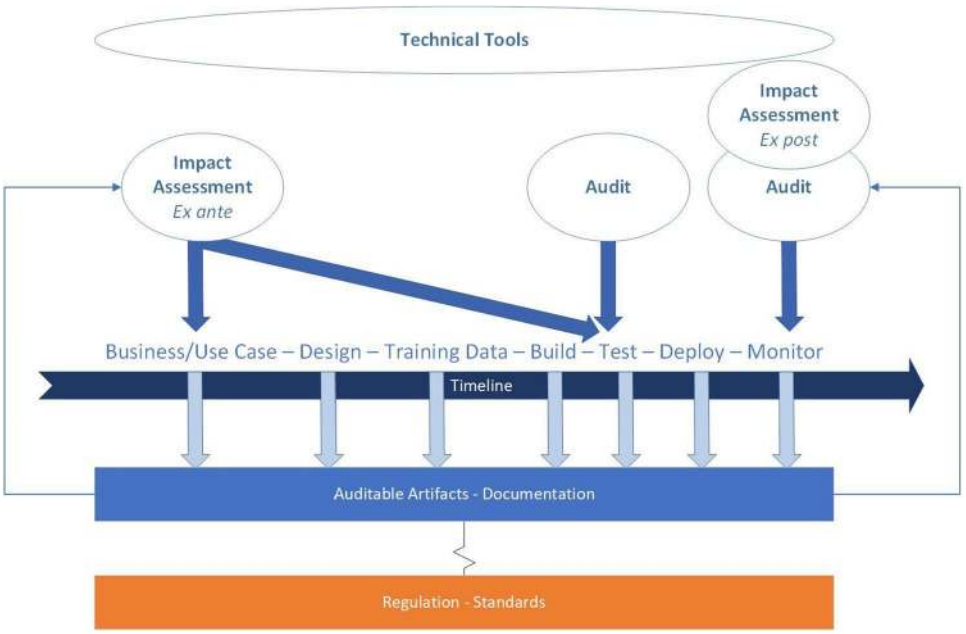

Fig. 10 Process model for application of tools

the possible impacts of its use. Ex post impact assessment is used as a post-deployment tool to capture the impacts of a system, often in comparison to a particular set of stakeholders, or issues like impact on human rights or democracy.

Audit tools showed an equal level of presence in our study to impact assessment which can be used for assurance of production and monitoring purposes. Audit processes traditionally follow well-defined systematic processes that require third-party verification. There is some confusion in the current landscape between a technical intervention (often called an audit e.g. for fairness or bias), and what is more generally understood business practice of formal auditing [98]. In our study, we have differentiated between those tools that more closely resemble other comparable audits, and categorised tools for specific aspects of the assessment of data training sets or models as technical tools—not audits. Technical tools do have an important role to play in addressing ethical issues in AI systems, but ultimately need to be part of a wider governance process. The documentation produced by these tools should form part of impact assessment and audit processes in order that all ethical aspects of a product can be captured (not just a focus on, e.g. metrics for fairness [102]). In Fig. 8, Technical and Design Tools have been incorporated into the model as an input to the category of auditable artifacts which are necessary for evidence in both impact assessments and audits.

This study contributes to the discussion about ethical AI by clarifying the different themes emerging in this landscape. It also serves to illustrate how complex this landscape is, and as others have noted [7, 19, 98, 102, 104, 105], this provides a barrier to those developing or purchasing AI systems as to which tool is appropriate for their purposes. Addressing ethical issues systematically requires resource and time, familiarity with assessment/audit regimes and the ability to use the outputs of these tools to make judgements. Even with the aid of procedures and processes to surface ethical risks, there are still difficult judgements to be made in the real world. Competing claims between different actors, balancing protection and benefits and differing ethical viewpoints mean that even the most rigorously applied tools will still require complex human judgements. As Floridi [106] observes 'there is no ethics without choices, responsibilities, and moral evaluations, all of which need a lot of relevant and reliable information and quite a good management of it.' Ethical tools can though, provide a reliable evidence base on which to make decisions, but without robust oversight may result in procedures that produce a checklist mentality and performative gestures that constitute 'ethics washing' [13, 107, 108].

An important finding from our research also puts in plain sight the fact that these tools are emerging in a landscape where there currently there are no specific regulatory regimes or legislation for AI systems. In Fig. 8, the top-level—Regulation and Standards—has no direct connection to the processes below. This means that these tools are for voluntary self-regulation without external governance mechanisms where third-party agents can interrogate the process and decisions. As Raab notes, 'an organisation or profession that simply marks its own homework cannot make valid claims to be trustworthy' [13, p. 13]. Impact assessment and audit practices in other domains as discussed above sit within national and international regulation and provide for external verification and assurance. Metcalf et al. [109] conclude that historically impact assessments

are tools for evaluation that operate within relationships of accountability between different stakeholder groups. As our analysis reveals there is currently a focus in AI ethics tools on a narrow group of internal stakeholders, with little transparency or accountability to wider stakeholders. In order for those who build AI products and services, and those who buy them, to provide credible and trustworthy governance of this technology, external verification, means of redress and contestation by different stakeholder groups, and methods of control for wrongdoing are required.

There are moves now to draft legislation to address the specific problems AI systems can produce with the EU leading the global pack with its recently published 'Proposal for a Regulation laying down harmonised rules on artificial intelligence' [110]. It proposes a risk-based approach to AI regulation, proposing an audit regime which will strengthen enforcement and sets out 'new requirements for documentation, traceability and transparency… The framework will envisage specific measures supporting innovation, including regulatory sandboxes and specific measures supporting small-scale users and providers of high-risk AI systems to comply with the new rules' [110, p. 10]. China is also working on these challenges with new regulation being proposed for data protection which includes processing using AI techniques, and specific new regulation for applications like facial recognition and autonomous vehicles [111, 112].

In the US, a surprisingly strongly worded blog by the Federal Trade Commission (FTC) [113] states that companies building or deploying AI should be 'using transparency frameworks and independent standards, by conducting and publishing the results of independent audits, and by opening your data or source code to outside inspection your statements to business customers and consumers alike must be truthful, non-deceptive, and backed up by evidence.' The post makes reference to a range of existing laws which might be applied to AI products and warns 'keep in mind that if you don't hold yourself accountable, the FTC may do it for you' [113]. As Bryson argues 'All human activity, particularly commercial activity, occurs in the context of some sort of regulatory framework' [114, p. 8]. Providing assurance of the safety, security and reliability of a project, product or system is the basis for the impact assessment and audit traditions discussed in this paper, the practices of which can be usefully applied to the domain of AI. It should also be noted that these traditions sit within established legal and regulatory frameworks. AI will need a similar regulatory ecosystem, which are being developed across jurisdictions, but yet to be formally adopted. Future work could usefully deploy typologies based on existing regulations and standards to map where gaps exist.

Our findings also serve to illustrate the confusion in language and approach to what are understood as the key features of impact assessment and audit. The latest thinking emerging from the UK Centre for Data Ethics and Innovation (CDEI) echoes the findings in our research, recognising the need for clarification around AI ethics methodologies in practice [115]. The CDEI categorises the difference between impact assessment and audit and assurance in a similar way to our mapping in Fig. 8, which they divide into compliance assurance (audit), and risk assurance (impact assessment) which are used at different stages of the process and meet different needs. 'The current discourse sometimes mistakenly calls on risk assurance tools like impact assessments to achieve the goals of Compliance, leading to complex and burdensome efforts to address common challenges. Meanwhile, sometimes compliance mechanisms like audits are discussed as if they can achieve loftier goals—an exercise which may be better suited to Risk Assurance tools like impact assessments' [115]. Clarifying the types of tools appropriate for which assessment and governance outcomes, and implementing well-regulated compliance regimes for producers of AI systems would be a great step towards effectively operationalising the ethical principles and concerns motivating the production of AI ethics tools. A note of caution though on how effective regulation might be, see for example, the recent European Parliament resolution on UK protection of personal data where concern is expressed 'about the lack and often non-existent enforcement of the GDPR by the UK when it was still a member of the EU; points, in particular, to the lack of proper enforcement by the UK Information Commissioner's (ICO's) Office in the past' [116, p. 6].

We also found gaps in the inclusion of a wide range of stakeholders in the process of AI ethics tools. As Fig. 2 illustrates, current tools, not surprisingly, are designed for use by those in the production process of AI systems and the key decision-makers around that process. Participation in these tools was found to be limited beyond these core stakeholders, except for tools explicitly focused on participation processes [78]. There is a long tradition in HCI of Participatory Design (PD), and Human/Ethically/Value-centred Design [117], which have been wrestling with the problem of inclusion and participation in the process of design and production of ICT systems [118]. Participatory processes have also been addressed in pTA where governance of emerging technology includes deliberative public forums [47, 48], and research organisations like the Ada Lovelace Institute enabling 'informed and complex public dialogue about technology, policy and values, and represent the voice of the public in debates around data and AI' [119].

Including wider stakeholders presents challenges at the level of companies producing AI systems, as it is time and resource heavy and requires particular sets of skills not necessarily present in developer teams [120]. Participation is also about power, who has the power to decide, who is invited to the table, whose views and goals take precedence. As Beck

pointed out in the field of PD 'rather than participation, concern with power and dominance needs to be stated as the core of the research field' [118, p. 77]. Who should decide on the design and use/non-use of AI systems is often framed as a 'project of expert oversight', giving little or no input to those stakeholders subject to AI systems [14], and where the process can become a form of 'participation washing' [121]. This is where informed public debate must feed into regulation and the law, to ensure appropriate governance is in place to protect rights and represent the views of all stakeholders in a society.

6 Conclusion

This work provides an analysis of the AI guidelines and frameworks that have practical tools to operationalise ethical concerns. By reviewing best practices from historical frameworks created to assess the effects of technology on the environment [24], information privacy [25], data protection [26] and human rights [27], we create a typology of concerns that previous generations of impact assessments and audits have found beneficial to consider. Using this typology, we examine the current crop of AI and data ethical guidelines and frameworks.

The available guidelines cluster around the product development phase of AI and are focused on being used by and documenting the concerns from mainly developers, delivery, and quality assurance roles. The reporting output from these tools is then used by management decision-makers as opposed to inform the developers of better practice, or any other stakeholders. Moreover, there is little participation in the assessment or audit process by certain stakeholder groups, particularly the voiceless, vested interests and users, who are not included in the process of applying the tools or interacting with the outputs as tools for transparency or decision-making. Nearly all of the tools available are for internal self-assessment, with only the IEEE standards requiring any kind of external verification, and the two examples of public registers providing explicit transparency. In addition to missing large stakeholder groups, the current set of AI Guidelines and tools do not fully utilize the full range of techniques available, including: participation process, baseline study, life-cycle assessment, change measurement or expert committees. Finally, we note that there is no regulatory requirement for any utilization of impact assessments or audits within this field at the moment, minimizing likely adoption and true application of them.

Appendix 1: Source documents

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
1	Risks, Harms and Benefits Assessment	2017	UN Global Pulse	UN Global Pulse	https://www.unglobalpulse.org/policy/risk-assessment/	27/06/2018
2	AI and Big Data: A blueprint for a human rights, social and ethical impact assessment	2018	Mantelero, Alessandro	Computer Law & Security Review	https://doi.org/10.1016/j.clsr.2018.05.017	17/05/2019
3	ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY	2018	Reisman, Dillon; Schultz, Jason; Crawford, Kate; Whittaker, Meredith	AI Now Institute	https://ainowinstitute.org/aiareport2018.pdf	24/06/2019
4	An Ethical Toolkit for Engineering/Design Practice	2018	Shannon, V; McKenna, D	Markkula Center for Applied Ethics, Santa Clara University	https://www.scu.edu/ethics-in-technology-practice/ethical-toolkit/	14/09/2019
5	Ethical Data and Information Management: Concepts, Tools and Methods	2018	O'Keefe, Katherine; Brien, Daragh O	Kogan Page Ltd	978-0-7494-8205-3	15/01/2020
6	Ethical OS	2018	Institute for the Future; Omidyar Network	Ethical.os	https://ethicalos.org/	13/06/2019

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
7	Ethics & Algorithms Toolkit (beta)	2018	GovEx; City and County of San Francisco; Harvard DataSmart; Data Community DC	Ethicstoolkit.ai	https://ethicstoolkit.ai/	27/01/2020
8	AI Fairness 360	2019	IBM Research	IBM	aif360.mybluemix.net/resources	12/01/2020
9	AI Procurement in a Box	2019	World Economic Forum	World Economic Forum	https://www.weforum.org/reports/ai-procurement-in-a-box/	13/10/2020
10	AI-RFX Procurement Framework	2019	The Institute for Ethical AI & Machine Learning		https://ethical.institute	18/06/2019
11	Algorithmic Impact Assessment (AIA)	2019	Secretariat, Treasury Board of Canada	Government of Canada	https://www.canada.ca/en/government/system/digital-government/modern-emerging-technologies/responsible-use-ai/algorithmic-impact-assessment.html	27/06/2019
12	Codex for Data- Based Value Creation	2019	Swiss Alliance for Data-Intensive Services Expert Group	Swiss Alliance for Data-Intensive Services	www.data-service-alliance.ch/codex	16/03/2020
13	Consequence Scanning – doteveryone	2019	Doteveryone	Doteveryone.org	https://doteveryone.org.uk/project/consequence-scanning/	18/06/2019
14	IBM Watson OpenScale	2019	IBM	IBM	https://www.ibm.com/uk-en/cloud/watson-openscale	13/11/2020
15	IEEE SA—The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)	2019	IEEE Standards Association	IEEE	https://standards.ieee.org/industry-connections/ecpais.html	30/08/2019
16	Judgment Call the Game: Using Value Sensitive Design and Design Fiction to Surface Ethical Concerns Related to Technology	2019	Ballard, Stephanie; Chappell, Karen M.; Kennedy, Kristen	Proceedings of the 2019 on Designing Interactive Systems Conference	https://doi.org/10.1145/3322276.3323697	16/11/2020
17	Model Cards for Model Reporting	2019	Mitchell, Margaret; Wu, Simone; Zaldivar, Andrew; Barnes, Parker; Vasserman, Lucy; Hutchinson, Ben; Spitzer, Elena; Raji, Inioluwa Deborah; Gebru, Timnit	asXiv Working Paper	https://doi.org/10.1145/3287560.3287596	25/09/2019
18	Model Ethical Data Impact Assessment	2019	IAF	Information Accountability Foundation	http://informationaccountability.org/publications/	08/12/2019
19	ODI Data Ethics Canvas	2019	ODI	ODI	https://theodi.org/article/data-ethics-canvas/	27/06/2019

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
20	Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector	2019	Leslie, David	The Alan Turing Institute	https://zenodo.org/record/3240529	13/01/2020
21	A Proposed Model AI Governance Framework—Second Edition	2020	PDPC Singapore	Personal Data Protection Commission Singapore	https://www.pdpc.gov.sg/resources/model-ai-gov	12/01/2020
22	AI Blindspot: A Discovery Process for preventing, detecting, and mitigating bias in AI systems	2020	Calderon, A; Taber, D; Qu, H; Wen, J	MIT	https://aiblindspot.media.mit.edu/	09/11/2020
23	Algorithm Register	2020	City of Amsterdam	City of Amsterdam	https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/grip-op-algoritmes/	27/11/2020
24	Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment	2020	EU HLEG AI	European Commission	https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence	30/08/2020
25	Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing	2020	Raji, Inioluwa Deborah; Smart, Andrew; White, Rebecca N; Mitchell, Margaret; Gebru, Timnit; Hutchinson, Ben; Smith-Loud, Jamila; Theron, Daniel; Barnes, Parker	FAT*'20 Barcelona	https://dl.acm.org/doi/pdf/10.1145/3351095.3372873	16/11/2020
26	Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI	2020	Madaio, Michael A.; Stark, Luke; Wortman Vaughan, Jennifer; Wallach, Hanna	Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems	https://doi.org/10.1145/3313831.3376445	08/10/2020
27	Corporate Digital Responsibility	2020	Lobschat, Lara; Mueller, Benjamin; Eggers, Felix; Brandimarte, Laura; Diefenbach, Sarah; Kroschke, Mirja; Wirtz, Jochen	Journal of Business Research	https://doi.org/10.1016/j.jbusres.2019.10.006	28/01/2020
28	Data Ethics Framework	2020	DCMS	Gov.uk	https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-legislation-and-codes-of-practice-for-use-of-data	13/10/2020

Key	Title	Year	Author	Publisher	URL DOI ISBN	Access date
29	Datasheets for Datasets	2020	Gebru, Timnit; Morgenstern, Jamie; Vecchione, Briana; Vaughan, Jennifer Wortman; Wallach, Hanna; Daumé III, Hal; Crawford, Kate	arXiv:1803.09010 [cs]	arXiv:1803.09010 [cs]	12/06/2020
30	Empowering AI Leadership	2020	World Economic Forum	World Economic Forum	https://spark.adobe.com/page/RsXNkZANwMLEf/	30/09/2020
31	Fairlearn: A toolkit for assessing and improving fairness in AI	2020	Bird, Sarah; Dudík, Miroslav; Edgar, Richard; Horn, Brandon; Lutz, Roman; Milan, Vanessa; Sameki, Mehrnoosh; Wallach, Hanna; Walker, Kathleen; Design, Allovus	IBM	https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf	13/10/2020
32	IEEE Draft Model Process for Addressing Ethical Concerns During System Design P7000/D3	2020	IEEE Standards Association	IEEE	https://standards.ieee.org/project/7000.html	04/06/2020
33	IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being Std 7010	2020	IEEE Standards Association	IEEE	https://standards.ieee.org/industry-connections/ec/autonomous-systems.html	30/08/2020
34	Responsible AI	2020	TensorFlow	Tensorflow.org	https://www.tensorflow.org/resources/responsible-ai	02/11/2020
35	Standard Clauses for Municipalities for Fair Use of Algorithmic Systems	2020	City of Amsterdam	City of Amsterdam	https://www.amsterdam.nl/wonen-leefomgeving/innovatie/de-digitale-stad/grip-op-algoritmes/	27/11/2020
36	Toward situated interventions for algorithmic equity: lessons from the field	2020	Katell, Michael; Young, Meg; Dailey, Dharma; Herman, Bernease; Guetler, Vivian; Tam, Aaron; Binz, Corinne; Raz, Daniella; Krafft, P. M	Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency	https://doi.org/10.1145/3351095.3372874	28/01/2020
37	Value-based Engineering for Ethics by Design	2020	Spiekermann, Sarah; Winkler, Till	IEEE pre-print	arXiv:2004.13676 [cs]	06/10/2020
38	Welcome to the Artificial Intelligence Incident Database	2020	Partnership on AI	The Partnership on AI	https://incidentdatabase.ai/	21/11/2020
39	White Paper on Data Ethics in Public Procurement of AI-based Services and Solutions	2020	Hasselbalch, Gry; Olsen, B; Tranberg, P	DataEthics.eu	https://dataethics.eu/wp-content/uploads/dataethics-whitepaper-april-2020.pdf	25/08/2020

Appendix 2: Abbreviations

CBA	Cost–Benefit Analysis
CDEI	UK Centre for Date Ethics and Innovation
DPIA	Data Protection Impact Analysis
EIA	Environmental Impact Assessment
ERA	Environmental Risk Assessment
FIA	Financial Impact Assessment
eTA	Ethical Technology Assessment
FIP	Fair Information Practices
FTC	US Federal Trade Commission
GAAP	Generally Accepted Accounting Principles
GDPR	EU General Data Protection Regulation
IFRS	International Financial Reporting Standards
PD	Participatory Design
pTA	Participatory Technology Assessment
SIA	Social Impact Analysis
TA	Technology Assessment

Funding Supported by EPSRC funding via Web Science CDT.

Availability of data and material All data contained within paper.

Declarations

Conflict of interest No conflicts.

References

1. Diakopoulos, N.: Accountability in algorithmic decision making. Commun. ACM **59**(2), 56–62 (2016). https://doi.org/10.1145/2844110
2. Eubanks, V.: Automating inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Publishing Group (2018)
3. Council regulation (EU) 2016/679: On the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Off. J. **L119/1** (2016) Available http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC. Accessed 23 Sep. 2017. **(Online)**.
4. Hagendorff, T.: The ethics of AI ethics—an evaluation of guidelines. Minds Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
5. Jobin, A., Ienca, M., Vayena, E.: The global landscape of AI ethics guidelines. Nat. Mach. Intell. **1**(9), 389–399 (2019). https://doi.org/10.1038/s42256-019-0088-2
6. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., Srikumar, M.: Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. In: Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3518482 (2020). Available https://papers.ssrn.com/abstract=3518482. Accessed 27 Jan. 2020. **(Online)**
7. 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 (2019). https://doi.org/10.1007/s11948-019-00165-5
8. Solove, D.J.: A taxonomy of privacy. Univ. Pa Law Rev. **154**(3), 477 (2006). https://doi.org/10.2307/40041279
9. Citron. D.K., Solove, D.J.: Privacy harms. In: Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3782222 (2021). https://doi.org/10.2139/ssrn.3782222.
10. Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2), 2053951716679679 (2016). https://doi.org/10.1177/2053951716679679
11. Hirsch, D., Bartley, T., Chandrasekaran, A., Norris, D., Parthasarathy, S., Turner, P. N.: Business data ethics: emerging trends in the governance of advanced analytics and AI. In: The Ohio State University, Ohio State Legal Studies Research Paper No. 628, 2020. Available https://cpb-us-w2.wpmucdn.com/u.osu.edu/dist/3/96132/files/2020/10/Final-Report-1.pdf. **(Online)**
12. Solove, D.J.: Privacy and power: computer databases and metaphors for information privacy. Stanford Law Rev. **53**, 71 (2001)
13. Raab, C.D.: Information privacy, impact assessment, and the place of ethics. Comput. Law Secur. Rev. **37**, 105404 (2020). https://doi.org/10.1016/j.clsr.2020.105404
14. Greene, D., Hoffmann, A.L., Stark, L.: Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Presented at the Hawaii International Conference on System Sciences (2019). https://doi.org/10.24251/HICSS.2019.258.
15. Kazim, E., Koshiyama, A.: AI assurance processes. In: Social Science Research Network, Rochester, SSRN Scholarly Paper ID 3685087 (2020). https://doi.org/10.2139/ssrn.3685087.
16. Kind, C.: The term 'ethical AI' is finally starting to mean something. VentureBeat (2020). https://venturebeat.com/2020/08/23/the-term-ethical-ai-is-finally-starting-to-mean-something/. Accessed 23 Aug. 2020
17. Ryan, M., Stahl, B.C.: Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. J. Inf. Commun. Ethics Soc. (2020). https://doi.org/10.1108/JICES-12-2019-0138
18. AlgorithmWatch: "AI Ethics Guidelines Global Inventory by AlgorithmWatch," AI Ethics Guidelines Global Inventory (2020). https://inventory.algorithmwatch.org. Accessed 11 Aug. 2020.
19. Schiff, D., Borenstein, J., Biddle, J., Laas, K.: AI ethics in the public, private, and NGO sectors: a review of a global document collection. IEEE Trans. Technol. Soc. (2021). https://doi.org/10.1109/TTS.2021.3052127
20. Bird, S., et al.: Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft (2020). Available https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf. Accessed 13 Oct. 2020. **(Online)**

21. Mitchell, M., et al.: Model cards for model reporting. Proc. Conf. Fairness Account. Transpar. FAT **19**, 220–229 (2019). https://doi.org/10.1145/3287560.3287596
22. Gebru, T., et al.: "Datasheets for Datasets" (2020). Available http://arxiv.org/abs/1803.09010. Accessed 03 Dec. 2020. **(Online)**
23. Crawford, K.: Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven (2021)
24. Morgan, R.K.: Environmental impact assessment: the state of the art. Impact Assess. Proj. Apprais. **30**(1), 5–14 (2012). https://doi.org/10.1080/14615517.2012.661557
25. Clarke, R.: Privacy impact assessment: Its origins and development. Comput. Law Secur. Rev. **25**(2), 123–135 (2009). https://doi.org/10.1016/j.clsr.2009.02.002
26. Information Commissioner's Office: Data protection impact assessments (2018). https://ico.org.uk/for-organisations/guide-to-the-general-data-protection-regulation-gdpr/accountability-and-governance/data-protection-impact-assessments/. Accessed 07 Jun. 2018
27. The Danish Institute for Human Rights: Human rights impact assessment guidance and toolbox - road-testing version. *The Danish Institute for Human Rights* (2016). https://www.humanrights.dk/business/tools/human-rights-impact-assessment-guidance-and-toolbox. Accessed 03 Feb. 2020
28. Renn, O.: Risk Governance: Coping with Uncertainty in a Complex World. Earthscan (2008)
29. Coates, J.F.: Some methods and techniques for comprehensive impact assessment. Technol. Forecast. Soc. Change **6**, 341–357 (1974). https://doi.org/10.1016/0040-1625(74)90035-3
30. IAIA: Technology Assessment (2009). https://www.iaia.org/wiki-details.php?ID=26. Accessed 26 Jan. 2021
31. Palm, E., Hansson, S.O.: The case for ethical technology assessment (eTA). Technol. Forecast. Soc. Change **73**(5), 543–558 (2006). https://doi.org/10.1016/j.techfore.2005.06.002
32. STOA: Centre for AI | Panel for the Future of Science and Technology (STOA) | European Parliament (2021). https://www.europarl.europa.eu/stoa/en/centre-for-AI. Accessed 11 Feb. 2021
33. Hennen, L.: Why do we still need participatory technology assessment? Poiesis Prax. **9**, 27–41 (2012). https://doi.org/10.1007/s10202-012-0122-5
34. CSPO: Participatory Technology Assessment | CSPO, Consortium for Science and Policy Outcomes (2021). https://cspo.org/areas-of-focus/pta/. Accessed 12 Feb. 2021
35. Kiran, A., Oudshoorn, N.E.J., Verbeek, P.P.C.C.: Beyond checklists: toward an ethical-constructive technology assessment. J. Respons Innov. **2**(1), 5–19 (2015). https://doi.org/10.1080/23299460.2014.992769
36. Suter, G.W., Barnthouse, L.W., O'Neill, R.V.: Treatment of risk in environmental impact assessment. Environ. Manage. **11**(3), 295–303 (1987). https://doi.org/10.1007/BF01867157
37. UN Environment: "Assessing Environmental Impacts A Global Review Of Legislation-UNEP-WCMC," In: UNEP-WCMC's official website—Assessing Environmental Impacts A Global Review Of Legislation (2018). https://www.unep-wcmc.org/assessing-environmental-impacts--a-global-review-of-legislation. Accessed 12 Feb. 2021.
38. Glucker, A.N., Driessen, P.P.J., Kolhoff, A., Runhaar, H.A.C.: Public participation in environmental impact assessment: why, who and how? Environ. Impact Assess. Rev. **43**, 104–111 (2013). https://doi.org/10.1016/j.eiar.2013.06.003
39. IMA Europe: "Life Cycle Assessment | IMA Europe," In: Industrial Mineral Association-Europe (2020). https://www.ima-europe.eu/eu-policy/environment/life-cycle-assessment. Accessed 06 May 2021
40. Aven, T.: Risk assessment and risk management: review of recent advances on their foundation. Eur. J. Oper. Res. **253**(1), 1–13 (2016). https://doi.org/10.1016/j.ejor.2015.12.023
41. Edwards, M.M., Huddleston, J.R.: Prospects and perils of fiscal impact analysis. J. Am. Plann. Assoc. **76**(1), 25–41 (2009). https://doi.org/10.1080/01944360903310477
42. Pearce, D.W.: Cost-Benefit Analysis, 2nd edn. Macmillan International Higher Education (2016)
43. Kemp, D., Vanclay, F.: Human rights and impact assessment: clarifying the connections in practice. Impact Assess. Proj. Apprais. **31**(2), 86–96 (2013). https://doi.org/10.1080/14615517.2013.782978
44. Kende-Robbe, C.: Poverty and social impact analysis : linking macroeconomic policies to poverty outcomes: summary of early experiences. IMF (2003). https://www.imf.org/en/Publications/WP/Issues/2016/12/30/Poverty-and-Social-Impact-Analysis-Linking-Macroeconomic-Policies-to-Poverty-Outcomes-16248. Accessed 12 Feb. 2021
45. Roessler, B.: New ways of thinking about privacy. (2008). https://doi.org/10.1093/oxfordhb/9780199548439.003.0038.
46. Westin, A.F.: Privacy and Freedom. Ig Publishing (1967)
47. Westin, A.F.: Information Technology in a Democracy. Harvard University Press (1971)
48. Stewart, B.: Privacy impact assessments. Priv. Law Policy Rep. **39**(4), 1996 (2021). Available http://www.austlii.edu.au/au/journals/PLPR/1996/39.html Accessed: Feb. 17, 2021. [Online]
49. Financial Reporting Council: Auditors I Audit and Assurance I Standards and Guidance for Auditors I Financial Reporting Council (2020). https://www.frc.org.uk/auditors/audit-assurance/standards-and-guidance. Accessed 26 Apr. 2021
50. Rusby, R.: The interpretation and evaluation of assurance cases. In: Computer Science Laboratory, SRI International, Menlo Park CA 94025, USA, Technical Report SRI-CSL-15–01 (2015)
51. Bloomfield, R., Khlaaf, H., Conmy, P.R., Fletcher, G.: Disruptive innovations and disruptive assurance: assuring machine learning and autonomy. Computer **52**(9), 82–89 (2019). https://doi.org/10.1109/MC.2019.2914775
52. International Organization for Standardization: "ISO-Standards," ISO (2021). https://www.iso.org/standards.html. Accessed 15 Jul. 2021
53. International Organization for Standardization: "ISO-Certification," ISO (2021). https://www.iso.org/certification.html. Accessed 15 Jul. 2021
54. PwC UK: "Understanding a financial statement audit," PricewaterhouseCooper, UK, (2013). Available https://www.pwc.com/gx/en/audit-services/publications/assets/pwc-understanding-financial-statement-audit.pdf. **(Online)**
55. Brundage, M., et al.: Toward trustworthy AI development: mechanisms for supporting verifiable claims (2020). Available http://arxiv.org/abs/2004.07213. Accessed 16 Nov. 2020. **(Online)**.
56. Mökander, J., Floridi, L.: Ethics-based auditing to develop trustworthy AI. Minds Mach. (2021). https://doi.org/10.1007/s11023-021-09557-8
57. Starr, C.: Social benefit versus technological risk. Science **165**(3899), 1232–1238 (1969)
58. Thompson, K.M., Deisler, P.F., Schwing, R.C.: Interdisciplinary vision: the first 25 years of the society for risk analysis (SRA), 1980–2005. Risk Anal. **25**(6), 1333–1386 (2005). https://doi.org/10.1111/j.1539-6924.2005.00702.x
59. Beck, P.U.: Risk Society: Towards a New Modernity. SAGE (1992)
60. Moses, K., Malone, R.: Development of risk assessment matrix for NASA Engineering and safety center NASA technical reports server (NTRS). In: NASA Technical Reports Server (NTRS) (2004). https://ntrs.nasa.gov/citations/20050123548. Accessed 27 May 2021

61. Hayne, C., Free, C.: Hybridized professional groups and institutional work: COSO and the rise of enterprise risk management. Account. Organ. Soc. **39**(5), 309–330 (2014). https://doi.org/10.1016/j.aos.2014.05.002
62. Lauterbach, A., Bonime, A.: Environmental risk social risk governance risk. Risk Manage, 3 (2018).
63. Floridi, L., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds Mach. **28**(4), 689–707 (2018). https://doi.org/10.1007/s11023-018-9482-5
64. High Level Expert Group on AI: "Ethics guidelines for trustworthy AI," European Commission, Brussels, Text (2019). Available https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed 23 May 2019. **(Online)**
65. Freeman, R.E.: Strategic Management: A Stakeholder Approach. Cambridge University Press (2010)
66. Donaldson, T., Preston, L.E.: The stakeholder theory of the corporation: concepts, evidence, and implications. Acad. Manage. Rev. **20**(1), 65 (1995). https://doi.org/10.2307/258887
67. Business Roundtable: "Our Commitment," *Business Roundtable—Opportunity Agenda* (2020). https://opportunity.businessroundtable.org/ourcommitment/. Accessed 05 Feb. 2021
68. TensorFlow: "Responsible AI," *TensorFlow* (2020). https://www.tensorflow.org/resources/responsible-ai. Accessed 02 Nov. 2020
69. Bantilan, N.: Themis-ml: a fairness-aware machine learning interface for end-to-end discrimination discovery and mitigation. (2017). Available http://arxiv.org/abs/1710.06921. Accessed 13 Nov. 2020. **(Online)**
70. Bellamy, R. K. E., et al.: AI Fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. (2018). Available http://arxiv.org/abs/1810.01943. Accessed 27 May 2021. **(Online)**
71. Lee, M.S.A., Floridi, L., Singh, J.: Formalising trade-offs beyond algorithmic fairness: lessons from ethical philosophy and welfare economics. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00067-y
72. Hutchinson, B., Mitchell, M.: 50 years of test (Un)fairness: lessons for machine learning. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, New York, pp. 49–58 (2019). https://doi.org/10.1145/3287560.3287600.
73. Veale, M., Van Kleek, M., Binns, R.: Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, New York, p. 440:1–440:14 (2018). https://doi.org/10.1145/3173574.3174014.
74. Hoffmann, A.L.: Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inf. Commun. Soc. **22**(7), 900–915 (2019). https://doi.org/10.1080/1369118X.2019.1573912
75. Radford, J., Joseph, K.: Theory in theory out: the uses of social theory in machine learning for social science. Front. Big Data. (2020). https://doi.org/10.3389/fdata.2020.00018
76. Institute for the Future and Omidyar Network, "Ethical OS," (2018). https://ethicalos.org/. Accessed 21 Jun. 2019
77. Doteveryone, Consequence Scanning—doteveryone (2019). https://doteveryone.org.uk/project/consequence-scanning/. Accessed 18 Jun 2019
78. Madaio, M. A., Stark, L., Wortman Vaughan, J., Wallach, H.: Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, pp. 1–14 (2020). https://doi.org/10.1145/3313831.3376445.
79. Stephanidis, C., et al.: Seven HCI grand challenges. Int. J. Hum Comput Interact. **35**(14), 1229–1269 (2019). https://doi.org/10.1080/10447318.2019.1619259
80. Krippendorff, K.: Content analysis. In: International encyclopedia of communication, vol. 1. Oxford University Press, New York, pp 8 (1989). Available http://repository.upenn.edu/asc_papers/22. Accessed 08 Jul 2020 **(Online)**
81. Smith, K.B.: Typologies, taxonomies, and the benefits of policy classification. Policy Stud. J. **30**(3), 379–395 (2002). https://doi.org/10.1111/j.1541-0072.2002.tb02153.x
82. Singh, A., et al.: PriMP visualization—principled artificial intelligence project. In: Harvard Law School, Berkman Klein Center for Internet and Society (2018). https://ai-hr.cyber.harvard.edu/primp-viz.html. Accessed 24 Jun. 2019
83. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., Bao, M.: The values encoded in machine learning research (2021). Available http://arxiv.org/abs/2106.15590. Accessed 25 Jul. 2021 **(Online)**
84. International Standardization Organisation: "ISO 14001:2015,". ISO (2021). https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/06/08/60857.html. Accessed 26 Jul. 2021
85. Bengtsson, M.: How to plan and perform a qualitative study using content analysis. NursingPlus Open **2**, 8–14 (2016). https://doi.org/10.1016/j.npls.2016.01.001
86. Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S.: The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, Honolulu, pp. 195–200 (2019). https://doi.org/10.1145/3306618.3314289.
87. Clarke, T.: Accounting for Enron: shareholder value and stakeholder interests. Corp. Gov. Int. Rev. **13**(5), 598–612 (2005). https://doi.org/10.1111/j.1467-8683.2005.00454.x
88. du Plessis, J.J., Hargovan, A., Harris, J.: Principles of Contemporary Corporate Governance. Cambridge University Press (2018)
89. Freeman, R. E.: Strategic management: a stakeholder approach. Pitman (1984)
90. Foden, C.: Our structure. City of Lincoln Council (2019). https://www.lincoln.gov.uk/council/structure. Accessed 10 Jan. 2021
91. Stanley, M.: UK Civil Service—Grades and Roles. In: Understanding Government (2020). https://www.civilservant.org.uk/information-grades_and_roles.html. Accessed 10 Jan. 2021
92. National Crime Agency: Our leadership. In: National Crime Agency (2021). https://www.nationalcrimeagency.gov.uk/who-we-are/our-leadership. Accessed 10 Jan. 2021
93. Badr, W.: Evaluating machine learning models fairness and bias. Medium (2019). https://towardsdatascience.com/evaluating-machine-learning-models-fairness-and-bias-4ec82512f7c3. Accessed 13 Nov. 2020
94. Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F.: Secure, privacy-preserving and federated machine learning in medical imaging. Mach. Intell Nat. (2020). https://doi.org/10.1038/s42256-020-0186-1
95. Chapman, A., Missier, P., Simonelli, G., Torlone, R.: Capturing and querying fine-grained provenance of preprocessing pipelines in data science. Proc. VLDB Endow. **14**(4), 507–520 (2020). https://doi.org/10.14778/3436905.3436911
96. Information Commissioner's Office: Guidance on the AI auditing framework Draft guidance for consultation p. 105 (2020)
97. Mayring, P.: Qualitative content analysis: demarcation, varieties, developments. Forum Qual. Sozialforschung Forum Qual. Soc. Res. (2019). https://doi.org/10.17169/fqs-20.3.3343
98. Carrier, R., Brown, S.: Taxonomy: AI Audit, Assurance, and Assessment. For Humanity (2021). https://forhumanity.center/

blog/taxonomy-ai-audit-assurance-and-assessment. Accessed 26 Apr. 2021
99. Ada Lovelace Institute and Data Kind UK: Examining the black box: tools for assessing algorithmic systems (2020). https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/. Accessed 23 Feb. 2021
100. Patton, M.Q.: Qualitative Research and Evaluation Methods: Integrating Theory and Practice. SAGE Publications (2014)
101. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. SAGE (2013)
102. Lee, M. S. A., Singh, J.: The landscape and gaps in open source fairness toolkits. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, New York, pp. 1–13 (2021) https://doi.org/10.1145/3411764.3445261.
103. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 115:1-115:35 (2021). https://doi.org/10.1145/3457607
104. Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M., Abrahamsson, P.: Ethically aligned design of autonomous systems: industry viewpoint and an empirical study, p. 18 (2019)
105. Mulgan, G.: AI ethics and the limits of code(s). In: nesta (2019). https://www.nesta.org.uk/blog/ai-ethics-and-limits-codes/. Accessed 16 Sep. 2019
106. Floridi, L.: Why Information Matters. In: The New Atlantis (2017). http://www.thenewatlantis.com/publications/why-information-matters. Accessed 14 Oct 2020
107. Kitchin, R.: The ethics of smart cities (2019). Available https://www.rte.ie/brainstorm/2019/0425/1045602-the-ethics-of-smart-cities/. Accessed 07 May 2019 **(Online)**
108. Bietti, E.: From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, pp 210–219 (2020) https://doi.org/10.1145/3351095.3372860.
109. Metcalf, J., Moss, E., Watkins, E. A., Singh, R., Elish, M. C.: Algorithmic impact assessments and accountability: the co-construction of impacts, p 19 (2021)
110. European Commission: Proposal for a Regulation laying down harmonised rules on artificial intelligence | Shaping Europe's digital future. In: European Commission, Brussels, Proposal (2021). Available https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. Accessed 21 May 2021 **(Online)**
111. Webster, G.: Translation: personal information protection law of the People's Republic of China (Draft) (Second Review Draft) | DigiChina. In: Stanford DigiChina Cyber Policy Unit (2021). https://digichina.stanford.edu/news/translation-personal-information-protection-law-peoples-republic-china-draft-second-review. Accessed 21 May 2021
112. Lee, A., Sacks, S., Creemers, R., Shi, M., Webster, G.: China's draft privacy law adds platform self-governance, solidifies CAC's Role | DigiChina. In: Stanford DigiChina Cyber Policy Unit (2021). https://digichina.stanford.edu/news/chinas-draft-privacy-law-adds-platform-self-governance-solidifies-cacs-role. Accessed 21 May 2021
113. Jillson, E.: Aiming for truth, fairness, and equity in your company's use of AI. In: Federal Trade Commission (2021). https://www.ftc.gov/news-events/blogs/business-blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai. Accessed 20 Apr. 2021
114. Bryson, J.J.: The artificial intelligence of the ethics of artificial intelligence: an introductory overview for law and regulation. In: Dubber, M.D., Pasquale, F., Das, S. (eds.) The Oxford Handbook of Ethics of AI, pp. 1–25. Oxford University Press (2020)
115. CDEI: Types of assurance in AI and the role of standards. In: Centre for Data Ethics and Innovation Blog (2021). https://cdei.blog.gov.uk/2021/04/17/134/. Accessed 26 May 2021
116. European Parliament: The adequate protection of personal data by the United Kingdom (2021) https://www.europarl.europa.eu/doceo/document/TA-9-2021-0262_EN.html. Accessed 26 May 2021.
117. Simonsen, J., Robertson, T.: Routledge International Handbook of Participatory Design. Routledge, London (2012)
118. Beck, E.: P for Political: Participation is not enough. Scand. J. Inf. Syst. **14**(1) (2002). Available at https://aisel.aisnet.org/sjis/vol14/iss1/1. **(Online)**
119. Ada Lovelace Institute: Our Strategy (2020). https://www.adalovelaceinstitute.org/about/. Accessed 26 May 2021
120. Thuermer, G., Walker, J., Simperl, E., Carr, L.: When data meets citizens: an investigation of citizen engagement in data-driven innovation programmes. In: Presented at the 2nd Data Justice Conference, Cardiff University Online (2021)
121. Sloane, M., Moss, E., Awomolo, O., Forlano, L.: Participation is not a design fix for machine learning (2020). Available http://arxiv.org/abs/2007.02423. Accessed 26 May 2021. **(Online)**