Working Paper

2021/04/DSC/TOM

Implementing AI Principles: Frameworks, Processes, and Tools

Pal Boza
INSEAD, pal.boza@insead.edu

Theodoros Evgeniou
INSEAD, theodoros.evgeniou@insead.edu

In recent years, a number of international organisations, regulators, governments, academics, and as well businesses have worked on developing principles of Artificial Intelligence (AI). Alongside the development of these principles, there is an on-going discussion on how to regulate AI in order to best align risk management with optimising potential value creation of these technologies. Risk managing AI systems will likely become a regulatory and social expectations requirement, for all sectors and for both business and government. However emphasis on how to implement the proposed AI principles and upcoming regulations in practice is more recent, and appropriate tools to achieve this still need to be identified and developed. For example, implementing so-called Responsible AI requires the development of new processes, frameworks and tools, among others. We review the current state and identify possible gaps.

Implementing AI Principles: Frameworks, Processes, and Tools

Pal Boza, Senior Research Associate, INSEAD, pal.boza@insead.edu
Theodoros Evgeniou, Professor, Decision Sciences and Technology Management, INSEAD
theodoros.evgeniou@insead.edu

Abstract

In recent years, a number of international organisations, regulators, governments, academics, and as well businesses have worked on developing principles of Artificial Intelligence (AI). Alongside the development of these principles, there is an on-going discussion on how to regulate AI in order to best align risk management with optimising potential value creation of these technologies. Risk managing AI systems will likely become a regulatory and social expectations requirement, for all sectors and for both business and government. However emphasis on how to implement the proposed AI principles and upcoming regulations in practice is more recent, and appropriate tools to achieve this still need to be identified and developed. For example, implementing so-called Responsible AI requires the development of new processes, frameworks and tools, among others. We review the current state and identify possible gaps.

1. Introduction

While organizations develop their data, Artificial Intelligence (AI) and Machine Learning (ML) capabilities, new business opportunities enabled by these technologies are continuously identified, with the overall value creation potential estimated to be in the trillions of dollars. For example, according to an earlier PwC estimate, AI technologies have the potential of increasing global GDP by 14% by 2030, contributing up to \$15.7 trillion – of which, \$6.6 trillion is likely to come from increased productivity and \$9.1 trillion from consumption-side effects for example due to better products and demand stimulation.[1] However, the deployment of AI systems will also lead to potential new risks[2,3] which, if not regulated and managed well, may not only delay adoption and slow down innovation but also partly offset the benefits of these technologies.[4]

The many past examples related to the malfunctioning of automated and earlier "intelligent" machines are reminders of what may come. For example, in March 2019, the entire fleet of Boeing MAX-737 was grounded following the crashes linked to the failure of one of its auto pilot related systems, resulting in an estimated \$18.6 billion financial loss[5]; Uber self-driving cars were involved in 37 accidents until 2018 when an autonomous Uber caused the death of a pedestrian[6], believed to be the world's first death by a self-driving car. Safety is also a key issue in healthcare. A survey[7] based on data from the US Food and Drug Administration (FDA), showed that between 2000 and 2013 the use of robotic systems in surgery has led to 144 deaths and 1391 patient injuries. The FDA is meanwhile developing regulations and frameworks for approving the marketing of AI and ML medical devices, having already done so for a number of them.[8,9]

Beyond safety, AI can reproduce and massively scale up other risks such as patterns of discrimination, namely being biased against certain groups of people.[10,11] A widely cited example is related to the software COMPAS[12,13], which estimates the risk of a person recommitting a crime and is currently being used to support judges in several US courts. A study[14] found that COMPAS is, on average, more likely to assign a higher risk score to

African-Americans. Gender bias in natural language processing[15] or in hiring[16], or skin shade bias in facial recognition systems data[17] are further examples of discrimination risks related to AI. In the future, regulatory as well as reputation risks and market (customer) acceptance and trust reasons may drive organizations to manage all these risks whenever they deploy AI enabled services and products.

Privacy has been another, earlier considered, issue with data driven technologies such as AI, as also reflected by the adoption of regulations such as the General Data Protection Regulation[18] (GDPR). While often associated with digital native companies such as online media, ecommerce firms, or search engines, privacy issues are pervasive across multiple sectors. For example, in the energy sector it is estimated that in 2020 there were around 150 million smart meters installed worldwide.[19] These collect energy consumption data from consumers, and in total could globally gather more than 5 trillion data records – equivalent to 420 petabytes - each year.[20] This raises privacy, discrimination, or even safety issues touching everyday lives of people in their homes. For example, the roll-out of smart meters had been delayed in the Netherlands because of privacy issues such as the possibility of predicting consumption patterns related to religion during Ramadan.[21]

As far as the regulation of AI and data privacy is concerned the European Union (EU) seems to be among the most advanced regions. In 2016, the EU has issued GDPR[18], considered as an important step towards regulating data privacy and usage. Article 83 of GDPR foresees strict enforcement rules and significant fines (as a function of turnover, it can be in the order of tens of millions of dollars to enforce compliance). GDPR will probably also be leveraged to regulate privacy issues of AI in the upcoming EU regulations. Several other countries have followed the EU's example on regulating data privacy. Brazil, India, South-Africa, Japan, South-Korea, New-Zealand, Chile, Thailand and some US states including California have adopted or are currently finalising legislations[22] along the lines of GDPR.

Although no country or region has adopted legislation for the moment to regulate the use of AI, several of them are actively considering AI regulations.[23] For example, based on the guidelines of the European Commission's High-Level Expert Group on AI[24], and as a first step to open public discussions on how to regulate these technologies, the European Commission has issued its White Paper on AI in February 2020.[24] The White Paper suggests a risk based approach towards AI regulation, identifying high-risk sectors and high-risk activities, where each obligation should be addressed to the actor who is best placed along the AI system's lifecycle (see Appendix II) to deal with the potential risks. As a next step, in October 2020, the European Parliament has adopted its recommendations[25] for the European Commission concerning the future regulations on AI. Accordingly, operators of high-risk AI systems would be financially liable for any damage and may need to be covered by mandatory insurance.

Privacy related regulations such as GDPR have already impacted businesses[26] and society, both in terms of (i) developing new processes, costly systems and teams, and in terms of (ii) impacting business decisions such as limiting choices of products or markets, which have, consequently, also influenced end consumers' available choices. For example, a number of US firms limited access to their services for European citizens for some time as the result of GDPR. While GDPR and related regulations focused on specific data issues, mainly privacy, current regulatory discussions focus on a number of other risks of AI algorithms with important implications for society and businesses.

Naturally, the development of these regulations relies on a clear and common understanding of what they aim to achieve. Regulators, academics and several companies have already formulated principles and values that help better understand and manage new AI risks and drive the development of related regulations and business practices. For now, value driven risk mitigation is considered one effective approach for companies to better align their AI strategies with responsible practices.[27,28] However, while identifying – and agreeing on – the principles and values AI systems need to adhere to across geographies and cultures has already been challenging. Implementing them in practice will likely prove to be significantly more so, impacting businesses and consumers for example in terms of costs, choices and new limitations, as the GDPR experience already illustrates.

While most organizations have developed well-established sector specific procedures to mitigate traditional risks[29,30], managing AI related risks will likely require new capabilities, processes and tools. The resources necessary to identify and manage these new risks may significantly increase in most organizations, and the ability to assess and manage AI related risks may also become a new source of competitive advantage.[31] However, while regulations and our understanding of these new risks evolve, there are still significant gaps[32] in the capabilities of organizations to implement AI principles and align with emerging regulations as well as social and market expectations.

The goal of this article is to provide an overview of the current state of resources and tools available to organizations in order to adhere to upcoming AI regulations and implement AI principles in practice. We start with reviewing the current landscape of AI principles, which can provide guidance in identifying, describing and linking AI related risks to core values and upcoming regulations. We then review a number of tools and frameworks, that complement existing methodologies for managing risks, focusing on implementing these AI principles and managing the related risks. Finally we discuss possible gaps as well as potential new service offerings to support responsible AI development and adoption by organizations while managing the new AI risks and adhering to upcoming regulations.

2. AI Principles: towards AI Regulations

Technology can fundamentally transform businesses, societies and our everyday lives and behavior, but can it also influence our core values? Some argue[33] that the arising ethical issues due to technologies such as AI are just a variation of existing ethical problems and the moral landscape itself has not changed because of behaviors made possible by these technologies. Others argue[34] that while technology shapes human experiences it is also, to some extent, reflecting and reciprocally affecting our human values, which is also one of the reasons why values have to be taken into account all along the technological design process. In any case, data intensive technologies and especially AI present new challenges and raise ethical issues and risks because of their transformative capabilities, scale, and impact on various societal[34,35] aspects. Consequently, a wide range of stakeholders are involved in developing principles that should guide the development and use of AI to maximize its value while minimizing its risks.

The number of AI principles, initiatives or guidelines has increased significantly in recent years. A broad range of stakeholders have participated in their development including international organizations, governments, but also several major private companies and academic institutions (see Appendix I). Since 2016 several national governments have developed[36] their AI national strategies, including their principles, such as the USA, South-Korea, France, Japan, Canada, Singapore, China, the UAE, the UK, Mexico. In 2019, the

OECD, a leading organization in this space, published its "AI Principles"[37], which were later adopted by the G20 countries.[38] The same year, the European Commission's High-Level Expert Group on AI published Guidelines for a trustworthy AI.[39] Other international organizations have been also working on these issues, such as the Council of Europe[40], UNESCO[41], or the UN (Secretary General's High-level Panel on Digital Cooperation[42]). Several major companies (e.g., Microsoft[43], Google[44], Salesforce[45], IBM[46], Intel[47], SAP[48]) have also published their AI principles, while academia and professional organizations have meanwhile significantly contributed to the framing of ethical issues (e.g., Future of Life Institute[49], IEEE[50,51], AI4People[52], etc.).

Based on the summary of 84 published documents regarding AI principles, Jobin et al.[34] indicate that both public bodies and private enterprises are equally active in shaping the field of AI principles, although contributions are provided mainly from economically developed countries. While mainly the same set of principles appears in most documents, there is also a substantive divergence on how AI principles should be interpreted, prioritized or implemented. Among all principles, transparency, justice and fairness, non-maleficence, responsibility and privacy are the 5 most often appearing ones in the published documents (see Table I.1 in Appendix I).

Floridi et al. also find[53] convergence to a few AI principles such as beneficence, non-maleficence, autonomy and justice, which are likewise commonly used in bioethics. Similarly, Mittelstadt[54] compares AI ethics principles with principles from medical ethics, but he points to several limitations in the AI field. These include the lack of common aims of parties developing AI systems, missing professional history and norms, the absence of proven methods to translate principles into practice, and the non-existence of legal and professional accountability mechanisms. Consequently, Mittelstadt argues that shared principles may have just limited impact on the design and governance of AI systems and are not sufficient on their own to guarantee trustworthy or ethical AI.

Related to the principle of responsibility, AI accountability is emerging as a key objective in order to ensure all stakeholders involved, from technology developers to deployers and users[55], contribute to a safe and responsible adoption and usage of these technologies. It is therefore expected that some form of regulation to foster accountability will be implemented to mitigate risks of AI systems. Algorithmic accountability provides means to ensure liability for any negative implications related to the use of AI.[56] It also implies an obligation to report or justify algorithmic decision-making as well to mitigate any negative social impacts or potential harms[57].

As principles are getting developed, implementing them in practice and managing businesses in the emerging regulatory landscape will require new capabilities and tools in order to find the right balance between risk management and value creation. Indeed, according to a recent study[58], while 62% of executives believe that AI technologies should be regulated, 57% of AI adopters have "major" or "extreme" concerns about how new and changing regulations could impact their AI initiatives by hampering research, innovation and competitive advantage. The discussion is only now shifting from developing AI principles and regulations, to how to best implement these and how organizations can best adhere to the upcoming regulations in practice. Doing so requires new frameworks, processes and tools, which we discuss next.

3. Frameworks, Processes and Tools to Implement AI Principles and Regulations

Various classifications of existing mechanisms supporting the verification and development of AI systems have been proposed in the past. For example, Brundage et al.[27] consider three different types of such mechanisms: institutional mechanisms related to incentives for allowing verifiable claims vis-à-vis AI systems, including third party auditing, considered also as an adequate incentive and a good example for providing external feedback; software mechanisms focusing on specific aspects of AI systems such as audit trails to collect information about the development and deployment, interpretability and privacy, and hardware mechanisms related to issues such as secure hardware or computing power. An implication of this classification is that while technical tools are important and widely used, a broader approach towards verification, including for instance institutional mechanisms, can provide additional benefits.

Similarly, Morley et al. [35] align AI principles identified by the European Commission's High-Level Expert Group – beneficence, non-maleficence, autonomy, justice, explicability – with each stage of the AI Lifecycle (Appendix II) – business and use-case development, design phase, training and test data procurement, building, testing, deployment, and monitoring. This was achieved through translating each principle into specific system requirements. However, the availability of tools and methods is not evenly distributed, neither in terms of the ethical principles nor in terms of the stages of development along the AI Lifecycle.

While there are other ways to organise the various tools, frameworks, and processes that can support organizations in adhering to regulations and implementing AI principles – which are also likely to evolve over time – in this paper we organize them into four broad categories: (i) methodologies and toolkits mainly focusing on implementing specific principles, (ii) documentation procedures, (iii) AI auditing processes, and (iv) standards and certifications. We discuss these separately next.

3.1 Methodologies and toolkits to implement principles in practice

As noted above, a key goal is to ensure that AI principles are considered and implemented at each step of the development and usage of AI systems, namely throughout the AI Lifecycle (Appendix II). AI needs to earn the trust of several stakeholders interacting with it and demonstrate responsible "behaviour" in verifiable ways throughout its lifecycle.[54] To achieve this, an iterative process has to be in place, involving all relevant stakeholders at all stages and allowing continuous adjustments as needed.[59] Reflecting ethical implications all through the design, development and implementation processes, also implies significant changes in the general design practices of algorithmic systems in order to embed AI principles.[54,59]

Accordingly, the Value Sensitive Design framework developed by Friedman et al.[60] takes values into consideration throughout the whole design process. A number of practical methodologies have been developed[61–63] based on this framework, privacy by design[64–66] being one of those. In this case, privacy is embedded into the design and architecture of organisational information management systems as well as business practices. As a result privacy becomes a fundamental part of the core functionality delivered throughout the whole lifecycle of the data involved.[64] The concept of privacy by design has also become an integral part of GDPR related implementations.

Beyond privacy, another area in which there is already some progress in terms of methodologies and toolkits is around AI fairness. For example, Madaio et al.[67] designed a checklist-based methodology about AI fairness, built on design processes that rely on previous checklists methodologies.[68] Checklists are constructed based on interviews and workshops defining elements to consider along the AI Lifecycle. Each stage contains between 6 and 14 items, such as “Envision system purpose and scrutinize for potential fairness issues,” “Define and scrutinize datasets for potential fairness issues,” “Define fairness criteria,” or “Assess fairness criteria.” The authors found that practitioners consider checklists to be beneficial both for formalising already existing but still ad-hoc processes and empowering individual advocates. Moreover, these goals are best achieved when aligned with teams' existing workflows and with the overall organizational culture.

Software toolkits, including open source ones, have also been developed to support AI fairness and manage other AI risks. Such toolkits will be essential both to demonstrate compliance with future regulations and to foster voluntary compliance, even if several questions remain concerning how to efficiently implement non-binding principles in practice.[28,35,54,69,70] Some example tools are shown in Table 1. We briefly discuss some next.

Table 1: Key Software Toolkits and Frameworks for Implementing AI Principles

Toolkit	Developer
Fairness Tool[71]	Accenture
Foolbox[72]	Bethge Lab
CleverHans[73]	CleverHans Lab
Model Guardian[74]	Deloitte
Digital Impact Toolkit[75]	Digital Civil Society Lab, Stanford Center on Philanthropy and Civil Society
Deon[76]	Driven Data
Fairness Flow[77]	Facebook
What-If Tool[78]	Google
Ethics & Algorithms Toolkit[79]	GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC
AI Fairness 360[80,81]	IBM
AI Explainability 360[82]	IBM
Adversarial Robustness Toolbox[83] (ART)	IBM
LinkedIn Fairness Toolkit[84] (LiFT)	LinkedIn
Fairlearn[85]	Microsoft
InterpretML[86]	Microsoft
Harms Modelling[87]	Microsoft
Community Jury[88]	Microsoft
Skater[89]	Oracle
REVISE: REvealing VIsual biaSEs[90]	Princeton University
Responsible AI Toolkit[91]	PwC
audit-AI[92]	Pymetrics
FAT Forensics[93]	University of Bristol
Aequitas[94]	University of Chicago Center for Data Science and Public Policy
Lime[95]	University of Washington

For example, IBM has mainly focused on 5 different areas,[80] explainability, fairness, robustness, transparency and privacy, supporting each one with a specific tool. An example toolkit is IBM's AI Fairness 360[88], which is an open-source toolkit that helps discover biases in datasets and machine learning models. It is an extensible architecture that incorporates dataset representations and algorithms for bias detection, bias mitigation, and bias metric explainability.

Another example of a toolkit to help implement AI principles is the Revealing Visual Biases (REVISE) one,[35] which investigates visual datasets and their annotations to determine model-agnostic patterns. It analyses object-based biases (size, context, or diversity of object representation), gender-based biases (stereotypical representation of genders) and geography based biases (representation of different geographic locations). An example of bias detected by this tool on the OpenImages dataset relates to gender bias. In this case it was found that images of people can be too small for human annotators to determine their gender, however annotators suppose they are male in 69% of the cases, especially in scenes of outdoor sports fields and parks. The toolkit provides further help to suggest specific action items, but once bias is detected it is the user's responsibility to determine whether bias is problematic based on the actual context.

As also noted by Morley et al.[24] there is still uneven progress in terms of methodologies and toolkits to support the implementation of AI principles. However, the ones mentioned in this section provide early indications of the directions such toolkits may evolve towards in the future. Meanwhile, beyond implementation methodologies and toolkits, another important aspect, related also to auditing and certification discussed later, is that of effective documentation. We turn to this topic next.

3.2 Documentation Tools and Processes

Provision of detailed documentation of AI systems is considered to be critical for accountability and the successful implementation of AI principles. Documentation procedures provide structured information and can also make algorithms and their development more auditable at the AI dataset or AI model levels. They identify and anticipate risks before deployment along several phases of the AI Lifecycle. Requirements for documentation (e.g., including a description of the main characteristics and how the data set was selected) and provision of information (e.g., concerning the AI system's capabilities and limitation) are also likely to be important elements in upcoming regulations, as the EU Commission's White Paper on AI also indicates.[96] Documentation has also been considered[50] as part of the "right to understanding" principle to combat biases in machine learning, while the IEEE recommends[15,97–101] the use of documentation procedures for businesses also as part of creating a culture of ethics.

Although there are currently no generally accepted standardized AI documentation procedures, several[102] have been developed recently. Some AI documentation procedures also rely on tools previously established in other industries such as electronics, food, telecommunication or transportation. They are also considered important for presenting metadata for ML models in a standardized way.[99] We discuss some recent ones related to AI datasets, and AI models – see Table 2, more details provided in Appendix III.

At the dataset level *Datasheets for Datasets*[15] are similar in spirit to tools used in the electronics industry, where every component is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, etc. Datasheets for ML datasets contain questions that support self-reflection for dataset creators and help to collect information along the dataset lifecycle to assist consumers of the data. *Data statements*[98] provide information and context about the datasets and their represented population especially in the settings of Natural Language Processing systems. For example, they help to fill the gaps in mitigating exclusion and bias in language technologies. The *Dataset Nutrition Label*[101] is another tool, inspired by the food industry and also built on experience from online privacy and algorithmic accountability. Like the other tools it also aims to foster more robust ML training datasets, but also to help improve data collection practices more generally. Dataset Nutrition Labels are modular thus allowing for bigger flexibility as each of its modules can be used for different types of datasets.

Model cards [100,103] are documentation procedures complementing the data ones, this time focusing at the AI model level. Their intended use is to accompany trained ML models detailing their performance characteristics, such as how the model was built, what assumptions were made during its development, or what type of model behaviour may be experienced by different cultural, demographic, or other population groups. Model reporting can provide information for most stakeholders along the AI Lifecycle, and can be valuable both for internal development purposes and for external third-party audits.

FactSheets[97] also provide information about how an AI model but also AI service was developed and deployed. They capture model or service facts about performance and reliability, safety, security and lineage across the entire AI Lifecycle. FactSheets are based on the idea of supplier's declaration of conformity, which is used in several industries including telecommunications and transportation. Since there is often an expertise gap between the producer and the consumer of an AI service it is important to communicate attributes in a standardised but flexible way. FactSheets can be adjusted to the specific AI model or service, and to the demands of a target audience or consumer, hence can differ in content and format.

Table 2: Key Documentation Processes and Tools

Name	Level of Documentation	Related earlier tools
Datasheets for datasets	Dataset level	Datasheets in electronics industry
Data statements for Natural Language Processing (NLP)	Dataset level	Practices from the fields of psychology and medicine
Dataset Nutrition Label	Dataset level	"Nutrition Facts" label from the food industry, "Nutrition Label" for Privacy[104] and "Nutritional Label" for Rankings[105]
Model cards for model reporting	AI models	Transparent Reporting of a prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement[88] in biomedical science
Factsheets	AI models	Supplier's declaration of conformity (SDoC) used in different industries including telecommunications, transportation

While documentation can play an important role in supporting transparency, accountability, and risk management, it is only part of the solution in ensuring that AI principles are rigorously implemented in practice. For example, it is often not possible to foresee biases hiding in data or models, and manual reviews are certainly not a feasible strategy given the scale of modern datasets.[106] Therefore further due diligence processes are necessary, auditing being a key one to which we turn next.

3.3 Auditing of AI systems

Auditing (e.g., of systems, processes or organizations) is used in many industries such as in finance, air travel, or software development. The IEEE standards for software development define an audit as "an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures".[27,107–109] Recently, auditing has been increasingly considered for assessing[110] whether AI development was performed in a way consistent with the affirmed principles of an organization. For example, Goodman[57] argues that GDPR already anticipated third party inspections of algorithms or 'algorithm audits' through distinct instruments such as data impact assessment, code of conducts and certification. However at the moment there exists no standard procedure for how to perform an AI audit – which are, however, under development. Moreover, since ML models can continuously evolve during usage through learning from data, their limitations may not be immediately clear at the onset, thus, as also argued before[111], ensuring accountability may require repeated auditing as well as continuous risk mitigation.

Various organizations are currently developing AI auditing frameworks. For example, the Information Commissioners' Office (ICO) in the UK has proposed a framework (Figure 1) exclusively to assess the challenges introduced or increased by the adoption of AI. A key component of that framework, the governance and accountability one, describes the measures an organisation must have in place to be compliant with data protection requirements. A second component focuses on the possible data protection risks that may occur in a number of AI specific fields and the adequate risk management practices to manage them.

Other auditing methodologies[108] ensure that externally, those who may consider an opposing view as to whether or not an AI system in use is safe and ethically-aligned, have a mechanism for questioning the rational of design decisions and requesting their change if necessary. Among proposed approaches, Raji et al. present[112] an "algorithmic audits mechanism to check that the development processes of AI system and their deployment meet declared ethical expectations and standards, such as organizational AI principles." Their approach is based on the framing of risk analyses centred on the failure to achieve AI principles objectives, outlining an internal audit practice that can help translate ethical principles into practice, prior to model deployment (Figure 2).

Another example framework is that of Mahajan et al., who present[24] an algorithmic audit framework to test and improve the performance of algorithms supporting the work of radiologists. The framework includes concepts of independent validation on data that the algorithm has not processed before, curating datasets for such testing, examination of false positives and false negatives and real-world deployment and testing of algorithms.

Figure 1: The ICO auditing framework to mitigate data protection risks related to AI

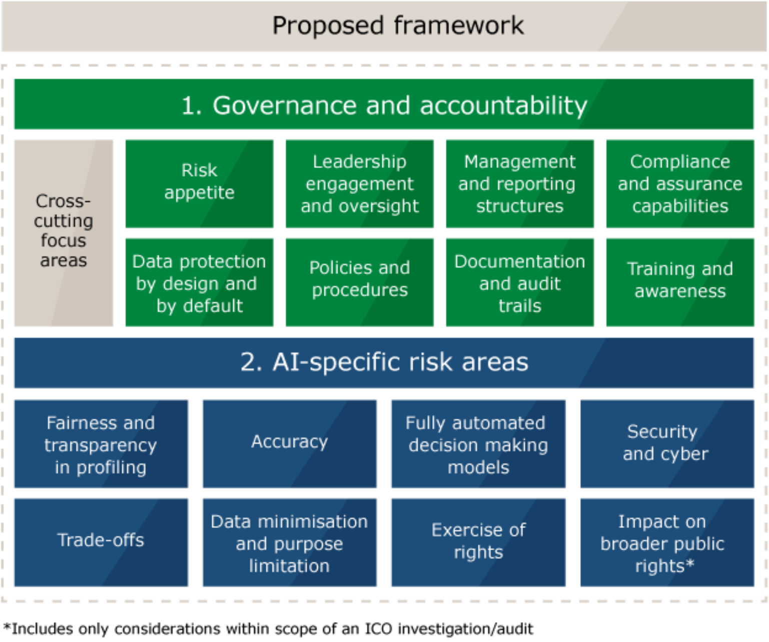
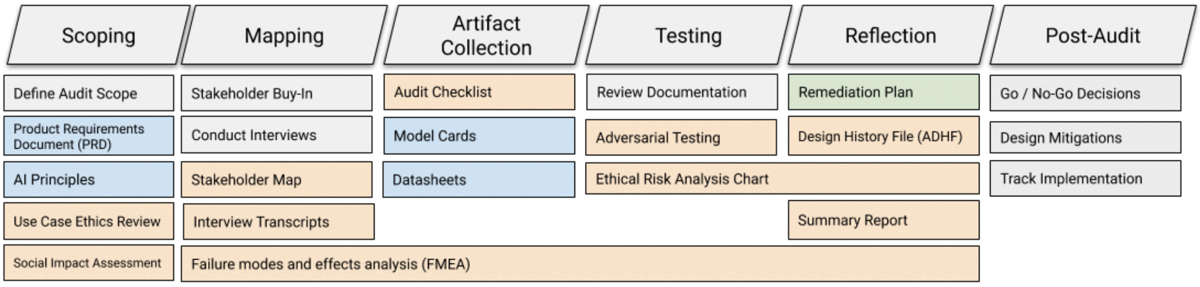

Proposed framework
1. Governance and accountability
Cross-cutting focus areas
Risk appetite
Leadership engagement and oversight
Management and reporting structures
Compliance and assurance capabilities
Data protection by design and by default
Policies and procedures
Documentation and audit trails
Training and awareness
2. AI-specific risk areas
Fairness and transparency in profiling
Accuracy
Fully automated decision making models
Security and cyber
Trade-offs
Data minimisation and purpose limitation
Exercise of rights
Impact on broader public rights*
*Includes only considerations within scope of an ICO investigation/audit

Figure 2: Overview of Raji et al. Internal Audit Framework. Gray indicates a process, and the coloured sections represent documents. Documents in orange are produced by the auditors, blue documents are produced by the engineering and product teams and green outputs are jointly developed.



Meanwhile, while auditing frameworks are getting developed, international organizations are working on developing standards that can support the development of AI systems in ways that are relatively uniform across businesses and markets, making auditing as well as certification – another ”tool” to support the deployment of risk managed AI systems that adhere to regulations to which we turn to next – easier to implement.

3.4 Standards and Certifications

Policy makers consider AI certification as a key component in ensuring responsible and well risk-managed AI systems. For example, as part of a prior conformity assessment, the European Commission’s White Paper on AI[113] considers including procedures for testing, inspection or certification to verify compliance with specific mandatory requirements applicable to high-risk AI applications. This could involve checks of the algorithms and of the data sets used in the AI development phase. The White Paper also emphasises the evolving nature of AI systems, which may necessitate repeated assessments over the lifetime of the system in question. Most of current certification processes are based on the notion that the behaviour of a system must be entirely specified and verified prior to operation. However adaptive intelligent systems such as AI can constantly alter their behaviour, which may not fit naturally into the context of current certification procedures.[114]

The ISO refers[115] to certification as the provision of written assurance by an independent body that a product, service or system meets specific requirements. The certification process relies on standards that provide guidance on proving compliance. Different standardisation organisations provide generic standards valid across many system domains.[51] Some of the most well-known organisations issuing standards are the International Organization for Standardization (ISO), the Institute of Electrical and Electronics Engineers (IEEE), the International Electrotechnical Commission (IEC) and the European Committee for Electrotechnical Standardization (CENELEC). The concept of Ethically Aligned Design[116], elaborated by members of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, refers to the involvement of ethics into the design of AI and autonomous systems. This has resulted[117] in different standardization projects (the IEEE P7000 series) which cover a broad range of features of autonomous and intelligent systems. In 2018, the IEEE has also launched its Certification Program for Autonomous and Intelligent Systems to create specifications for certification to foster transparency, accountability and reduction of algorithmic bias.[108]

An idea gaining some traction is that because certifying AI systems may be challenging, doing so for organizations developing these systems may be an alternative. For example, the proposed ISO 37000 standard considers “the system by which the whole organization is directed, controlled and held accountable to achieve its core purpose over the long term”. As argued before, if the responsible development of AI is a core purpose of an organization, then a governance system by which the whole organization is held accountable should be established.[9,118] Similarly, in the context of software as a medical device, the US FDA is planning to assess the culture of quality and organizational excellence of a particular company.[118]

In each industry, fundamentally new and sector specific regulatory approach might be necessary. For example, to support AI/ML based software as a medical device, that learn and adapt over time to improve patient care, the FDA has proposed[119] a total product lifecycle approach. This approach aims to facilitate a rapid cycle of product improvement and promote a mechanism for medical device manufacturers to continuously maintain the safety and effectiveness of their AI embedded products and services. In the autonomous vehicle sector, Schmid is proposing[120] a model-based safety approach, which provides a systematic process to analyse interactive effects and identify unsafe control and demonstrates how this can be implemented through certification.

Certification of data driven, continuously evolving, complex systems based on technologies such as AI is arguably difficult. The challenges of data driven technologies and AI certification are already well demonstrated by GDPR, which foresees the possibility of certification, although for the moment there is no dedicated standard available. For example, the ISO/IEC 27701 standard for privacy information management provides guidance to support compliance with GDPR but does not cover its full spectrum. For instance it is difficult to create a global standard for data portability due to divergent cultures of law across countries. It is thus unclear, whether there will ever be a fully compliant and globally adopted certification with GDPR.[28] The development of certification methodologies for AI systems is a complex issue, even more so in light of the certification issues related to GDPR.

4. Discussion

While a lot of progress has been made to support the development of AI systems that are properly risk-managed and adhere to regulations and principles, a lot of work still remains to be done. To begin with AI principles and the goals of AI regulations, let alone the regulations themselves, are still under development – and may take a few years before many of the regulations are put in practice. Meanwhile, there are already many gaps in the portfolio of frameworks, processes, and tools to implement AI principles. As argued before[9,118], some of the key reasons for these gaps are the high complexity of the AI systems, the high number of stakeholders affected, multiple disciplines along the design process, the abundance of available tools and the functional separation of technical (e.g., software engineers) and non-technical experts (e.g., C-levels, Corporate Social Responsibility staff) in organizations. These can also limit the potential to communicate effectively, understand issues robustly, and may eventually lead to holes in responsibility and accountability. Building an organizational culture focusing on AI principles will likely also prove critical, as it is unlikely that any set of tools or frameworks (or "ticking a box to comply with regulation") will be enough to ensure AI risks are properly considered and managed inside – and outside – organizations. In addition to ensuring an appropriate organizational culture, adopting good practices throughout the organization, such as Good Machine Learning Practices – as, for example, also considered by regulators such as the US FDA[24] – will be needed, and possibly even required by regulators as the FDA example indicates.

Implementation of AI principles and adherence to upcoming regulations therefore requires for organizations to overcome multiple challenges. These create opportunities for new tools, new research, as well as for new services and business models to be developed. We close by discussing some possibilities.

4.1 Potential Tools and Research to Develop: The importance of Monitoring

While there is a number of existing tools and frameworks focusing on the development and deployment phases of the AI Lifecycle, discussed above, more focus may be needed on the monitoring phase of the lifecycle (Appendix II) – namely during usage. As also noted in the EU White paper on AI[121], "particular account should be taken of the possibility that certain AI systems evolve and learn from experience, which may require repeated assessments over the life-time of the AI systems in question". Continuous monitoring may prove necessary to ensure that divergence between the expected and actual behaviour of a system is captured early and promptly, and addressed adequately.[122]

There may be multiple ways to achieve this. For example Taddeo et al.[123] proposes that, users or providers of AI systems should maintain cloned systems as control systems, which is in fact different from a 'digital twin'.[122] The clone would be the same system as the deployed one in controlled environmental conditions and would be the benchmark against which the behaviour of the original system is assessed.

Similarly, standards and certification procedures focusing on the robustness of AI systems will be effective only insofar as they will take into account the dynamic and self-learning nature of AI systems, and start envisaging forms of monitoring and control that span from the design to the development and usage stages. This point has also been stressed in the OECD principles on AI, which refer explicitly to the need for continuous monitoring and assessment of threats for AI systems.[124]

Beyond the monitoring, other types of more specific tools and technologies may need to be developed. Kroll et al. considers[125] for instance cryptographic commitments and zero-knowledge proofs. Cryptographic commitments can be understood as the digital equivalent of a sealed document, similarly to an envelope that can provide certainty that an object was not changed. These can be important tools for automated systems to ensure for instance that the same decision policy was used for each of many decisions. In relation to cryptographic commitments zero-knowledge proofs provide certainty about the property of a specific policy decision without having to reveal the content of how that property is known or what the decision policy actually is. A related challenge is that of traceability, for example of the various components possibly used in an AI system, including traceability of the data used to train AI models. In a sense "data supply chain" traceability frameworks and tools may need to be developed, similar in spirit to what is used for non-digital products (e.g., agricultural, food and beverages, etc.).

Of course a related question is who bears the burden for the risks and for the costs of their management. Some of the frameworks, processes, and tools may require special skillsets, while regulation may also require the involvement of third parties. As the space matures, new services and businesses may need to be built and offered to assist organizations implement their AI principles and, as important, adhere to upcoming regulations in ways that regulators can also supervise and confirm. We briefly discuss these next.

4.2 Potential Services and Business Models

The European regulation for privacy has already led to a number of new business offerings. Indeed, based on the Global GDPR Services Market Research Report[125] the market for global GDPR services is expected to reach USD 4 billion by 2025, resulting in a CAGR of 23.4% during the forecast period, 2019–2025. Examples of possible solutions and service offerings related to GDPR include data management, data discovery and mapping, data governance, API management, GDPR readiness assessments, data protection and risk assessments, data officer-as-a-service, trainings and certification. Main players providing GDPR related solutions and services include among others[126] IBM (US), AWS (US), Micro Focus (UK), Veritas (US), Capgemini (France), Microsoft (US), Absolute Software (Canada), Mimecast (UK), Informatica (US), Iron Mountain (US), Proofpoint (US), Oracle (US), and Trustwave (US). Top tier management consulting firms (e.g., The Boston Consulting Group, McKinsey, the Big Four, etc.) also provide service offerings in the GDPR space, sometimes in coordination with the above mentioned solution providers. Various start-ups have also developed new offerings[127] on the privacy regulation related market such as InCountry, OneTrust, TrustArc, Privitar, BigID, and others.

Like for GDPR, new business offerings will likely also emerge to support businesses to adhere to upcoming AI regulations and in general manage their AI related risks. Not only regulatory reasons and related sanctions will require companies to do so, but investors and stakeholders – e.g., customers – will likely also push companies towards this direction. Indeed, one can argue that the value creation potential of AI embedded products and services, much like for any other product and service, is closely related to the ability of businesses to manage the associated risks.

Various business models may emerge. For example, new consulting services, as also indicated by current work of large consulting firms, will likely be developed (see Table 3). Software and other tools may also be developed and marketed, possibly as a service (e.g., SaaS business models). Auditing and certification based offerings and business models may be created, especially as regulators may gear towards requiring some form of third party involvement for these. Finally, possibly depending on the level of risk, insurance based schemes may also be needed. Indeed, given the complexity of AI systems and other aforementioned reasons, it may prove impossible to fully manage risks for any individual organization adopting or developing AI solutions or AI embedded products and services. Some form of risk pooling, justifying the development of insurance type products, may eventually also be needed.

Table 4: Example business offerings developed by major consulting firms

Name of Issuer	**AI Principles**	**Description**
EY	Trusted AI Platform[128]	The platform provides an integrated approach to evaluate, quantify and monitor the impact and trustworthiness of artificial intelligence
KPMG	KPMG Ignite[129]	Ignite is a portfolio that includes methods, tools, approaches and resources that focus on improving the consistency, efficiency and time to make decisions and take action
PWC	Responsible AI Toolkit[107]	Customizable frameworks, tools and processes designed to help the use of AI in an ethical and responsible manner, from strategy through execution
BCG	Different tools to support fast scaling of AI solutions[103]	A set of different tools and frameworks, for instance: Data and Digital platform, Build-Operate-Transfer methodology, High-frequency data and analytics platform (Lighthouse), etc.

As our understanding of AI, related regulations, and tools, frameworks, and processes are developed, it is clear given the current status and complexity of the issues that many opportunities and challenges remain. In this paper we provided an overview of the various aspects currently under development or that may need to be developed. However, much like the ability of AI and ML algorithms to evolve through learning, the issues we discussed will also be continuously evolving.

Appendix I: AI Principles

1. AI principles published by major companies

#	Name of Issuer	AI Principles	Source	Accessed
1	Microsoft	(1) Fairness, (2) Reliability & Safety, (3) Privacy & Security, (4) Inclusiveness, (5) Transparency, (6) Accountability	https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6	06 November 2020
2	Google	(1) Socially beneficial, (2) Avoiding unfair bias, (3) Built and tested for safety, (4) Accountable to people, (5) Incorporating privacy design principles, (6) Scientific excellence, (7) Uses of the technologies are in accordance with these principles	https://ai.google/principles/	06 November 2020
3	Salesforce	(1) Being of benefit, (2) Aligns with human values, (3) Open debate between AI researchers and policymakers, (4) Cooperation, trust and transparency in systems and among the AI community, (5) Safety and Responsibility	https://www.salesforce.org/blog/ai-good-principles-believe/	06 November 2020
4	IBM	(1) Explainability, (2) Fairness, (3) Robustness, (4) Transparency, (5) Privacy	https://www.ibm.com/artificial-intelligence/ai-ethics-focus-areas	06 November 2020
5	Intel	(1) Foster Innovation and Open Development, (2) Create New Human Employment Opportunities and Protect People's Welfare, (3) Liberate Data Responsibly, (4) Rethink Privacy, (5) Require Accountability for Ethical Design and Implementation	https://blogs.intel.com/policy/2017/10/18/naveen-rao-announces-intel-ai-public-policy/	06 November 2020
6	SAP	(1) Driven by their values, (2) Designed for people, (3) Enables businesses beyond bias (mitigating biases), (4) Transparency and integrity, (5) Quality and safety, (6) Data protection and privacy, (7) Engage with the wider societal challenges of AI	https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/	06 November 2020
7	Tencent	(1) Available, (2) Reliable, (3) Comprehensible, (4) Controllable	https://www.tisi.org/13747	06 November 2020
8	Sony	(1) Supporting Creative Life Styles and Building a Better Society, (2) Stakeholder Engagement, (3) Provision of Trusted Products and Services, (4) Privacy Protection, (5) Respect for Fairness, (6) Pursuit of Transparency, (7) The Evolution of AI and Ongoing Education	https://www.sony.net/SonyInfo/sony_ai/responsible_ai.html	06 November 2020
9	Workday	(1) We Put People First, (2) We Care about	https://blog.workday.	06 November 2020

		Our Society, (3) We Act Fairly and Respect the Law, (4) We Are Transparent and Accountable, (5) We Protect Data, (6) We Deliver Enterprise-Ready ML Technologies	com/en-us/2019/workdays-commitments-to-ethical-ai.html	
10	Sage	(1) AI should reflect the diversity of the users it serves, (2) AI must be held to account—and so must users, (3) Reward AI for 'showing its workings', (4) AI should level the playing field , (5) AI will replace, but it must also create	https://www.sage.com/~/media/group/files/business-builders/business-builders-ethics-of-code.pdf	06 November 2020
11	Philips	(1) Well-being, (2) Oversight, (3) Robustness, (4) Fairness, (5) Transparency	https://www.philips.com/a-w/about/artificial-intelligence/philips-ai-principles.html	06 November 2020
12	Facebook	(1) Openness, (2) Collaboration, (3) Excellence, (4) Scale, (5)	https://ai.facebook.com/research#fundamental-and-applied	06 November 2020
13	Tieto	(1) Responsibility, (2) Human rights, (3) Fairness & equality, (4) Safety & security, (5) Transparency	https://www.tietoevry.com/contentassets/b097de43d84d4c84832f1fff2cb6a30d/tieto-s-ai-ethics-guidelines.pdf	10 November 2020
14	OP Group	(1) People-first approach, (2) Transparency and openness, (3) Impact evaluation, (4) Ownership, (5) Privacy protection	https://www.op.fi/op-financial-group/corporate-social-responsibility/commitments-and-principles	10 November 2020
15	Deutsche Telekom	(1) Responsible, (2) Careful, (3) Careful, (4) Transparent, (5) Secure, (6) Reliable, (7) Trustworthy, (8) Cooperative, (9) Illustrative	https://www.telekom.com/en/company/digital-responsibility/details/artificial-intelligence-ai-guideline-524366	10 November 2020
16	Telefónica	(1) Fair, (2) Transparent and explainable, (3) Human-centric, (4) Respect people's right to privacy and their personal data	https://www.telefonica.com/en/web/responsible-business/our-commitments/ai-principles	10 November 2020
17	DeepMind Ethics & Society	(1) Social purpose, (2) Privacy, (3) Transparency, (4) Fairness, (5) Accountability	https://deepmind.com/about/ethics-and-society	10 November 2020
18	PricewaterhouseCoopers UK	(1) Governance, (2) Interpretability & Explainability, (3) Bias & Fairness, (4) Robustness & Security, (5) Ethics & Regulation	https://www.pwc.com/gx/en/issues/artificial-intelligence/responsible-ai-placemat.pdf	10 November 2020
19	Accenture UK	(1) Decision making and liability, (2) Transparency, (3) Bias, (4) Human values, (5) Data protection and IP, (6) Social	https://www.accenture.com/gb-en/company-	10 November 2020

		dislocation, (7) Cybersecurity	responsible-ai-robotics	
20	Unity Technologies	(1) Unbiased, (2) Accountable, (3) Fair, (4) Responsible, (5) Honest, (6) Trustworthy	https://blogs.unity3d.com/2018/11/28/introducing-unitys-guiding-principles-for-ethical-ai/	10 November 2020

Table I.1: Ethical principles identified in existing AI guidelines (Source: Jobin et al.)

Ethical principle	**Number of documents**	**Included codes**
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice and fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom and autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

2. AI principles published by research institutes

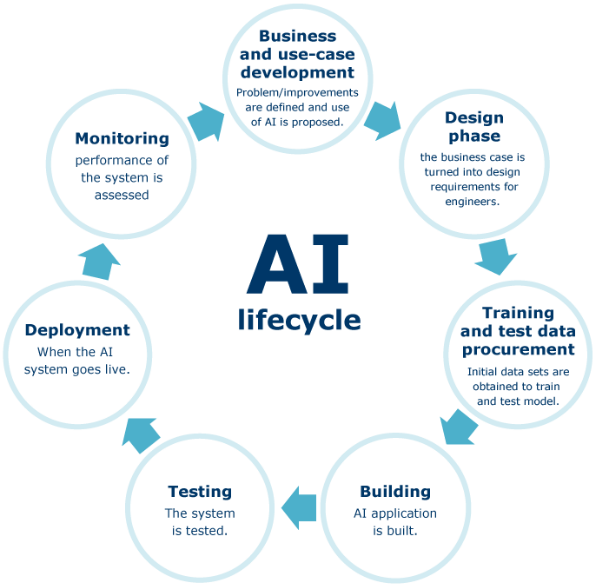
#	Name of Issuer	AI Principles	Source	Accessed
1	AI4People	(1) Beneficence: Promoting Well-Being, Preserving Dignity, and Sustaining the Planet, (2) Non-maleficence: Privacy, Security and "Capability Caution", (3) Autonomy: The Power to Decide (Whether to Decide), (4) Justice: Promoting Prosperity and Preserving Solidarity, (5) Explicability: Enabling the Other Principles Through Intelligibility and Accountability	https://doi.org/10.1007/s11023-018-9482-5	26 November 2020
2	Future of Life	(1) Safety, (2) Failure Transparency, (3) Judicial Transparency, (4) Responsibility, (5) Value Alignment, (6) Human Values, (7) Personal Privacy, (8) Liberty and Privacy, (9) Shared Benefit, (10) Shared Prosperity, (11) Human Control, (12) Non-subversion, (13) AI Arms Race should be avoided	https://futureoflife.org/ai-principles/	26 November 2020
3	IEEE	(1) Human Rights, (2) Well-being, (3) Accountability, (4) Transparency, (5) Extending benefits and minimizing risks of misuse	https://doi.org/10.1007/978-3-030-12524-0_2	26 November 2020
4	Institute for Information and Communications Policy (IICP), The Conference toward AI Network Society	(1) Collaboration, (2) Transparency, (3) Controllability, (4) Safety, (5) Security, (6) Privacy, (7) Ethics, (8) User assistance, (9) Accountability	https://www.soumu.go.jp/main_content/000507517.pdf	25 November 2020
5	Mission Villani	(1) Transparency and auditability, (2) Protection of our rights and freedoms, (3) Legal responsibility for any damages caused, (4) Architects of our digital society act responsibly, (5) Create a diverse and inclusive social forum for discussion	https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf	25 November 2020

3. AI principles published by national and international level institutions

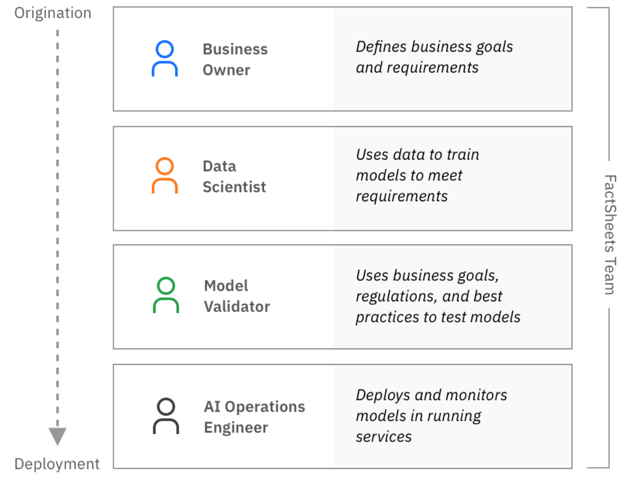
#	Name of Issuer	AI Principles	Source	Accessed
1	Australia, Department of Industry Innovation and Science	(1) Human, social and environmental wellbeing, (2) Human-centred values, (3) Fairness, (4) Privacy protection and security, (5) Reliability and safety, (6) Transparency and explainability, (7) Contestability, (8) Accountability	https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles	25 November 2020
2	Canada, Université de Montréal	(1) Well-being, (2) Respect for autonomy, (3) Protection of privacy and intimacy, (4) Solidarity, (5) Democratic participation, (6) Equity, (7) Diversity inclusion, (8) Prudence, (9) Responsibility, (10) Sustainable development	https://www.montrealdeclaration-responsibleai.com/the-declaration	25 November 2020
3	Finland, Ministry of Economic Affairs and Employment	(1) Transparency, (2) Responsibility, (3) Extensive societal benefits	https://julkaisut.valtioneuvosto.fi/handle/10024/160980	25 November 2020
4	France, French Data Protection Authority (CNIL)	(1) Fairness, (2) Continued attention and vigilance	https://www.cnil.fr/en/how-can-humans-keep-upper-hand-report-ethical-matters-raised-algorithms-and-artificial-intelligence	25 November 2020
5	India, National Institution for Transforming India (NITI Aayog)	(1) Fairness, (2) Transparency	https://niti.gov.in/writereaddata/files/document_publication/NationalStrategy-for-AI-Discussion-Paper.pdf	25 November 2020
6	Japan, Japanese Society for Artificial Intelligence	(1) Contribution to humanity), (2) Abidance of laws and regulations), (3) Respect for the privacy of others, (4) Fairness, (5) Security, (6) Act with integrity, (7) Accountability and Social Responsibility, (8) Communication with society and self-development, (9) Abidance of ethics guidelines by AI	http://ai-elsi.org/wp-content/uploads/2017/05/JSAI-Ethical-Guidelines-1.pdf	25 November 2020
7	Netherlands, Special Interest Group on Artificial Intelligence (SIGAI), ICT Platform Netherlands (IPN)	(1) Socially-aware, (2) Explainable, (3) Responsible	http://ii.tudelft.nl/bnvki/wp-content/uploads/2018/09/Dutch-AI-Manifesto.pdf	25 November 2020
8	Singapore,	(1) Explainable, (2) Transparent, (3) Fair,	https://www.pdpc.go	25 November 2020

	Personal Data Protection Commission Singapore	(4) Beneficence or “Do no harm”	v.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/Discussion-Paper-on-AI-and-PD---050618.pdf	
9	UAE, Smart Dubai	(1) Fair, (2) Accountable, (3) Transparent, (4) Explainable	https://www.smartdubai.ae/docs/default-source/ai-principles-resources/ai-ethics.pdf?sfvrsn=d4184f8d_6	25 November 2020
10	UK, UK House of Lords, Select Committee on Artificial Intelligence	(1) Developed for the common good, (2) Intelligibility and fairness, (3) Should not be used to diminish the data rights or privacy of individuals, families or communities, (4) All citizens have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence, (5) Autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence	https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf	25 November 2020
11	World Economic Forum	(1) Active inclusion, (2) Fairness, (3) Right to understanding, (4) Access to remedy	https://www.weforum.org/whitepapers/how-to-prevent-discriminatory-outcomes-in-machine-learning	26 November 2020
12	OECD	(1) Inclusive growth, sustainable development and well-being, (2) Human-centred values and fairness, (3) Transparency and explainability, (4) Robustness, security and safety, (5) Accountability	https://doi.org/10.1017/ilm.2020.5	26 November 2020
13	EU – European Commission’s High-Level Expert Group	(1) Human agency and oversight (2) Technical Robustness and safety (3) Privacy and data governance (4) Transparency (5) Diversity, non-discrimination and fairness (6) Societal and environmental well-being (7) Accountability	https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai	26 November 2020

Appendix II: An AI Lifecycle Framework


AI
lifecycle
Business and use-case development
Problem/improvements are defined and use of AI is proposed.
Design phase
the business case is turned into design requirements for engineers.
Training and test data procurement
Initial data sets are obtained to train and test model.
Building
AI application is built.
Testing
The system is tested.
Deployment
When the AI system goes live.
Monitoring
performance of the system is assessed

Source: ICO[99]



Business Owner	*Defines business goals and requirements*
Data Scientist	*Uses data to train models to meet requirements*
Model Validator	*Uses business goals, regulations, and best practices to test models*
AI Operations Engineer	*Deploys and monitors models in running services*

Source: Richards et al.[15]

Appendix III: Documentation paradigms

AI Dataset level[1]

Datasheets – Gebru[98]

- Based **on the electronics industry**, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy every dataset should be accompanied with a datasheet based on questions relating to the lifecycle of the dataset.
- Elements of datasheets:
    - **Motivation**: to encourage dataset creators to clearly articulate their reasons for creating the dataset and to promote transparency about funding interests.
    - **Composition**: intended to provide dataset consumers with the information they need to make informed decisions about using the dataset for specific tasks (e.g. is GDPR relevant)
    - **Collection process**: provide information that allow others to reconstruct the dataset without access to it.
    - **Preprocessing/cleaning/labelling**: to provide dataset consumers with the information they need to determine whether the "raw" data has been processed in ways that are compatible with their chosen tasks
    - **Uses**: intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used
    - **Distribution**: to ask related questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties
    - **Maintenance**: intended to encourage dataset creators to plan for dataset maintenance and communicate this plan with dataset consumers

Data statements for NLP- Bender [101]

- Data statements will help alleviate issues related to exclusion and bias in language technology,
- Data statements should be included in system documentation and in academic papers presenting new datasets, based on the following aspects:
    - **Curation Rationale**: Which texts were included and what were the goals in selecting texts, both in the original collection and in any further sub-selection
    - **Language Variety**: Languages differ from each other in structural ways that can interact with NLP algorithms. Within a language, regional or social dialects can also show great variation
    - **Speaker Demographic**: Sociolinguistics has found that variation (in pronunciation, prosody, word choice, and grammar) correlates with speaker demographic characteristics

[1] Some explanations pasted from original sources.

    - **Annotator Demographic**: What are the demographic characteristics of the annotators and annotation guideline developers? Their own "social address" influences their experience with language and thus their perception of what they are annotating.
    - **Speech Situation**: Characteristics of the speech situation can affect linguistic structure and patterns at many levels.
    - **Text Characteristics**: Both genre and topic influence the vocabulary and structural characteristics of texts, and should be specified.
    - **Recording Quality**: For data that include audio-visual recordings, indicate the quality of the recording equipment and any aspects of the recording situation that could impact recording quality.
    - **Other**

The Dataset Nutrition label - Holland[100]

- **Increase the overall quality of AI models as a result of more robust training datasets** and the ability to check for issues at the time of model development
- Label creates an **expectation of explanation**, which will **drive better data collection practices**
- The Label is designed in an extensible fashion with multiple distinct components: "modules". The modules are stand-alone, allowing for greater flexibility as arrangements of **different modules can be used for different types of datasets**.
- Different modules can be used for different types of datasets:
    - **Meta data**: Meta information. This module is the only required module. It represents the absolute minimum information to be presented. Filename, file format, URL, domain, keywords, type, dataset size, % of missing cells, license, release date, collection range, description
    - **Provenance**: Information regarding the origin and lineage of the dataset. Source and author contact information with version history
    - **Variables**: Descriptions of each variable (column) in the dataset. Textual descriptions
    - **Statistics**: Simple statistics for all variables, in addition to stratifications into ordinal, nominal, continuous, and discrete. Least/most frequent entries, min/max, median, mean, etc
    - **Pair Plots**: Distributions and linear correlations, between 2 chosen variables. Histograms and heatmaps
    - **Probabilistic Model**: Synthetic data generated using distribution hypotheses from which the data was drawn - leverages a probabilistic programming backend. Histograms and other statistical plots
    - **Grand Truth Correlations:** Linear correlations between a chosen variable in the dataset and variables from other datasets considered to be "ground truth", such as Census Data. Heatmaps

AI Model level

Model cards for model reporting[103]

- Need to have detailed documentation accompanying trained machine learning models, including metrics that capture bias, fairness and inclusion considerations.
- Structure of the model:
  - **Model Details.** Basic information about the model
    - Person or organization developing model
    - Model date
    - Model version
    - Model type
    - Information about training algorithms, parameters, fair- ness constraints or other applied approaches, and features
    - Paper or other resource for more information
    - Citation details
    - License
    - Where to send questions or comments about the model
  - **Intended Use.** Use cases that were envisioned during development
    - Primary intended uses
    - Primary intended users
    - Out-of-scope use cases
  - **Factors.** Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others
    - Relevant factor
    - Evaluation factors
  - **Metrics**. Metrics to be chosen to reflect potential real-world impacts of model
    - Model performance measure
    - Decision thresholds
    - Variation approaches
  - **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
    - Datasets
    - Motivation
    - Preprocessing
  - **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
  - **Quantitative Analyses**
    - Unitary results
    - Intersectional results
  - **Ethical Considerations**
  - **Caveats and Recommendations**

Factsheets - Arnold [103]

- Despite active research and development to address these issues, there is no mechanism yet for the creator of an AI service **to communicate how it addresses trustworthiness**.

- Can be tailored to the **particular AI model or service**, to the needs of their **target audience or consumer**, and thus **can vary in content and format**,
- Capture model or service facts from the **entire AI Lifecycle**, are compiled with **inputs from multiple roles** in this lifecycle
- Methodology is motivated by **user-centred design principles**, where user input from multiple stakeholders is collected **to inform design**.
- Modelled after a supplier's declaration of conformity **(SDoC**) used in many different industries and sectors including telecommunications and transportation
- Elements of Trust in AI systems: (1) **Basic Performance and Reliability**, (2) **Safety** , (2a) **Dataset Shift**, (2b) **Fairness** (2c) **Explainability**. (3) **Security** (4) **Lineage**

Factsheets construction methodology – Richards [103]

- Step 1: Know Your FactSheet Consumers
- Step 2: Know Your FactSheet Producers
- Step 3: Create a FactSheet Template
- Step 4: Fill In FactSheet Template
- Step 5: Have Actual Producers Create a FactSheet
- Step 6: Evaluate Actual FactSheet With Consumers
- Step 7: Devise Other Templates and Forms For Other Audiences and Purposes

References

1. Rao, A. S. & Verweij, G. *Sizing the prize: What's the real value of AI for your business and how can you capitalise? PWC* https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html (2017).
2. FERMA. *Artificial Intelligence Applied To Risk Management.* https://www.ferma.eu/publication/artificial-intelligence-ai-applied-to-risk-management/ (2019).
3. Babic, B., Cohen, G., Evgeniou, T. & Gerke, S. When Machine Learning Goes Off the Rails. *Harvard Business Review* (2021).
4. Allianz. Allianz Risk Barometer - Identifying the major business risks for 2020. **25**, 1–11 (2020).
5. Bushey, C. & Meyer, G. Boeing expects 737 Max crisis costs to reach $18.6bn. *Financial Times* (2020).
6. Wakabayashi, D. Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. *New York Times* (2018).
7. Alemzadeh, H., Raman, J., Leveson, N., Kalbarczyk, Z. & Iyer, R. K. Adverse Events in Robotic Surgery: A Retrospective Study of 14 Years of FDA Data. *PLOS ONE* **11**, e0151470 (2016).
8. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *npj Digital Medicine* **3**, 1–8 (2020).
9. U.S. Food & Drug Administration. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan.* https://www.fda.gov/media/145022/download (2021).
10. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A Survey on Bias and Fairness in Machine Learning. (2019).
11. Baer, T. & Kamalnath, V. Controlling machine-learning algorithms and their biases. *McKinsey Quarterly* 1–7 (2017).
12. Brennan, T., Dieterich, W. & Ehret, B. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior* **36**, 21–40 (2009).
13. Desmarais, S. L. & Singh, J. P. Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States: An Empirical Guide. 1–59 (2013).
14. Angwin, U., Larson, J., Mattu, S. & Kirchner, L. Machine Bias. *ProPublica* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (2016).
15. Bender, E. M. & Friedman, B. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* **6**, 587–604 (2018).
16. Cohen, L., Lipton, Z. C. & Mansour, Y. Efficient candidate screening under multiple tests and implications for fairness. (2019).
17. Buolamwini, J. & Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. in *Conference on Fairness, Accountability, and Transparency* PMLR 81:77-91 (2018).
18. *General Data Protection Regulation. Regulation (EU) 2016/679 of the European Parliament and of the Council.*
19. Scully, P. Smart Meter Market 2019: Global penetration reached 14% – North America, Europe ahead. https://iot-analytics.com/smart-meter-market-2019-global-penetration-reached-14-percent/ (2019).

20. Zhang, Y., Huang, T. & Bompard, E. F. Big data analytics in smart grids: a review. *Energy Informatics* **1**, 1–24 (2018).
21. Webborn, E. & Oreszczyn, T. Champion the energy data revolution. *Nature Energy* vol. 4 624–626 (2019).
22. Simmons, D. 10 Countries with GDPR-like Data Privacy Laws. https://insights.comforte.com/countries-with-gdpr-like-data-privacy-laws.
23. Kathleen Walch. AI Laws are coming. https://www.forbes.com/sites/cognitiveworld/2020/02/20/ai-laws-are-coming/ (2020).
24. European Commission. *White Paper on Artificial Intelligence - A European approach to excellence and trust*. *COM(2020) 65 final* (2020).
25. *European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL))*. (2020).
26. Tankard, C. What the GDPR means for businesses. *Network Security* **2016**, 5–8 (2016).
27. Brundage, M. *et al.* Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. (2020).
28. Schiff, D., Rakova, B., Ayesh, A., Fanti, A. & Lennon, M. Principles to Practices for Responsible AI: Closing the Gap. (2020).
29. Chen, J., Sohal, A. S. & Prajogo, D. I. Supply chain operational risk mitigation: a collaborative approach. *International Journal of Production Research* **51**, 2186–2199 (2013).
30. Wade, C. E. Method and apparatus for processing risk assessment and operational oversight framework. (2008).
31. The Economist, I. U. *Staying ahead of the curve The business case for responsible AI*. https://pages.eiu.com/rs/753-RIQ-438/images/EIUStayingAheadOfTheCurve.pdf (2020).
32. Tavani, H. T. *Ethics and technology: contreversies, questions, and strategies for ethical computing*. *Journal of Chemical Information and Modeling* vol. 53 (Wiley, 2015).
33. Friedman, B. & Hendry, D. G. *Value Sensitive Design Shaping Technology with Moral Imagination*. (MIT Press, 2019).
34. Jobin, A., Ienca, M. & Vayena, E. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* **1**, 389–399 (2019).
35. 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. (2019).
36. Martinho-Truswell, E. & Mont, C. G. Mexico leads Latin America as one of the first ten countries in the world to launch an artificial intelligence strategy.
37. OECD. Recommendation of the Council on Artificial Intelligence (OECD). (2020).
38. OECD. OECD AI Principles. https://search.oecd.org/going-digital/ai/principles/.
39. High-Level Independent Group on Artificial Intelligence (AI HLEG). Ethics Guidelines for Trustworthy AI. *European Commission* 1–39 (2019).
40. Council of Europe. Council of Europe and Artificial Intelligence. https://www.coe.int/en/web/artificial-intelligence.
41. UNESCO. Elaboration of a Recommendation on the ethics of artificial intelligence. Elaboration of a Recommendation on the ethics of artificial intelligence.
42. United Nations. Secretary-General's High-level Panel on Digital Cooperation.
43. Microsoft AI Principles. https://www.microsoft.com/en-us/ai/responsible-ai.
44. Google AI principles. https://ai.google/principles/.
45. Salesforce AI Principles. https://www.salesforce.org/blog/ai-good-principles-believe/.
46. IBM AI Focus areas. https://www.ibm.com/artificial-intelligence/ai-ethics-focus-areas.

47. Intel. *The Public Policy Opportunity Intel and Artificial Intelligence*. (2017).
48. SAP AI Principles. https://news.sap.com/2018/09/sap-guiding-principles-for-artificial-intelligence/ (2018).
49. Asilomar AI Principles. https://futureoflife.org/ai-principles/.
50. IEEE. A Call to Action for Businesses Using AI Ethically Aligned Design for Business. (2020).
51. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, IEEE*. https://standards.ieee.org/content/ieee-standards/en/industry-connections/ec/ autonomous-systems.html (2019).
52. Floridi, L. *et al. The AI4People's Ethical Framework for a Good AI*. (2018).
53. Floridi, L. & Cowls, J. A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review* 1–15 (2019) doi:10.1162/99608f92.8cd550d1.
54. Mittelstadt, B. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* **1**, 501–507 (2019).
55. Gerke, S., Babic, B., Evgeniou, T. & Cohen, I. G. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. *npj Digital Medicine* **3**, 53 (2020).
56. Diakopoulos, N. & Friedler, S. How to Hold Algorithms Accountable. *MIT Technology Review* (2016).
57. Unceta, I., Nin, J. & Pujol, O. Risk mitigation in algorithmic accountability: The role of machine learning copies. *PLoS ONE* **15**, 18–20 (2020).
58. Ammanath, B., Hupfer, S. & Jarvis, D. *The state of artifical intelligence in business. Deloitte Insights* https://www2.deloitte.com/us/en/insights/focus/cognitive-technologies/state-of-ai-and-intelligent-automation-in-business-survey.html?id=us:2ps:3gl:aisurvey:awa:con:071420:ad2:kwd-932327896302:%2Bai in the %2Benterprise&gclid=EAIaIQobChMIpsyytr2V6wIVBeiGCh (2020).
59. Peters, D., Vold, K., Robinson, D., Calvo, R. A. & Member, S. Responsible AI-Two Frameworks for Ethical Design Practice. *Ieee Transactions on Technology and Society* **1**, 34–47 (2020).
60. Friedman, B. & Hendry, D. G. *Value Sensitive Design: Shaping Technology with Moral Imagination*. (Cambridge, MA, USA: MIT Press, 2019).
61. Friedman, B., Hendry, D. G. & Borning, A. A Survey of Value Sensitive Design Methods. *Foundations and Trends® in Human–Computer Interaction* **11**, 63–125 (2017).
62. Xu, H., Crossler, R. E. & Bélanger, F. A Value Sensitive Design Investigation of Privacy Enhancing Tools in Web Browsers. *Decision Support Systems* **54**, 424–433 (2012).
63. Stahl, B. C. & Wright, D. Ethics and Privacy in AI and Big Data: Implementing Responsible Research and Innovation. *IEEE Security and Privacy* **16**, 26–33 (2018).
64. Cavoukian, A., Taylor, S. & Abrams, M. E. Privacy by Design: essential for organizational accountability and strong business practices. *Identity in the Information Society* **3**, 405–413 (2010).
65. Alshammari, M. & Simpson, A. Towards a principled approach for engineering privacy by design. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **10518 LNCS**, 161–177 (2017).
66. Yanisky-Ravid, S. & Hallisey, S. 'Equality and Privacy by Design': Ensuring Artificial Intelligence (AI) Is Properly Trained and Fed: A New Model of AI Data Transparency Via Auditing; Certification and Safe Harbor Procedures. *SSRN Electronic Journal* 1–65 (2018) doi:10.2139/ssrn.3278490.

67. Madaio, M. A., Stark, L., Wortman Vaughan, J. & Wallach, H. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. 1–14 (2020) doi:10.1145/3313831.3376445.
68. Cramer, H., Garcia-Gathright, J., Reddy, S., Springer, A. & Takeo Bouyer, R. Translation, Tracks & Data: An Algorithmic Bias Effort in Practice. *CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland Uk* 1–8 (2019) doi:10.1145/3290607.3299057.
69. Hagendorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines* **30**, 99–120 (2020).
70. Vakkuri, V., Kemell, K.-K., Kultanen, J., Siponen, M. & Abrahamsson, P. Ethically Aligned Design of Autonomous Systems: Industry viewpoint and an empirical study. (2019).
71. Peters, A. This tool lets you see and correct the bias in an algorithm. *Fastcompany* https://www.fastcompany.com/40583554/this-tool-lets-you-see-and-correct-the-bias-in-an-algorithm (2018).
72. Bethge Labs. Foolbox. *GitHub repository* https://github.com/bethgelab/foolbox (2020).
73. Clever Hans. Cleverhans. *GitHub repository* https://github.com/cleverhans-lab/cleverhans (2018).
74. Deloitte. The new AI Fairness Paradigm. https://www2.deloitte.com/de/de/pages/risk/solutions/ai-fairness-with-model-guardian.html.
75. Digital Civil Society Lab & Society, at the S. C. on P. and C. Digital Impact Toolkit. https://digitalimpact.io/toolkit/.
76. Driven Data. Deon. *GitHub repository* https://github.com/drivendataorg (2020).
77. Pesenti, J. AI at F8 2018: Open frameworks and responsible development. *Facebook Engineering* (2018).
78. Wexler, J. The What-If Tool: Code-Free Probing of Machine Learning Models. *Google AI Blog* https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html (2018).
79. Ethicstoolkit.ai. Ethics & Algorithms Toolkit.
80. Bellamy, R. K. E. *et al.* AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* **63**, 4:1-4:15 (2019).
81. IBM. AI Fairness 360. *GitHub repository* https://github.com/Trusted-AI/AIF360.
82. IBM. AI Explainability 360. *GitHub repository* https://github.com/Trusted-AI (2019).
83. IBM. Adversarial Robustness Toolbox. *GitHub repository* (2021).
84. Linkedin. LinkedIn Fairness Toolkit. *GitHub repository* https://github.com/linkedin/LiFT (2020).
85. Bird, S. *et al. Fairlearn: A toolkit for assessing and improving fairness in AI.* https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/ (2020).
86. Microsoft. InterpretML. *GitHub repository* https://github.com/interpretml (2021).
87. Microsoft. Community Jury. *Azure* (2020).
88. Wang, A., Narayanan, A. & Russakovsky, O. REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets. (2020).
89. Oracle. Skater. *GitHub repository* https://github.com/oracle/Skater (2018).
90. IBM AI focus areas. https://www.ibm.com/artificial-intelligence/ai-ethics-focus-areas.
91. PWC. PwC's Responsible AI. https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html.
92. Pymetrics. audit-ai. *GitHub repository* https://github.com/pymetrics/audit-ai (2020).

93. Sokol, K. *et al.* FAT Forensics: A Python Toolbox for Implementing and Deploying Fairness, Accountability and Transparency Algorithms in Predictive Systems. *Journal of Open Source Software* **5**, 1904 (2020).
94. Center for Data Science and Public Policy - University of Chicago. Aequitas. http://aequitas.dssg.io/ (2018).
95. Ribeiro, M. T. C. Lime. *GitHub repository* https://github.com/marcotcr/lime (2020).
96. World Economic Forum. *How to prevent Discriminatory Outcomes in Machine Learning (White Paper).* (2018).
97. Kelley, P. G., Bresee, J., Cranor, L. F. & Reeder, R. W. A " Nutrition Label " for Privacy. **1990**, (2009).
98. Holland, S., Hosny, A., Newman, S., Joseph, J. & Chmielinski, K. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. (2018).
99. Gebru, T. *et al.* Datasheets for Datasets. (2018).
100. Arnold, M. *et al.* FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* **63**, (2019).
101. Mitchell, M. *et al.* Model cards for model reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* 220–229 (2019) doi:10.1145/3287560.3287596.
102. Benjamin, M. *et al.* Towards Standardization of Data Licenses: The Montreal Data License. 1–16 (2019).
103. Richards, J., Piorkowski, D., Hind, M., Houde, S. & Mojsilović, A. A Methodology for Creating AI FactSheets. *IBM Research* (2020).
104. Yang, K. *et al.* A Nutritional Label for Rankings. in *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18* 1773–1776 (ACM Press, 2018). doi:10.1145/3183713.3193568.
105. Moons, K. G. M. *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine* **162**, W1-73 (2015).
106. IEEE. IEEE Standard for Software Reviews and Audits. *IEEE Std 1028-2008* 1–53 (2008).
107. ICO. Guidance on the AI auditing framework: Draft guidance for consultation. *Information Commissioner's Office* (2020).
108. Raji, I. D. *et al.* Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* 33–44 (2020) doi:10.1145/3351095.3372873.
109. Clark, A. The machine learning audit- CRISP-DM Framework. *ISACA Journal* **1**, 42–47 (2018).
110. Goodman, B. A Step Towards Accountable Algorithms?: Algorithmic Discrimination and the European Union General Data Protection. *29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.* 1–7 (2016).
111. Diakopoulos, N. Algorithmic Accountability: Journalistic investigation of computational power structures. *Digital Journalism* **3**, 398–415 (2015).
112. Mahajan, V., Venugopal, V. K., Murugavel, M. & Mahajan, H. The Algorithmic Audit: Working with Vendors to Validate Radiology-AI Algorithms—How We Do It. *Academic Radiology* **27**, 132–135 (2020).
113. Bhattacharyya, S., Cofer, D., Musliner, D., Mueller, J. & Engstrom, E. Certification considerations for adaptive systems. *2015 International Conference on Unmanned Aircraft Systems, ICUAS 2015* 270–279 (2015) doi:10.1109/ICUAS.2015.7152300.

114. International Organization for Standardization - Certification and Conformity. https://www.iso.org/certification.html.

115. Fisher, M. *et al.* Towards a Framework for Certification of Reliable Autonomous Systems. (2020).

116. Havens, J. C. & Hessami, A. From Principles and Standards to Certification. *Computer* **52**, 69–72 (2019).

117. IEEE Launches Ethics Certification Program for Autonomous and Intelligent Systems. https://standards.ieee.org/news/2018/ieee-launches-ecpais.html (2018).

118. FDA. *Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback*. *U.S Food & Drug Administration* (2019).

119. Schmid, M. Model-Based Certification of Automated Vehicles. (2020).

120. Peter Katko. How GDPR compliance demands have shifted the focus to certification. https://www.ey.com/en_gl/tax/how-gdpr-compliance-demands-have-shifted-the-focus-to-certification (2019).

121. Babic, B., Gerke, S., Evgeniou, T. & Cohen, I. G. Algorithms on regulatory lockdown in medicine. *Science* **366**, 1202–1204 (2019).

122. Taddeo, M., McCutcheon, T. & Floridi, L. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence* **1**, 557–560 (2019).

123. Purdy, M., Eitel-Porter, R., Krüger, R. & Deblaere, T. How Digital Twins Are Reinventing Innovation. *MIT Sloan Management Review* (2020).

124. JOSHUA A. KROLL, JOANNA HUEY, SOLON BAROCAS, EDWARD W. FELTEN, JOEL R. REIDENBERG, D. G. R. & YU, & H. Accountable Algorithms. *University of Pennsylvania Law Review* **165**, (2016).

125. Global GDPR Services Market Research Report. 161 https://www.marketresearchfuture.com/reports/gdpr-services-market-7189 (2020).

126. Sawers, P. 5 data privacy startups cashing in on GDPR. *VentureBeat* https://venturebeat.com/2019/07/23/5-data-privacy-startups-cashing-in-on-gdpr/ (2019).

127. Burgess, B. EY announces the first solution designed to help gauge impact and trustworthiness of artificial intelligence systems. https://www.ey.com/en_gl/news/2019/04/ey-announces-the-first-solution-designed-to-help-gauge-impact-and-trustworthiness-of-artificial-intelligence-systems (2019).

128. KPMG. *KPMG Ignite*. https://assets.kpmg/content/dam/kpmg/uk/pdf/2018/09/kpmg-ignite.pdf (2018).

129. PWC. *A practical guide to Responsible Artificial Intelligence (AI)*. https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html (2019).