Filling the gap between principle and practice: building an ethical and human rights-based tool-kit for AI development.

Dr. Nicola Palladino
Trinity College Dublin
palladin@tcd.ie

Abstract

Over the past few years, the awareness that the full potential of artificial intelligence (AI) could be attained only through the establishment of a trustworthy and human-centric framework has expanded, thereby prompting demand for greater regulation as well as engendering a flourish of initiatives that set ethical codes and good governance principles for AI development. In this context, developers and deployers could play a crucial role because they have the capabilities to address ethical issues in a concrete, timely, and effective manner. In so doing, their organizations may therefore contribute to shaping the regulatory environment of the near future. However, many studies raise concerns about a "principle-to-practice" gap: organizations rather often fall short in providing enforcement of the principles they claim to adhere to. This project aims to fill this void mapping, reviewing, and combining principles, requirements, and tools in terms of both technical and governance arrangements, in order to provide a repertoire of already existing instruments, highlights holes and shortcoming in the current scenario and outline possible ways forward.

1. Introduction

"Artificial Intelligence" is a label used as shorthand for an expanding 'family' of software (and hardware) systems capable of performing specific tasks by collecting, analysing, and interpreting data, sometimes perceiving the environment in which they operate, to make decisions and take actions with a certain degree of autonomy (Russel and Norving 2003). AI is increasingly crucial in everyday life and social relations, which raises both expectations on AI's capacity to foster human well-being as well as concerns about the risks for human autonomy and integrity (Renda 2019, Boiler 2018). On the one side, they can help us address the complexity of the modern world, optimizing our decision-making processes and resource allocation with relevant effects on different aspects of social life, including production, transport, crisis management, environment, and healthcare. On the other side, AI systems raise concerns ranging from privacy and data protection, to discrimination, manipulation, misinformation or the endangerment of democratic institutions and the effects on jobs and rights on the workplace. As testified by recent initiatives such as the EU Ethical Guidelines for Trustworthy AI, the Ethical Aligned Design report drafted by Institute of Electrical and Electronics Engineers (IEEE), and the OECD Recommendations on AI, in the past few years, governments, the private sector, civil society, and the technical community reached the awareness that the full potential of this technology is attainable only by building a trustworthy and human-centric framework. In this view, AI systems must be aligned with societal values and governed through accountable arrangements to avoid both misuse of AI applications capable of endangering people and underuse because of a lack of public acceptance (Floridi et al 2019)[1]. Scholars also pointed out how cooperation among stakeholders is needed to achieve regulation capable of ensuring predictability and legal certainty even if the debate remains

[1] See also European Commission (2020) Public Consultation on the AI White Paper Final Report, https://ec.europa.eu/digital-single-market/en/news/white-paper-artificial-intelligence-public-consultation- towards-european-approach-excellence

open regarding the role and responsibilities of different actors and the proper degree of mandatory requirements needed to safeguard people without hindering innovation (Turner 2019, Brown and Mardsen 2013, Brownsword and Yeung 2008).

Not by chance, in the past few years, we have witnessed to a flourish of initiatives setting ethical codes and good governance principles for AI development that usually converge around a common set of guiding principles, including respect for human autonomy, prevention of harm, fairness, privacy, transparency, and explicability (Whittaker et al. 2018, Jobin et al. 2019, Berkman Center 2020).

In particular, the European Union has been active in developing a regulatory framework grounded in fundamental rights to position trustworthy and human-centric AI as the "distinctive trademark for Europe and its industry as a leader in cutting-edge AI" (EU COM 2019:9) and set the global standard for future AI. As stated in the White Paper on Artificial Intelligence (EU COM 2020), and confirmed in the recent European Commission proposal for an AI regulation (EU COM 2021,'Artificial Intelligence Act'), new initiatives are expected to address some aspects not specifically covered under the current legislation. Mandatory requirements are being asked for high-risk applications and encouraged for low-risk applications through labelling or other voluntary schemes.

However, many studies raise concerns about a **"principle-to-practice" gap**, noting that AI's developers and deployers (mostly private companies) often fall short in ensuring effectiveness and enforcement of the principles they adhere to (Mittelstadt 2019, Schiff et al 2020). On the one side, the principles-to-practices gap testifies of "ethical washing" practices put in place by companies to delay or soften state regulation (Greene 2019, van Dijck and Casiraghi 2020). On the other side, some scholars also identify factors related to AI systems' productive process to explain the poor impact of many AI ethical initiatives (Schiff et al. 2019, Hallensleben et al. 2020). These scholars comment that AI comprises complex "socio-technical" systems, meaning on the one side digital codes and architectures embed specific values and rules (consciously or unconsciously), which discipline people's behavior, impacting on their integrity and autonomy (Musiani et al. 2016; Lessig 2006; DeNardis 2013). On the other side that they are processes of technical design involving different professionals from different backgrounds in a long production cycle (Kitchin 2017, Hildebrandt 2019) influenced by factors such as division of labor, organizational culture, operational routine, governance arrangements, and the broader regulatory environment. This complexity may result in a lack of awareness by developers of the social implication of their job, different interpretations of the same principles, functional separation and lack of communication between more technical or social-oriented components in the process, unclear accountability mechanisms, and attribution of responsibilities (Mittelstadt 2019).

Initiatives currently on the table do not provide developers and deployers with sufficient details on how to implement principles for trustworthy and human-centric AI within concrete contexts. Even the recent European Commission proposal for an AI regulation ('Artificial Intelligence Act'), after setting a series of requirements for high risk AI applications relating to risks management, data governance, transparency, human oversight, robustness, and cybersecurity, stated that: "The precise technical solutions to achieve compliance with those requirements may be provided by standards or by other technical specifications or otherwise be developed in accordance with general engineering or scientific knowledge at the discretion of the provider of the AI system". On the other hand, standard-setting initiatives such as the International Organization for Standardization (ISO)'s projects on a 'management system standard for AI and other related issues', could provide more operative suggestions (Lewis 2020), but they are still in their infancy. Without guidance, developers and their companies' ethical approach may result into PR strategy or checklist attitude resembling ethical washing operations.

2. Objective, Data, Methodology

This paper aims at providing some insights to close the "principle to practices gap" within operational contexts. For this purpose, I mapped and reviewed principles and tools already developed to address the ethical challenge posed by AI technologies.
The intention in presenting this research is to advance the discussion on how to embed ethical and human rights standards within the AI systems' socio-technical design, by:
1) outlining the backbone of a comprehensive tool-kit capable of providing developer, deployer, and other practitioners and stakeholders with practical guidance to 'operationalize' ethical principles and human rights standards into technical, organizational, and governance arrangements.
2) highlighting the main features of the current landscape, identifying shortcomings and holes that could undermine or the efforts to build an ethical and human-rights based approach to AI development;
3) define possible ways forward to improve the development of AI Ethics tools.
For these purposes, this study will move from the structure outlined by the VCIO (Value, Criteria, Indicator, Observables) model, developed by AI Ethics Impact Group, led by VDE Association for Electrical, Electronic & Information Technologies and Bertelsmann Stiftung (Hallensleben et al. 2020). Their framework starts from the definition of 'values' such as "general ethical concern, something that should guide our actions" (Idem: 6) in developing AI systems, which should be chosen to maximize the benefit and minimize the risk of AI technologies. To assess and measure if a determinate value has been fulfilled or violated, is necessary to formulate 'criteria' that break down values into less abstract dimensions and define the conditions of principles compliance or infringement. Furthermore, 'indicators' are required to monitor whether specific criteria are met or not, with related 'observables', which measure the extent to which indicators are satisfying criteria.
It is worth noting that the VCIO model has been designed to be applied in specific and concrete AI systems development processes, and it is based on a strong context-dependence assumption, according to which "how we implement and prioritise values […] depends to some extent on the field of application and the cultural context an AI system operates in" (Idem:10). For this reason, in this model, values are not given, and the upper layers could not serve to deduct either criteria, indicators, and observable logically, but all these elements must be set through deliberative processes involving different kinds of stakeholders.
Although this framework looks well-grounded and promising, I will revise its structure to serve a slightly different purpose in this paper.
A structure values-criteria-indicators will be employed to map and organize already existing initiatives put in place to tackle the ethical challenge to AI development.
In this view, values will identify ethical and human rights-based principles proposed in ethical codes, policy papers, and recommendations to ensure that AI technologies will contribute to human well-being, avoiding risks and harms to the integrity and autonomy of individuals and communities.
Following a common approach recently adopted also by the European Union, especially in its Ethical Guidelines for Trustworthy AI and the recent Artificial Intelligence Acts, criteria will define more specific requirements that must be put in place to ensure that a determinate value in terms of ethical or human-rights based principles has been fulfilled. In this context, indicators refer to technical and governance arrangements, which 'indicate' that proper actions have been implemented to meet specific criteria/requirements. For the level of detail of this paper observables are not taken into account. However, in this context, they could be conceived as monitoring systems associated with each technical and governance arrangement, assessing to what extent they reached their expected goals.
Thus, this paper will:

1) Map and review principles and requirements: to this purpose I resort to the AI Ethics Guidelines Global Inventory[2], a participated initiative launched by Algorithm Watch, through which, at the date of 20 April 2020, 173 different items have been collected ad stored in a public accessible repository. These documents have been scrutinized using a qualitative content analysis methodology assisted by the software Nvivo (Mayring 2019, Kaefer et al. 2015) to realize a classification of principles and requirements. To carry out this task, I relied on the schemes elaborated by the European Commission Hight Level Expert Group on Artificial Intelligence and the Berkman Center (2020), which already distinguishes between general principles and more specific requirements. These initial inputs have been revisited to meet the purpose of this investigation excluding items that cannot be employed in the operative development of AI system; integrating new entry and insights coming from the document analysed and relevant literature; paying attention to minimizing redundancy and maximizing hierarchical classification.
2) Similarly, tools implementing AI ethical and human right based principles have been mapped and reviewed. In the first instance I relied on list or repository of tools made publicly available by previous research and initiatives[3], integrated with items derived from relevant literature or monitoring the website of organizations operating in the field. The items so gathered has then been filtered according to the following criteria: 1) they must be developed within the last five years (2016-2020), except for quoted previous works that have had a significant impact in the field; 2) they must provide guidance, instructions, methods appliable in the context of concrete AI systems development. Even in this case, tools have been analysed through qualitative content analysis conducted through NVIVO to iteratively identify main features, key concepts, and class or family of techniques and methods at different levels of abstraction that could be employed to implement AI ethical and human right based principles.
3) Finally, principles, requirements and tools have been combined to: 1) outline clusters of value-criteria-indicators offering pictures of the available instruments for the most relevant ethical issues in the development of AI system; 2) identify more cross-cutting issues and get some insight for potential ways forward.

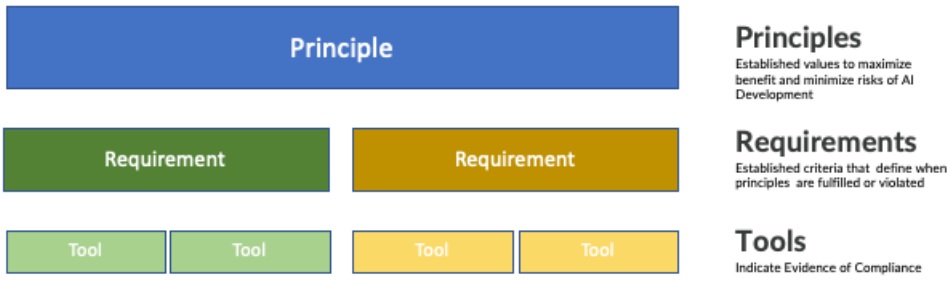

Fig.1 Principles/Requirements/Tools Structures

[2] https://inventory.algorithmwatch.org/about

[3] The following lists of AI Ethics Tools have been accessed: AI Ethics Tool Landscape (https://edwinwenink.github.io/ai-ethics-tool-landscape/); Morley et al. 2020 (https://docs.google.com/document/d/1h6nK9K7qspG74_HyVlT0Lx97URM0dRoGbJ3ivPxMhaE/edit); Ayling, J., Chapman 2021; http://algorithmtips.org/resources/; AI NOW Algorithmic Accountability Toolkit (https://ainowinstitute.org/aap-toolkit.html).

3. Mapping and Reviewing Principles and Requirements

The analysis of the ethical codes, guidelines and recommendations' corpus of stored in the AI Ethics Guidelines Global Inventory, allowed to identify the following structure of principles and requirements:

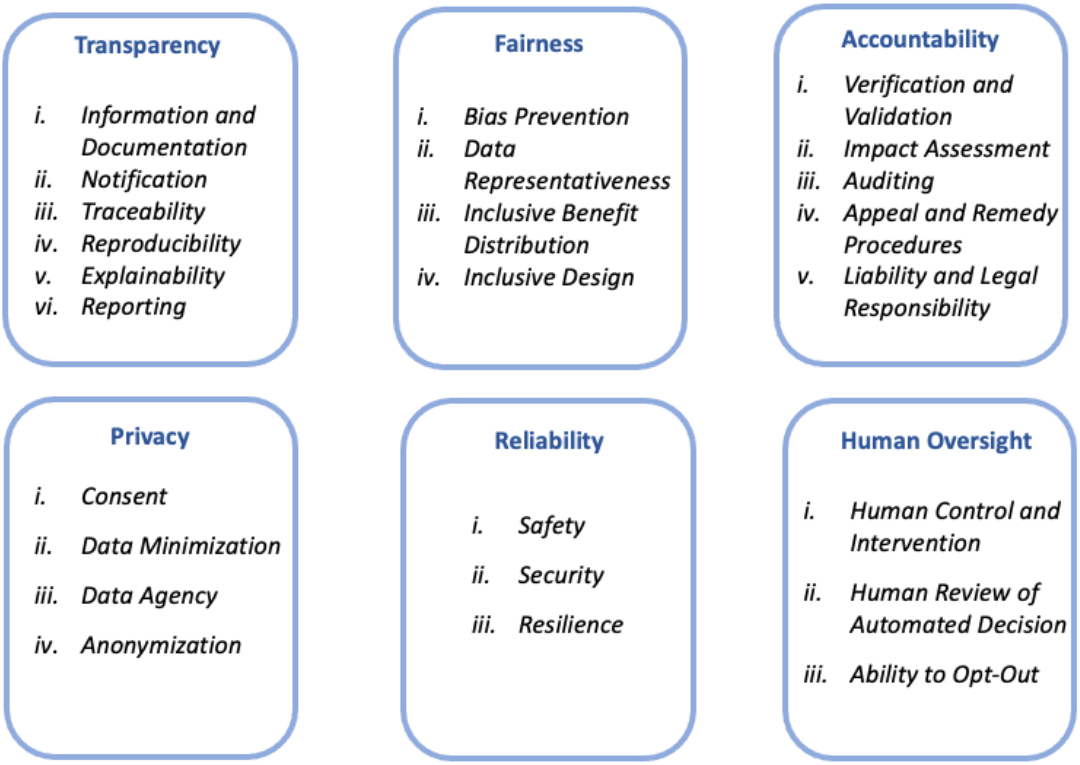

Fig.2 Principles and Requirement

1) Transparency: Transparency is one of the most referred principle in AI governance, even because it figures as a pre-condition for fulfilling of other principles such as accountability, privacy and fairness. As the Toronto Declaration states, recalling the UN Guiding Principles on Business and Human Rights: "Transparency is a key component of human rights due diligence, and involves communication, providing a measure accountability to individuals or groups who may be impacted and to other relevant stakeholders" (Amnesty International and Access Now, 2018:14).
Although, in the AI context transparency tends to be reduced to the explainability and traceability requirements, focusing on the understanding of automated decisions, it is worth noting that the safeguard of human rights and the creation of a trustworthy environment also rely on the delivery of information not directly related to AI system operations, but rather on its governance structure. Indeed, we can breakdown the transparency principle in the request to provide six different kinds of information to the public:
i. *Provide General Information*: this requirement, generally neglected in ethical codes, has been enshrined by the recent European Commission proposal for an Artificial Intelligence Act, which establishes that users should be informed about "the identity and the contact details of the provider and, where applicable, of its authorised representative", the "intended purpose" of the AI system, the "level of accuracy, robustness, and cybersecurity against which the high-risk AI system has been tested and validated"[4],
ii. *Notification*: according to Smart Dubai AI Ethical Principles and Guidelines, "people should be informed of the extent of their interaction with AI systems", both "when a significant decision

[4] The proposal for an Artificial Intelligence Act set out many other disclosure obligation, which nevertheless are included in the other requirement in the list.

affecting them has been made by an AI system" (Smart Dubai 2018:18) and when they are interacting with an AI system impersonating a human agent (chatbot, personal assistant, customer care services).
iii. *Traceability*: it could be considered a key requirement for trustworthiness, related to the need "to maintain a complete account of the provenance of data, processes, and artifacts involved in the production of an AI model" (Mora-Cantallops et al. 2021:1). It is often considered a pre-condition for explainability.
iv. *Reproducibility:* As stated in the EU Ethical guidelines for Trustworthy AI*: "*It is critical that the results of AI systems are reproducible […] Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions. This enables scientists and policy makers to accurately describe what AI systems do" (EU HLEG 2019: 17). According to The Responsible Machine Learning Principles drafted by the Institute for Ethical AI & Machine Learning, to obtain reproducibility is necessary abstract its constituent components (data, configuration/environment, computational graph) and adopt open standards "to abstract multiple machine learning libraries with specific data input/output formats"[5].
v. *Explainability*: it is probably the transparency requirement toward which most of efforts and resources are devoted. The Berkman Center report, summarizes this requirement in terms of "translation of technical concepts and decision outputs into intelligible, comprehensible formats suitable for evaluation" (Berkman Center 2020:42). Instead the EU Ethical guidelines, stressing the socio-technical nature of AI systems, specify how "Explainability concerns the ability to explain both the technical processes of an AI system and the related human decisions" (EU HLEG 2019: 18). In any case, the explainability requirements means that it should be always possible traceback and understands how AI systems came to a determinate decision, especially if they have a relevant impact on people life.
vi. *Regular Reporting*: this requirement implies that organizations developing or deploying AI systems should periodically disclose information about "operating errors, unexpected or undesirable effects, security breaches, and data leaks" (University of Montreal 2018:12)

2) Fairness and Non-discrimination
Broadly speaking, fairness considers whether people are treated equally in the context of a decision-making process (Palladino and Santaniello 2021). It can be divided into formal and substantial fairness (Hooker 2005). The latter refers to the concrete opportunity stemming from a decision-making process, and the former considers if rules and procedures have been applied impartially to all subjects (Scholte and Tallberg 2018; Schmidt and Wood 2019). Both formal and substantial fairness could apply to the participation and the effects of a decision-making process.
For the most part, AI systems are autonomous decision systems, which could produce biased decisions violating rights systematically and undermining the opportunities of determinate groups of subjects because the data employed for their training was biased, or some previous assumption. However, fairness in the AI context is not just a question of bias. It also involves issues concerning the distribution of the benefit produced and its influence in shaping this technology. To realize fairness, the following requirement should be satisfied:
i. *Non-discrimination and the Prevention of Bias*: it is one of the most quoted requirements, especially by private sectors actors, since it is likely to be addressed by technological means, and more than others could led to legal and reputational costs. It means that "AI must be designed to minimize bias and promote inclusive representation" (IBM 2019:34); avoiding "unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief"[6].
ii. *Data Representativeness*: according to this requirement AI system developers must employ accurate and high-quality data properly representing the population of interest, to avoid decisions

[5] https://ethical.institute/principles.html
[6] https://ai.google/principles/

and operations flawed by lack of accuracy and misrepresentations. In this regard, it is important to pay attention to path dependency effects and historical biases, which could crystalize the proprieties and the evaluations of subgroups across time.
iii. *Inclusive Benefit Distribution*: it requires that AI systems will not become instruments advantaging a narrow élite, already privileged peoples and country, or increasing existing inequalities, but they will serve as means to "benefit" and "empower" "as many people as possible" (Future Life Institute 2017, Microsoft 2018, Partnership on AI 2016, Smart Dubai 2018, Think 20 2018, UNI 2017).
iv. Inclusiveness in design: according to the Berkman Center, fairness also "requires more diverse participation in the development process for AI systems" (Berkman Center 2020: 52). This participation could take two different forms: 1) increasing the diversity within AI design teams, involving more woman, people with disabilities and minorities; 2) increasing broad multistakeholder collaboration during the design process.

3) Accountability: it could be understood as "a relationship in which a decision-maker is asked to report on their activities, and likely involving sanctions in the case of misconduct" (Palladino and Santaniello 2021: 34, see also Buchanan and Keohane 2006; Schmidt and Wood 2019). Usually scholars distinguish between internal accountability, which refers to "the authorization and support which principals give to agents who are institutionally linked to one another" (Risse 2006: 185) and external accountability, which requires decision-makers to justify their behavior "to people or groups outside the acting entity who are nevertheless affected by it" (ibid.).
In the field of AI, accountability requires:
i. *Verification and Validation*: AI systems' developers and deployers must provide evidence that their application function correctly, according with expected performance. Fallowing IEEE (2019: 269) "verification is a demonstration that a given application meets a narrowly defined requirement; validation is a demonstration that the application answers its real-world use case".
ii. *Impact Assessments*: potential risks and harms, especially if related to human rights, should be investigated and identified in the very early stage of AI application development and subsequent prevention and mitigation plan put in place (Access Now 2018, Amnesty International and Access Now 2018, EU HLEG 2019, Japanese Cabinet Office 2019, Public voice 2018).
iii. *Auditing*: AI system should be scrutinized by independent third-party.
iv. *Appeal and Remedy Procedures*: decision made by AI systems should be always disputable within appropriate bodies, and processes to redress adverse impacts should be established (Amnesty International and Access Now 2018)
v. *Liability and Legal Responsibility*: as stated by the Chinese AI Industry Code of Conduct, it is necessary clarify "the rights and obligations of parties at each stage in research and development, design, manufacturing, operation and service of AI, to be able to promptly determine the responsible parties when harm occurs."[7]

4) *Privacy*: privacy is a widely recognized right both in international human rights law and national legislation in almost all democratic countries. The recent European General Data Protection Regulation established higher international standards, further enhancing privacy protection. AI systems rely on huge amount of data, and then, if not designed properly, they are likely to result in systemic and pervasive privacy violation. To avoid this risk, AI system must satistfy the following requirements:
i. *Consent:* personal data should not be used without the explicit permission by the concerned subjects. Most of the documents call for informed consent, a "more robust principle – derived from the medical field – which requires individuals be informed of risks, benefits, and alternatives" (Berkman Center 2020:22).

[7] https://www.secrss.com/articles/11099

ii. *Data Minimization*: personal data collected should be adequate, proportionate and no more than minimum amount necessary to fulfil the specific purpose consented by the user. This requirement should be coupled with the 'storage limitation' principle, according to which AI systems should not keep personal data or no longer than is necessary for the purposes of the processing.
iii. *Data Agency/Control*: according to IEEE Ethical Aligned Design, the respect of human rights in the digital sphere requires that individuals are empowered with means ensuring "their dignity through some form of sovereignty, agency, symmetry, or control regarding their identity and personal data" (IEEE 2019:23). This also include provide individuals with the ability "to request that an entity stop using or limit the use of personal information" (Access Now 2018:31), or their rectification or erasure (Think 20 2018, Monetary Authority of Singapore 2019)
iv. *Anonimization*: anonymisation or 'de-identification', requires that "datasets are processed in order to remove as much data which relates to individuals as possible, while retaining the usefulness of the dataset for the desired purpose" (UK HoL 2018:31). Satisfying this requirement allows to maximize the usage of data, while minimizing risks for privacy infringements, however many experts points out that AI technologies make re-identification possible, for example cross-referring different datasets.

5) Reliability: An AI system is expected to be reliable, meaning that proper arrangements should be put in place to ensure an ongoing correct functioning for the purpose it was created, avoiding unintended harms to people, regardless of whether they could be caused by design or manufacturing faults, malfunctioning, external threats or misuses.
Reliability could be broken down in the following requirements:
i. Safety: "the system will do what it is supposed to do without harming living beings" (EU HLEG 2019:17):
ii. Security: refers to the capacity of the system to prevent or resist to external attacks or threats
iii. Resilience: resilience requires that AI systems to keep carrying out their core function without harming people, even in the case of internal or external damage
iv. Predictability: according to European Commission High Level Expert Group on Artificial Intelligence, "it must be ensured that the outcome of the planning process is consistent with the input, and that the decisions are made in a way allowing validation of the underlying process" (EU HLEG 2019: 22).

6) Human Oversight: The Human Oversight principle entails that AI systems should be designed in a way that safeguards human autonomy and human control on the system, ensuring the faculty of choosing whether or not delegating decision to AI systems, and to intervene on their operation (Future life Institute 2017, University of Montreal 2018). This principle could be broken down in the following more specific requirement:
i. Ensuring human control and intervention over operations: AI systems should "seek human input during critical situation" and "transfer control to a human in a manner that is meaningful and intelligible" (Microsoft 2018: 65, see also IBM 2019).
ii. Ensuring Human Review of Automated Decision: as the Berkman Center summarized, proper procedures should be put in place in order to ensure that "people who are subject to their decisions should be able to request and receive human review of those decisions" (Berkman Center 2020: 55)
iii. Ability to Opt out of Automated Decisions: this requirement is conceivable as a "a natural corollary of the right to notification when interacting with an AI system" (Berkman Center 2020: 55), and requires to establish procedures allowing people to drop out AI systems (UK HoL 2018, EU HLEG 2019, Smart Dubai 2019).

4. Mapping and Reviewing Tools for implementing ethical and human-rights based principles in AI Development

Following the rise of concerns, awareness, and public commitment on the social and political implication of AI technologies, in the last few years, a lot of attention and resources have been devoted to developing tools, methods, and techniques capable of dealing with the principles and requirements discussed in the previous paragraph. The field is rapidly expanding, even if fragmented and chaotic, pushed by business companies' research of competitive advantages (and their need to be ready for incoming regulatory frameworks), suffering the lack of agreed standards. The research presented in this paper has been carried out without claims of comprehensiveness, which will be unrealistic due to the complexity and the pace of technological innovation in this field. Instead, this works aims to identify some categories to order this messy scenario, distinguishing 'class' or 'families' of tools, methods, and techniques, highlighting their relationship with principles and requirements, and advancing some reflections about trends, shortcomings, and desirable future trajectories/development.
In this paragraph, I will present a taxonomy of AI ethical tools, that will be combined with the principles/requirements scheme discussed in the previous section in order to provide a first systematization of currently available instruments to reach a ethical and human-rights based AI development.
In the first instance, it is possible to distinguish between 'governance' and 'technical' arrangements. Recalling the socio-technical nature of AI systems, governance arrangements refer to the 'social' side of the productive cycle of AI systems and identify instruments setting organizational procedures to control AI development. Instead, technical arrangements approach the compliance with ethical and human-rights requirements in terms of technical performance problems, which could be solved by intervening on coding, data, and other feature of the digital architectures. Of course, the distinction between governance and technical arrangements is supposed to be purely analytical. It does not fit all the cases, which could present both aspects in different degrees. The same conception of AI as 'socio-technical' systems looks at the interpenetration of social and technical aspects, pointing out how digital infrastructures are means of governance in themselves embedding social norms and values. As recently noted, it is recommendable that technical tools will be integrated and shaped by a "wider governance process" (Ayling and Chapman2021: 16). However, as detailed in the next section, currently available tools tend to focus primarily or quite exclusively on governance or technical features. Hence, the distinction also has empirical validity.

Among technical arrangements we can distinguish the following different categories:

a) Metrics: Many tools consist in the development of metrics that quantify ethical and human rights-based principles through some mathematical or definition and measure to what extent the actual performance of an AI system diverts from some acceptable threshold.
For example, the fairness of an AI system, in terms of non-discrimination, is often calculated as the ratio of false-positive and false-negative between different social categories on a given protected attribute such as sex, race, religion (Fedelman et al. 2015). However, this method requires to compare algorithms' predictions with data on the actual behavior of the cases processed (typically, has been the subject able to pay back the loan, has the subject committed another crime, within a given period), which are not always available (Agarwal 2018). For this reason other approaches have been developed based on the ratio of favourable labelling between different social categories on a given protected attribute. In this case, a favorable label indicates that the classification under a label value corresponds to an outcome providing an advantage to the recipient (s/he will pay back the loan, s/he will not commit crimes in the future). To this purpose, different conceptions of fairness have been quantified into measurable terms. Thus, we can refer to 'demographic parity' fairness, according to which the rate of favourable labelling should be as

close as possible to the percentage of each social category in the population; 'equal opportunity' fairness, which requires that favourable labelling occurs in the same percentage within each social category; 'individual fairness' meaning that the system should returns the same output for subjects with same features except the protected attribute (Fraenkel 2020, Lee et al. 2021, Bellamy et al. 2018). Metrics have been developed also for explainability, (Hofmann et al. 2008, Guidotti et al 2018); safety (Cheng 2021), security (Nicolae 2019) or more broader key ethical indicator (KEI) (Lee et al. 2021).

b) Model Exploration: Other tools could be grouped since they allow to explore the model underlying system decisions. In so doing, they enable a better understanding of the relationship between the input and output of the system, and in some cases, overcome the 'black box' effect characterizing deep neural networks machine learning models. Within this 'family' of tools, we can distinguish two major subgroups.
The first one gathers tools resorting to metrics to better understand the role of each feature in the model. In this regard, many tools rely on 'Shapley values' to calculate the contribution of each variable in the dataset to the prediction of the target variable (Lundberg and Lee 2017, Datta et al. 2017). Sometimes, these metrics are integrated into interactive exploratory tools, which facilitate the understanding of the machine learning model also offering the possibility to alter the value of one or more features on a case to test the effect on the final classification, as in the case of Google's What If Toolkit .
The second subgroup relies on 'reverse engineering' methods, that is a "process of articulating the specifications of a system through a rigorous examination drawing on domain knowledge, observation, and deduction to unearth a model of how that system works" (Detta et al. 2017:1, Epstein 2018).

c) Pre-processing Tools: Pre-processing Tools refer to the manipulation of data occurring before they are processed to perform AI system tasks. They could consist in Re-weighing (Kamiran & Calders, 2012); Optimized pre-processing (Calmon et al., 2017); Learning fair representations (Zemel et al., 2013); Disparate impact remover (Feldman et al., 2015) techniques to comply with fairness requirements or anonymizations or pseudo-anonymization, data minimization methods to address privacy concerns (Antignac et al 2016).

d) Post-processing Tools: Post-Processing Tools intervene on AI systems after that a model is trained, and predictions are made. Talking about fairness, post-processing tools identify techniques such as Equalized odds post-processing (Hardt et al., 2016), Calibrated equalized odds postprocessing (Pleiss et al., 2017) Reject option classification (Kamiran et al., 2012), which modify algorithm's predictions so that they could comply with pre-established fairness thresholds. Further, postprocessing tools could be employed to enforce differential privacy formal definition by adding noise to queries and prevent de-anonymization attack (Dwork and Roth 2014).

e) *Open Standards:* The adoption of open standards could be crucial in terms of reproducibility and explainability of AI systems. Tools such as Open Neural Network Exchange, Neural Network Exchange Format, and Predictive Model Markup Language, even if developed to increase the interoperability of machine learning models, provide abstractions and descriptions of the main components (structures, operations, parameters) of such models. In so doing, they can facilitate the interpretability of AI systems.

Among governance arrangements we can distinguish:

a) *Information and Documentation Provision:* Many of the tools analyzed require preparing accurate information and documentation to be delivered to the general public, users, customers, and internal or external oversight bodies. In most cases, the tools do not provide detailed

instructions or methodologies on how to acquire the necessary data and structure the report, especially if they consist of a broad framework embracing the AI system as a whole.
In the last few years, several attempts have been made to formalize instruments to deliver information and documentation on AI systems (o some their component) and increase custumers and public trust, such as Datasheet (Gebru et al. 2018), Dataset Nutrition Label (Holland et al. 2018); Model Cards (Mitchell 2019), Factsheet (Arnold et al. 2019, Richards et a.l 2020). This latter appears to be the most promising tool for the purpose of this paper. The Factsheet, indeed are designed to provide "a collection of information about how an AI model or service was developed and deployed", summarizing "key characteristics […] for use by a variety of stakeholders" (Richards et a.l 2020:2), including information on some of the requirements discussed above such as safety, security, fairness, and explainability. This methodology provides practical guidance on delivering proper documentation, even if, stressing contextuality is far from providing a standard template.

b) *Assessing and Reviewing Procedures*: The last few years have witnessed a flourishing of initiatives setting mechanisms to assess and review what organizations developing AI systems have done to comply with ethical and human rights-based principles. For the most part, they consist of checklists asking organizations if they have put in place proper arrangements to address a wide range of requirements, as in the case of the Canada Algorithmic Impact Assessment[8] or the European Assessment List for Trustworthy Artificial Intelligence[9]. These instruments are extremely useful to highlight weak points and shortcomings in organizations' ethical approach. However, they provide little guidance on how to set up proper mechanisms.
More detailed methodologies to assess and review ethical and human-right based compliance could be found in more specific instruments such as the ICO's Guide to the General Data Protection Regulation (GDPR)[10], Oetzel and Speikermann (2014) privacy impact assessments, which support organization in building a clear structure of role, responsibilities and processes.

c) *Oversights Procedures*: Tools aimed at setting oversight procedures could be distinguished in those establishing mechanisms to intervene and take control of AI system operations and those seeking to supervise AI systems' development and management processes.
Among the first group, the European Union The Assessment List For Trustworthy Artificial Intelligence (ALTAI) identifies three primary governance mechanisms ensuring human oversight, namely human-in-the-loop (HITL), human-on-the-loop (HOTL), or a human-in-command (HIC). HITL refers to "the capability for human intervention in every decision cycle of the system"; HOTL stand for the "capability for human intervention during the design cycle of the system and monitoring the system's operation", and HIC means "the ability to decide when and how to use the system in any particular situation" (EU HLEG 2020:8), including the decision to not use an AI system in a particular situation. ALTAI provides some further clues asking for example, about the training on human oversight, the establishment of detection and response mechanisms for undesirable adverse effects of the AI system or procedures to abort unsafe operations safely. Although this level of detail is perfectly appropriate with the function of an assessment list, it is not sufficient to guide practitioners in establishing a proper governance mechanism ensuring human oversight.
Regarding tools supervising AI systems' ethical development and management, establishing an ethical committee is one of the most common measures to ensure the alignment of AI systems development with ethical principles and human rights-based principles. Even in this case, however, little details are given on how effectively implement this instrument. In this regard, the Accenture's Building Data And Ai Ethics Committees guide figure as a notable exception. This

[8] https://open.canada.ca/aia-eia-js/?lang=en
[9] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence
[10] https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/

document clarifies the main options concerning key aspects such as the composition, scope, organizational position, powers and procedures of an ethical committee, highlighting pros and cons for each of them.

d) Participation Procedures: very often, ethical codes, guidelines, and recommendations stress the relevance of involving stakeholders in the development of AI systems to achieve trustworthiness and ensure the AI application corresponds to the real needs of social actors. This aspect is caught by the requirement '*Inclusiveness in Design*' under the fairness principle, but it also has evident implications regarding accountability and human oversights. Despite this declared relevance, and despite a long tradition of studies and practices in the close field of Human and Computer Interaction on inclusive design, such as Participatory Design, Value-centered Design, etc., stakeholder participation in the development and management of AI system figure as an overlooked dimension.

5. Combining Principles, Requirements and Tools

Table 1 summarizes the review carried out in this paper by combining principles, requirements and tools in terms of values, criteria and indicators, providing an overview of methods and instruments to deal with ethical and human rights-based requirements, organized at different levels of abstraction, from general kind of action to more granular techniques and methodologies, to specific ready-to-use tools.
This mapping exercise could be the first step to develop an overall and more detailed framework to support practitioners in compliance with ethical and human-rights-based AI development principles, offering them a repertoire of instruments to adopt, or at least suggesting ways forward to develop their own solutions.
Furthermore, the outcome of this research allows us to move forward the conversation on how to translate AI principles into organizational and practices, embedding them into the socio-techncal design of AI system:

1) A first consideration concerns the role of documentation and information. Providing accurate information and documentation on as many aspects of an AI system is the simpler and more direct way to realize a trustworthy AI environment and deal with ethical and human rights-based requirements. Documentation serves transparency purpose, but it is also a precondition to reach other principles. For example, consider how vital information about governance structures and decision-making processes is for accountability, informed consent for privacy, or regular reporting on vulnerability to increase security. The first e foremost relevant condition to ensure compliance with human rights and ethical principles in AI environment is the possibility to scrutinize AI system in order to "understand what it was designed to do, how it was designed to do that, and why it was designed in that particular way instead of some other way" (Kroll 2018). As recently noted, "this kind of scrutiny will only be possible through a combination of tools or processes that facilitate auditing, transparent development, education of the public, and social awareness of developers" (Morley and Floridi 2020: 2155).

Table 1 (1/6) Requirements and Tools for Transparency Principle

Principle	Transparency					
Requirements	*Information*	*Notification*	*Traceability*	*Reproducibility*	*Explainability*	*Reporting*
Tools (General)	Documentation	Documentation	Documentation Open Standard	Documentation Open Standard	Documentation Open Standard Metrics Model Exploration	Documentation
Tools (Techniques, Methods)	Model Card Data/fact Sheet		Provenance Data Model Logging System		Shapley Value Accumulated Local Effects (ALE) Contrastive Explanation Reverse Engineering Counterfactual explanation	
Tools (Examples of specific cases)	Model Cards Dataset Nutrition Label Datasheet Factsheet		PROV W3C Open Provenance Model Open ML Whole Tale	The Turing Way The Machine Learning Reproducibility Checklist XAI Library	What If (Google) Alibi Turing Box Ai Explainability 360 (IBM)	

Table 1 (2/6) Requirements and Tools for Fairness Principle

Principle	Fairness			
Requirements	*Bias Prevention (Non Discrimination)*	*Data Representativeness*	*Inclusive Benefit Distribution*	*Inclusive Design*
Tools (General)	Documentation Metrics Model Exploration Data Pre-Processing Data Post - Processing	Documentation Metrics	Documentation Metrics Participatory Procedures	Participatory Procedures
Tools (Techniques, Methods)	Disparate impact Statistical parity difference Average odds difference Equal opportunity difference Re-weighing Optimized pre-processing Learning fair representations Adversarial debiasing Equalized odds post-processing Calibrated eq. odds postprocessing Reject option classification Reverse Engineering	Model Card Data/fact Sheet	Well-Being Metrics Impact Assessment Value Based Design	Inclusive Design Team Participatory Design
Tools (Examples of specific cases)	What If Turing Box Fairlearn Aequitas	Datasheet Data Nutrition Label	IEEE 7010 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being	Co-designing checklists to understand organizational challenges and opportunities around fairness in AI

Table 1 (3/6) Requirements and Tools for Accountability Principle

Principle	Accountability				
Requirements	*Verification and Validation*	*Impact assessment*	*Auditing*	*Appeal and Remedy Procedures*	*Liability and Legal Responsibility*
Tools (General)	Documentation Metrics Dynamic Testing	Documentation Review Procedures	Documentation Review Procedures	Review Procedures Oversight Procedures	Documentation Review Procedures Oversight Procedures
Tools (Techniques, Methods)	Deductive Verification Algorithmic verification	Checklist Risk Assessment Cost-Benefit Analysis Life-cycle Assessment Procurement Process	Expert Committee Code Inspection Data Inspection	Ethical Committee	
Tools (Examples of specific cases)	Model Cards Model Checking	AI Procurement in a Box Algorithmic Impact Assessment (Canada) Algorithmic Impact Assessment (AI NOW) Model Ethical Data Impact Assessment	End-to- End Framework for Internal Algorithmic Auditing Auditing Algorithms	Building Data And Ai Ethics Committees (Accenture)	

Table 1 (4/6) Requirements and Tools for Privacy Principle

Principle	Privacy			
Requirements	*Consent*	*Data Minimization*	*Data Agency*	*Anonymization*
Tools (General)	Documentation	Design Data Pre-Processing	Documentation Design	Data Pre-Processing Data Post-Processing
Tools (Techniques, Methods)		Data Minimizer Federated Learning		Differential Privacy Multi-party Computation Homomorphic encryption
Tools (Examples of specific cases)	ICO's Guide to the General Data Protection Regulation	Data Minimisation: a Language-Based Approach ICO's Guide to the General Data Protection Regulation OpenMined Agile Ethics for AI	ICO's Guide to the General Data Protection Regulation	Uber Differential Privacy OpenMined ICO's Guide to the General Data Protection Regulation Model-Agnostic Private Learning via stability

Table 1 (5/6) Requirements and Tools for Reliability Principle

Principle	Reliability		
Requirements	*Safety*	*Security*	*Resilience*
Tools (General)	Metrics Others	Metrics Data Pre Processing Data Post Processing	Review Procedures Oversight Procedures
Tools (Techniques, Methods)	Dataset Optimization Robust Training	Outlier detection Adversarial detection Drift Detection	Risk assessment Risk Management Plan
Tools (Examples of specific cases)	OpenMined	ART: Adversial Robustness 360 Toolbox Alibi Detect	CERT Resilience Management Model (CERT-RMM) Cyber Resilience Review (CRR)

Table 1 (6/6) Requirements and Tools for Human Oversight Principle

Principle	Human Oversight		
Requirements	*Human Control and Intervention*	*Human Review of Automated Decision*	*Ability to Opt-out*
Tools (General)	Oversight Procedures Metrics	Oversight Procedures Review Procedures	Oversight Procedures
Tools (Techniques, Methods)	Human-in-the-loop (HITL) Human-on-the-loop (HOTL) Human-in-command (HIC) Ethical Committee	Ethical Committee Dedicated Staff Training	
Tools (Examples of specific cases)	Assessment List for Trustworthy Artificial Intelligence Building Data And Ai Ethics Committees (Accenture)	Assessment List for Trustworthy Artificial Intelligence Building Data And Ai Ethics Committees (Accenture)	Assessment List for Trustworthy Artificial Intelligence Building Data And Ai Ethics Committees (Accenture)

2) However, it is worth noting that governance and technical arrangements appear very poorly integrated. Typically, governance arrangements, such as an assessment toolkit or an ethical committee, may require some metrics or the adoption of technical arrangements, such as anonymization. Still, nothing is said about the governance and organizational processes underlying the development of this technical tool (who decides about the technical specification and the unavoidable trade-off between different values? Based on what input, provided by whom? How this choice could be put into question and reviewed). Similarly, technical tools move from some definition of the principle they are called to tackle, but then the task is carried out as a purely technical problem. Once again, most of the time, nothing is said about how to choose the most proper definition of a principle in relation to the effective context in which the application will be used or how to monitor unintended effects and consequences.

3) Furthermore, most of them are technical arrangements aimed at addressing bias prevention, anonymization, or explicability requirements. This finding could be due to the limit of data collection strategy of this work; however, it seems confirmed by the observation advanced in other studies. Morley and his colleagues noted that "post hoc explanations […] seeking to meet the principle of explicability during the testing phase having the greatest range of tools and methods from which to choose" and argued that "the 'problem' of 'interpreting' an algorithmic decision seems tractable from a mathematical standpoint, so the principle of explicability has come to be seen as the most suitable for a technical fix" (Morley et al. 2020:2153). Also, Ayling and Chapman noted that "much attention and research has been focused on metrics like fairness, accountability, explainability and transparency" (Ayling and Chapman 2021). These instruments promise to face ethical challenges in a concrete, quantifiable and relatively easy to implement way, attracting companies' investment and resources. However, this approach reflects a "reductionist understanding" of ethical and human rights-based principles as "mathematical conditions" (Lee et al. 2021), with the risk to create a gap between formalized statistical definition of principles and the real needs of society, law and politics in concrete context (Hutchinson and Mitchell 2019).

4) Another point of concern is the poor participation of external stakeholders. By and large, clear rules and procedures capable of giving stakeholders a real say in the shaping of AI applications are still missing. Furthermore, participation seems focused on a limited set of internal stakeholders, or stakeholders involved in the production or procurement of AI systems. Even users and customers are often included in some phase of the AI development cycle just to provide input that will be used by management and senior staff. As has been noted, even when participatory initiatives are settled, "AI project owners exert executive authority in deciding tactics of participation" and "who are considered stakeholders, what role each stakeholder plays, how they

interact, whether they need to reach a consensus at the end" (Degaldo et al 2021:4, see also Palladino and Santaniello 2021).

6. Conclusions

The emphasis on ethical and human rights-based approach to AI development is motivated primarily by strategic and pragmatical considerations. Governments, and especially the European Union, are afraid that lack of public trust could contain the adoption of this new technology as already happened in the case of GMO and nuclear power, with relevant loss of opportunities in terms of economic growth and geopolitical influence. Private companies, instead, are concerned about the legal, economic, and reputational costs stemming from malfunctioning of their products and the damages they can cause. Despite the claim for an ethical and human rights- based approach is supported by concrete and interested motivations, the path from principles to practices is filled of pitfalls. As seen, companies have focused their efforts and resources in the development of technical instruments capable to return measurable and ready-to-use output. In so doing, they are delegating to their research departments complex decisions related to the values, preferences and needs of affected social actors, as well as their fundamental rights, rule of law and democracy. Also, governments in their effort to set standards and requirement for AI system, are leaving broad discretion to developers. As already mentioned, the European Commission's recent proposal of an Artificial Intelligence Act does not indicate specific arrangements to implement, rather entrust providers of AI systems with the task to identify or develop proper solutions according with most up-to-date and validated scientific knowledge and agreed standards. This is not necessary a flawed strategy. Entrusting developers with the responsibility to identify the more feasible to satisfy specific requirements could ensure the flexibility to follow the technological and scientific pace, without imposing measures that could result outdated or disproportionate in a relatively short time. Developers and deployers can play a crucial role in the development of a trustworthy and human-centric AI, since "the technical community will not only be best placed but will have the sole ability to protect human rights standards [...] precisely because they are the only community able to see the human rights issues that have been hard-wired into" digital technologies (Liddicoat & Doria, 2012, p. 15) and have the necessary know-how to translate ethical and human rights principles in workable technical specifications (Palladino 2021).

The problem rather relies in the fact that the development of technical solutions to ethical issues is not integrated within governance mechanisms ensuring oversights and accountability as well as a meaningful participation of that external stakeholders. If governance and technical arrangements are not coupled according to the socio-technical nature of AI systems, the production of AI ethical tools may result into a mere "technologist" approach, undermining the social dimension of AI development, (Shilton 2015, Floridi 2013), which could create a disturbing confusion between the mathematical check of machine learning algorithms and the needs for social control.

Although a "by design" approach is deemed crucial to ensure the effective safeguarding of human rights and ethical concerns in the digital realm (Suzor et al. 2019, Cath and Floridi 2017), it is increasingly clear that the design dimension is not limited to codes and digital architectures and that it should involve the social dimension of AI development, which includes corporate governance and organizational practices, besides the broader regulatory environment (Shilton 2015). Principles, organizational practices, and technical requirements are all essential elements to reach an effective ethical and human rights-based AI approach, and they require consistent development within a unitary framework.

The mapping, reviewing and combination of principles, requirements and tools already elaborated to deal with AI ethical challenges that has been carried out in this paper allowed to identify the above mentioned hurdles and provides some insights to move forward.

It could be developed further to realize a comprehensive tool-kit supporting practitioner and stakeholders offering them: 1) detailed description of most relevant alternative instruments and

methods to deal with specific AI ethical and human right requirements, highlighting pros and cons of each of them in terms of both social and political implication of technical choices and feasibility; 2) operational guidelines holding together technical, organizational, and governance arrangements; and a system of criteria, indicators, and observables ensuring the effective compliance with ethical and human rights standards in the design of AI systems.

7. Funding And Acknowledgment

The author has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the HUMAN+ COFUND Marie Skłodowska-Curie grant agreement No. 945447.

8. References

Aaron Fraenkel, 2020, Fairness and Algorithmic Decision Making, https://afraenkel.github.io/fairness-book/intro.html

Access Now, 2018, Human Rights in the Age of Artificial Intelligence, https://www.accessnow.org/cms/assets/uploads/2018/11/AI-and-Human-Rights.pdf

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR.

Amnesty International and Access Now, 2018, The Toronto Declaration, available at https://www.torontodeclaration.org/

Antignac, T., Sands, D., & Schneider, G. (2017, May). Data minimisation: a language-based approach. In IFIP International Conference on ICT Systems Security and Privacy Protection (pp. 442-456). Springer, Cham.

Artificial Intelligence Industry Alliance, 2019, Artificial Intelligence Industry Code of Conduct (Consultation Version)' (2019) (See Principle 8, English translation available upon request) https://www.secrss.com/articles/11099

Ayling, J., Chapman, A. Putting AI ethics to work: are the tools fit for purpose?. AI Ethics (2021). https://doi.org/10.1007/s43681-021-00084-x

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.

Berkman Center (2020_ Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI: https://cyber.harvard.edu/publication/2020/principled-ai

Boiler, G., (2018) Artificial Intelligence: The Great Disruptor. Washington, DC: The Aspen Institute.

Bryson, J.J., and Theodorou, A. (2019) How Society Can Maintain Human-Centric Artificial Intelligence. In Toivonen-Noro, M., and Saari., E. (eds.) Human-Centered Digitalization and Services. Springer.

Cath, C., and Floridi, L. (2017) The Design of the Internet's Architecture by the Internet Engineering Task Force (IETF) and Human Rights. Sci Eng Ethics, 23, 449–468.

Cheng, C. H., Knoll, A., & Liao, H. C. (2021). Safety Metrics for Semantic Segmentation in Autonomous Driving. arXiv preprint arXiv:2105.10142.

Cookson, C. (2018, June 9). Artificial intelligence faces public backlash, warns scientist. Financial Times. Retrieved from https://www.ft.com/content/0b301152-b0f8-11e8-99ca-68cf89602132

Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (pp. 598-617). IEEE.

Delgado, F., Yang, S., Madaio, M., & Yang, Q. (2021). Stakeholder Participation in AI: Beyond" Add Diverse Stakeholders and Stir". arXiv preprint arXiv:2111.01122.

DeNardis, L. (2013). Protocol Politics: The Globalization of Internet Governance. Boston: MIT Press.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci., 9(3-4), 211-407.

Epstein, Z., Payne, B. H., Shen, J. H., Hong, C. J., Felbo, B., Dubey, A., ... & Rahwan, I. (2018). TuringBox: An experimental platform for the evaluation of AI systems. In IJCAI 2018 (pp. 5826-5828). International Joint Conferences on Artificial Intelligence.

EU HLEG, 2019, Ethical guidelines for Trustworthy AI, https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

European Commission (2019) Communication 168 "Building Trust in Human-Centric Artificial Intelligence." https://digital-strategy.ec.europa.eu/en/library/communication-building-trust-human-centric-artificial-intelligence

European Commission (2020) Communication 65 "White Paper on Artificial Intelligence – A European Approach to Excellence and Trust.", https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en

European Commission (2021) proposal for an "Artificial Intelligence Act", https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).

Floridi, L. (2013) Distributed Morality in an Information Society, Science and Engineering Ethics, 19(3), 727–743

Floridi, L., Cowls, J., Beltrametti, M., et al. (2018) AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines 28, 689–707

Future of Life Institute, 2017, Asilomar AI Principles https://futureoflife.org/ai- principles/?cn-reloaded=1

Google, 2018, AI at Google: Our Principles, https://www.blog.google/technology/ai/ ai-principles/

Governance: The example of the European Commission High-Level Expert Group on Artificial

Greene, D., Hoffmann, A.L., and Stark, L. (2019) Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In Proc. 52nd Hawaii International Conference on System Sciences, 2122–2131

Hallensleben et al. (2020) From Principles to Practice. An Interdisciplinary Framework to Operationalise AI Ethics. Report, AI Ethics Impact Group: https://www.ai-ethics-impact.org/resource/blob/1961130/c6db9894ee73aefa489d6249f5ee2b9f/aieig---report---download-hb-data.pdf

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems 29, pp. 3315–3323, Barcelona, Spain, December 2016.

Hildebrandt, M. (2019) Law for Computer Scientists, Oxford University Press.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Hutchinson, B., Mitchell, M.: 50 years of test (Un)fairness: les- sons 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.

IBM, 2019, Everyday Ethics for AI, https://www.ibm.com/watson/assets/duo/pdf/ everydayethics.pdf

Intelligence, Telecommunications Policy, Volume 45, Issue 6.

Japanese Cabinet Office, Council for Science, 2019, Technology and Innovation, 'Social Principles of Human-Centric Artificial Intelligence, https://www8.cao.go.jp/cstp/ english/humancentricai.pdf

Jobin, A., Ienca, M., and Vayena, E. (2019) The Global Landscape of AI Ethics Guidelines. Nat Mach Intell, 1, 389–399;

Kaefer, F., Roper, J., & Sinha, P. (2015). A software-assisted qualitative content analysis of news articles: example and reflections. In Forum Qualitative Sozialforschung/Forum: Qualitative Social Research (Vol. 16, No. 2, p. 20). DEU.

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In IEEE Interna- tional Conference on Data Mining, pp. 924–929, 2012. doi: https://doi.org/10.1109/ICDM.2012.45.

Kitchin, R. (2017) Thinking Critically About and Researching Algorithms. Information, Communication & Society, 20(1), 14–29.

Lessig, L. (2006). Code 2.0. New York: Basic Books.

Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).

M.Arnold,R.K.E.Bellamy,M.Hind,S.Houde,S.Mehta,A.Mojsilović,R.Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. FactSheets: Increasing trust in AI services through supplierâĂŹs declarations of conformity. IBM Journal of Research & Development, 63(4/5), Sept. 2019.

M.Mitchell,S.Wu,A.Zaldivar,P.Barnes,L.Vasserman,B.Hutchinson,E.Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, Atlanta, USA, Jan. 2019.

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

Microsoft, 2018, AI Principles https://www.microsoft.com/en-us/ai/our-approach-to- ai

Mittelstadt, B. (2019) Principles Alone Cannot Guarantee Ethical AI. Nat Mach Intell 1, 501–507;

Monetary Authority of Singapore, 2019, Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector, <http://www.mas.gov.sg/~/media/MAS/ News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>.

Morley, J., Floridi, L., Kinsey, L. et al. (2020) From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Sci Eng Ethics 26, 2141–2168.

Musiani, F., Cogburn, D.L., DeNardis, L., and Levinson, N.S. (2016) The Turn to Infrastructure in Internet Governance. New York: Palgrave MacMillan.

Nicolae, M. I., Sinn, M., Tran, M. N., Buesser, B., Rawat, A., Wistuba, M., ... & Edwards, B. (2018). Adversarial Robustness Toolbox v1. 0.0. arXiv preprint arXiv:1807.01069.

Palladino N. (2021) The role of epistemic communities in the "constitutionalization" of Internet

Palladino N. and Santaniello M. (2021) Legitimacy, Power and Inequalities in Multistakeholder Internet Governance. Palgrave McMillan: Cham.

Partnership on AI, 2016, Tenets https://www.partnershiponai.org/tenets/

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Wein- berger, K. Q. On fairness and calibration. In Advances in Neural Information Processing Systems, pp. 5680–5689, 2017.

Renda, A. (2019) Artificial Intelligence, Ethics Governance and Policy Challenges. Brussels: CEPS

Rieke, A., Bogen, M., and Robsinson, D.G. (2018) Public Scrutiny of Automated Decisions: Early Lessons and Emerging Methods, an Upturn and Omidyar Network Report.

Russell, S., and Norvig, P. (2003) Artificial Intelligence: A Modern Approach. Prentice Hall Series in Artificial Intelligence. Upper Saddle River, NJ: Prentice Hall, Pearson Education.

S.Holland,A.Hosny,S.Newman,J.Joseph,andK.Chmielinski.The dataset nutrition label: A framework to drive higher data quality standards. arXiv:1805.03677, May 2018.

Schiff, D., Rakova, B., Ayesh, A., Fanti, A., and Lennon, M. (2020) Principles to Practices for Responsible AI: Closing the Gap: https://arxiv.org/abs/2006.04707v1.

Shilton, K. (2015) "That's Not an Architecture Problem!": Techniques and Challenges for Practicing Anticipatory Technology Ethics. In iConference 2015 Proceedings 7. iSchools.

Smart Dubai, 2018, AI Ethical Principles and Guidelines, https://smartdubai.ae/initiatives/ai-principles-ethics

Suzor, N., Dragiewicz, M., Harris, B., Gillett, R., Burgess, J., and Van Geelen, T. (2019) Human Rights by Design: The Responsibilities of Social Media Platforms to Address Gender-Based Violence Online. Policy & Internet, 11, 84–103.

T.Gebru,J.Morgenstern,B.Vecchione,J.W.Vaughan,H.Wallach,H.Daumé, III, and K. Crawford. Datasheets for datasets. In Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Workshop, Stockholm, Sweden, July 2018.

The Public Voice Coalition, 2018, Universal Guidelines for Artificial Intelligence, https://thepublicvoice.org/ai-universal-guidelines/

Think 20, 2018, Future of Work and Education for the Digital Age, https://www.g20- insights.org/wp-content/uploads/2018/07/TF1-1-11-Policy-Briefs_T20ARG_Towards-a- G20-Framework-For-Artificial-Intelligence-in-the-Workplace.pdf

UK House of Lords, 2018, Select Committee on Artificial Intelligence, 'AI in the UK: Ready, Willing and Able?' (2018) Report of Session 2017-19 <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf>

UNI Global Union, 2017, Top 10 Principles for Ethical Artificial Intelligence, http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf

University of Montreal, (2018) 'Montreal Declaration for a Responsible Development of Artificial Intelligence' https://www.montrealdeclaration-responsibleai.com/the- declaration

van Dijck, N., and Casiraghi, S. (2020) The "Ethification" of Privacy and Data Protection Law in the European Union. The Case of Artificial Intelligence. Brussels Privacy Hub Working Paper Vol. 6, N. 22

Whittaker, M., et al. (2018) AI Now Report 2018, https://ainowinstitute.org/AI_Now_2018_Report.pdf;

Yeung, K., Howes, A., and Pogrebna, G. (2019) AI Governance by Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing, In Dubber, M.D., Pasquale, F., and Das, S., Oxford Handbook on AI Ethics. Oxford: Oxford University Press.

ANNEX 1: LIST OF TOOLS ANALYZED

1. Aaron Fraenkel, 2020, Fairness and Algorithmic Decision Making, https://afraenkel.github.io/fairness-book/intro.html
2. Aequitas [https://github.com/dssg/aequitas; Saleiro, P., Kuester, B., Hinkson, L., London, J., Stevens, A., Anisfeld, A., ... & Ghani, R. (2018). Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:1811.05577]
3. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018, July). A reductions approach to fair classification. In International Conference on Machine Learning (pp. 60-69). PMLR.
4. Agile Ethics for AI [https://trello.com/b/SarLFYOd/agile-ethics-for-ai-hai]
5. AI and Big Data: A blueprint for a human rights, social and ethical impact assessment [AI and Big Data: A blueprint for a human rights, social and ethical impact assessment]
6. AI Explainability 360 [https://aix360.mybluemix.net/; Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012]
7. AI Procurement in a Box [https://www.weforum.org/ reports/ai-procurement- in-a-box/]
8. Algorithmic Impact Assessment (AIA) (Canandian Government) [https://www.canada.ca/en/ government/system/digit al-government/modern- emerging-technologies/ responsible-use-ai/algorithmic-impact-assessment.html
9. Algorithmic Impact Assessments: A Practical Framework For Public Agency Accountability (AI NOW Institute) [https://ainowinstitute.org/ aiareport2018.pdf]
10. Alibi [https://github.com/SeldonIO/alibi]
11. Alibi Detect [https://github.com/SeldonIO/alibi-detect]
12. Antignac, T., Sands, D., & Schneider, G. (2017, May). Data minimisation: a language-based approach. In IFIP International Conference on ICT Systems Security and Privacy Protection (pp. 442-456). Springer, Cham.
13. ART: Adversial Robustness 360 Toolbox [https://github.com/Trusted-AI/adversarial-robustness-toolbox]
14. Assessment List for Trustworthy Artificial Intelligence [https://futurium.ec.europa. eu/en/european-ai-allia nce/pages/altai-asses sment-list-trustworthy- artificial-intelligence]
15. Assessments: Cyber Resilience Review (CRR) [https://us-cert.cisa.gov/resources/assessments]
16. Auditing Algorithms website [http://auditingalgorithms.science/?page_id=89]
17. Bassily, R., Thakurta, A. G., & Thakkar, O. D. (2018). Model-agnostic private learning. Advances in Neural Information Processing Systems.

18. Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2018). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv preprint arXiv:1810.01943.
19. CERT Resilience Management Model (CERT-RMM) [https://resources.sei.cmu.edu/library/asset-view.cfm?assetID=514489]
20. Cheng, C. H., Knoll, A., & Liao, H. C. (2021). Safety Metrics for Semantic Segmentation in Autonomous Driving. arXiv preprint arXiv:2105.10142.
21. Clarke, E.M., Grumberg, O., Peled, D.: Model Checking. MIT Press, Cambridge, MA (1999)
22. CleverHans [https://github.com/cleverhans-lab/cleverhans]
23. Closing the AI Accountability Gap: Defining an End-to- End Framework for Internal Algorithmic Auditing [https://dl.acm.org/doi/pdf/ 10.1145/3351095.33728 73]
24. Contrastive Explanation Method (CEM) [Dhurandhar, A., Chen, P. Y., Luss, R., Tu, C. C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. arXiv preprint arXiv:1802.07623.
25. Dataset Nutrition Label [S.Holland,A.Hosny,S.Newman,J.Joseph,andK.Chmielinski.The dataset nutrition label: A framework to drive higher data quality standards. arXiv:1805.03677, May 2018]
26. Datasheet [T.Gebru,J.Morgenstern,B.Vecchione,J.W.Vaughan,H.Wallach,H.Daumé, III, and K. Crawford. Datasheets for datasets. In Proceedings of the Fairness, Accountability, and Transparency in Machine Learning Workshop, Stockholm, Sweden, July 2018].
27. Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In 2016 IEEE symposium on security and privacy (SP) (pp. 598-617). IEEE.
28. DeepExplain [https://github.com/marcoancona/DeepExplain]
29. Dennis, L. A., Fisher, M., Lincoln, N. K., Lisitsa, A., & Veres, S. M. (2016). Practical verification of decision-making in agent-based autonomous systems. Automated Software Engineering, 23(3), 305–359. https://doi.org/10.1007/s10515-014-0168-9
30. Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital journalism, 3(3), 398-415.
31. Equity Evaluation Corpus (EEC) [https://github.com/algorithmicbiaslab/expanded-equity-corpus]
32. FactSheet [M.Arnold,R.K.E.Bellamy,M.Hind,S.Houde,S.Mehta,A.Mojsilović,R.Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, and K. R. Varshney. FactSheets: Increasing trust in AI services through suppliers declarations of conformity. IBM Journal of Research & Development, 63(4/5), Sept. 2019]
33. Fairlearn [https://github.com/fairlearn/fairlearn]
34. Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015, August). Certifying and removing disparate impact. In proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 259-268).
35. Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
36. IEEE 7010 Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-being [Schiff, D., Ayesh, A., Musikanski, L., & Havens, J. C. (2020, October). IEEE 7010: A new standard for assessing the well-being implications of artificial intelligence. In 2020 IEEE international conference on systems, man, and cybernetics (SMC) (pp. 2746-2753)]
37. Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).

38. Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).
39. 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.
40. 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.
41. Model Card [M.Mitchell,S.Wu,A.Zaldivar,P.Barnes,L.Vasserman,B.Hutchinson,E.Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, Atlanta, USA, Jan. 2019].
42. Model Ethical Data Impact Assessment [http://informationaccountability.org/publications/]
43. Neural Network Exchange Format (NNEF) [https://www.khronos.org/api/nnef]
44. Oetzel, M. C., & Spiekermann, S. (2014). A systematic methodology for privacy impact assessments: a design science approach. European Journal of Information Systems, 23(2), 126-150.
45. Open Neural Network Exchange [https://onnx.ai/]
46. OpenMined [https://www.openmined.org/]
47. OpenML [Vanschoren, J.; Van Rijn, J.; Bischl, B.; Torgo, L. OpenML: Networked science in machine learning. SIGKDD 2014, 15, 49–60].
48. Predictive Model Markup Language (PMML) [http://dmg.org/]
49. Prov W3C [Souza, R.; Azevedo, L.; Lourenço, V.; Soares, E.; Thiago, R.; Brandão, R.; Civitarese, D.; Brazil, E.; Moreno, M.; Valduriez, P. Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering. In Proceedings of the 2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS), Denver, CO, USA, 17 November 2019; pp. 1–10]
50. Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms.
51. The Machine Learning Reproducibility Checklist [https://www.cs.mcgill.ca/~jpineau/ReproducibilityChecklist.pdf]
52. The open provenance model [Moreau, L.; Freire, J.; Futrelle, J.; McGrath, R.E.; Myers, J.; Paulson, P. The open provenance model: An overview. In International Provenance and Annotation Workshop; Springer: Berlin/Heidelberg, Germany, 2008; pp. 323–326].
53. The Turing Way [https://github.com/alan-turing-institute/the-turing-way]
54. Turing Box [Epstein, Z., Payne, B. H., Shen, J. H., Hong, C. J., Felbo, B., Dubey, A., ... & Rahwan, I. (2018). TuringBox: An experimental platform for the evaluation of AI systems. In IJCAI 2018 (pp. 5826-5828). International Joint Conferences on Artificial Intelligence.]
55. Uber Differential Privacy [https://github.com/uber-archive/sql-differential-privacy]
56. Visser, W., Havelund, K., Brat, G.P., Park, S., Lerda, F.: Model checking programs. Autom. Softw. Eng. 10(2), 203–232 (2003)
57. What If [https://github.com/pair-code/what-if-tool]
58. Whole Tale [https://wholetale.org/; Brinckman, A.; Chard, K.; Gaffney, N.; Hategan, M.; Jones, M.B.; Kowalik, K.; Kulasekaran, S.; Ludäscher, B.; Mecum, B.D.; Nabrzyski, J.; et al. Computing environments for reproducibility: Capturing the “Whole Tale”. Future Gener. Comp. Syst. 2019, 94, 854–867]
59. XAI Library [https://github.com/EthicalML/xai]

60. Corporate Digital Responsibility [Lobschat, L., Mueller, B., Eggers, F., Brandimarte, L., Diefenbach, S., Kroschke, M., & Wirtz, J. (2021). Corporate digital responsibility. Journal of Business Research, 122, 875-888.
61. Data Ethics Framework [https://www.gov.uk/government/publications/data-ethics-framework/data-ethics-framework-legislation-and-codes-of-practice-for-use-of-data]
62. White Paper on Data 2020 Ethics in Public Procurement of AI- based Services and Solutions [https://dataethics.eu/wp- content/uploads/datae thics-whitepaper-april-2020.pdf]
63. Improving Social Responsibility of Artificial Intelligence by Using ISO 26000 [Zhao, W. W. (2018, September). Improving social responsibility of artificial intelligence by using ISO 26000. In IOP Conference Series: Materials Science and Engineering (Vol. 428, No. 1, p. 012049). IOP Publishing]