Z-Inspection®: A Process to Assess Trustworthy AI

Roberto V. Zicari *(1)*, John Brodersen *(4)(9)*, James Brusseau *(8),* Boris Düdder (*6)*, Timo Eichhorn *(1)*, Todor Ivanov *(1)*, Georgios Karagiras *(3)*, Pedro Kringen *(1)*, Melissa McCullough *(1),* Florian Möslein *(7),* Naveed Mushtaq *(1)*, Gemma Roig *(1),* Norman Stürtz *(1)*, Karsten Tolle *(1)*, Jesmin Jahan Tithi *(2)*, Irmhild van Halem *(1)*, Magnus Westerlund *(5)*.

Abstract—The ethical and societal implications of artificial intelligence systems raise concerns. In this paper we outline a novel process based on applied ethics, namely Z-inspection®, to assess if an AI system is trustworthy. We use the definition of trustworthy AI given by the high-level European Commission's expert group on AI. Z-inspection® is a general inspection process that can be applied to a variety of domains where AI systems are used, such as business, healthcare, public sector, among many others. To the best of our knowledge, Z-inspection® is the first process to assess trustworthy AI in practice.

Index Terms—***Artificial Intelligence, Ethics, Law, Society, Machine Learning, Deep Learning, AI-audit, AI ethics, AI policy, Corporate social responsibility, Algorithmic audits, Accountability, Responsible innovation, Z-Inspection.***

I. INTRODUCTION

ARTIFICIAL INTELLIGENCE (AI) is becoming a sophisticated tool in the hands of a variety of stakeholders, including political leaders. Some AI applications, e.g. applications based on Machine Learning (ML) and/or Deep Learning (DL), raise new ethical and legal questions, and in general have a significant impact on society (for the good or for the bad or for both).

Roberto V. Zicari *(1)*, John Brodersen *(4)*, James Brusseau *(8),* Boris Düdder (*6)*, Timo Eichhorn *(1)*, Todor Ivanov *(1)*, Georgios Karagiras *(3)*, Pedro Kringen *(1)*, Melissa McCullough *(1),* Florian Möslein *(7),* Naveed Mushtaq *(1)*, Gemma Roig *(1),* Norman Stürtz *(1)*, Karsten Tolle *(1)*, Jesmin Jahan Tithi *(2)*, Irmhild van Halem *(1)*, Magnus Westerlund *(5)*.

*(1) Frankfurt Big Data Lab, Goethe University Frankfurt, Germany; (2) Intel Labs, Santa Clara, CA, USA; (3) Department of Physiology, Faculty of Medicine, University of Iceland, Reykjavik, Iceland; (4) Section of General Practice and Research Unit for General Practice, Department of Public Heal*th, *Faculty of Health and Medical Sciences, University of Copenhagen, Danemark; (5)Arcada University of Applied Sciences, Helsinki, Finland; (6) Department of Computer Science (DIKU), University of Copenhagen (UCPH), Denmark; (7) Institute of the Law and Regulation of Digitalization, Philipps-University Marburg, Germany; (8) Philosophy Department, Pace University, New York, USA; (9) Primary Health Care Research Unit, Region Zealand, Denmark*

Trust between humans and AI systems is integral to promoting the development and deployment of socially beneficial and responsible AI. Therefore, building trust in AI ranks highly on current political, business, social and legal agendas [1], [2], [3],[4],[5],[6],[7],[8],[19],[10],[11]. In principle as stated in [12] "it is essential to provide evidence of trust that an organization operates with ethical values, to support independent judgment on whether an expectation of ethical behavior is warranted. Mere claims by a company that it can be trusted will clearly not suffice. Mechanisms should be designed to produce reliable evidence of trust." A number of researchers, including Turing Award winner Yoshua Bengio [13], describe the problem when dealing with AI systems as follows: "the process of AI development is often opaque to those outside a given organization, and various barriers make it challenging for third parties to verify the claims being made by a developer. As a result, claims about system attributes may not be easily verified**. "** Assessing trustworthy AI is difficult. In fact "the real-life ethical impact that a technology will have on people, their communities and the planet, can only be fully understood once the product or service is in real-world use." [14]

When considering the Western perspective on ethics, the main problem is that most of the principles proposed for AI ethics are not specific enough to be used in practice. Bengio and colleagues [13] note that: **"**for a sufficiently complex AI system or development process, a wide variety of mechanisms will likely need to be brought to bear in order to adequately substantiate a high-level claim such as "this system was developed in accordance with our organization's ethical principles and relevant laws." To solve this problem, they recommend that "a coalition of stakeholders should create a task force to research options for conducting and funding third party auditing of AI systems". Bengio and colleagues [13] further state that, "techniques and best practices have not yet been established for auditing AI systems". This is confirmed by the high-level European Commission's expert group on AI [15], who recommend that work be done to "develop auditing mechanism for AI systems to allow public enforcement authorities as well as independent third-party auditors to identify potentially illegal outcomes or harmful consequences generated by AI systems, such as unfair bias or discrimination" [16]. This is echoed by a recent White Paper of the European Commission [17] and a report of The

Organization for Economic Co-operation and Development (OECD) [18].

The main contribution of this paper is the definition of a process, Z-Inspection, which can be useful for auditing an AI system. The process can also be used before production of an AI system, helping relevant actors to be aware of the ethical, social, technical and legal risks and pitfalls when implementing an AI system. The rest of the paper is structured as follows: Section II gives the motivation and compares our approach with relevant similar work. Section III presents the main contribution of this paper, the Z-Inspection process; Section IV illustrates some key element of the process with a use case; Section V review related work; and Section VI presents some concluding remarks.

II. Z-Inspection: Motivation

TRUSTWORTHY AI

Z-Inspection takes into account the "Framework for Trustworthy AI" and the seven key requirements that AI systems should meet in order to be deemed trustworthy [15], as defined by the independent High-Level Expert Group of Artificial Intelligence, set by the European Commission, and also confirmed by a recent report of The Organization for Economic Co-operation and Development (OECD) [18]. Z-Inspection applies these principles and requirements by proposing a practical implementable assessment process that can be adapted to specific use cases and domains in practice.

There are already some toolkits available to assess some of the aspects of applied ethics. The toolkit introduced by Anderson et al. [19], is basically a checklist aimed specifically at Government leaders and staff to help them assess and manage algorithm risks. Some of the questions related to assessing risks have been relevant for our work as well. Another example is the toolkit introduced by the Open Roboethics Institute [20] designed for the early design and deployment process. In contrast to the aforementioned, Z-Inspection can be used for auditing and performing an ethical evaluation over time of a deployed AI system. Raji et al. [21] recently introduced a framework for algorithmic auditing that supports artificial intelligence system development end-to-end, to be applied throughout the internal organization development life cycle. Their framework is intended to be used for internal audit, and encompasses five stages— Scoping, Mapping, Artifact Collection, Testing and Reflection (SMACTR). Their framework, however, does not address the Post Audit Stage, i.e. after the AI System has been deployed. Z-Inspection, in contrast, covers the 'Post Audit Stage', and can also be used by investigators from outside the organizations deploying the algorithms. Outcomes, or claim-based "assurance frameworks" widely use in safety-critical auditing contexts such as the Claims, Arguments and Evidence (CAE) framework, do not focus on AI [13] and do not have a specific auditing process/framework for AI in place [22]. The Institute of Internal Auditors (IIA) has recognized that the internal audit profession must understand AI basics, the roles that internal audit can and should play, and AI risks and opportunities [22]. However, in contrast to Z-Inspection, they do not have a specific auditing process/framework for AI in place.

The core idea of our assessment is to create an orchestration process to help teams of skilled experts assessing the ethical, technical and legal implications of using an AI-product/service within a given context. Wherever possible, Z-Inspection allows us to use existing frameworks, checklists, and to "plug in" existing tools to perform specific parts of the verification. The goal is to customize the assessment process for AIs deployed in different domains and in different contexts.

The Trustworthy AI Assessment List defined in [23] presents a self-assessment checklist of questions classified into the seven requirements defined by the independent High-Level Expert Group of Artificial Intelligence, set by the European Commission. Prior to self-assessing an AI system with this Assessment List, a fundamental rights impact assessment (FRIA) should be performed [23], drawing on specific articles in the Charter and the European Convention on Human Rights (ECHR) [24] its protocols and the European Social Charter [25]. We believe that self-assessment is a welcome first step, but we do not believe that self-assessment is entirely sufficient. The risk of conflict of interest within the organization performing the self-assessment is very high. We therefore take this into account in our assessment process as described in the next section.

Our approach is inspired by both theory and practices ("learning by doing"). We have used and tested Z-Inspection by evaluating a non-invasive AI medical device designed to assist medical doctors in the diagnosis of cardiovascular diseases. We will describe the use case in Section IV. We use this case study for illustration of some key points of the Z-Inspection process.

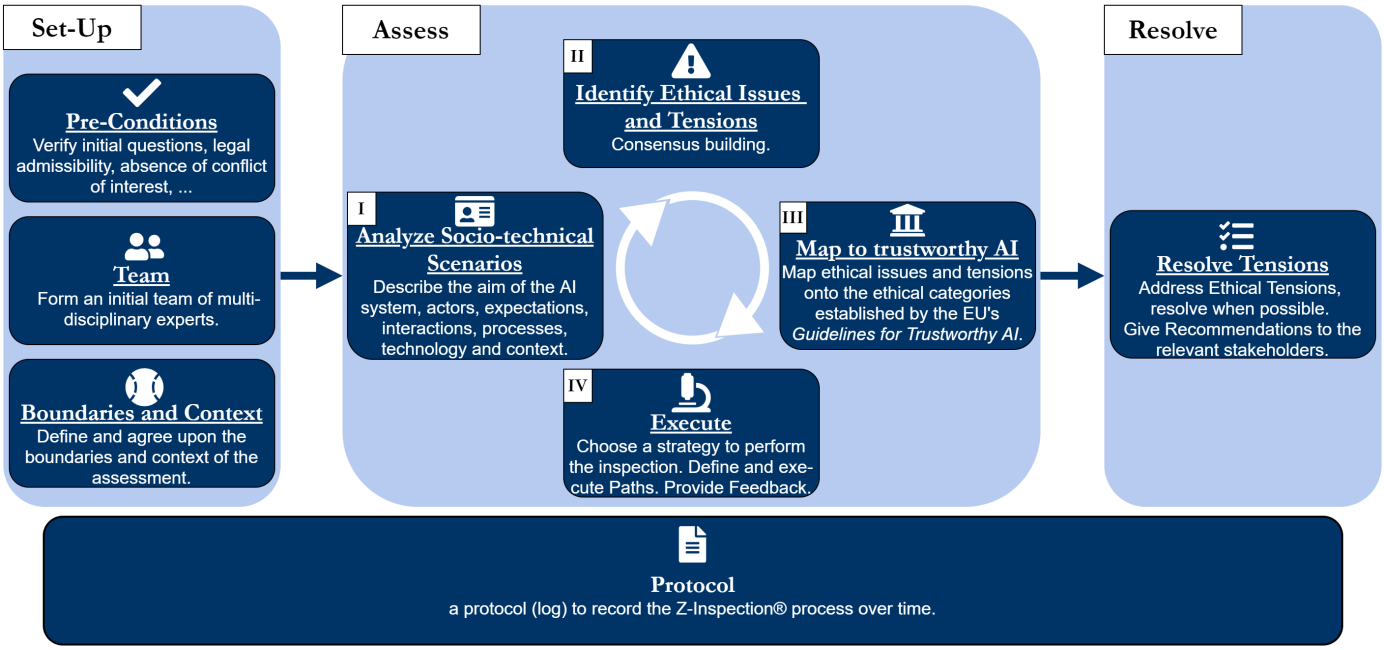

Figure 1. Z-Inspection process in a nutshell

III. Z-INSPECTION METHODOLOGY: ASSESSING TRUSTWORTHY AI IN PRACTICE

Z-Inspection is designed by integrating two well-known approaches: 1) A holistic approach, that aims grasping the whole without consideration of the various parts; and 2) An analytic approach, that aims to consider each part of the problem domain. This is in keeping with the views expressed by Peters et al [14] that "evaluating the ethical impact of a technology's use, [is] not just on its users, but often, also on those indirectly affected, such as their friends and families, communities, society as a whole, and the planet."

The Z-Inspection process in a nutshell is composed of three main phases: The Set Up phase, the Assess phase and the Resolve phase [Fig 1].

The Set Up phase [Chart I] clarifies some pre-conditions, set the team of investigators, help define the boundaries of the assessment, and create a protocol.

The Assess phase [Chart II] is composed of four tasks: I. The Analysis of the usage of the AI system; II. The identification of possible ethical issues, as well as technical and legal issues; III mapping of such issues to the Trustworthy AI ethical values and requirements; IV. The verification of such requirements.

The Resolve phase [Chart III] addresses resulting ethical, technical and legal issues, address when possible ethical tensions, and produce recommendations, when required prescribe a so called ethical AI maintenance over time. Each of these phases is detailed in the rest of this section.

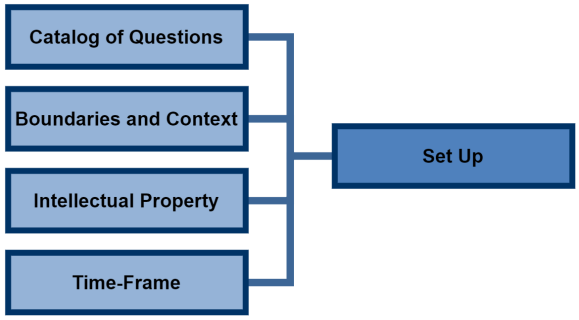

Chart I. The Set Up Phase

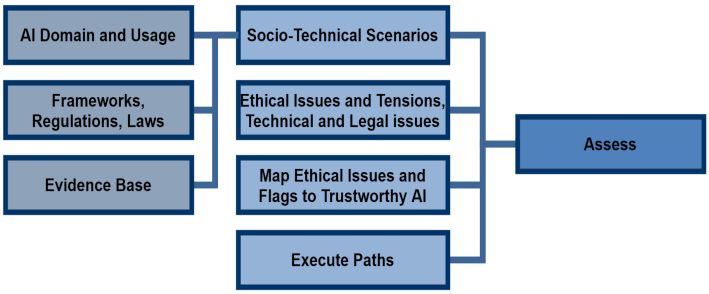

Chart II. The Assess Phase

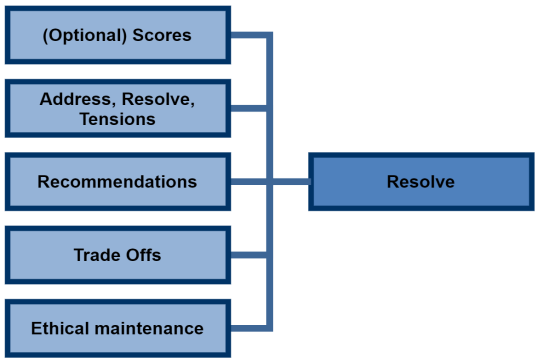

Chart III. The Resolve Phase

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TTS.2021.3066209, IEEE Transactions on Technology and Society

A. THE SET UP PHASE

We developed a catalog of questions (Table 1) that we recommend should be used before starting an assessment, in order to define a clear understanding of the expectations of the various stakeholders involved in the assessment.

Table 1. Catalog of questions for the Set Up Phase

1. Who requested the inspection?
2. Why carry out an inspection?
3. For whom is the inspection relevant?
4. Is it recommended or required (mandatory inspection)?
5. What are the sufficient vs. necessary conditions that need to be analyzed?
6. How are the inspection results to be used? There are different, possible uses of the results of the inspection: e.g. verification, certification, and sanctions (if illegal).
7. Will the results be shared (public) or kept private? In the latter case, the key question is why is it being kept private?

An initial team of multi-disciplinary experts is formed. As with any process, to perform an effective assessment it is necessary to choose team members with required skills, define how many resources will be used, for how long, and at what costs. The key is to engage different stakeholders in adopting a common multi-perspective view, yet focusing on a systematic discussion during various phases of the assessment from AI design to ethical maintenance. It is very important that each member builds up relevant solid knowledge of the other involved disciplines. The composition of the team is a dynamic process and the choice of the experts, their skills, background and roles, have a significant ethical implication for the overall process; the ethical verification of the team is part of the catalog of questions.

In our opinion, one cornerstone of being able to conduct an independent AI Ethical assessment is the absence of conflict of interests both direct and indirect. Therefore, we recommend before starting an assessment that an external, neutral stakeholder verifies the absence of conflicts of interests (Table 2). During the assessment, conflicts of interests may arise and therefore there is need for a constant monitoring. For the entire duration of the assessment, a protocol (log) of the process is created and maintained to record information of the inspection process over time. The protocol can be shared with relevant stakeholders at any time to ensure transparency of the process and the possibility of re-doing actions.

Table 2. Checklist for Conflicts of Interest

Q. Are there any conflict of interests?

This requires assurance that:
a) no conflict of interests exists between the inspectors and the entity/organization to be examined;
b) no conflict of interests exists between the inspectors and vendors of tools/toolkits/frameworks/data platforms to be used in the inspection;
c) any potential bias of the team of inspectors is assessed. If teams are too homogeneous, the likelihood of "group-thinking" and one-dimensional perspectives arises – thereby increasing the risk of leaving the whole assessment vulnerable to inherent biases and unwanted discrimination.
Answers to a),b),c) result in a **'GO'** if all three above are satisfied**,** a **'Still GO'** with restricted use of specific tools, if b) above is not satisfied, and a **'NoGO'** if a) or c) are not satisfied.

Definition of boundaries and context.

The Set Up phase continues with the definition of the boundaries and the context of the assessment. The responsible use of AI, (processes and procedures, protocols and mechanisms and institutions to achieve it) inherits properties from the wider political and institutional contexts. In this respect, the following aspects need to be taken into consideration: AI is not a single element; AI is not in isolation; AI is dependent on the domain where it is deployed; AI is part of one or more (digital) ecosystems; AI is part of processes, products, services, etc.; AI is related to people and data.

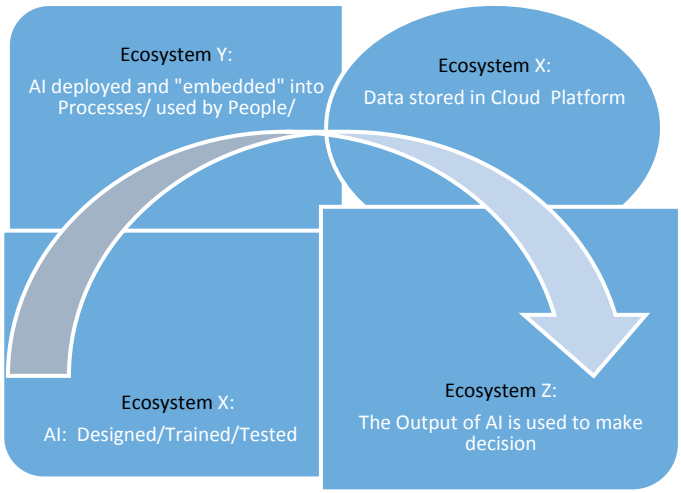

Figure 2. An Example of Ecosystems

In our process, the concept of ecosystems plays an important role in defining the boundaries of the assessment. We define an ecosystem, as applied to our work, as a set of sectors and parts of society, level of social organization, and stakeholders

within a political and economic context. An example of ecosystems is depicted in Figure 2. It is important to note that illegal and unethical are not the same thing, and that both law and ethics are context dependent for each given ecosystem. The legal framework is dependent on the geopolitical boundaries of the assessment.

How to handle Intellectual Property (IP)

When assessing an AI system, it is key to clarify what constitutes Intellectual Property (IP) and how to handle it during the assessment. This may also require looking at the definition of the IP for other parts of the ecosystem(s), and for the entity/company where the AI has been designed/deployed. A word of warning here, this might provoke some resistance, therefore, it is important to identify possible restrictions to the inspection process due to the IP and assess the consequences, if any. Further, we need to assess on a case by case basis if and when code reviews are needed/possible, for example, by checking if there are no risks to the security of the system; the privacy of underlying data is ensured; and no risk of undermining of intellectual property exists [26]. There is an inevitable trade-off to be made between disclosing all activities of the inspection vs. delaying them to a later stage (or not disclosing them at all).

Defining which time-frame for the assessment

A useful framework that can be used for deciding on a time-frame when assessing an AI system is defined in [27], and is formulated around three different time-scales, adapted here for our process:

Present risks "What are the risks of using the AI system that we are already aware of and already facing today?"

Near-future risks: "What risks might we face in the near future when using the AI system, and assuming the current technology and contexts?"

Long-run risks: "What risks might we face in the longer-run when using the AI system, as technology becomes more advanced?"

B. THE ASSESS PHASE

Socio-technical Scenarios

The Assess phase of the process begins with the analysis of Socio-technical scenarios. Usage scenarios are a useful tool to describe the aim of the system, the actors, their expectations, the goals of actors' actions, the technology, and the context. Socio-technical scenarios can also be used to broaden stakeholder understanding of one's own role in understanding technology, as well as awareness of stakeholder interdependence [23], [28]. The basic idea is to analyze the AI system by using socio-technical scenarios with relevant stakeholders including designers (when available), domain, technical, legal, and ethics experts [28]. In Z-Inspection, the scenarios are used by a team of inspectors, to identify a list of potential ethical, technical and legal issues that need to be further deliberated. Scenarios can be used as a participatory design tool if the AI is in design phase; or as a part of the assessment of an AI system already deployed. One possible way to use socio-technical scenarios is within discussion workshops, where expert groups work together to systematically examine, and elaborate the various tasks in respect to different contexts of AI. In Table 4 we present a simplified scenario for the AI product we assessed. The assessment took place with the AI system already deployed.

In our process, analyzing scenarios of the possible usage of an AI System involves the following tasks:

a) Classify the AI System by Domain and Usage.

For this task we classify the AI system by domain and the context in which it is used. Certain domains are already highly regulated (e.g. health care, finance), while others are only loosely regulated. This has to be taken into consideration during the assessment.

b) Review Domain specific Frameworks, Regulation and Laws.

This task includes reviewing the relevant regulations and laws for the specific domain, and the contracts in place for the entities, which are relevant for the assessment.

c) Develop of an evidence base.

This task consists of reviewing and creating an evidence base to verify/support any claims made by producers of the AI system and other relevant stakeholders. This task is domain specific. A key question here is who is qualified and absent from personal bias to define what is a relevant evidence base. This question can only be answered taking into account the context and the specific domain.

e) Discover and list potential Ethical Issues and Tensions, and Technical and Legal issues.

This step of the Assess Phase of the process consists of identifying possible ethical, technical and legal issues for the use of the AI within the given the boundaries and context. For some Ethical issues a tension may occur. A tension as defined in [27] refers to different ways in which values can be in conflict, i.e. tensions between the pursuit of different values in technological applications rather than an abstract tension between the values themselves. The scenarios representing different usage situations of the system are discussed with a number of experts and if necessary other stakeholders, and examined phase by phase according to the selected ethical values in order to define potential ethical issues. As suggested in [28], the selected ethical principles are cross-checked against each phase of a scenario and any ethical issues arising are discussed and reported at each step and documented. The process used to reach consensus is made transparent, so that it is possible to go back and re-assess possible relevant changes in the ecosystems. We recall that all the information should be included in the protocol (log) of the Z-Inspection process being sure to make clear who is making decisions, why and who has the authority to decide.

The output of analyzing scenarios is a list of ethical issues that we call 'flags'. A flag identifies an area that needs further

investigation. Experts review flags and indicate if any can be classified as ethical issues and if so, describe the tensions between values (when possible). Some of the flags may indicate a more technical issue and or a legal issue, rather than an ethical issue. If this is the case, for this phase of the process no further action is required.

Describe Ethical issues and Tensions

The next step in the process is to describe and classify if such ethical issues represent ethical tensions and if so, to describe them. This is done by a selected number of members of the inspection team, with inter-disciplinary skills, e.g. experts in ethics, philosophy, policy, law, domain experts, machine learning. Such a variety of backgrounds is necessary to identify all aspects of the ethical implications of the use of the AI. Whilst the interdisciplinary nature of the team is essential, it can pose a challenge on how to reach a consensus among the various experts. Our method consists of reviewing the applied ethical frameworks relevant for the domain, asking the experts to classify the ethical issues discovered with respect to a pre-defined catalog of ethical tensions [Table 3], and use the following classification of ethical tensions, defined in [27]:

- True dilemma, i.e. "a conflict between two or more duties, obligations, or values, both of which an agent would ordinarily have reason to pursue but cannot";
- Dilemma in practice, i.e. "the tension exists not inherently, but due to current technological capabilities and constraints, including the time and resources available for finding a solution";
- False dilemmas, i.e. "situations where there exists a third set of options beyond having to choose between two important values".

Where some ethical issues do not clearly fit into one or more pre-defined examples of ethical tensions, then experts describe the tension by using free text. If some of the ethical issues are not really such, then experts describe and explain why.

Table 3. Examples of Tensions between Values from [27]

EXAMPLES OF TENSIONS BETWEEN VALUES
Quality of services *versus* privacy;
Personalisation *versus* solidarity;
Convenience *versus* dignity;
Privacy *versus* transparency;
Accuracy *versus* explainability;
Accuracy *versus* fairness;
Satisfaction of preferences *versus* equality;
Efficiency *versus* safety and sustainability.

Trustworthy AI Areas of Investigation.

The basic idea of the process in this step is to identify from the list of ethical issues and flags which areas require inspection. We therefore map ethical issues and flags to some or all of the seven requirements for trustworthy AI. From this mapping, we then create a plan of investigation. Once the ethical issues and tensions have been agreed upon among the experts, and a number of flags have been raised, the consensus building process among experts continue by asking them to map ethical issues and tensions, and flags onto the four the four ethical principles based on fundamental rights are [15], that is: Respect for human autonomy, Prevention of harm, Fairness, and Explicability

We then ask the experts to map the ethical issues and flags to the seven requirements established by the EU High Level Experts Guidelines for Trustworthy AI (HLEG), that is:

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

These are general AI principles and requirements. Each domain in which AI is applied has, in addition, its own principles and requirements / values, some of them correspond to the above, others need to be added. We added to the seven above the following two requirements:

-Assessing if the ecosystems respect values of Western European democracy.

-Avoiding concentration of power;

Although these two categories seem implicitly already incorporated into the HLEG principles, we believe that there is a need to explicitly mention them as separate indicators for Trustworthy AI, to ensure they are not underestimated or lost during an assessment. The rational for this is as follows. Ethical values are "embedded" often by software designers when they design an AI system, resulting in the AI system setting the bar for what is defined as "good" or "bad", "correct" or "incorrect" "healthy" or "unhealthy" for a given domain and context. If we do not trust the ecosystems -where the AI has been designed/produced/used, we may not trust the AI system [29]. As for avoiding concentration of (illegitimate or monopoly) power, this could be considered a component of societal well-being or accountability, but in the definition of trustworthy AI by the HLGE, this is not explicitly spelled out.

Map Ethical issues and Flags to Trustworthy AI Areas of Investigation.

For the mapping, we have used the following consensus method (but other methods are possible): We use a factor M weighting of relevance 0,1,2,3,4 ... with 4 being of high relevance. i.e. min relevance = X(0) , X(1), X(2), X(3), max. relevance = X(4). An Accountable Person [ACP] prepares and sends template tables to each of the team members to score from 0 to 4 the relation between Ethical Issues and

Flags, and their mapping to the predefined Cluster of Areas. After independently filling the template tables, each team member sends them back to the ACP, who is responsible for keeping the results private, and secure. The process of assigning weights is done independently to avoid bias among team members. The ACP then prepares the final tables by calculating the average score for each relation:

- Sim - score S for each member i for mapping m
- The difference between each member score Si should not be bigger than 2, else there should be a "consensus" discussion between all members to agree on the final average mapping score.

We report three examples of such mapping of Ethical Issues and Ethical tensions for the use case that we assessed in Section IV.

Execution

Verification of the AI system can now start. Two scenarios apply:

i) The team of inspectors have not signed a Non-Disclosure Agreement (NDA), and a part of the AI is under IP. In this case, no deep dive (top down) analysis up to code is possible. The AI is verified as a "black box".

ii) The team of inspectors have signed a NDA and have access to the data/ the AI model and other components under IP. In this case a deep dive (top down) investigation up to code is possible (e.g. to verify Training, Test Data, ML Model, Output). Depending on the particular AI system considered, a "white box" -approach might be possible.

In what follows we describe the part of the process, which can be used to coordinate the verification.

Pre-Check

In some cases, it could already be possible to come up with an initial pre-assessment with no need to go deeper into technical levels. This can be seen as a kind of pre-check, and depends on the domain. For example, if the AI systems is used in the domain of healthcare, the result of assessing, and reviewing best available evidence to decide whether evidence is trustworthy, considering patients' unique predicament and values and preferences, could already provide enough information for decision makers to determine a trade-off decision between the benefits and risks, burden, and costs associated with the use of the AI system. In this case the process would stop here and would provide recommendations.

Creating Paths

It is important to note that we do not prescribe the way the verification of the requirements for trustworthy AI is done. The process allows for coordination and synchronization of the various activities performed by teams of human investigators. Similar to a due diligence, verification of the AI systems corresponds to create sub-teams of experts (the investigators) whose task is to verify the trustworthy indicators mapped during the previous step. Investigators follow what we call Paths. Different paths of investigation can be created to verify Ethical, Technical and Legal issues. Therefore, the execution of the inspection means execution of paths. A path in this work describes the dynamic of the inspection. It differs on a case-by-case basis and depends on the domain. By following paths, the inspection can then be traced and reproduced. Execution of a path here means that a series of steps are performed, either by individuals, e.g. via workshops, interviews, etc. and/or via software tools. Each step is a self-contained unit of the overall execution of a path. The steps of a path can be executed by different teams of inspectors with special expertise.

Execution of a Path, occur at various level of abstractions and require a number of ad hoc decisions to be made depending on the area of investigation. Execution of this path provides a feedback related to selected areas of investigation (See Section IV for an example). In general, after validation, a flag may disappear or may turn into an ethical issue if the validation confirmed the issue being flagged. When executing more then one path, the various feedback needs to be stored in a protocol and can be assessed by an authority who will have the final word in deciding on the appropriate (or not) use of the AI in the given context. The protocol may also be shared and used for engaging other stakeholders involved in the decision-making process.

Looking for Paths

An analogy is to think of a path like a waterfall - water finding its way (case-by-case). One can start with a predefined set of paths and then follow the flow, or just start randomly. An interesting part is to discover the missing parts (what has not been done) [30], [31].

Ethical issues and Flags are re-assessed

The end of the execution of each path produces a so-called final feedback. An example of a path final feedback is a set of values of measurable indicators, together with a textual description of the investigation for those indicators that are not easily measurable. An example of the result of the execution of a Path is given in Section IV. The process continues by sharing the feedback of all Paths executed to the domain and ethics experts. Since the feedback of the execution of the various paths may be too technical-specific, it is useful to "explain" the meaning to the rest of the team (e.g. domain and ethical experts) who may not have prior knowledge of ML. This process may require the team to re-analyze the socio-technical scenarios in an iterative way, until a "consensus" is reached and experts might revise the set of Ethical issues and Flags

C. RESOLVE PHASE

(Optional): Scores are defined

If required, a labeling scheme could be offered to quantify the level of trust and risk for various measurable indicators. One possibility is to use the web based online tool (ALTAI) developed by the HLEG, which gives a visualization of the self-assessed level of adherence of an AI system and its use with the seven requirements for Trustworthy AI. Scores may be static and/or may be revised over time if an Ethical maintenance is required.

Address, Resolve Tensions

In this step we address and possibly resolve tensions. To do this we proceed as follows: Once the revised list of Ethical issues and Flags have been defined, the next step is to prioritize them [20]: What are the most pressing challenges for the use case? After the revised the set of Ethical issues and Flags has been defined and prioritized, we use a framework, suggested by [27], which lists three elements.

1."Deepening understanding of technological capabilities and limitations in areas particularly relevant to key ethical and societal issues". This is accomplished in our process with the execution of the various Paths, the results of which yield a better understanding of the possibilities and limitations of the AI being investigated.
2."Applying evidence to resolve tensions". Building a stronger evidence base on the current uses and impacts of the AI in our process is performed executing a horizontal Path (*build evidence*). "Evidence on current societal impacts of AI provides a stronger basis on which to assess the risks, and to predict possible future impacts. This is especially important for addressing key tensions as they may affect marginalized or underrepresented groups". One lesson we have learned in assessing a commercial AI-based product in healthcare is that building a stronger evidence base of the current uses and impacts of AI-based technologies is not a trivial task and it is dependent on the domain considered. In healthcare, for example, there is an ongoing discussion of who the qualified non-biased experts are for building such medical evidence: domain experts (in our case cardiologists) or domain methodologists? The goal is to reduce the risk of (social) bias for the experts building medical evidence, therefore, to avoid making assumptions based on biased medical evidence that might influence in one way or another the process of resolving ethical tensions. This is an ethical issue per se.
3. "Building on existing public engagement work to better understand the perspectives of different members of society on important issues and trade-offs." Different stakeholders will be affected differently by the AI system, and may hold different values. "Resolving these tensions requires us to understand varied public opinions on questions related to tensions and trade-offs". In our process this is considered during the analysis of socio-technical scenarios where we identify the various stakeholders and agree on ways to involve them (when possible) in the evaluation. Several questions need be addressed:
- How to engage the public?
- What information is given, and to whom?
- How to select stakeholders, since each stakeholder might have a social/political/economic/knowledge bias?
- How to take into account their feedback?

4. "Identify the extent to which key tensions involve true dilemmas, dilemmas in practice or false dilemmas" In our process we classify (when possible) the resulting revised list of ethical issues in two categories: ethical issues that can be solved (dilemma in practice), and the ones that are true ethical dilemmas.

Recommendations

Depending on the list of classified ethical issues (dilemmas in practice and true ethical dilemmas), the team of inspectors may give recommendations. Such recommendations should be considered as a source of qualified information that help decision makers make good decisions, and that help the decision-making process for defining appropriate trade-offs (See discussion on how to handle Trade Off below). When possible, this would also help improve the design of the AI. Developers could be using the feedback and results derived from the areas of the assessment and from the ethical aspects identified to improve the technical aspects of the AI and/or to better match the ethical challenges. In Section IV. we give an example (simplified) of some of the recommendations we have given for AI medical device we assessed.

Trade Offs

We envision that the results of the investigation would be useful for relevant stakeholders who are responsible for making final decisions on the appropriate use or not of the AI in the given context. They would also help continue the discussion by engaging additional stakeholders in the decision-process. We envision four possible scenarios:
Appropriate use: The AI system is appropriate to use for the purpose anticipated and perception of use.
Remedies: If risks are identified, then define ways to mitigate them (when possible).
Ability to redress: If the AI harms, redress could involve reparation, restitution or giving indemnification for the wrong.
Request for an AI Ethical Maintenance over time: Given the dynamic nature of how some AI systems are trained and improved over time, and given the changes in the context, a continuous ethical maintenance may be requested.

Ethical maintenance

It is crucial to monitor that the AI system that fulfilled the Trustworthy AI requirement at launch, continues to do so over

time. We call this "Ethical Maintenance". The objective of AI Ethical maintenance is to monitor within the dynamic context of given legal and contractual frameworks, the deployed AI system performance over time with respect to a set of defined ethical principles to ensure robustness from a technical and social perspective until it is decommissioned. The ISO/IEC 14764 standard [32] defines forms of software maintenance that are also reasons for conducting an Ethical AI maintenance. In [33] we have defined an AI ethical maintenance process based on an adapted version of the Reliability-centered maintenance (RCM) model [34]. The maintenance process is used to analyze possible "failure modes" for some or all of the all seven trustworthy AI requirements, and to develop a customized maintenance plan and strategy. Our approach assumes that there must be an owner of the AI ethical maintenance process who assures that the RCM plan is established and that maintenance according to plan is being carried out. RCM is a costly and complex process. To perform an assessment of cost, respectively the budget needed for inspections and maintenance, an organization should analyze the trade-off between the risks entered, their economic impact, and the cost of mitigation. The following provide some pragmatic guidance:

- Use existing or define frameworks considering operational, financial, and reputational risk;
- Perform the risk assessment of AI and identify the "opportunity cost of harm" for the organization (cost, revenue loss);
- Determine the appetite of the organization for that risk in terms of money and conclude an appropriate budget level from a risk/benefit view;
- Based on that assessment use this budget for an inspection of the AI asset in an RCM approach;
- After inspection, feedback results, and potential impact and the likelihood of AI failures, e.g., false positives on the economic result;
- Assess whether the current budget and depending factors are sufficient to manage the potential "opportunity cost of harm" and derive if further benefits/mitigations can be realized by expanding the inspection budget.

The peril of inaccurate inspection

There is a danger that a false or inaccurate inspection will create natural skepticism by the recipient, or even harm them and, eventually, backfire on the inspection method. This is a well-known problem for all quality processes. We alleviated it using an open development and incremental improvement to establish a process and brand ("*Z-Inspected*").
Importantly, the potential liability of those who monitor AI needs to be considered. Inaccurate inspection may (and indeed should) give rise to such legal consequences, but the standard of care requires careful consideration and adjustment.
While liability regimes of certifiers are already well established in other contexts, they need to be adjusted in order to suit the specific characteristics of ethical maintenance of AI, in particular to its dynamic, revolving character [35].

IV. Assessing Trustworthy AI. Best Practice: AI for Predicting Cardiovascular Risk

The Use Case
We have used and tested Z-Inspection by evaluating a non-invasive AI medical device designed to assist medical doctors in the diagnosis of cardiovascular diseases. Z-Inspection was conducted with the AI system already deployed and in use in several countries in Europe and in other parts of the world. The work we performed was conducted purely for research purposes and did not involve any compensation or personal benefits. The product in question was a non-invasive AI medical device that used machine learning to analyze sensor data (i.e. electrical signals of the heart) of patients to predict the risk of cardiovascular heart disease.

The Problem Domain

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year [36]. Over the past decade, several ML techniques have been used for cardiovascular disease diagnosis and prediction. The potential of AI in cardiovascular medicine is high; however, ignorance of the challenges may overshadow its potential clinical impact [37]–[40].

AI Medical Device

The product we assessed was a non-invasive AI medical device of class 1, according to the European Commission Medical Device Directives (MDD) [41], that uses ML to analyze sensor data (i.e. electrical signals of the heart) to predict the risk of cardiovascular heart disease. The company uses a traditional machine learning pipeline approach, which transforms raw data into features that better represent the predictive task. The features are interpretable and the role of ML is to map the representation to output. The mapping from input features to output prediction is done with a classifier based on several neural networks that are combined with an Ada boost [42] ensemble classifier. The output of the network is an Index (range -1 to 1), a scalar function dependent on the input measurement, classifying impaired myocardial perfusion.

We Assembled a Harmonious Team

Our team included philosophers, AI engineers, and domain experts, which in this case, meant medical doctors. Because of the diversity, our project proved demanding in two senses: the transacted concepts were sophisticated and specialized, and working together required exchanging them across the humanistic and scientific sides of knowledge. So, besides having people from the required backgrounds and in the right numbers, our team's congealing depended on members embracing interdisciplinarity and wielding significant cognitive power [43].

Illustration of Actors and Socio-Technical Scenarios (simplified) based on model's prediction

An overview of the ML pipeline can be summarized as follows:
- Measurements, Data Collection (Data acquisition, data annotation with ground truth, Signal processing);
- Feature extraction, feature selection;
- Training of the Neural Network-based classifier using the annotated examples;
- Once the model is trained, actions are taken for new data, based on the model's prediction and interpreted by an expert and discussed with the person.

The initial output of the device was visualized as "Red" for prediction of a risk of a cardiovascular disease, and "Green" for prediction of the absence of a cardiovascular disease. In a later stage, the company, based on feedback from customers, a third color, "Yellow", to indicate a generic non specified cardiovascular health issue.
-The AI-system predicts a "Green" score for the person. The medical doctor agrees. No further actions are taken, and the patient does nothing;
- The AI-system predicts a "Green" score for the person. The person and/or the medical doctor do not trust the prediction. The person is asked to perform a further invasive test;
- The AI-systems predicts a "Red"/ "Yellow" score for the person. The medial doctor agrees, the person does a further invasive test
- The AI-systems predicts a "Red"/ "Yellow" score for the person. The person and/or the medical doctor do not trust the prediction. Patient is asked for a further invasive test.

Reaching Consensus

To maximally utilize the practical medical experience of our domain experts, as well as the theoretical expertise of our philosophers, our team located and described ethical issues by working in two directions. Starting from the empirical, we asked the healthcare team members to describe problems they saw – or expected to see – arising from the AI's use. Then we went to the other extreme by consulting the requirements listed in the *Ethics Guidelines for Trustworthy AI*. From those abstract principles, we formed specific question to ask about the technology and human autonomy, privacy, and similar concerns [43].

To reach consensus, we use the following process. We had each team member, commit their personal thoughts to a short rubric. First, we narrated each discussed ethical dilemma and tension, and technical and legal issues ("flags" in our terminology) in our own words. Then we mapped each one onto our ethics principles and tensions. Concretely this meant taking the four pillars of the *Ethics Guidelines for Trustworthy AI* (Respect for human autonomy, Prevention of harm, Fairness, Explicability) and selecting the one we individually found the most apt. But to make it working in practice, we considered that each of those pillars supports a number of requirements for Trustworthy AI, from which we selected and then each requirement contains sub- requirements, which we also selected. For Ethical issues, the idea was to capture the dilemma in structured ethical terms. For "flags" the mapping helps transforming them into requirements that needs to be further validated.

We then gathered all contributions, and their clean articulations allowed easy comparisons across the group and so enabled a smooth move to agreement on final issues, mappings, and tensions. The use of rubrics facilitated a conclusion by forcing stark descriptions and categorizations that funneled our team toward consensuses. For the use case, we started with the claim of the company that their AI increases patients' quality of life. Getting notified before a heart attack instead of *by* one is an improvement. During the analysis of the socio-technical scenarios of usage of this AI a number of questions arise.

Here are some of them (simplified): When a patient is faced by only a green or red result, ambiguities surge. What level of risk does red actually imply? Why? When it comes to living, is it sometimes better to *not* know? Does the addition of a yellow, intermediary score – which the company did add – resolve any of these questions and dilemmas, or just make them worse? Because of the geography, it seemed possible that some races may have been over- or underrepresented in the training data. Should patients from the overrepresented race(s) wait to use the technology until other races have been verified as fully included in the training data?

Illustration of Ethical Issues and Mappings to the Trustworthy AI, Areas of Investigation.

We list below three examples of Ethical issues we identified for this use case. In each case we organized our findings as a description, followed by a mapping of the issues onto the taxonomy proposed by the *Ethics Guidelines for Trustworthy AI*. Finally, we described ethical tensions that arose along with the principle concerns. In some cases, multiple mappings and tensions emerged. We use a template rubric, as described in Table 4.

Table 4 Template Rubric.

Ethical Issue Ei. Description: (open vocabulary) Map to Trustworthy AI ethical Pillars and Requirements: (closed vocabulary) Ethical Tensions: Kind of tensions:

Ethical Issue E1. Description: When the AI is applied to asymptomatic people who are then *notified* of a "minor" Cardiovascular problem, people may restrict their lifestyle. However, these restrictive changes may be unnecessary.

Map to ethical Pillars and Requirements: **Respect for human autonomy > Human agency and oversight > Human agency and Autonomy**; **Prevention of Harm > Technical Robust and Safety > Accuracy.**

Ethical Tensions: **Over-diagnosis vs. correct diagnosis. Benefits vs. harms: Saving the few and harming others who are falsely diagnosed as vulnerable to severe cardiac problems.**

Ethical "Issue", E2. Description: Using a black-box algorithm might impair the trust of the doctor in the diagnostic app, especially if the app / algorithm has not been verified by independent studies. This complicates and obscures the attribution of *accountability* [44]. Ultimate responsibility for decisions is not clearly defined for the variety of stakeholders involved (e.g. clinicians, healthcare institutions, etc.).

Map to ethical Pillars and Requirements: **Explicability > Transparency > Traceability, Explainability.**

Ethical Tensions: **Accuracy *versus* Transparency/ Explainability.**

Kind of tensions: **This is a dilemma in practice because the AI could, potentially, be rendered explainable.**

Ethical "Issue", E7. Description: The data used to optimize the ML predictive model is from a limited geographical area, and no background information on race, gender or ethnicity. Because all clinical data to train and test the ML Classifier was received from three hospitals near to each other, there is a risk that the ML predictions' effectiveness will be tilted toward a certain population segment.

Map to ethical Pillars and Requirements: **Fairness > Diversity, non-discrimination and fairness > Avoidance of unfair bias.**

Ethical Tensions: **Accuracy *versus* Fairness.**

Kinds of tensions: **This is a dilemma in practice because the AI training data could be diversified.**

Illustration of Creation of Paths from Ethical Issues.

We illustrate here an example of how we create a Path for execution.

The Ethical issue E7, described above, has been mapped by a consensus process by experts onto the ETHICAL Pillar: Fairness, and to the 7 Trustworthy AI REQUIREMENTS: Diversity, non-discrimination and fairness > Avoidance of unfair bias.

The Ethical Tension identified was: Accuracy *versus* Fairness.

The experts created a Path1 to investigate the ethical tension identified in E7: Algorithm Fairness vs. Usability to assess Accuracy, Bias, Fairness and Discrimination.

Path 1 Accuracy, Bias, Fairness, Discrimination

This path has been created to analyse accuracy, bias, fairness and discrimination. It also takes into account unfair bias avoidance, accessibility and universal design, stakeholder participation. The execution of this path was conducted by teams of investigators, using a mix of interviews, and by analyzing the data sets used for training and testing, and verifying the corresponding output. Part of this execution was done using available software tools ([45], [46]).
Actions performed in Path1 includes (simplified): Measure key metrics for Bias and Fairness and allocation across groups; Compare deployment data with training data to ensure comparability; Assess the usefulness of predictions to clinicians.

Let´s consider how the implementation of Path 1, help verify "Fairness" for our use case.

Example Execution of a Path: Verify Fairness.

The domain is healthcare. The task here is to verify if we are having a classical case of ethical tensions (E7 issue):
Accuracy *versus* fairness: what if the ML outcome is most accurate on average, but systematically discriminate against a specific minority?

If we were to use the Assessment List for Trustworthy AI (ALTAI), the checklist recommended by the HLEG [23], we were first asked to perform a fundamental rights impact assessment (FRIA) answering questions based on specific articles in the Charter and the European Convention on Human Rights (ECHR) its protocols and the European Social Charter. For our use case, one relevant question is (adapted from [23]):

Q. Does the AI system potentially negatively discriminate against people on the basis of any of the following grounds (non-exhaustively): sex, race, color, ethnic or social origin, genetic features, membership of a national minority, birth, disability, age or sexual orientation?

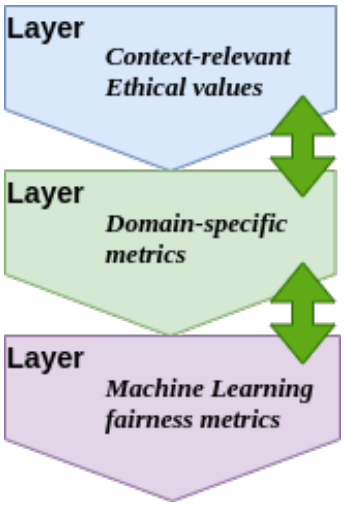

Figure 3. Various Layers of Abstractions for "Fairness"

Moreover, following the same check list, we have to verify Requirement #5: Diversity, Non-discrimination and Fairness of Trustworthy AI, where we are asked if the AI system has unfair bias (both for training and operation), leading to unintended direct or indirect prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalization. When using the HLGE requirements in practice, the challenge is that the HLEG is a general check list, applicable for several different domains, but it does not give any specific indication of which definition(s) of "fairness" might be more appropriate for a specific domain and for a specific use case.
In fact, "Fairness" can be defined at different level of abstractions as illustrated in (Figure 3).
From top to bottom:
- Layer Context-Relevant Ethical values (Philosophy, Social science): Fairness here is not measured in metrics, but with a narrative description.
- Layer Domain-specific (e.g. healthcare): Fairness is expressed with domain-specific metrics;
- Layer: Machine Learning: Fairness expressed by ML fairness metrics.

If we consider Machine Learning, in the literature common fairness definitions focus on balancing notions of disparity and utility, such as such as Demographic Parity, Equality of Odds, Equality of Opportunity (see Section on Related Work). While thesed efinition of fairness might be appropriate in many domains, in healthcare - where quality of service is very important—it is argued that it is necessary to define a different notion of "fairness" that does not introduce "unnecessary" harm to any subgroup [47]. However, examples of what "unnecessary" harm means is unclear.

Let´s consider our use case. The goal was to verify "fairness" with the ML model already in production. The first decision experts made when executing this path was to clarify what kind of algorithmic “fairness” was most important for the domain (healthcare) and for the specific use case. This meant choosing a "fairness criteria" which apply for the use of the AI medical device for predicting cardio vascular diseases. There is no unique solution for that. Experts have chosen *Distributive justice* (from philosophy and social sciences) as suggested in [21], as a context-relevant ethical value (Figure 3).

The second decision experts made, was to define who the protected groups were for the use case. This is a key decision which has an ethical implication on how we would "measure" fairness for the use case. In our use case, protected groups were defined as non-caucasian people who undergo a test using the AI system. This was based on the information derived from the analysis of the scenarios of usage: the ML model was trained and tested using exclusively patients who stayed in three hospitals, each one near to the other, in a particular geographical areas in Germany. So the training and test data was potentially "biased" towards a specific patient ethnicity. However, this was only an assumption, since patient data were anonymized, and no further information on ethnicity was provided.

Having chosen *Distributive justice*, what follows were two possible domain-specific fairness criteria [21] applicable for our use case:
Equal patient outcomes: protected groups have equal benefit in terms of patient outcomes from the deployment of the ML model.
Equal performance: the ML model is equally accurate for patients in the protected and non-protected groups.

To verify these *fairness* criteria we need to have access to the ML Model in order to map Domain specific “Fairness” to ML metrics. We sketch the steps of execution:
- A sub team of our Experts had to sign a NDA to have access to the ML model, to the training and data sets;
- ML specialists together with medical experts and experts in ethical reasoning decided which type of fairness was most appropriate for the given application and what level of it is satisfactory. This also meant identifying well known Trade Offs, that is, choosing between incompatible types of fairness: *Equal patient outcomes and Equal performance* [21].

Several different approaches in ML define "Fairness, resulting in different metrics and formal “non-discrimination” criteria. [48]. This required choosing appropriate ML metrics.
The access to the training and testing data and the ML model was possible, and part of the verification was conducted using appropriate open source tools, i.e. WhatIf [45], Fairness 360 [46] to name a few. As a general consideration, some of the ML metrics depend on the training labels and raise questions, e.g.:
- Is the training data trusted?
- Do we have negative legacy?
- Are labels unbiased?
These questions are highly related to *the context* (e.g. ecosystems) in which the AI has been designed/ tested and deployed, and cannot always be answered technically. It goes back to the initial pre-condition if we trust the ecosystem(s) addressed initially in the Set Up phase.

Illustration Paths Feedback (simplified)

We give an example of the result of the execution of the Path1 executed for the use case.

Path 1 Accuracy, Bias, Fairness, Discrimination
This path mainly analysis accuracy, bias, fairness and discrimination. It also takes into account unfair bias avoidance, accessibility and universal design, stakeholder participation.

Path 1 Execution Feedback:
- For the data sets used by the AI, a correlation with age was identified. The analysis of the data used for training, indicates that there are more positive cases in certain age segments than others, and this is probably the reason for a bias on age.

- A higher accuracy prediction for male than female patients was identified. The dataset is biased in having more male than female positive cases, and this could be the reason.
- The size of the datasets for training and testing is small (below 1,000) and not well balanced (with respect to gender, age, and with unknown ethnicity). This may increase the bias effects mentioned above.
- *Sensitivity* was discovered to be lower than *specificity*, i.e. not always detecting positive cases of cardiovascular risks.

Illustrations of Recommendations

After ethical issues and flags have been captured and described, the execution of the corresponding Paths have been completed, a response is written and presented to relevant stakeholders to complete the evaluation. It may include concrete solutions to identified problems.

We present a simplified version here for the use case:

Findings:
Accuracy, sensitivity and specificity deviate in part strongly from the published values.
Not sufficient medical evidence exists to support the claim that the device is accurate for all gender and different ethnicities.
This poses a risk of non-accurate prediction when using the device with patients of various ethnicities.
There is no clear explanation on how the model is being medically validated when changed, and how the accuracy performance of the updated model compares to the previous model.

Recommendations:
1. Consider continuously evaluate metrics with automated alerts.
2. Consider a formal clinical trial design to assess patient outcomes.
3. Consider periodically collect feedback from clinicians and patients.
4. Consider establishing an evaluation protocol that is clearly explained to users.
5. We recommend that feature importance for decision making should be given, providing valuable feedback to the doctor to explain the reason of a decision to the model (healthy or not). At present, this is not provided, giving only the red/green/yellow flag with the confidence index.

The Value Added is the Engagement

We conclude this brief description of the use case with a quote from one of the philosopher who helped with the assessment [43]: "Companies that commission an ethics evaluation will find that the primary value added lies outside of solved problems, and instead in questions that succeed primarily because they get posed."

VI. CONCLUSION AND LINKS TO RELATED WORK

The ethical and societal implications of AI have been discussed across a range of academic disciplines (e.g. computer science, machine learning, philosophy, ethics, law, human-machine interaction, political, and social sciences), in policy and civil reports, and in popular science and media [27], [49]–[56]. Proposals in support of ethical and trusted AI are emerging ([15], [17], [57]–[61]). From a Western perspective, the terms context, trust, and ethics are closely related to our concept of democracy. The German Data Ethics Commission (DEK) [62] recommends the examination of the extent to which the function of an AI system can affect the function of democracy, fundamental rights, secondary law, and the basic rules of law.

Whittlestone et al. [27] provide a literature review and a list of common ways organizations group and structure ethical principles that can be applicable to AI. The Artificial Intelligence Index Report 2019 [49] also lists a number of ethical challenges and documents the key topics discussed in global news media on AI and ethics. However, the ethics of AI is still an underexplored area in practice.
The Z-Inspection process was inspired by a number of initiatives.

Whittlestone et al. [27] defined a roadmap for work on the ethical and societal implications of algorithms, data and AI. The original report was aimed at those involved in planning, funding and pursuing research in policy. Our work is inspired by their research in numerous areas: We use their “concept building” approach for defining our area of investigation; we use their approach for “exploring and addressing tensions” between the ways technology (in our case AI) may simultaneously threaten and support different values. We follow their recommendation of applying principles to concrete cases, as a way to learn what obstacles and challenges that arise in practice. In our analysis of a real use case we support their proposal to build “a rigorous evidence base for discussion and societal issues”, and to apply evidence to resolve tensions.

Our notion of “ecosystems” as part of the ethical and societal assessment of AI resembles and generalizes the notion of sectors and parts of society, level of social organization, and publics defined in [27], by adding the political and economic dimensions. It also takes into account the recommendation of The German Data Ethics Commission (DEK) [62] who propose to use what they call the "overall socio-technical system" as "context", including all components of an algorithmic application including all human actors, from the development phase to implementation in an application environment and the phase of evaluation and correction.

Our research work goes along the line suggested by Webb and Chou [26], of creating "due diligence best practices" for developing and applying AI.

A number of proposals have been defined for implementing ethical self-assessments, e.g. [15], [20], [28], [63]–[67] with the aim to look for possible examples of bias, ethical issues, or

other safety risks as a result of using AI, and self-assessments for increased transparency in AI for data sets ([48], [64], [68]–[73]) models ([74]), and services ([68]). A corresponding number of software tools have been implemented to help with specific aspects of AI self-assessment e.g. [45], [46], [75]. In Madaio et al. [76], they identify that the beneficial outcome of implementing an AI ethics checklist may be to prompt discussion and reflection that might otherwise not take place. Checklists themselves may not be sufficient to influence practitioners' decisions, rather, checklists need to address the stakeholder roles and organizational procedures, thus empowering practitioners while following certain processes. Empowerment may be particularly important in areas where paternalism is an unwanted behavior, such as in the design and implementation of evidence-based medical software. One problem with self-assessment is that the data and/or the AI model details may be proprietary and part of the IP and therefore not publicly available. In [69], it is reported that it is often not clear where to draw the line between providing enough information for a model to be adopted while not revealing information that threatens competitive advantage.

Dorian Peters et al. recently published two frameworks (Responsible Design Process and Spheres of Technology Experience) that are supposed to help the design of what they call responsible AI [14]. Robertson et al. propose an ethics-based co-design approach for complex systems, such as biotechnologies [77]. Their approach does not explicitly consider AI, but they also propose to involve a variety of stakeholders in the system design, toward a greater consideration of end-user exposure.

A recent report [78] introduced a model for the operationalization and measurements of ethical principles, for algorithmic-decision-making systems (ADM), introducing a concept of ethical labeling (taking the energy efficiency label scheme as a guide), and classifying ADM systems using a so-called risk matrix, for potential societal effects and degree of potential harm. The approach is interesting and is intended primarily for regulators and consumers. However, the framework in its present form does not specify how to identify ethical issues, how to perform measurable observations, and does not take into account post deployment ethical monitoring (i.e. how to cope with dynamic changes during operation).

An additional source of inspiration for our work is what is referred to in the corporate world as Environmental, Social and Governance (ESG) Due Diligence. It is a different kind of due diligence which focuses more on environmental, social and governance, and not the standard due diligence related to the finance or business models of the company. Our approach differs from ESG due diligence in a number of ways. Our metrics are focused on trustworthy AI, whereas the "ESG metrics" are focusing on general or specific social/environmental/governance issues, such as whether a company's level of carbon emissions compare favorably to its industry peers, etc. [79]. Another source of inspiration for our work is the so-called, Human Rights Due Diligence [80]. Human rights is included in the EU Trustworthy AI framework, reflected in the major principles / requirements.

We have presented Z-Inspection, a novel holistic and analytic processes, to assess Ethical AI that can be applied to a variety of domains where AI systems are used. To the best of our knowledge, Z-Inspection is the first process that assesses Trustworthy AI, as defined by the HLEG, in practice.

Acknowledgment

Many thanks to Matthew Eric Bassett, Kathy Baxter, Aleks Berditchevskaia, Stefano Bertolo, Jörg Besier, Vint Cerf, Virginia Dignum, Yvonne Hofstetter, Alan Kay, Graham Kemp, Romeo Kienzler, Stephen Kwan, Robert Madelin, Abhijit Ogale, Jeffrey S. Saltz, Miroslaw Staron, Alex (Sandy) Pentland, Dragutin Petkovic, Michael Puntschuh, Andrea Renda, Francesca Rossi, Lucy Suchman, John Spohrer, Clemens Szyperski, Pieter van Halem, Dennis Vetter, Armin Wunder, and Andrej Zwitter for proving valuable comments.

In memoriam

Our team member, colleague and friend Naveed Mushtaq has passed away on December 27, 2020, after he suffered a sudden cardiac arrest a few weeks before. This work is dedicated to him.

REFERENCES

[1] R. V. Zicari, "On the Future of AI in Europe. Interview with Roberto Viola," *ODBMS Industry Watch*, Oct. 09, 2018. http://www.odbms.org/blog/2018/10/on-the-future-of-ai-in-europe-interview-with-roberto-viola/ (accessed Mar. 10, 2021).

[2] R. V. Zicari, "On Artificial Intelligence, Machine Learning, and Deep Learning. Interview with Pedro Domingos," *ODBMS Industry Watch*, Jun. 18, 2018. http://www.odbms.org/blog/2018/06/on-artificial-intelligence-machine-learning-and-deep-learning-interview-with-pedro-domingos/ (accessed Mar. 10, 2021).

[3] D. Nitkin and L. J. Brooks, "Sustainability Auditing and Reporting: The Canadian Experience," *J. Bus. Ethics*, vol. 17, no. 13, pp. 1499–1507, Oct. 1998, doi: 10.1023/A:1006044130990.

[4] J. Tashea and N. Economou, "Be competent in AI before adopting, integrating it into your practice," *ABA Journal*, Apr. 23, 2019. https://www.abajournal.com/lawscribbler/article/before-lawyers-can-ethically-adopt-and-integrate-ai-into-their-practices-they-must-first-be-competent (accessed Mar. 11, 2021).

[5] L. Eileen M. and N. Economou, "INSIGHT: Four Principles for the Trustworthy Adoption of AI in Legal Systems," *Bloomberg Law*, Mar. 29, 2019. https://news.bloomberglaw.com/tech-and-telecom-law/insight-four-principles-for-the-trustworthy-adoption-of-ai-in-legal-systems (accessed Mar. 11, 2021).

[6] J. Marciano, "Automating The Law: A Landscape of Legal AI Solutions," *TOPBOTS*, Jun. 10, 2017. https://www.topbots.com/automating-the-law-a-landscape-of-legal-a-i-solutions/ (accessed Mar. 11, 2021).

[7] A. Renda, "Artificial Intelligence - Ethics, governance and policy challenges. Report of a CEPS Task Force," Centre for European Policy Studies (CEPS), Brussels, Feb. 2019. [Online]. Available: https://www.ceps.eu/download/publication/?id=10869&pdf=AI_TFR.pdf.

[8] Personal Data Protection Commission, "Model Artificial Intelligence Governance Framework," Singapore, Jan. 2020. Accessed: Mar. 11, 2021. [Online]. Available: https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

[9] Treasury Board of Canada, "Directive on Automated Decision-Making," Government of Canada, Feb. 2020. Accessed: Mar. 11, 2021. [Online]. Available: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[10] R. Austin, "AI on the Case: Legal and Ethical Issues," May 17, 2019, Accessed: Mar. 11, 2021. [Online]. Available:

https://www.dww.com/sites/default/files/ai_on_the_case_-_legal_and_ethical_issues_may_17_2019.pdf.

[11] J. E. Stiglitz, *The price of inequality: how today's divided society endangers our future*, 1st ed. New York: W.W. Norton & Co, 2012.

[12] C. Hodges, "Ethical Business Regulation: Understanding the Evidence," Department for Business Innovation & Skills, 2016. Accessed: Oct. 26, 2020. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/497539/16-113-ethical-business-regulation.pdf.

[13] M. Brundage *et al.*, "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," *ArXiv200407213 Cs*, Apr. 2020, Accessed: Oct. 20, 2020. [Online]. Available: http://arxiv.org/abs/2004.07213.

[14] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two Frameworks for Ethical Design Practice," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 34–47, Mar. 2020, doi: 10.1109/TTS.2020.2974991.

[15] (AI HLEG) High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Text, Apr. 2019. Accessed: Oct. 26, 2020. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[16] (AI HLEG) High-Level Expert Group on Artificial Intelligence, "Policy and investment recommendations for trustworthy Artificial Intelligence," European Commission, Text, Jun. 2019. Accessed: Mar. 10, 2021. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

[17] European Commission, "On Artificial Intelligence - A European approach to excellence and trust," European Commission, Brussels, Text COM(2020) 65 final, Feb. 2020. Accessed: Oct. 27, 2020. [Online]. Available: https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en.

[18] OECD, *Artificial Intelligence in Society*. OECD, 2019.

[19] D. Anderson, J. Bonaguro, M. McKinney, A. Nicklin, and J. Wiseman, "Ethics & Algorithms Toolkit (beta)," 2018. https://ethicstoolkit.ai/ (accessed Mar. 10, 2021).

[20] J. Adamson, J. Charles, A. Darden, F. Lee, and M. Lowe, "Foresight into AI Ethics in Healthcare (FAIE-H): A toolkit for creating an ethics roadmap for your healthcare AI project," Open Roboethics Institute, Jan. 2020. Accessed: Mar. 11, 2021. [Online]. Available: https://openroboethics.org/wp-content/uploads/2020/02/FAIE-H-Final-to-Upload.pdf.

[21] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring Fairness in Machine Learning to Advance Health Equity," *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866–872, Dec. 2018, doi: 10.7326/M18-1990.

[22] The Institute of Internal Auditors, "Global Perspectives and Insights. The IIA's Artificial Intelligence Auditing Framework: Practical Applications, Part B," p. 8, 2018.

[23] (AI HLEG) High-Level Expert Group on Artificial Intelligence, "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment," European Commission, Text, Jul. 2020. Accessed: Feb. 04, 2021. [Online]. Available: https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[24] European Court of Human Rights and Council of Europe, *European Convention on Human Rights*. 2013, p. 34.

[25] Council of Europe, "The European Social Charter," *European Social Charter*. https://www.coe.int/en/web/european-social-charter/home (accessed Mar. 11, 2021).

[26] L. Webb and C. Chou, "Perspectives on Issues in AI Governance," Google, Jan. 2019. Accessed: Mar. 10, 2021. [Online]. Available: https://ai.google/static/documents/perspectives-on-issues-in-ai-governance.pdf.

[27] J. Whittlestone, R. Nyrup, A. Alexandrova, K. Dihal, and S. Cave, "Ethical and societal implications of algorithms, data, and artificial intelligence: a roadmap for research," *Lond. Nuffield Found.*, 2019.

[28] J. Leikas, R. Koivisto, and N. Gotcheva, "Ethical framework for designing autonomous intelligent systems," *J. Open Innov. Technol. Mark. Complex.*, vol. 5, no. 1, p. 18, 2019.

[29] D. Helbing, Bruno S. Frey,Gerd Gigerenzer,Ernst Hafen,Michael Hagner,Yvonne Hofstetter,Jeroen van den Hoven,Roberto V. Zicari, Andrej Zwitter, "Will Democracy Survive Big Data and Artificial Intelligence?," *Scientific American*, Feb. 25, 2017. https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/ (accessed Mar. 10, 2021).

[30] B. Edwards, *The new drawing on the right side of the brain*, 2nd rev. ed. New York: Jeremy P. Tarcher/Putnam, 1999.

[31] A. Dhurandhar *et al.*, "Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives," *ArXiv180207623 Cs*, Oct. 2018, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1802.07623.

[32] ISO/IEC 14764:2006, "Software Engineering - Software Life Cycle Processes - Maintenance." https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/90/39064.html (accessed Mar. 11, 2021).

[33] B. Düdder, F. Möslein, N. Stürtz, M. Westerlund, and R. V. Zicari, "Ethical Maintenance of Artificial Intelligence Systems," in *Artificial Intelligence for Sustainable Value Creation*, M. Pagani and R. Champion, Eds. Edward Elgar Publishing, 2021.

[34] J. Moubray, *Reliability-centered Maintenance*. Industrial Press Inc., 2001.

[35] F. Möslein and R. V. Zicari, "Certifying Artificial Intelligence Systems," in *Research Handbook on Big Data Law*, R. Vogt, Ed. Edward Elgar Publishing, 2021.

[36] WHO, "Cardiovascular diseases." https://www.who.int/westernpacific/health-topics/cardiovascular-diseases (accessed Oct. 27, 2020).

[37] J. Stegenga, *Medical nihilism*. 2020.

[38] E. Sanz, J. P. Steger, and W. Thie, "Cardiogoniometry," *Clin. Cardiol.*, vol. 6, no. 5, pp. 199–206, May 1983, doi: 10.1002/clc.4960060502.

[39] W. M. M. Schüpbach *et al.*, "Non-invasive diagnosis of coronary artery disease using cardiogoniometry performed at rest," *Swiss Med. Wkly.*, vol. 138, no. 15–16, pp. 230–238, Apr. 2008, doi: 2008/15/smw-12040.

[40] T. Braun *et al.*, "Detection of myocardial ischemia due to clinically asymptomatic coronary artery stenosis at rest using supervised artificial intelligence-enabled vectorcardiography – A five-fold cross validation of accuracy," *J. Electrocardiol.*, vol. 59, pp. 100–105, Mar. 2020, doi: 10.1016/j.jelectrocard.2019.12.018.

[41] European Parliament and Council of European Union, "Council Directive 93/42/EEC of 14 June 1993 concerning medical devices," *Off. J. Eur. Communities*, vol. L 169, pp. 1–43, Jul. 1993.

[42] R. E. Schapire, "Explaining AdaBoost," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, B. Schölkopf, Z. Luo, and V. Vovk, Eds. Berlin, Heidelberg: Springer, 2013, pp. 37–52.

[43] J. Brusseau, "What a Philosopher Learned at an AI Ethics Evaluation," *AI Ethics J.*, vol. 1, no. 1, Dec. 2020, doi: 10.47289/AIEJ20201214.

[44] T. Grote and P. Berens, "On the ethics of algorithmic decision-making in healthcare," *J. Med. Ethics*, vol. 46, no. 3, pp. 205–211, Mar. 2020, doi: 10.1136/medethics-2019-105586.

[45] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson, "The What-If Tool: Interactive Probing of Machine Learning Models," *IEEE Trans. Vis. Comput. Graph.*, pp. 1–1, 2019, doi: 10.1109/TVCG.2019.2934619.

[46] R. K. E. Bellamy *et al.*, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," *ArXiv181001943 Cs*, Oct. 2018, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1810.01943.

[47] N. Martinez, M. Bertran, and G. Sapiro, "Fairness With Minimal Harm: A Pareto-Optimal Approach For Healthcare," *ArXiv191106935 Cs Stat*, Nov. 2019, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1911.06935.

[48] A. Beutel *et al.*, "Putting Fairness Principles into Practice: Challenges, Metrics, and Improvements," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA, Jan. 2019, pp. 453–459, doi: 10.1145/3306618.3314234.

[49] R. Perrault *et al.*, "The AI Index 2019 Annual Report," AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA, Dec. 2019.

[50] V. Eubanks, *Automating inequality: how high-tech tools profile, police, and punish the poor*, First Edition. New York, NY: St. Martin's Press, 2017.

[51] S. U. Noble, *Algorithms of oppression: how search engines reinforce racism*. New York: New York University Press, 2018.

[52] J. Casey *et al.*, "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."

[53] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[54] I. A. Cristea, E. M. Cahan, and J. P. A. Ioannidis, "Stealth research: Lack of peer-reviewed evidence from healthcare unicorns," *Eur. J. Clin. Invest.*, vol. 49, no. 4, p. e13072, 2019, doi: https://doi.org/10.1111/eci.13072.

[55] C. Ball, "What Is Transparency?," *Public Integr.*, vol. 11, no. 4, pp. 293–308, Sep. 2009, doi: 10.2753/PIN1099-9922110400.

[56] C. Stohl, M. Stohl, and P. M. Leonardi, "Digital Age | Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age," *Int. J. Commun.*, vol. 10, no. 0, Art. no. 0, Jan. 2016.

[57] H. Hilligoss and J. Fjeld, "Introducing the Principled Artificial Intelligence Project," *Cyberlaw Clinic, Harvard Law School, Berkman Klein Center for Internet & Society*, Jun. 07, 2019. https://cyber.harvard.edu/story/2019-06/introducing-principled-artificial-intelligence-project (accessed Mar. 10, 2021).

[58] Partnership on AI, "About ML," *The Partnership on AI*, 2019. https://www.partnershiponai.org/about-ml/ (accessed Mar. 11, 2021).

[59] D. Leslie, "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector," Zenodo, Jun. 2019. doi: 10.5281/ZENODO.3240529.

[60] IEEE P7006, "Standard for Personal Data Artificial Intelligence (AI) Agent." https://standards.ieee.org/project/7006.html (accessed Mar. 11, 2021).

[61] "Artificial Intelligence: How to get it right. Putting policy into practice for safe data-driven innovation in health and care," NHSX, London, United Kingdom, Oct. 2019. Accessed: Mar. 11, 2021. [Online]. Available: https://www.nhsx.nhs.uk/media/documents/NHSX_AI_report.pdf.

[62] Datenethikkommission, "Opinion of the Data Ethics Commission," Federal Ministry of Justice and Consumer Protection, Berlin, Germany, Oct. 2019. Accessed: Oct. 27, 2020. [Online]. Available: https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3.

[63] D. Reisman, J. Schultz, K. Crawford, and M. Whittaker, "Algorithmic Impact Assessments: A practical framework for public agency accountability," AI Now Institute, Apr. 2018. Accessed: Mar. 10, 2021. [Online]. Available: https://ainowinstitute.org/aiareport2018.pdf.

[64] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards," *ArXiv180503677 Cs*, May 2018, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1805.03677.

[65] A. Xiang and I. D. Raji, "On the Legal Compatibility of Fairness Definitions," *ArXiv191200761 Cs Stat*, Nov. 2019, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1912.00761.

[66] U. Bhatt *et al.*, "Explainable Machine Learning in Deployment," *ArXiv190906342 Cs Stat*, Jul. 2020, Accessed: Nov. 05, 2020. [Online]. Available: http://arxiv.org/abs/1909.06342.

[67] Partnership on AI, "Explainable AI in Practice Falls Short of Transparency Goals," *The Partnership on AI*, Jan. 14, 2020. https://www.partnershiponai.org/xai-in-practice/ (accessed Mar. 11, 2021).

[68] M. Arnold *et al.*, "FactSheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM J. Res. Dev.*, vol. 63, no. 4/5, p. 6:1-6:13, Jul. 2019, doi: 10.1147/JRD.2019.2942288.

[69] M. Hind *et al.*, "Experiences with Improving the Transparency of AI Models and Services," *ArXiv191108293 Cs*, Nov. 2019, Accessed: Mar. 10, 2021. [Online]. Available: http://arxiv.org/abs/1911.08293.

[70] M. Du, N. Liu, and X. Hu, "Techniques for interpretable machine learning," *Commun. ACM*, vol. 63, no. 1, pp. 68–77, Dec. 2019, doi: 10.1145/3359786.

[71] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, May 2010, doi: 10.1093/bioinformatics/btq134.

[72] E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Trans. Assoc. Comput. Linguist.*, vol. 6, pp. 587–604, Dec. 2018, doi: 10.1162/tacl_a_00041.

[73] T. Gebru *et al.*, "Datasheets for Datasets," *ArXiv180309010 Cs*, Mar. 2020, Accessed: Jan. 11, 2021. [Online]. Available: http://arxiv.org/abs/1803.09010.

[74] M. Mitchell *et al.*, "Model Cards for Model Reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, NY, USA, Jan. 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[75] V. Arya *et al.*, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," *ArXiv190903012 Cs Stat*, Sep. 2019, Accessed: Mar. 11, 2021. [Online]. Available: http://arxiv.org/abs/1909.03012.

[76] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "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*, New York, NY, USA, Apr. 2020, pp. 1–14, doi: 10.1145/3313831.3376445.

[77] L. J. Robertson, R. Abbas, G. Alici, A. Munoz, and K. Michael, "Engineering-Based Design Methodology for Embedding Ethics in Autonomous Robots," *Proc. IEEE*, vol. 107, no. 3, pp. 582–599, Mar. 2019, doi: 10.1109/JPROC.2018.2889678.

[78] S. Hallensleben *et al.*, "From Principles to Practice : An interdisciplinary framework to operationalise AI ethics," Bertelsmann Stiftung, Gütersloh, 2020. [Online]. Available: https://www.bertelsmann-stiftung.de/fileadmin/files/BSt/Publikationen/GrauePublikationen/WKIO_2020_final.pdf.

[79] J. Gonas, "Is Wall Street endorsing, or even dictating, a moral standard in the capital markets? Implications of the popularity of ESG (environmental, social, and governance) mutual funds and ETFs," *EnLightening Talks*, Apr. 2019, [Online]. Available: https://repository.belmont.edu/enlightening/Spring_2019/multimedia_hall_spring_2019/11.

[80] G. Holly, L. Smit, and R. McCorquodale, "Making sense of managing human rights issues in supply chains - 2018 report and analysis," British Institute of International and Comparative Law, 2018. Accessed: Mar. 11, 2021. [Online]. Available: https://www.biicl.org/documents/1939_making_sense_of_managing_human_rights_issues_in_supply_chains_-_2018_report_and_analysis_-_full_text.pdf?showdocument=1.