Digital Society (2023) 2:35
https://doi.org/10.1007/s44206-023-00063-1

ORIGINAL PAPER

Lessons Learned from Assessing Trustworthy AI in Practice

Dennis Vetter · Julia Amann · Frédérick Bruneault · Megan Coffee · Boris Düdder · Alessio Gallucci, et al. ***[full author details at the end of the article]***

Received: 5 January 2023 / Accepted: 22 August 2023 / Published online: 9 September 2023

Abstract

Building artificial intelligence (AI) systems that adhere to ethical standards is a complex problem. Even though a multitude of guidelines for the design and development of such trustworthy AI systems exist, these guidelines focus on high-level and abstract requirements for AI systems, and it is often very difficult to assess if a specific system fulfills these requirements. The Z-Inspection® process provides a holistic and dynamic framework to evaluate the trustworthiness of specific AI systems at different stages of the AI lifecycle, including intended use, design, and development. It focuses, in particular, on the discussion and identification of ethical issues and tensions through the analysis of socio-technical scenarios and a requirement-based framework for ethical and trustworthy AI. This article is a methodological reflection on the Z-Inspection® process. We illustrate how high-level guidelines for ethical and trustworthy AI can be applied in practice and provide insights for both AI researchers and AI practitioners. We share the lessons learned from conducting a series of independent assessments to evaluate the trustworthiness of real-world AI systems, as well as key recommendations and practical suggestions on how to ensure a rigorous trustworthiness assessment throughout the lifecycle of an AI system. The results presented in this article are based on our assessments of AI systems in the healthcare sector and environmental monitoring, where we used the framework for trustworthy AI proposed in the *Ethics Guidelines for Trustworthy AI* by the European Commission's High-Level Expert Group on AI. However, the assessment process and the lessons learned can be adapted to other domains and include additional frameworks.

Keywords Artificial intelligence · Trustworthy AI · AI ethics · AI assessment · Z-Inspection®

1 Introduction

The ever-growing capabilities of artificial intelligence (AI) systems have the potential to transform many parts of society. At the same time, the increasing complexity of these systems makes it more and more difficult to determine whether their use improves or preserves the desired social outcomes. Another concern is how the adoption of these new algorithms and the lack of precise knowledge and control over their inner workings may impact the people that use them to support their decisions (AI HLEG, 2019; Bommasani et al., 2021; Zicari et al., 2021b). This leads to specific and mostly unintentional risks that are considered within AI ethics, and as a consequence, the quest for ethical and trustworthy AI has become a central issue for governance and technology impact assessments efforts in the last years when implementing AI systems (Bommasani et al., 2021). These efforts have resulted in the existence of nearly 100 high-level guidelines for the development of ethical AI (Datenethikkommission, 2019; Hagendorff, 2020; Jobin et al., 2019; Mittelstadt et al., 2016; Morley et al., 2021; Schiff et al., 2020; Zeng et al., 2018). However, at the same time, there seems to be a significant mismatch between the high-level ethical guidelines and practical implications for AI research and development, as there also is no shortage of reports describing unethical applications of AI (Angwin et al., 2016; Hamilton, 2018; Morley et al., 2021; Thorbecke, 2019). One of the main reasons is that the current frameworks are very abstract, have limited practical application for researchers and developers, and offer only very few practical insights into algorithms and AI systems (Bélisle-Pipon et al., 2022; Morley et al., 2021).

Z-Inspection® (Zicari et al., 2021b) alleviates such limitations by offering a process to assess the trustworthiness of an AI system at any stage of its lifecycle. The Z-Inspection® process was initially based on the *Ethics Guidelines for Trustworthy AI*, which were proposed by the European Commission's High-Level Expert Group on AI (AI HLEG; AI HLEG, 2019) and are also currently used as the foundation of the requirements for AI systems under the upcoming AI Act (European Commission, 2021). Z-Inspection® is a holistic process for evaluating new technologies, where the ethics of specific use-cases are discussed by elaborating socio-technical scenarios. In particular, Z-Inspection® can be used to co-design, self-assess, or conduct independent audits of AI systems together with the stakeholders owning the use-case.

This article is a methodological reflection on Z-Inspection® by the original developers and users of the Z-Inspection® process. It illustrates for both researchers and practitioners how the AI HLEG's guidelines for trustworthy AI can be applied in practice. The reflections and recommendations are based on recent past assessments for trustworthy AI, primarily in the field of medicine, conducted using the Z-Inspection® process. In addition, we were also successful in applying the process to assess a remote sensing system for environmental monitoring. Therefore, we are convinced that the process can also be easily used in other domains and adapted to other ethical frameworks.

1.1 The EU Ethics Guidelines as a Framework for Trustworthy AI

We mainly base Z-Inspection® on the framework for trustworthy AI proposed in the EU *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019). It is important to note that, in these guidelines, trustworthiness concerns not only the development, deployment, and use of AI, but also the complete socio-technical systems involving the AI applications, which include, for example, also the humans, corporations, infrastructure, standards, or existing laws relevant to the use of the AI (AI HLEG, 2019). Trustworthy AI is seen as a path to reap benefits in a way that is aligned with European foundational values of respect for human rights, democracy, and the rule of law. The guidelines propose that an AI system must respect all applicable laws and regulations, respect ethical principles and values, and be robust from a technical and social perspective. Furthermore, the guidelines define four ethical principles for trustworthiness: respect for human autonomy, prevention of harm, fairness, and explicability. Building on these ethical principles, the guidelines propose seven key requirements (each further subdivided into several aspects called sub-requirements) for the practical implementation of the ethical principles. These key requirements are (1) human agency and oversight; (2) technical robustness and safety; (3) privacy and data governance; (4) transparency; (5) diversity, non-discrimination, and fairness; (6) societal and environmental well-being; and (7) accountability (AI HLEG, 2019). While considering the seven requirements comprehensive, Z-Inspection® also proposed two new/additional requirements: (8) "*assessing if the ecosystems respect values of Western Modern democracy*" and (9) "*avoiding concentration of power*" (Zicari et al., 2021b).

1.2 Relevance, Challenges, and Limitations of the EU Framework for Trustworthy AI

The AI HLEG trustworthy AI guidelines were formulated as non-legal and non-binding guidance to direct the development of AI towards the consideration of a wide range of ethical principles in a bid to balance innovation with safety (AI HLEG, 2019; Hickman & Petrin, 2021). Given the broad scope of AI systems and the fact that the definition of the term AI itself is still a matter of debate, the seven requirements laid out in the guidelines have not been anchored to a specific context (Zicari et al., 2021c).

In connection with the guidelines, the EU framework offers a static checklist and a web tool, the Assessment List for Trustworthy Artificial Intelligence (ALTAI) (AI HLEG, 2020; Insight Centre, n.d.), which is designed to enable self-assessment of the trustworthiness of AI systems. However, the ALTAI assessment provides no validation of claims regarding trustworthiness and does not take into account changes in AI technology over time. In addition, due to its very broad and general nature, some of the questions and recommendations are likely to not apply to a specific AI system. For example, a large part of the questions regarding technical robustness is about the AI system's resilience to cyber attacks, a part we found inapplicable

during our assessments, as the devices running the AI system were not connected to the internet. Another set of questions is concerning universal design and accessibility. This is especially relevant if the AI system under assessment will be used by a variety of users with different backgrounds. However, the systems we had under assessment were mainly targeted at a small, homogeneous group of specialized users, which made the questions less relevant to our assessments. In addition to the inapplicability of some questions, recommendations such as "You should ensure that the AI system corresponds to the variety of preferences and abilities in society," while useful points for consideration, are very abstract and it is not obvious how to implement them. Nonetheless, the ALTAI questionnaire can serve well as a starting point to identify areas on which future assessments should focus.

The AI HLEG trustworthy AI guidelines have also formed the foundation upon which a proposed AI Act has been built (European Commission, 2021). At the time of writing, the AI Act is not yet enacted, but as drafted it categorizes AI systems (defined broadly) into those that bear unacceptable risk, high risk, limited risk, and low risk (Madiega, 2022). While AI systems with unacceptable risks are prohibited, those AI systems categorized as high risk will need to address several requirements before being used or placed on the market. The requirements mentioned in the AI Act build on the AI HLEG guidelines, for example, transparency and human oversight (Mökander et al., 2022). Assessment as to whether those requirements are fulfilled by specific AI systems will be carried out by any third-party bodies who are already responsible under pre-existing product safety legislation (such as the Medical Device Regulation (MDR) for any medical devices being brought to market in the EU) during the process of awarding the CE marking (European Commission, 2021, art. 44; Mökander et al., 2022). Currently, the AI Act proposes that AI systems that are not subject to pre-existing legislation will need to have their conformity with the requirements self-assessed by those responsible for it or by an independent third-party auditor (European Commission, 2021; Mökander et al., 2022). However, while the AI Act is being developed, and probably even after its finalization, there will still be aspects of AI systems that will not be covered by any law but need to be governed by ethics, including the central question for all AI systems, "is an AI system the most appropriate and ethical solution for the problem at hand?" High-level guidelines and principles alone do not guarantee the ethical use of AI. What is needed is a thorough reflection on the relevance and implications of ethical concepts and principles in the context of specific AI systems.

2 The Z-inspection® Process to Assess Trustworthy AI in Practice

Despite the challenges and limitations of the Ethics Guidelines for Trustworthy AI, the underpinning principles are very useful in organizing a systematic assessment of AI systems, especially compared to other guidelines which do not detail specific requirements. That is why we build on the four ethical principles and the seven key requirements for trustworthy AI in the Z-Inspection® process by applying them in practice. Implementing the EU guidelines and identifying potential ethical tensions and concerns which can affect the trustworthiness of an AI system in

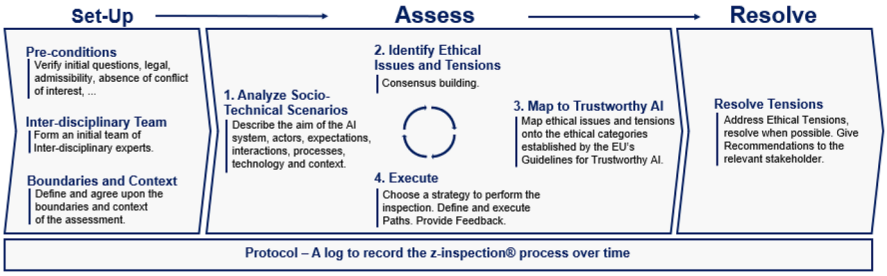

Fig. 1 Z-Inspection® process flow chart describing the main steps of the *set-up*, *assess*, and *resolve* phases. In parallel to the phases, a log is kept in which the process and events of the assessment are tracked. Adapted from Zicari et al. (2021b)

a given socio-technical context require more than a one-size-fits-all approach. For a thorough consideration of the various aspects of a use-case and its subject area environment, independent interdisciplinary experts with specific knowledge bases relevant to the respective context are needed. We created a *participatory process*, called Z-Inspection®, to help teams of skilled experts assess the *ethical*, *technical*, *legal*, and *domain-specific* implications of using an AI product or service within given contexts (Zicari et al., 2021b).

The Z-Inspection® process offers a voluntary, non-binding assessment of potential ethical concerns and issues that may surface in the use of an AI system. It is designed to complement, rather than supplant, other assessments focused on the AI system's compliance with relevant laws, standards, and regulations.

Z-Inspection® started as the university research project of a team of computer science researchers that were trying to use the EU trustworthy AI guidelines to perform an assessment of an actual AI system. While the researchers could assess the technical properties of the system, they were not equipped to assess the domain-specific properties of the system, or sufficiently identify the ethical dilemmas or implications of issues they discovered. Therefore, domain experts, ethicists, and other use-case-specific experts were contacted to make their expertise available to the assessment process. In addition, a process was developed to streamline the collection of information, onboarding of team members, and identification of ethical issues. Today, the Z-Inspection® initiative consists of an international group of researchers and practitioners from many different areas that are all interested in ethical, societal, legal, and trustworthy AI. They either work on assessments as part of their research or volunteer their time. They do not receive compensation for their participation in the assessment, and there are no financial contributions from the stakeholders of the system under assessment to any participants or the initiative.

The process comprises three phases: (1) set-up, (2) assess, and (3) resolve. A schematic description of them is presented in temporal order in Fig. 1.

The *set-up phase* consists of the validation of several pre-conditions before the assessment starts including the legal admissibility and absence of conflict of interest, the setup of an interdisciplinary team of experts working together with the key stakeholders owning the specific AI use-case, and finally, the definition of

the boundaries and context where the assessment takes place. The *assess phase* is an iterative process that includes the creation and analysis of socio-technical scenarios, the identification of ethical issues and tensions, the validation of claims by providing evidence (if any), and the mapping to the EU trustworthy AI framework using a mapping from "open to closed vocabulary" as a consensus-based approach. The *resolve phase* addresses the ethical tensions identified during the assess phase, here possible trade-off solutions are proposed, possible risks and remedies are identified, and recommendations are made to the key stakeholders. A detailed description of the three phases can be found in Zicari et al. (2021b).

The Z-Inspection® process can be applied to the entire AI lifecycle, typically including (1) design, (2) development, (3) deployment, (4) monitoring, and (5) decommissioning. In the *design phase*, the process can provide insight, adjustments, and recommendations on how to design a trustworthy system. During the *development phase*, the process can be used to specify test cases, e.g., to verify the absence of certain biases, especially when requirements change or the original requirements from the design phase are refined. During *deployment*, the process can be integrated into the acceptance test if trustworthiness is a specified user requirement. It can also be used to provide recommendations for the mitigation of ethical risks and the handling of potential tensions and trade-offs. Since AI systems evolve over time due to updated models, algorithms, data, or environments, the trustworthiness of a system needs to be assessed as a continuous *monitoring* process. The *decommissioning* of systems and replacement by other systems is a critical activity due to the required compatibility for other systems using the functionality of the soon-to-be-replaced AI system. Here, the protocols and logs (recording document) of the process over the full lifecycle of the old trustworthy AI system can facilitate the lifecycle of the new product.

Z-Inspection® uses a holistic, interdisciplinary, and dynamic approach, rather than monolithic and static ethical checklists. Using a holistic process means that the Z-Inspection® assessment brings different considerations together to provide an assessment of the system as a whole and as part of a functioning socio-technical unit. This interdisciplinarity ensures that a variety of expert methodologies, cultural ontologies, and disciplinary interpretations are expressed while assessing the trustworthiness of an AI system. The holistic and dynamic approach determines which issues are central to the use-case at different stages of the process, and moves back and forth between intradisciplinary and interdisciplinary discussions of which aspects of the case are most significant. In this way, the process has a certain degree of plasticity, which means that its assessments will be tailored to the use-case at hand. It also means that the assumptions that guide the AI system's creation and deployment—as well as the assumptions of the researchers conducting the Z-Inspection®—are exposed and evaluated during the inspection. Finally, the dynamic lens reflects the commitment to an ongoing and iterative investigation of the harms and benefits of a particular AI system. While each round of the Z-inspection® process is necessarily limited to a particular development phase, it provides space for participants to openly reflect and document what is known (and unknown) about the system's capabilities as a baseline for subsequent evaluations.

3 Reflections on the Z-inspection® Process

In the two years since its initial inception, we have used Z-Inspection® to assess trustworthy AI for four healthcare use-cases, and an environmental monitoring use-case, which included AI models in different stages of their development and with different requirements. This has provided us the opportunity to further improve and develop the initial Z-Inspection® process using complex real-world examples. Our previous collaborations include evaluating (1) a deployed AI pipeline estimating the risk of cardiovascular disease (Brusseau, 2020; Zicari et al., 2021b); (2) a deployed AI system for the detection of cardiac arrest in emergency calls (Amann et al., 2022; Blomberg et al., 2019; Zicari et al., 2021c); (3) co-design of a deep learning-based tool that helps dermatologists understand AI predictions of skin lesion malignancy (Lucieri et al., 2020; Zicari et al., 2021a); and (4) a deep learning-based system for support of radiologists in the pulmonary analysis of COVID-19 patients (Allahabadi et al., 2022; Signoroni et al., 2021). In the environmental monitoring use-case, the AI system was used to automatically monitor the health of heather fields via satellite images (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2022). All assessments were conducted as self-assessments together with the stakeholders owning the use-cases i.e., the ones responsible for the development and/or deployment of the AI system in their organization. In the rest of the paper, we will use the term *use-case owners* to refer to these stakeholders.

In the following, we will reflect on our use of the EU Trustworthy AI guidelines, lessons learned, and how involvement in these assessment cases helped us with improvements and refinements of the original Z-Inspection® process as introduced in Zicari et al. (2021b)[1]. Therefore, the presented reflections and refinements are informed by the subjective experiences and opinions of the authors as academic researchers and participants in previous assessments. Due to the interdisciplinary nature of our assessment team with participants from areas such as computer science, machine learning, and deep learning, healthcare ethics, AI ethics, healthcare, and policy and governance, we can include a multitude of different viewpoints. We developed the changes to the original process through internal post-hoc evaluations of previous assessments, interviews with external stakeholders of past assessments, and new ideas from participants during the assessments. We are only including changes whose usefulness we could verify in at least one other assessment. And finally, while our learnings were obtained during assessments of AI systems in the healthcare sector, we believe that our findings are general enough to translate well to other sectors. The refinements will be discussed in order of the different steps introduced in the original process.

[1] An extended version of this article is available in Zicari et al. (2022).

3.1 The Set-up Phase

3.1.1 Verification of Pre-conditions

The first step of the assessment process is the definition of pre-conditions and boundaries of the assessments. The pre-conditions include information on who requested the assessment, what happens with the results of the assessment, whether important stakeholders will be left out of the assessment, or if the inspection will be revisited at a later date with potentially different participants. This is also the part in the assessment where it is clarified what intellectual property the use-case owner holds and how the assessment should work with it, as well as a discussion of possible conflicts of interest from the use-case owners. We found that initially establishing these pre-conditions helps in setting the expectations for the assessment from both the assessment team and use-case owners and it influences what experts will be needed. We also found that this is the right moment to discuss how to handle intellectual property from the use-case owners. Depending on the use-case, this could include requests for signing non-disclosure agreements and restrictions to the information we could include in publications.

3.1.2 The Choice of Experts

The choice of experts required for each use-case has an ethical implication since the quality of the analysis and the results depend on the diligent selection and quality of experts. This includes the experts not being biased or in a position of conflict of interest. Domain experts may need to include several classes of expertise and practice, especially as we found that AI systems often impact multiple categories of professionals and the impact on each of these affected groups can be different.

We also encountered cases where the design, implementation, and/or management of the AI system, or at least of parts of the AI system, were outsourced to a third-party vendor. In such cases, we adopted the policies that the third-party vendor would not be part of the assessment team to avoid any conflict of interest, and the main use-case owners would need to declare that they do not have any involvement with the third-party vendor that could lead to a conflict of interest. In some cases, we additionally decided that the use-case owners agree to act as the sole communication channel with the third-party vendor so that no communication between the vendor and the assessment team is happening. We decided on these policies in particular after an assessment, where the AI system was already being sold as a commercial product. In that case, we deemed it too big of a conflict of interest for the original developers to participate in the assessment, as their interests, in addition to their concerns regarding their intellectual property, could conflict with that of the process.

Furthermore, over the course of our assessment, we found that we could group the participants of the assessment process into different roles with different responsibilities towards the success of an assessment:

1. A *team lead* to coordinate the process and manage the creation of reports,
2. A *rapporter*, who writes minutes, manages the meetings with stakeholders, and organizes provided information
3. *Domain experts*, ideally more than one, to assist inter alia with their knowledge of the problem domain, and to verify the problem specification of the task the AI system is aimed at solving,
4. *Legal experts*, ideally with a specialization in the problem domain as well as an understanding of the legal aspects of data protection and human rights, to give an early warning for possible liabilities and privacy or discrimination issues,
5. *Technical experts*, with a specialty in machine learning, deep learning, and data science to assess the technical dimension of the AI system,
6. *Representatives of end-users*, to include the perspective of users of the AI system,
7. *Representatives of affected populations*, as the users of the system and the people affected the most by the system can be different; in our assessments of AI systems in healthcare, this role was filled by patient representatives,
8. *Ethicists/philosophers* to help other team members identify ethical aspects, dilemmas, and tensions, and to give guidance towards possible solutions.

In addition to these required expertises, we found that legal experts specialized in the problem domain, social scientists, policymakers, and communication specialists can provide valuable input.

Team members should be selected based primarily on required skills and expertise—availability, motivation, and interest in the case are essential but should not be the primary criteria for involvement. To ensure the quality of the inspection process, it is important that all team members respect the specific areas of competency of each other. Once experts from all relevant areas are present, later additions to the team should be limited, or preferably avoided. We found that adding experts at later stages could lead to imbalances in the team's viewpoints and destabilize the assessment workflow. However, sometimes specific knowledge requirements were only identified later in the process. In such cases, the addition of new expertise was important.

We also found it important to limit the number of team members to 10–20 experts, as the team size correlates to the complexity of the AI assessment. In one use-case, we had a large team including over 40 interdisciplinary experts and we had to split the work into parallel working groups. In other use-cases, we did not have to split the work into parallel working groups since we had a midsize team including around 20 interdisciplinary experts. If the team is too small and it does not reflect the truly interdisciplinary nature of the assessment work, it will likely be incomplete. However, if the team is too big with too much overlap of knowledge and expertise, the assessment process may become cumbersome and delayed due to communication overhead. During the assessments, we learned that the most efficient teams were the ones of small to medium size (i.e., 10–20 experts) that included experts from all fields involved. An efficient assessment team could consist of the following experts:

1 Team lead

1 Rapporter
2 Use-case owners
3 Domain experts (i.e., general physicians, radiologists, …)
3 Technical experts
1 Representative of end users
2 Ethicists
2 Legal experts

The most important aspect is to include specialists from all fields. A larger span of domains provides various advantages since the team can draw conclusions and analogies on a broader spectrum of real-world similar use-cases or problems. The benefit of having multiple experts from the same field on the assessment team is that they can provide vastly different perspectives, depending on their exact specializations and working conditions. For example, during our assessments of healthcare-related AI systems, we saw that different clinicians will work with different populations and can therefore have different concerns regarding equity and diversity, and may recognize important gaps. In addition, different clinicians can have different appreciations of the measurements of the effectiveness of the AI system. For example, a general physician can have slightly different objectives and expectations from AI compared to a dermatologist. Finally, while there can be merit in having representatives of those who will be impacted by the AI, if, and how, to involve such end users in a self-assessment process of an AI system requires further experimentation to allow for lessons learned.

3.1.3 Definition of the Boundaries

The set-up phase also includes the definition of the boundaries of the assessment, taking into account that we do not assess the AI system in isolation but rather consider the social-technical interconnection with the ecosystem(s) where the AI is developed and deployed. Explicitly defining the boundaries helped us in keeping the assessment process focused by aligning on which parts of the system to include and which not, a decision that also heavily influences the considerations of the later assessment (Dobbe et al., 2021). However, while keeping the assessment focused, it is at the same time important to expand the assessment as far as feasible to reflect the inherent diversity in societies and the need for equity, particularly, for medical devices. The explicit definition of boundaries also includes clearly recognizing and stating any limits of the assessment. In our experience, the aim of having a "trustworthy AI system" and the fact that our assessments are voluntary, non-binding self-assessments, contributed positively to the use-case owners' interest in expanding the process beyond a narrow or compliance-based focus. Expanding the assessment as much as possible has the benefit that the ethical issues identified are more comprehensive and better reflect the ethical perspectives present in pluralistic and diverse societies.

3.2 The Assess Phase

3.2.1 Analysis of Socio-technical Scenarios

One of the specific aspects of the methodology involves using so-called socio-technical scenarios (Leikas et al., 2019; Lucivero, 2016), in order to anticipate possible usage and problems of the system under review. The scenarios are built around experiences that result from the intended use of the AI system. A trustworthy AI assessment can not be performed on the technical components alone. What can be considered ethical or problematic strongly depends on the broader societal context where the AI system is used, with consideration of people, institutions, and cultures (Chopra & Singh, 2018; Selbst et al., 2019). Anticipating different experiences helps with this, as one can then "look" at the situation from different points of view, highlighting different approaches and appraisals of the technology at hand and its usage implications. The team draws from their diverse experiences in technological and ethical assessments to debate the specific context of the situation. This prevents abstract opposition between general principles (Lucivero, 2016). By collecting the relevant resources, a team of interdisciplinary experts creates socio-technical scenarios and analyzes them to describe the aim of the AI systems, the actors and their expectations and interactions, the processes where the AI systems are used, and the technology and the context (*ecosystem*). For past use-cases, the socio-technical scenarios were developed over a number of workshops, where the use-case owners presented the AI system to the assessment team and answered clarifying questions regarding system specifications, actors, and intended use. This way, the stakeholders and our interdisciplinary assessment team could develop a shared understanding of the objective of the system, its capabilities, and its limitations (Dean et al., 2021). We found it useful to collect a written summary of these workshops, a template structure for which can be found in Appendix 1[2]. Team members are then encouraged to go through the materials again and ask questions in the document where they need additional clarifications from the use-case owners.

We found it important to differentiate between information provided by the use-case owners and information provided by the assessment team. The information coming from the use-case owners is absolutely crucial to the process. However, we also observed tendencies of use-case owners to control a narrative about how the AI system would be used and affect users and subjects, even though they might not have had experience with larger downstream effects. Therefore it is important that the use-case owners produce precise, verifiable statements, "verifiable claims" (Brundage et al., 2020), on the capabilities of the AI system, for which they can show supporting facts or evidence. This helps the team get an overview of the AI system's actual capabilities and avoids speculation and analysis of hypothetical scenarios. This also includes "*Concept Building*" (Whittlestone et al., 2019), where possible vagueness and ambiguities in the description of the AI system are addressed and

[2] The template and other resources developed for the assessments are also available on GitHub at https://github.com/dennisrv/z-inspection-toolkit..

clarified. This could be, for example, a focus on abstract concepts such as "fairness" without stating explicit definitions for this concept, or the use of domain-specific terms that might carry different meanings in other disciplines and might therefore be misunderstood by parts of the assessment team (Whittlestone et al., 2019). Systematic organization of evidence and concept building also benefits the use-case owners by showing them where the information they provide can be misunderstood and what kinds of evidence they need to produce, as well as a way to assess if the evidence they produce is sufficient. We found the Claims, Arguments, and Evidence (CAE) Framework (Bloomfield & Netkachova, 2014) useful for this organization process, as it provides detailed guidance on how to disseminate complex claims into easier verifiable ones.

However, we also learned that there may be tensions when considering what the relevant existing evidence to support a claim is, and when managing the different viewpoints between experts composing the assessment team. Opposite points of view regarding both the claims of what the AI system can do and the assessment of the issues it poses may both have good arguments in favor of them, and they might both be supported by peer-reviewed scientific literature. In such cases, both viewpoints likely have their merits, and it is, therefore, useful to be aware of and articulate them both. Recognizing these disagreements can even lead to recommendations for the use-case owners on how they can mitigate potential risks or manage concerns in practice. We encountered one such disagreement during the co-design of a deep learning-based tool to help dermatologists detect malignancy in skin lesions (Lucieri et al., 2020; Zicari et al., 2021c). This disagreement between two valid expert opinions and the open discussion around it helped the use-case owners re-evaluate and refine the main goal of their AI system. Managing different viewpoints also made evident that the process requires researchers to bring their own ideas and arguments to the discussion of aspects of the case while at the same time understanding that their input is a contribution to teamwork, not a matter of "winning the argument." Having a team leader who is tasked, empowered, and capable of bringing the different viewpoints together holds a space for dialog, and steering the processes toward the CAE approach is essential for the success of the assessment.

3.2.2 Identification of Ethical Issues and Tensions

Initially, the description of ethical issues and tensions was performed by an interdisciplinary sub-team (Zicari et al., 2021b). However, in later assessments with larger teams and shared backgrounds between team members, we found it more successful to separate the interdisciplinary assessment team into different working groups with a common background, for example, a working group of radiologists, one of machine learning experts, and one of the social scientists. This allowed for more efficient communication within working groups and a way to identify issues with the AI system from multiple perspectives (Allahabadi et al., 2022; Vetter et al., 2022; Zicari et al., 2021a, c).

Each of these working groups then analyzes the socio-technical scenarios and produces a preliminary report, independently from the other working groups. In this report, they summarize the potential problems they see with the system. This

allows the inspection work to proceed in parallel and also avoids cognitive biases while taking advantage of the different unique perspectives of experts from different fields. The preliminary reports are then shared with the entire team for feedback and comments to allow a common insight into the viewpoints of experts from all fields. Based on this feedback, the working groups develop their final report, listing the identified ethical, technical, domain-specific, and legal issues they see with the system. We found it useful to include the use-case owners early in this identification phase, as their input on perceived issues was frequently needed. Furthermore, this helped to avoid that important evidence was being provided by the use-case owners only at the end of the assessment process.

With the exemplary team composition introduced in Sect. 3.1.2, the team could be split into four working groups: (1) domain experts and end-users, (2) technical experts, (3) legal experts, and (4) affected people. The use-case owners and ethicists support the different groups by providing them with information as needed, and the project lead is coordinating information exchange. The domain experts discuss the domain-specific points of the AI system, such as the following questions: Is the problem the AI is trying to solve well-stated? Is it based on the most recent knowledge? Do the outputs of the AI system reflect decisions a domain expert would make? Is the training data reflecting domain agreed relevance and soundness for the output sought? The technical experts focus on the implementation of the AI system, such as whether it uses state-of-the-art technology, whether the used datasets are large and diverse enough, and if the evaluation procedure is sound. The legal experts look at issues concerning privacy, data protection regulations, decision-making, and good administration, as well as other requirements, and the representative of affected people express all their concerns regarding the system, as these concerns will need to be addressed by the use-case owners in the future. In our past use-cases in the healthcare sector, the end-users could often be included with the domain experts, as the systems under assessment were often aimed at supporting specialists, so the domain experts and end-users were the same group of people or at least people of a similar background.

Since some experts in the assessment team may not have a background in ethics, we use a predefined catalog of ethical tensions as examples to help the identification of "issues" or to help articulate why and how, for example, technical issues could lead to ethical issues. Ethical experts in the group can then provide some additional feedback regarding the theoretical aspects of the discussion while participating in the assessments of the socio-technical scenarios. Specifically, we used the catalog of tensions defined by Whittlestone et al. (2019), namely:

Accuracy vs. fairness
Accuracy vs. explainability
Privacy vs. transparency
Quality of services vs. privacy
Personalization vs. solidarity
Convenience vs. dignity
Efficiency vs. safety and Sustainability
Satisfaction of preferences vs. equality

Table 1 Example of an issue and its mapping to the ethical principles (bold) and key requirements (italics) of the EU *Ethics Guidelines for Trustworthy AI.* Adapted from (Allahabadi et al., 2022)

Issue	Low system transparency
Description	It can be difficult to establish a link between input image and output severity score. The system is not easily explainable due to its many blocks and complexities
Mapping	**Respect for Human Autonomy** > *Human Agency and Oversight* **Prevention of Harm** > *Technical Robustness and Safety* **Explicability** > *Transparency*

In addition, Whittlestone et al. point out the three "conceptual lenses" of power, time, and locus, which can help with the identification of additional tensions: *winners versus losers*, *short-term versus long-term*, and *local versus global*. For example, an AI system might benefit a specific local population in the short term while penalizing the overall population's well-being in the long run (Whittlestone et al., 2019, p. 23).

Once the ethical tensions have been identified as part of the case study assessment, the next question is how—if at all—these ethical tensions can be resolved. Thus, the next step in the assessment process consists in deciding which of the options is available to choose from. For example, in the case of an identified ethical tension between accuracy and fairness, whether to choose the option that maximizes the overall statistical accuracy of the system or the option that minimizes disparate impact on minority groups. In this step of the assessment, the distinction between *true dilemmas*, *dilemmas in practice*, and *false dilemmas*, as suggested by Whittlestone et al. (2019), proved to be very useful. The classification of tensions into one of these three categories was performed via group consensus and then later used to inform the recommendations given to the use-case owners (see Sect. 3.3).

3.2.3 Mapping of Ethical Issues to Trustworthy AI Requirements

The mapping of Ethical Issues to the Ethical Principles and Requirements for Trustworthy AI proposed in the EU *Ethics Guidelines* (AI HLEG, 2019) is an essential step of the Z-Inspection® process. Until this point, the issues found by the different working groups are described in an open vocabulary. However, as the experts in the different working groups do not necessarily have a background in ethics, the issues described so far are, for example, technical, social, or domain-specific issues. In the mapping step, these issues are mapped to the ethical principles, key requirements, and sub-requirements of the EU *Ethics Guidelines for Trustworthy AI* they are found to be in conflict with. This, in turn, helps the non-ethicists in the working groups to vocalize how, and why, a specific issue might manifest as an ethical issue that impacts the trustworthiness of the AI system under assessment. Table 1 below provides an example of such a mapping and the corresponding issue from (Allahabadi et al., 2022).

An important lesson from the mapping process was that a mapping in terms of the four principles of trustworthy AI turned out to be too coarse, whereas mapping to the sub-requirements presented the group with a multitude of options and may make the mapping too difficult. Focusing the mapping on the seven requirements proved to be a useful conceptual middle ground for the mapping process. Additionally, sometimes it is not obvious which of the ethical principles or key requirements best applies to an issue and multiple pillars or requirements can apply. Thus, the exact mapping of an issue strongly depends on the background of the person performing the mapping. This demonstrates that experts from different backgrounds can shed light on different perspectives on the underlying issues. We consider this an advantage and a strength of the interdisciplinary nature of the Z-Inspection® process. Another relevant strength of the Z-Inspection® process is that it can be used with any framework for trustworthy AI that proposes ethical values that AI systems should respect. Instead of the ethical principles and key requirements for trustworthy AI proposed by the EU, the mapping could also be used to map the issues to the principles and values proposed in the UNESCO *Recommendation on the Ethics of Artificial Intelligence* (UNESCO, 2021) or the principles proposed in the OECD *Recommendation of the Council on Artificial Intelligence* (OECD, 2019). This allows the Z-Inspection® process to flexibly adapt to a variety of different contexts.

The mapping of issues is followed by a consolidation. The consolidation aims to produce one final list of issues that captures the findings of the different groups, as well as the related ethical principles and key requirements for trustworthy AI. For this, we established a separate working group, the *mappers*, that consists of members from every working group. The mappers identify overlaps in the issues described by the different working groups and their mappings.

However, this proved to be a non-trivial task. An explicit goal of the consolidation phase is that the mappers should agree upon the final result. When the number of issues is relatively small (e.g., less than 10), it is feasible to consolidate them manually (Brusseau, 2020), and have these results evaluated by other experts. However, in one of the most complex use-cases we assessed (Allahabadi et al., 2022), we had a very large team of experts (over 40) who identified over 50 issues. Due to a large number of participants, the manual approach was not feasible, as it proved too demanding for a single person to be aware of all issues for consolidation. This, in turn, also made it difficult to get expert consensus, as we had no initial version to discuss, but only a large list of issues. We first tried to separate the issues into smaller groups according to the trustworthy AI key requirements they were mapped to. Still, we quickly found that this approach was not working well, as different WGs assigned similar issues to vastly different key requirements.

The main challenge was that the evaluation work was split into different working groups that often used different terminologies and jargons popular in their sub-fields, such as artificial intelligence, ethics, or medicine. This led to situations where the different working groups described, with free text, similar issues using different terminologies or from different perspectives. This made the consolidation task of mappers difficult since identifying overlaps between such issues was quite challenging. Indeed, performing the consolidation manually was both cognitively challenging and time-consuming. Therefore, we

developed a *natural language processing* (NLP) tool to scan the issue descriptions for semantic similarities and identify clusters of issues that describe similar problems. The identification of such clusters then helped us in managing the large number of issues and getting an overview of common topics between them. And while the result was not perfect, it drastically reduced the complexity of the mapping step. This in turn allowed for more active contribution and discussion by the mappers, who then produced the final consolidation of issues based on group consensus. (Vetter et al., 2022). However, our experience with this AI-supported approach is limited to one use-case. While the AI could support this case, where the number of issues was very high, we still consider the human component of the mapping process, which is based on group consensus, essential to our assessment.

3.3 The Resolve Phase

The resolve phase completes the process by addressing ethical tensions and by giving recommendations to the key use-case owners. The recommendations might, for example, concern appropriate use, remedies for mitigating risks, and the ability to redress.

One way in which trade-offs and recommendations have been developed in practice is through discussion in the working groups. To give an example, in one of the ethics working groups, we listed a set of concrete recommendations and our reasons for highlighting them. For instance, when considering the development of an AI system during a pandemic, we must consider how to trade off standard procedures for securing informed consent against the need for speedy training of an algorithm. We found that it was recommendable that a policy was put in place making sure to protect patient rights during a pandemic, where very high societal costs are at stake, and such rights can come under pressure. This recommendation grew out of a more general discussion about ethical issues relating to the use-case. One of the main challenges here is how to motivate and engage with the main use-case owners to ensure that they act upon (some of) the given recommendations. This is an open area of practical research.

However, we also found that different types of tensions identified during the assessment phase can inform different recommendations. If a tension was identified as a *true dilemma*, the main recommendation to the use-case owners was to be aware of this dilemma and to openly communicate and motivate what trade-off they are pursuing. For *dilemmas in practice*, which are often technical in nature, the recommendations often concern technical aspects of the system. These recommendations can, for example, include additional data collection efforts, additional model validation efforts, or better ways for end-users to provide feedback to the developers. And finally, if a tension is classified as a *false dilemma*, this is because an option was found for the AI system where the norms or principles are not in conflict. In such cases, this third option is presented and recommended to the use-case owners.

3.4 Monitoring AI Systems Over Time

It is crucial to monitor that the AI system that fulfilled the trustworthy AI requirement after its initial deployment continues to do so over time. Multiple factors can evolve, both regarding the AI system and in its use and context. For example, the learned model changes its behavior with updated training data or new data for inference, the software and hardware of the execution environment change, used machine learning libraries are updated, or the human decision makers are not using the AI system's outputs as the AI engineers expected. Similarly, the decision-making processes or contexts where the AI system is deployed may change and there may be unwanted results from the use of the system. A system once considered trustworthy cannot be guaranteed to remain trustworthy for its lifetime, given the multitude of possible changes. It is also possible that the "ground truth" or "gold standard" changes if the knowledge in that field evolves. Therefore, when required, the resolve phase includes conducting a trustworthy monitoring of the AI system over time, which we call "ethical maintenance" (Düdder et al., 2020). One initial benefit of this ethical maintenance is consistently updated documentation about the deployed system reflecting its current and past states. This regular documentation can also be beneficial during system maintenance and in the decommissioning phase when the system needs to be replaced or shut down (Gilbert et al., 2022).

4 AI Certification and Fundamental Rights Assessment

The certification of AI-based products and services is a growing need for companies wanting to sell products in safety–critical areas (IEEE SA—The IEEE Standards Association, n.d.). A related requirement for assessment will likely soon be required under the forthcoming AI Act (Mökander et al., 2022). Our process can assist in the certification process by providing a trustworthy AI assessment of company claims for the system, or by providing a structured process for the companies to perform self-assessments of their AI systems. However, the Z-Inspection® initiative is not a certifying body under any jurisdiction, nor is the process aimed at compliance or complete in terms of certification. A fundamental rights assessment for AI systems used, or considered to be used, by the government (Gerards et al., 2022) has recently been proposed as a law by the Dutch parliament. Our process nicely complements such an impact assessment as a dynamic counterpart that helps verify claims and identifies ethical dilemmas and tensions from different interdisciplinary perspectives, and we are currently working on a pilot project in cooperation with the Dutch government where we combined the fundamental rights assessment with the Z-Inspection® process for an even more inclusive view on the ethical implications of AI systems (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, 2022; Z-Inspection® Initiative, 2023).

5 Limitations

Z-Inspection® is a voluntary, non-binding assessment complementing audits for legal compliance and technical robustness. One inherent limitation of this process is that its success depends on good-faith cooperation from the use-case owners that go "beyond compliance" (Selbst, 2021). For our assessments, we rely on use-case owners to request an assessment, provide us with the relevant information, and implement our recommendations for mitigating the discovered ethical issues. Therefore, use-case owners could have incentives to ignore the results or to withhold relevant information to avoid negative results (Costanza-Chock et al., 2022; Mökander et al., 2021). In a similar fashion, the process contains many subjective steps that depend on team members' backgrounds, experience, and opinions, especially for the mapping and consolidation steps. And while this clearly adds value to our process, it bears the risk of falsely disregarding information, due to missing knowledge or biases in the assessment team (Costanza-Chock et al., 2022). One such area of missing knowledge is the lack of inclusion of experts on human behavior, human–machine interaction, or user interface design in previous use-cases. The assessments surfaced ethical issues related to human behavior and the interaction with AI. Through the involvement of domain experts, experts on ethical dilemmas, and governance experts, human behavior issues were identified and highlighted, sometimes with recommendations for further assessment or study (Allahabadi et al., 2022; Zicari et al., 2021c). However, considering the importance of this area, specific expertise in human behaviors and ethical dilemmas could provide valuable new insights for future assessment teams. Furthermore, future assessments of AI systems in healthcare would also likely benefit from including representatives of patients, as their trust in the system can also play a big role in its successful adoption. Another limitation comes from the fact that the assessment team consists of experts that volunteer their time and receive no compensation for their work in the assessment. And while such independent third-party assessments are highly desirable (Costanza-Chock et al., 2022), this also imposes limits regarding the invested time and can lead to uneven contributions in both quality and quantity. We were experiencing these limitations during past assessments, where we were not able to perform potentially desirable tasks such as code reviews, security reviews, or red teaming exercises with the assessed AI systems due to the limitations in the availability of qualified experts. Overall, a successful assessment requires a balanced act and special effort to select a group of experts that are motivated, have the required knowledge, and are willing to contribute quality time (Mökander et al., 2021).

6 Discussion

When comparing Z-Inspection® with other frameworks for assessing trustworthy or ethical AI, such as *Independent Audit of AI Systems* by Falco et al. and ForHumanity (Falco et al., 2021; ForHumanity, 2021), Brown et al.'s framework

for ethical algorithm audits (Brown et al., 2021), the *Reviewability* framework presented by Cobbe et al. (Cobbe et al., 2021), Felländer et al.'s *DRESS-eAI* (Felländer et al., 2022), or the *capAI* procedure developed by Floridi et al. (Floridi et al., 2022), multiple similarities arise. They all share the notion that the assessment should not only cover the technical parts of the AI system, but the complete socio-technical context where the AI is employed. In addition, due to the complexity and interdisciplinarity of these socio-technical contexts, AI assessments need expertise from a multitude of related areas, and the inclusion of a wide range of stakeholders is important. Special focus is also put on the need for "verifiable claims" (Brundage et al., 2020) and how to put forth and organize evidence for these claims. Furthermore, all frameworks are intended for assessments during any stage of the AI lifecycle.

However, among these frameworks, Z-Inspection® is the least rigid. This has the downside that assessments with the Z-Inspection® process require familiarity with the process, as it does not provide explicit lists or surveys of common pitfalls or required evidence. At the same time, this also enables the Z-Inspection® process to dynamically react to context-specific assessment needs. Furthermore, the open approach of identifying issues and mapping them to ethical principles enables participants from different domains to contribute, even without much prior knowledge of AI or ethics. In addition, the assessment process allows flexibility with regard to different sets of norms and values. While we were using the EU Trustworthy AI guidelines (AI HLEG, 2019), the principles and values presented in these guidelines are not directly part of the assessment process and can be exchanged for other values without changes to the assessment process and the process could also easily be used in different geopolitical areas of the world. One consequence of this, however, is that Z-Inspection® does not take a stance on rival theories of justice in the context of AI system development. Important philosophical distinctions, such as the tension between distributive justice (just outcomes) and procedural justice (just decision processes) (Colquitt & Rodell, 2015; Kordzadeh & Ghasemaghaei, 2022), are out of scope for Z-Inspection® as they deal in basic definitions of what is ethical or unethical rather than a method for enacting a given definition. Z-Inspection® could be used to support such a definition by applying it in the context of assessment. Furthermore, as the Trustworthy AI guidelines include both procedural and substantial aspects in their definition of fairness (AI HLEG, 2019, p. 12), we could identify issues related to both principles in past use-cases. This included, for example, differing performance of the AI system for different sub-populations (Allahabadi et al., 2022; Zicari et al., 2021a, c), or lack of informed consent from patients regarding the use of the AI system (Allahabadi et al., 2022; Zicari et al., 2021c).

Z-Inspection® puts the assessment of trustworthiness at its center, of which legal compliance is only one pillar. So while other dedicated auditing tools, such as *capAI* (Floridi et al., 2022), are better suited to assess compliance with legal regulations, the Z-Inspection® process provides a tool for assessing how to make an AI more trustworthy. Z-Inspection® is also less suited for use as an internal continuous auditing tool, a near-real-time support system for auditors that continuously and automatically audits an AI system to assess its consistency (Minkkinen et al., 2022). Its role is much more comparable to that of a scientific peer review, where an

interdisciplinary group of external experts provides non-binding recommendations based on their analysis of the AI system, complementing other audits that ensure compliance with legal and technical regulations.

An important distinction, however, is that Z-Inspection® is not solely founded in theory, but mainly informed by practical experience. It was also already used to carry out multiple independent third-party assessments of real-world AI systems. Finally, the description of the process (Zicari et al., 2021b), as well as extensive use-case descriptions, and the results of the assessments (Allahabadi et al., 2022; Zicari et al., 2021a, c) are readily available as open-access peer-reviewed publications. This open availability of the assessment procedure, use-cases, and results of independent third-party assessments make it a valuable contribution to the growing literature on AI assessments (Costanza-Chock et al., 2022).

7 Conclusions

In evaluating specific use-cases, we developed Z-Inspection®, a participatory process for the assessment of the trustworthiness of AI systems. The process begins by describing the tensions between ethical values using an open vocabulary and gradually narrows the options down to finally agree on the closed vocabulary description of an ethics framework, in our case the EU *Ethics Guidelines for Trustworthy AI* (AI HLEG, 2019). Our process allows for the inclusion of various experts from different backgrounds and provides a structured way for them to find an agreement on ethical issues with AI systems while also including their viewpoints.

While gaps in AI regulation remain, the Z-Inspection® process can provide important validation and ethical considerations in accordance with soft ethics guidelines that go beyond hard legal requirements. With increasing regulation efforts, the Z-Inspection® process will necessarily evolve alongside the regulatory environments, and once regulation is in place, the lessons learned from assessments with Z-Inspection® can assist AI systems developers and users in navigating the legal and ethical requirements. Overall, broad and interdisciplinary subject matter expertise will be critical to making a valuable assessment of trustworthiness, something Z-Inspection® is able to efficiently provide, either in self-assessment or in supplement to third-party assessments of bodies whose remit will be more focused.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44206-023-00063-1.

Acknowledgements We would like to thank Emma Ruttkamp-Bloem, Mikael Boesen, Helga Brogger, and Wonki Min for providing us with valuable feedback.

Funding Open Access funding enabled and organized by Projekt DEAL. DV received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 101016233 (PERISCOPE), and from the European Union's Connecting Europe Facility program under grant agreement no. INEA/CEF/ICT/A2020/2276680 (xAIM). MC received funding from The National Institutes of Health's Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) Fellowship Program in Leadership. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability We did not generate or analyze new datasets. As this article is an analysis of previous work, all relevant data can be found with the related publications.

Declarations

Conflict of Interest The authors declare no competing interests.

References

(AI HLEG) High-Level Expert Group on Artificial Intelligence. (2019). *Ethics guidelines for trustworthy AI* [Text]. European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

(AI HLEG) High-Level Expert Group on Artificial Intelligence. (2020). *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment* [Text]. European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342

Allahabadi, H., Amann, J., Balot, I., Beretta, A., Binkley, C., Bozenhard, J., Bruneault, F., Brusseau, J., Candemir, S., Cappellini, L. A., Chakraborty, S., Cherciu, N., Cociancig, C., Coffee, M., Ek, I., Espinosa-Leal, L., Farina, D., Fieux-Castagnet, G., Frauenfelder, T., & Zicari, R. V. (2022). Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients. *IEEE Transactions on Technology and Society*, *3*(4), 272–289. https://doi.org/10.1109/TTS.2022.3195114

Amann, J., Vetter, D., Blomberg, S. N., Christensen, H. C., Coffee, M., Gerke, S., Gilbert, T. K., Hagendorff, T., Holm, S., Livne, M., Spezzatti, A., Strümke, I., Zicari, R. V., Madai, V. I., & on behalf of the Z-Inspection Initiative. (2022). To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health*, *1*(2), e0000016. https://doi.org/10.1371/journal.pdig.0000016

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine Bias*. ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing?token=l0i8JndZRzf9U7hmG1DlFV6RjLJo1zYf

Bélisle-Pipon, J.-C., Monteferrante, E., Roy, M.-C., & Couture, V. (2022). Artificial intelligence ethics has a black box problem. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01380-0

Blomberg, S. N., Folke, F., Ersbøll, A. K., Christensen, H. C., Torp-Pedersen, C., Sayre, M. R., Counts, C. R., & Lippert, F. K. (2019). Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation, 138*, 322–329. https://doi.org/10.1016/j.resuscitation.2019.01.015

Bloomfield, R., & Netkachova, K. (2014). Building Blocks for Assurance Cases. *IEEE International Symposium on Software Reliability Engineering Workshops, 2014*, 186–191. https://doi.org/10.1109/ISSREW.2014.72

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., & Liang, P. (2021). *On the Opportunities and Risks of Foundation Models*. *ArXiv:2108.07258 [Cs]*. http://arxiv.org/abs/2108.07258

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society, 8*(1), 2053951720983865. https://doi.org/10.1177/2053951720983865

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J.,

O'Keefe, C., Koren, M., & Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. *ArXiv:2004.07213 [Cs]*. http://arxiv.org/abs/2004.07213

Brusseau, J. (2020). What a Philosopher Learned at an AI Ethics Evaluation. *AI Ethics Journal*, *1*(1). https://doi.org/10.47289/AIEJ20201214

Chopra, A. K., & Singh, M. P. (2018). Sociotechnical Systems and Ethics in the Large. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 48–53. https://doi.org/10.1145/3278721.3278740

Cobbe, J., Lee, M. S. A., & Singh, J. (2021). Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 598–609. https://doi.org/10.1145/3442188.3445921

Colquitt, J. A., & Rodell, J. B. (2015). Measuring Justice and Fairness. In R. S. Cropanzano & M. L. Ambrose (Eds.), *The Oxford Handbook of Justice in the Workplace* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199981410.013.0008

Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583. https://doi.org/10.1145/3531146.3533213

Datenethikkommission. (2019). *Opinion of the Data Ethics Commission* (p. 238). Federal Ministry of Justice and Consumer Protection. https://www.bmjv.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN_lang.pdf?__blob=publicationFile&v=3

Dean, S., Gilbert, T. K., Lambert, N., & Zick, T. (2021). Axes for Sociotechnical Inquiry in AI Research. *IEEE Transactions on Technology and Society, 2*(2), 62–70. https://doi.org/10.1109/TTS.2021.3074097

Dobbe, R., Krendl Gilbert, T., & Mintz, Y. (2021). Hard choices in artificial intelligence. *Artificial Intelligence*, *300*, 103555. https://doi.org/10.1016/j.artint.2021.103555

Düdder, B., Möslein, F., Stürtz, N., Westerlund, M., & Zicari, R. V. (2020). Ethical Maintenance of Artificial Intelligence Systems. In M. Pagani & R. Champion (Eds.), *Artificial Intelligence for Sustainable Value Creation*. Edward Elgar Publishing.

European Commission. (2021). *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union legislative Acts* (COM(2021) 206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C., Jirotka, M., Johnson, H., LaPointe, C., Llorens, A. J., Mackworth, A. K., Maple, C., Pálsson, S. E., Pasquale, F., Winfield, A., & Yeong, Z. K. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*(7), Article 7. https://doi.org/10.1038/s42256-021-00370-7

Felländer, A., Rebane, J., Larsson, S., Wiggberg, M., & Heintz, F. (2022). Achieving a Data-Driven Risk Assessment Methodology for Ethical AI. *Digital Society, 1*(2), 13. https://doi.org/10.1007/s44206-022-00016-0

Floridi, L., Holweg, M., Taddeo, M., Amaya Silva, J., Mökander, J., & Wen, Y. (2022). *CapAI - A Procedure for Conducting Conformity Assessment of AI Systems in Line with the EU Artificial Intelligence Act* (SSRN Scholarly Paper No. 4064091). https://doi.org/10.2139/ssrn.4064091

ForHumanity. (2021). *Independent Audit of AI Systems*. https://forhumanity.center/independent-audit-of-ai-systems/

Gerards, J., Schäfer, M. T., Vankan, A., & Muis, I. (2022). *Impact Assessment—Fundamental rights and algorithms* (p. 99). Ministry of the Interior and Kingdom Relations. https://www.government.nl/binaries/government/documenten/reports/2021/07/31/impact-assessment-fundamental-rights-and-algorithms/fundamental-rights-and-algorithms-impact-assessment-fraia.pdf

Gilbert, T. K., Dean, S., Lambert, N., Zick, T., & Snoswell, A. (2022). *Reward Reports for Reinforcement Learning. (arXiv:2204.10817)*. arXiv. https://doi.org/10.48550/arXiv.2204.10817

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Hamilton, I. A. (2018). *Amazon built an AI tool to hire people but had to shut it down because it was discriminating against women*. Business Insider. https://www.businessinsider.com/amazon-built-ai-to-hire-people-discriminated-against-women-2018-10

Hickman, E., & Petrin, M. (2021). Trustworthy AI and Corporate Governance: The EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective. *European Business Organization Law Review, 22*(4), 593–625. https://doi.org/10.1007/s40804-021-00224-0

IEEE SA - The IEEE Standards Association. (n.d.). *IEEE CertifAIEd—The Mark of AI Ethics*. Retrieved November 23, 2021, from https://engagestandards.ieee.org/ieeecertifaied.html

Insight Centre. (n.d.). *How to complete ALTAI - ALTAI*. Retrieved March 2, 2022, from https://altai.insight-centre.org/Home/HowToComplete

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), Article 9. https://doi.org/10.1038/s42256-019-0088-2

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems, 31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Leikas, J., Koivisto, R., & Gotcheva, N. (2019). Ethical Framework for Designing Autonomous Intelligent Systems. *Journal of Open Innovation: Technology, Market, and Complexity*, *5*(1), Article 1. https://doi.org/10.3390/joitmc5010018

Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., Dengel, A., & Ahmed, S. (2020). On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors. *International Joint Conference on Neural Networks (IJCNN), 2020*, 1–10. https://doi.org/10.1109/IJCNN48605.2020.9206946

Lucivero, F. (2016). *Ethical Assessments of Emerging Technologies: Appraising the moral plausibility of technological visions* (1st ed. 2016). Springer International Publishing : Imprint: Springer. https://doi.org/10.1007/978-3-319-23282-9

Madiega, T. (2022). *Briefing—EU Legislation in Process. Artificial intelligence act.* (p. 12). European Parliamentary Research Service. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2021)698792

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties. (2022). *Pilot: Assessment voor verantwoorde Artificial Intelligence - Rijks ICT Gilde - UBRijk* [Webpagina]. Ministerie van Algemene Zaken. https://www.rijksorganisatieodi.nl/rijks-ict-gilde/mycelia/pilot-kunstmatige-intelligentie

Minkkinen, M., Laine, J., & Mäntymäki, M. (2022). Continuous Auditing of Artificial Intelligence: A Conceptualization and Assessment of Tools and Frameworks. *Digital Society, 1*(3), 21. https://doi.org/10.1007/s44206-022-00022-2

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 205395171667967. https://doi.org/10.1177/2053951716679679

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines, 32*(2), 241–268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations. *Science and Engineering Ethics, 27*(4), 44. https://doi.org/10.1007/s11948-021-00319-4

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2021). Operationalising AI ethics: Barriers, enablers and next steps. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01308-8

OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (C/MIN(2019)3/FINAL). Organisation for Economic Co-operation and Development (OECD). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Schiff, D., Biddle, J., Borenstein, J., & Laas, K. (2020). What's Next for AI Ethics, Policy, and Governance? A Global Overview. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 153–158. https://doi.org/10.1145/3375627.3375804

Selbst, A. D. (2021). An Institutional View of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology (harvard JOLT), 35*, 117.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. https://doi.org/10.1145/3287560.3287598

Signoroni, A., Savardi, M., Benini, S., Adami, N., Leonardi, R., Gibellini, P., Vaccher, F., Ravanelli, M., Borghesi, A., Maroldi, R., & Farina, D. (2021). BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Medical Image Analysis*, *71*, 102046. https://doi.org/10.1016/j.media.2021.102046

Thorbecke, C. (2019). *New York probing Apple Card for alleged gender discrimination after viral tweet*. ABC News. https://abcnews.go.com/US/york-probing-apple-card-alleged-gender-discrimination-viral/story?id=66910300

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence* (SHS/BIO/PI/2021/1). United Nations Educational, Scientific and Cultural Organization (UNESCO). https://unesdoc.unesco.org/ark:/48223/pf0000381137

Vetter, D., Tithi, J. J., Westerlund, M., Zicari, R. V., & Roig, G. (2022). *Using Sentence Embeddings and Semantic Similarity for Seeking Consensus when Assessing Trustworthy AI* (arXiv:2208.04608). arXiv. https://doi.org/10.48550/arXiv.2208.04608

Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., & Cave, S. (2019). *Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research*. Nuffield Foundation. https://www.nuffieldfoundation.org/wp-content/uploads/2019/02/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

Zeng, Y., Lu, E., & Huangfu, C. (2018). *Linking Artificial Intelligence Principles. (arXiv:1812.04814)*. arXiv. https://doi.org/10.48550/arXiv.1812.04814

Zicari, R. V., Ahmed, S., Amann, J., Braun, S. A., Brodersen, J., Bruneault, F., Brusseau, J., Campano, E., Coffee, M., Dengel, A., Düdder, B., Gallucci, A., Gilbert, T. K., Gottfrois, P., Goffi, E., Haase, C. B., Hagendorff, T., Hickman, E., Hildt, E., & Wurth, R. (2021a). Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. *Frontiers in Human Dynamics*, *3*, 40. https://doi.org/10.3389/fhumd.2021.688152

Zicari, R. V., Brodersen, J., Brusseau, J., Düdder, B., Eichhorn, T., Ivanov, T., Kararigas, G., Kringen, P., McCullough, M., Möslein, F., Mushtaq, N., Roig, G., Stürtz, N., Tolle, K., Tithi, J. J., van Halem, I., & Westerlund, M. (2021b). Z-Inspection®: A Process to Assess Trustworthy AI. *IEEE Transactions on Technology and Society, 2*(2), 83–97. https://doi.org/10.1109/TTS.2021.3066209

Zicari, R. V., Brusseau, J., Blomberg, S. N., Christensen, H. C., Coffee, M., Ganapini, M. B., Gerke, S., Gilbert, T. K., Hickman, E., Hildt, E., Holm, S., Kühne, U., Madai, V. I., Osika, W., Spezzatti, A., Schnebel, E., Tithi, J. J., Vetter, D., Westerlund, M., & Kararigas, G. (2021c). On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Frontiers in Human Dynamics*, *3*, 30. https://doi.org/10.3389/fhumd.2021.673104

Zicari, R. V., Amann, J., Bruneault, F., Coffee, M., Düdder, B., Hickman, E., Gallucci, A., Gilbert, T. K., Hagendorff, T., van Halem, I., Hildt, E., Holm, S., Kararigas, G., Kringen, P., Madai, V. I., Mathez, E. W., Tithi, J. J., Vetter, D., Westerlund, M., & Wurth, R. (2022). *How to Assess Trustworthy AI in Practice (arXiv:2206.09887)*. arXiv. https://doi.org/10.48550/arXiv.2206.09887

Z-Inspection® Initiative. (2023). *Conference Reader*. First World Z-Inspection Conference, Venice, Italy. https://z-inspection.org/wp-content/uploads/2023/05/World-Z-inspection-conference-reader-.pdf

Authors and Affiliations

Dennis Vetter[1,2] · Julia Amann[3,4] · Frédérick Bruneault[5,6] · Megan Coffee[7] · Boris Düdder[8] · Alessio Gallucci[2] · Thomas Krendl Gilbert[9] · Thilo Hagendorff[10] · Irmhild van Halem[2] · Eleanore Hickman[11] · Elisabeth Hildt[12,13] · Sune Holm[14] · Georgios Kararigas[15] · Pedro Kringen[2] · Vince I. Madai[16,17] · Emilie Wiinblad Mathez[2] · Jesmin Jahan Tithi[2,18] · Magnus Westerlund[13,19] · Renee Wurth[2] · Roberto V. Zicari[2,13,20] · Z-Inspection® initiative (2022)

✉ Dennis Vetter
vetter@em.uni-frankfurt.de

[1] Computational Vision and Artificial Intelligence Lab, Goethe University, Frankfurt, Frankfurt Am Main, Germany

[2] Z-Inspection® Initiative, Venice, Italy

[3] Health Ethics and Policy Lab, ETH Zurich, Zurich, Switzerland

[4] Strategy and Innovation, Careum Foundation, Zurich, Switzerland

[5] Philosophie Departement, Collège André-Laurendeau, Montréal, Canada

[6] École Des Médias, Université du Québec À Montréal, Montréal, Canada

[7] Department of Medicine, Division of Infectious Diseases and Immunology, New York University Grossman School of Medicine, New York City, NY, USA

[8] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

[9] Digital Life Initiative, Cornell Tech, New York City, NY, USA

[10] Cluster of Excellence "Machine Learning: New Perspectives for Science", University of Tuebingen, Tuebingen, Germany

[11] School of Law, University of Bristol, Bristol, UK

[12] Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, IL, USA

[13] Department of Business Management and Analytics, Arcada University of Applied Sciences, Helsinki, Finland

[14] Department of Food & Resource Economics, University of Copenhagen, Copenhagen, Denmark

[15] Department of Physiology, Faculty of Medicine, University of Iceland, Reykjavik, Iceland

[16] QUEST Centre for Responsible Research, Berlin Institute of Health, Charité Universitätsmedizin Berlin, Berlin, Germany

[17] Faculty of Computing, Engineering and the Built Environment, School of Computing and Digital Technology, Birmingham City University, Birmingham, UK

[18] Parallel Computing Labs, Intel, Santa Clara, CA, USA

[19] School of Economics, Innovation and Technology, Kristiania University College, Oslo, Norway

[20] Data Science Graduate School, Seoul National University, Seoul, South Korea