frontiers in Human Dynamics

ORIGINAL RESEARCH
published: 13 July 2021
doi: 10.3389/fhumd.2021.688152

Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier

Roberto V. Zicari[1,2,3], Sheraz Ahmed[4], Julia Amann[5], Stephan Alexander Braun[6,7], John Brodersen[8,9], Frédérick Bruneault[10], James Brusseau[11], Erik Campano[12], Megan Coffee[13], Andreas Dengel[4,14], Boris Düdder[15], Alessio Gallucci[16], Thomas Krendl Gilbert[17], Philippe Gottfrois[18], Emmanuel Goffi[19], Christoffer Bjerre Haase[20,21], Thilo Hagendorff[22], Eleanore Hickman[23], Elisabeth Hildt[24], Sune Holm[25], Pedro Kringen[1], Ulrich Kühne[26], Adriano Lucieri[4,14], Vince I. Madai[27,28,29], Pedro A. Moreno-Sánchez[30], Oriana Medlicott[31], Matiss Ozols[32,33], Eberhard Schnebel[1], Andy Spezzatti[34], Jesmin Jahan Tithi[35], Steven Umbrello[36], Dennis Vetter[1], Holger Volland[37], Magnus Westerlund[2] and Renee Wurth[38]*

[1]Frankfurt Big Data Lab, Goethe University Frankfurt, Frankfurt, Germany, [2]Department of Business Management and Analytics, Arcada University of Applied Sciences, Helsinki, Finland, [3]Data Science Graduate School, Seoul National University, Seoul, South Korea, [4]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany, [5]Health Ethics and Policy Lab, Swiss Federal Institute of Technology (ETH Zurich), Zurich, Switzerland, [6]Department of Dermatology, University Clinic Münster, Münster, Germany, [7]Department of Dermatology, Medical Faculty, Heinrich-Heine University, Düsseldorf, Germany, [8]Section of General Practice and Research Unit for General Practice, Department of Public Health, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Danemark, [9]Primary Health Care Research Unit, Region Zealand, Denmark, [10]École des médias, Collège André-Laurendeau, Université du Québec à Montréal and Philosophie, Montreal, QC, Canada, [11]Philosophy Department, Pace University, New York, NY, United States, [12]Department of Informatics, Umeå University, Umeå, Sweden, [13]Department of Medicine and Division of Infectious Diseases and Immunology, NYU Grossman School of Medicine, New York, NY, United States, [14]Department of Computer Science, TU Kaiserslautern, Kaiserslautern, Germany, [15]Department of Computer Science (DIKU), University of Copenhagen (UCPH), Copenhagen, Denmark, [16]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands, [17]Center for Human-Compatible AI, University of California, Berkeley, CA, United States, [18]Department of Biomedical Engineering, Basel University, Basel, Switzerland, [19]The Global AI Ethics Institute, Paris, France, [20]Section for Health Service Research and Section for General Practice, Department of Public Health, University of Copenhagen, Copenhagen, Denmark, [21]Centre for Research in Assessment and Digital Learning, Deakin University, Melbourne, VIC, Australia, [22]Ethics & Philosophy Lab, University of Tuebingen, Tuebingen, Germany, [23]Faculty of Law, University of Cambridge, Cambridge, United Kingdom, [24]Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, IL, United States, [25]Department of Food and Resource Economics, Faculty of Science, University of Copenhagen, Copenhagen, Denmark, [26]Hautmedizin Bad Soden, Bad Soden, Germany, [27]Charité Lab for AI in Medicine, Charité Universitätsmedizin Berlin, Berlin, Germany, [28]QUEST Center for Transforming Biomedical Research, Berlin Institute of Health (BIH), Charité Universitätsmedizin Berlin, Berlin, Germany, [29]School of Computing and Digital Technology, Faculty of Computing, Engineering and the Built Environment, Birmingham City University, Birmingham, United Kingdom, [30]School of Healthcare and Social Work Seinäjoki University of Applied Sciences (SeAMK), Seinäjoki, Finland, [31]Freelance researcher, writer and consultant in AI Ethics, London, United Kingdom, [32]Division of Cell Matrix Biology and Regenerative Medicine, The University of Manchester, Manchester, United Kingdom, [33]Human Genetics, Wellcome Sanger Institute, England, United Kingdom, [34]Industrial Engineering & Operation Research, UC Berkeley, CA, United States, [35]Intel Labs, Santa Clara, CA, United States, [36]Institute for Ethics and Emerging Technologies, University of Turin, Turin, Italy, [37]Z-Inspection® Initiative, New York, NY, United States, [38]T. H Chan School of Public Health, Harvard University, Cambridge, MA, United States*

OPEN ACCESS

Edited by:
Remo Pareschi, University of Molise, Italy

Reviewed by:
Rocco Oliveto, University of Molise, Italy
Maria Antonietta Grasso, Naver Labs Europe, France

***Correspondence:**
Roberto V. Zicari
roberto@zicari.de

Specialty section:
This article was submitted to Digital Impacts, a section of the journal Frontiers in Human Dynamics

Received: *30 March 2021*
Accepted: *09 June 2021*
Published: *13 July 2021*

Citation:
Zicari RV, Ahmed S, Amann J, Braun SA, Brodersen J, Bruneault Fé, Brusseau J, Campano E, Coffee M, Dengel A, Düdder B, Gallucci A, Gilbert TK, Gottfrois P, Goffi E, Haase CB, Hagendorff T, Hickman E, Hildt E, Holm S, Kringen P, Kühne U, Lucieri A, Madai VI, Moreno-Sánchez PA, Medlicott O, Ozols M, Schnebel E, Spezzatti A, Tithi JJ, Umbrello S, Vetter D, Volland H, Westerlund M and Wurth R (2021) Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier. Front. Hum. Dyn 3:688152. doi: 10.3389/fhumd.2021.688152

This paper documents how an ethically aligned co-design methodology ensures trustworthiness in the early design phase of an artificial intelligence (AI) system component for healthcare. The system explains decisions made by deep learning networks analyzing images of skin lesions. The co-design of trustworthy AI developed here used a holistic approach rather than a static ethical checklist and required a multidisciplinary team of experts working with the AI designers and their managers.

Ethical, legal, and technical issues potentially arising from the future use of the AI system were investigated. This paper is a first report on co-designing in the early design phase. Our results can also serve as guidance for other early-phase AI-similar tool developments.

Keywords: artificial intelligence, healthcare, trustworthy AI, ethics, malignant melanoma, Z-inspection®[1], ethical co-design, trustworthy AI Co-design

TRUSTWORTHY ARTIFICIAL INTELLIGENCE CO-DESIGN

Our research work aims to address the need for co-design of trustworthy AI in healthcare using a holistic approach, rather than monolithic ethical checklists. This paper summarizes the initial results of using an ethically aligned co-design methodology to ensure a trustworthy early design of an AI system component. The system is aimed to explain the decisions made by deep learning networks when used to analyze images of skin lesions. Our approach uses a holistic process, called Z-inspection® (Zicari, et al., 2021b), to help assisting engineers in the early co-design of an AI system to satisfy the requirement for Trustworthy AI as defined by the High-Level Expert Group on AI (AI HLEG) set up by the European Commission. One of the key features of the Z-inspection® is the involvement of a multidisciplinary team of experts co-creating together with the AI engineers, their managers to ensure that the AI system is trustworthy. Our results can also serve as guidance for other similar early-phase AI tool developments.

Basic Concepts

Z-inspection® can be considered an ethically aligned co-design methodology, as defined by the work of Robertson et al. (2019) who propose a design process of robotics and autonomous systems using a co-design approach, applied ethics, and values-driven methods. In the following, we illustrate some key concepts.

Co-Design

Co-design is defined as a collective creativity, applied across the whole span of a design process, that engages end-users and other relevant stakeholders (Robertson et al., 2019). In their methodology, Robertson et al. (2019) suggest that the design process is open, in the sense that within this process "interactions occur in a broader socio-technical context"; this is the reason why "stakeholder engagement should not be restricted to end-user involvement but should encourage and support the inclusion of additional stakeholder groups" which are part of the design process or which are impacted by the designed product. The ethical aspects of the process and product must also be considered in relation to the "existing regulatory environment (. . .) to facilitate the integration of such provisions in the early stages" of the co-design.

Vulnerability

Robertson et al. (2019) mention that "within a socio-technical system where humans interact with partially automated technologies, an end-user is *vulnerable* to failures from both humans and the technology". These failures and the risks associated with them are symptomatic of power asymmetries embedded in these technologies. This stresses the importance of an "exposure analysis that employs a metric of end-user exposure capable of attributing variations across measurements to specific contributors (which) can aid the development of designs with reduced end-user vulnerability".

Exposure

Exposure represents an evaluation of the contact potential between a hazard and a receptor (Robertson et al., 2019), for this reason, the authors state that: "A threat to an end-user, engaging with a technological system is only significant if it aligns with a specific weakness of that system resulting in contact that leads to exposure". Conversely, every weakness can potentially be targeted by a threat—either external or arising from a component's failure to achieve "fitness for purpose"—and so the configuration of the system's weaknesses influences the end-user's "exposure." They accordingly emphasize that, in the case of autonomous systems, the "analysis of the "exposure" of the system provides a numerical and defensible measure of the weaknesses" of that system, and thus must be an integral part of the co-design process.

In this paper, we focus on the part of the co-design process that helps to identify the possible exposures when designing a system. In the framework for Trustworthy AI, exposures are defined as ethical, technical, and legal issues related to the use of the AI system.

TRUSTWORTHY ARTIFICIAL INTELLIGENCE

Our process is based on the work of the High-Level Expert Group on AI (AI HLEG) set up by the European Commission who published ethics guidelines for trustworthy AI in April 2019 (AI HLEG, 2019). According to the AI HLEG, for an AI to be trustworthy, it needs to be:

Lawful–respecting all applicable laws and regulations,
Robust–both from a technical and social perspective, and
Ethical–respecting ethical principles and values.

The framework makes use of four ethical principles rooted on fundamental rights (AI HLEG, 2019): Respect for human autonomy, Prevention of harm, Fairness, and Explicability.

[1]Z-inspection® is a registered trademark

Acknowledging that these ethical principles cannot give solutions to AI practitioners, the AI HLEG suggested, based on the above principles, seven requirements for Trustworthy AI (AI HLEG, 2019) that should enable the self-assessment of a AI System, namely:

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 wellbeing,
7) Accountability.

These guidelines are aimed at a variety of stakeholders, especially guiding practitioners towards more ethical and more robust applications of AI. The interpretation, relevance, and implementation of trustworthy AI, however, depends on the domain and the context where the AI system is used. Although these requirements are a welcome first step towards enabling an assessment of the societal implication of the use of AI systems, there are some challenges in the practical application of requirements, namely:

- The AI guidelines are not domain specific.
- They offer a static checklist and do not offer specific guidelines during design phases.
- There are no available best practices to show how to implement how such requirements can be applied in practice.

Particularly in healthcare, discussions surrounding the need for trustworthy AI have been soaring. In the next two sections, we illustrate the process of co-design that we use for a specific use case in healthcare. *Related Work* reviews some relevant related work and *Conclusion* presents some conclusions.

CO-DESIGN OF A TRUSTWORTHY ARTIFICIAL INTELLIGENCE SYSTEM IN HEALTHCARE: DEEP LEARNING BASED SKIN LESION CLASSIFIER

In recent years, AI systems statistically reached human-level performances in the diagnosis of malignant melanomas from dermatoscopic images in a visual based experimental setting. Such Computer-Aided Diagnosis (CAD) systems have already yielded higher sensitivity and specificity in diagnosing malignant melanoma analyzing dermatoscopic pictures compared to well-trained dermatologists (Brinker T. J. et al., 2019; Brinker et al., 2019 TJ.). However, the acceptance of these CAD systems in real clinical setups is severely limited primarily because their decision-making process remains largely obscure due to the lack of explainability (Lucieri, et al., 2020a). Moreover, the images used are specific (dermoscopy images), whereas dermatologists will usually palpate, as well as look at the region, to determine the position of the lesion, age, sex, etc.

A team led by Prof. Andreas Dengel at the German Research Center for Artificial Intelligence (DFKI) developed a framework for the domain-specific explanation of arbitrary Neural Network (NN)-based classifiers. Dermatology has been chosen as a first use case for the system. They developed a prototype called Explainable AI in Dermatology (named exAID) (Lucieri, et al., 2020b). exAID combines existing high-performing NNs designed for the classification of skin tumors with concept-based explanation techniques, providing diagnostic suggestions and explanations conforming with the definition of expert approved diagnostic criteria. exAID can therefore be considered in its current status a "trust-component" for existing AI systems. The designers of the exAID hoped to provide dermatologists with an easy-to-understand explanation that can help to guide the diagnostic process (Lucieri, et al., 2020b).

Status: AI System in the early design phase.

The Research Questions

How do we help engineers to design and implement a trustworthy AI system for this use case? What are the potential pitfalls of the AI system and how might they be mitigated at the development stage?

Motivation: This is a co-design conducted by a team of independent experts with the engineering team that performed the initial design of the AI component. The main goal of this research work is to help create a Trustworthy AI roadmap for the design, implementation, and future deployment of AI systems.

Co-Design Framework

A diagrammatic representation of the proposed co-design framework is offered in **Figure 1**

For this use case, three Zoom workshops were organized with 35 experts with interdisciplinary background [including philosophers, ethicists, policy makers, social scientists, medical doctors, legal and data protection specialists, computer scientists and machine learning (ML) engineers] where the initial aim of the early prototype AI system was analyzed.

The initial outcome of this co-design process, as described in this paper, is the redefinition of the AI designs goal and purpose. This was achieved by discussing a number of socio-technical scenarios using an approach inspired by Leikas et al. (2019) and modified for the healthcare domain, consisting of the following phases: 1) Definition the boundaries for the AI system; 2) Identification of the main stakeholders; 3) Identification of the needs based on several different viewpoints; 4) Consideration of the Aim of the system as a claim; 5) Literature review and creation of evidence base; 6) Usage situations, Look for similarities; 7) Measure of risks of harms with respect to different viewpoints; 8) Consideration of ethical and legal aspects.

After this initial discussion with the complete team, the co-design process was conducted in parallel with a number of smaller working groups of three to five experts each. The groups worked independently to avoid cognitive bias of the members, followed by a general meeting to assess and merging of the various results

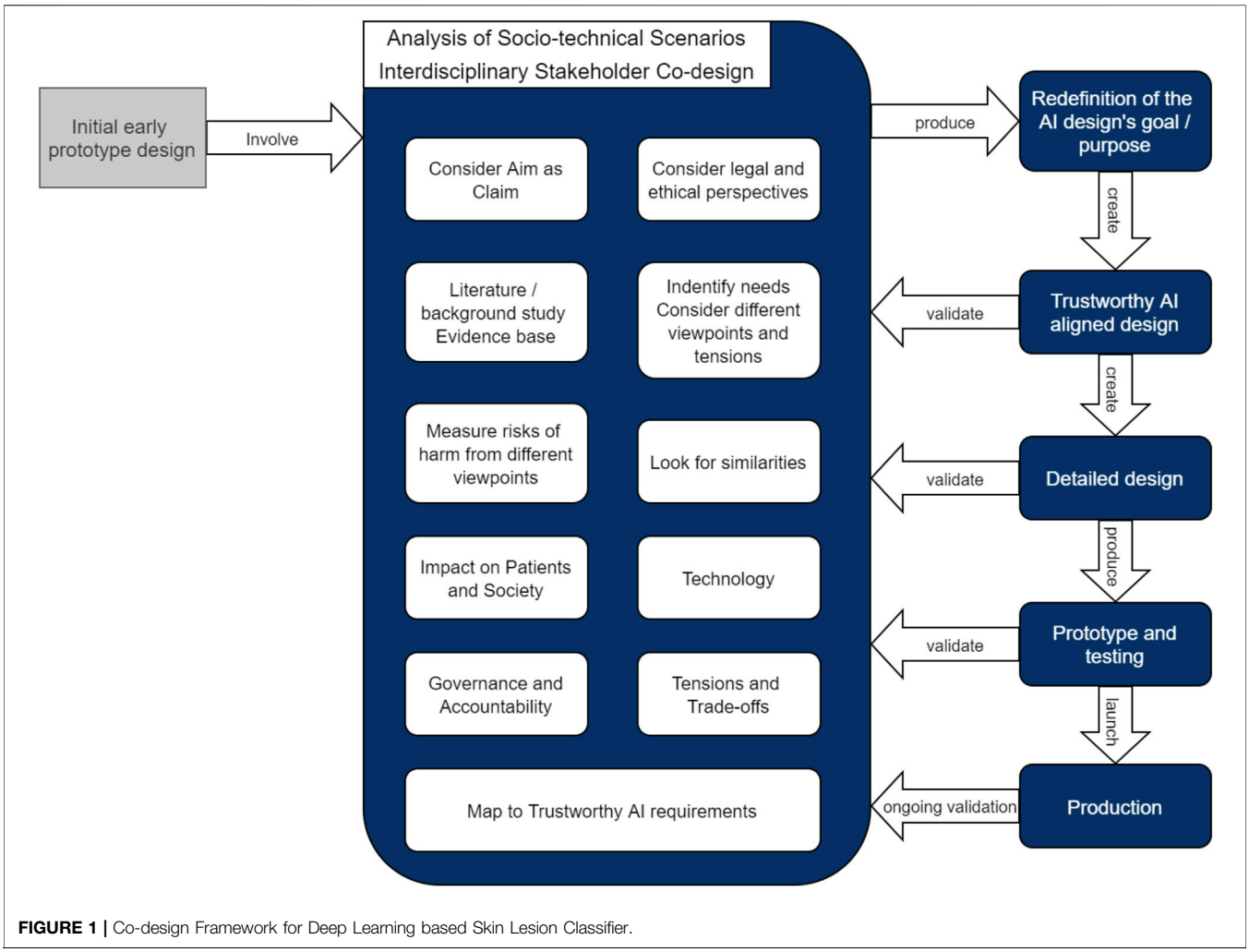

FIGURE 1 | Co-design Framework for Deep Learning based Skin Lesion Classifier.

from the different working groups, taking into account the Impact on the Patients and Society, the Technology, Governance and Accountability. This resulted in the Identification of Tensions, Trade-offs and then the Mapping to the Trustworthy AI requirements, as presented in *Co-Design: Think Holistically*.

In the future, we plan to address the other phases of the co-design process depicted in **Figure 1**, namely: 1) the creation of the initial Trustworthy AI aligned design; 2) the validation of the design by iterating the co-design process; 3) the creation of a detailed design, and the validation of the detailed design by iterating the co-design process; followed by 4) the implementation of the prototype with testing and validation; and finally 5) putting the system in production and ongoing validation.

Socio-Technical Usage Scenarios of the Artificial Intelligence System

Socio-technical usage scenarios is a participatory design tool for achieving a trustworthy AI design and implementation (Leikas et al., 2019). Socio-technical usage scenarios are also a useful tool to describe the aim of the system, the actors, their expectations, the goals of actors' actions, the technology, and the context, while consequentially fostering moral imagination and providing a common ground where experts from different fields can come together (Lucivero, 2016, p. 160). We use socio-technical scenarios within discussion workshops, where expert groups work together to systematically examine, and elaborate the various tasks with respect to different contexts of AI under consideration.

Socio-technical scenarios can also be used to broaden stakeholders' understanding of one's own role in the technology, as well as awareness of stakeholders' interdependence. The theoretical background behind the socio-technical scenarios as a way to trigger moral imagination and debate is linked to the pragmatist philosophical tradition, which states that ethical debates must be both principles-driven and context-sensitive (Keulartz et al., 2002; Lucivero, 2016, p. 156). This is why these tools are especially interesting for the holistic approach of the Z-inspection®.

In Z-Inspection®, socio-technical usage scenarios are used by a team of experts, to identify a list of potential ethical, technical, and legal issues that need to be further deliberated (Zicari, et al.,

2021a). Scenarios are used as part of the assessment of an AI system already deployed, or as a participatory design tool if the AI is in the design phase (as in this case). We have been using socio-technical scenarios within discussion workshops, where expert groups work together to systematically examine and elaborate the various tasks with respect to different contexts of AI.

In this early phase of the design, multiple usages for the same AI technology are possible. Each use of the same technology may pose different challenges. In our approach, we work together with the designers and the prime stakeholders to identify such possible usage of the technology. We consider the pros and cons and evaluate which of the various possible usages is the primary use case. The basic idea is to analyze the AI system using the socio-technical scenarios with relevant stakeholders including designers (when available), domain, technical, legal, and ethics experts (Leikas et al., 2019).

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. 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.

CO-DESIGN: THINK HOLISTICALLY

We report in this section some of the main lessons we learned so far for this use case when assisting the engineers in making the AI design trustworthy. These lessons learned can be considered a useful guideline when assessing how trustworthy an AI design is for similar tool developments.

In the process we did not prioritize discussing different viewpoints or similarities. We began our discussion by building common concepts and discussing the used terminology. In our example we explored our definitions of: 1) early stage vs. localized stage, 2) early diagnosis vs. timely diagnosis, 3) survival vs. mortality, and 4) overdiagnosis vs underdiagnosis. From these discussions similarities and different viewpoints were revealed.

An important aspect of this co-design process is a good balance between sequentiality of actions and freedom of discussion during the workshops. Recurrent, open-minded, and interdisciplinary discussions involving different perspectives of the broad problem definition turned out as extremely valuable core components of the process. Their outcomes turned out to be fertile soil for the identification of various aspects and tensions that could then be transferred to the current state of development of the system within self-contained focus groups to compile action recommendations, streamlining the whole process.

The early involvement of an interdisciplinary panel of experts broadened the horizon of AI designers which are usually focused on the problem definition from a data and application perspective. The co-creation process with its different perspectives highlighted different important aspects, which aided the AI design in a very early phase, benefiting the final system and thus the users. Successful interdisciplinary research is challenging and requires participants to articulate in detail exactly how concepts—including the most basic—are theoretically and practically understood. Using "accuracy of a diagnostic test", for example, not only requires clarification of how each of the words are understood as scientific concepts but also clarification regarding the accuracy for whom, for what purpose, which context, and under what presumptions, etc.

One of the key lessons learned is that by evaluating the design of trustworthy AI with a holistic co-creative approach, we are able to identify a number of problems that were not possible with traditional design engineering approaches. We list some of them in the rest of this section.

Initial Aim of the ML System

The initial aim of the exAID framework was to act as an add-on component to support dermatologists in clinical practice. Given an existing AI system trained for skin melanoma detection, the add-on component goal is to explain the system's decisions in terms that dermatologists can understand. With these explanations, the AI system can support the clinician's decision-making process by providing a qualified second opinion on relevant features in a dermatoscopic image and, therefore, potentially improve diagnostic performance (Tschandl et al., 2020).

The Initial Prototype

Image-based ML algorithms use pattern recognition to derive abstract representations that reveal structures in the image which can be used for automatic classification. This abstract representation can significantly deviate from the human perception of the problem. However, first efforts in decoding the ML representation in skin lesion classification indicate that the process is partially transferable to human understandable domains (Lucieri, et al., 2020a).

The system explanation module transfers the abstract representation of the ML algorithm into expert entrusted concepts to make the ML algorithm's diagnostic suggestion more understandable by a clinician (Lucieri, et al., 2020a; Lucieri, et al., 2020b). These explanatory concepts are borrowed from the established 7-point checklist algorithm (Argenziano et al., 1998) and include the following: 1) Typical Pigment Network, 2) Atypical Pigment Network, 3) Streaks, 4) Regular Dos & Globules, 5) Irregular Dots & Globules, 6) Blue Whitish Veil.

The presence or absence of each of these concepts is quantified. These quantitative results are leveraged by a rule-based method to generate textual explanations and complemented by heatmaps that localize the criteria in the original dermatoscopic image.

Limitations of the Initial Prototype

The project was in the early design phase and no final target groups which can benefit from it were identified. So far, the use by dermatologists in clinic or practice and self-screening by patients

were considered. Furthermore, no clinical trials have been conducted so far.

The ML model currently used for skin lesion classification has been trained on a limited number of publicly available skin images (Mendonça et al., 2013; Codella et al., 2018; Kawahara et al., 2019; Rotemberg et al., 2021). As most of the data was acquired from a small number of distinct research facilities, the dataset suffers from variations in image quality and the occasional presence of artifacts and it cannot be guaranteed that the data is representative of patients' backgrounds (e.g., age, sex, skin tone distribution). The ISIC dataset (Codella et al., 2018) has also been shown to have significant bias in its images (Bissoto et al., 2019, 2020).

Darker skin tones are absent from the training data. As the system should only be used with skin types that were represented in the training datasets, it is at this stage limited to a population of Fitzpatrick skin phototypes I—III (Fitzpatrick, 1988) and is therefore not appropriate to use for those with darker skin tones.

The training data also only contains lesions from certain body regions and excludes others, e.g. there is no training data from moles on the genitals, nails or the lining of the mouth. As the dermatoscopic features of melanoma/naevi differ between body regions, the system should only be used in areas that the underlying AI was trained for.

The designers chose to use dermatoscopic images and therefore, only images from dermatoscopes should be used, excluding clinical images. In addition, the chosen dataset excludes common skin alterations (e.g. piercings, tattoos, scars, burns, etc. The system's use should consequently be restricted to cases where none of these alterations are present. What guardrails are in place to ensure that this model is never used outside of its scope? This may require that all users of the model be well informed of its limitations and scope.

Re-Evaluate and Understand What is the "Aim" of the System.

Deciding and defining the aim of the system is obviously important yet surprisingly ambiguous in a research setting of different disciplines and research traditions.

The original motivation for the design of the AI system was to try solving a general problem given the lack of acceptance of deep neural network enabled Computer-Aided Diagnosis (CAD), as its decision-making process remains obscure. The designers wanted to demonstrate that it is possible to explain the results of a deep neural network used as a deep learning based medical image classifier. They have chosen publicly available data sets and open access neural network to classify skin tumors (Lucieri, et al., 2020a).

The first question we raised at the first multi-disciplinary workshop was: Is there a real need for an AI system as initially presented by the designers?

The first thing we did during the first two workshops was to clarify the motivation and the aim of such an AI system between experts. Successful interdisciplinary research is challenging and requires participants to articulate in detail exactly how concepts—including the most basic—are theoretically and practically understood. Using "accuracy of a diagnostic test", for example, not only requires clarification of how each of the words are understood as scientific concepts but also clarification regarding the accuracy from whom, for what purpose, which context, and under what presumptions, etc.

A relevant aspect we identified up front is whether this AI system is what clinicians and patients want, and if this AI system results in more good than harm. Furthermore, the interdisciplinary exchange brought up several tensions described in detail in subsequent sections.

Consider Different Viewpoints

By considering various viewpoints together with the AI engineers, it is possible to re-evaluate the aim of the system. We present different viewpoints discussed among experts in this subsection.

During the co-creation process, the discussion with different experts in our team, including dermatologists, experts in public health, evidence-based diagnosis, ethics, healthcare, law, and ML, prompted the main stakeholder and owner of the use case, the team of DFKI, to redefine their stated main aim of the system.

When evaluating the design of the use case, it soon became clear that different stakeholders have different scopes, timeframe, and the population in mind. Thanks to the heterogeneity of our team, such differences and tensions were confronted. We present here a summary of different points of view. For each viewpoint, intensive exchange and communication took place between various domain experts, with different knowledge and backgrounds, all of whom form part of our team.

The Dermatologist's View

We first present the dermatologist's view as defined by two of the dermatologists in our team. This viewpoint helps clarify what kind of AI tool could help dermatologists during their daily practice.

The dermatologist's daily routine is to examine skin lesions and to determine if they are benign or if there is a risk of malignancy, thus needing further diagnostic or therapeutic measures. Currently, the diagnostic algorithm mainly consists of patient history, assessment of risk factors, and inspection of the skin with the naked eye. The next step is dermoscopy of suspicious pigmented or not pigmented lesions, sometimes followed by videodermoscopy, which yields higher magnification and high resolution images. Further non-invasive diagnostic measures could be *in-vivo* confocal microscopy or electric impedance spectroscopy (Malvehy et al., 2014).

The aim of this diagnostic algorithm is to determine if the examined lesions bear a risk of malignancy or not. If this can be ruled out, the lesions will be left in place. If the risk is determined to be high, excisional biopsy or rarely incisional biopsy will be performed. In case of uncertainty, video-dermatoscopic follow-up after 3 months is another option. In the end, the final question is always: is there an indication to remove the lesion or not (a functional diagnosis). As malignancy frequently cannot be ruled out by non-invasive measures this leads to the excision of benign lesions. In the particular case of melanocytic lesions, the ratio of benign nevi detected for each malignant melanoma diagnosed, i.e.

the Number Needed to Treat (NNT), is regularly used as an indicator of diagnostic accuracy and efficacy (Sidhu et al., 2012). However, this value depends on the prevalence of the disease and varies according to physician and lesion-related variables (English et al., 2004; Baade et al., 2008; Hansen et al., 2009).

The ratio of total excisions to find one melanoma varies from approximately 6–22 (number needed to treat, NNT) (Petty et al., 2020). A tool that would allow for less unnecessary excisions would thus be very helpful.

The acceptance and use of a diagnostic aid is dependent on many factors at different levels. First there is the medical level. During their training, physicians have learned to seek a second opinion from a colleague if they are uncertain. Here it is usual not only to exchange opinions, but also to include an explanation of why one decides the way one does. Therefore, explainable AI might be very beneficial for the acceptance of CADs since physicians are used to this. Of course, doctors also want to know about the quality of the second opinion. In order to evaluate the meaningfulness of the results of AI for clinical decisions, it is therefore necessary to know about the performance of a technical device (sensitivity, specificity, positive predictive value, negative predictive value, likelihood ratios, ROC (receiver operating characteristic) curve, and Area under the curve (AUC) plus the degree of overdiagnosis (Brodersen et al., 2014).

When it comes to the technical details of the presented prototype, in particular the explainability of dermoscopy criteria, it is necessary to note that uniformly accepted criteria for dermoscopy do not exist. Dermatologists often just attend weekend courses, acquire knowledge from various books or are even self-taught. This can be a problem for the wide acceptance of the discussed device. It would therefore be important that recognized expert groups first establish broadly accepted diagnostic criteria and validate the tool on an expert level. It would also be necessary to check whether existing criteria can be simply transferred to AI-devices.

Then there is a legal level. Many dermatologists fear that they could miss malignant lesions and be sued for them. As a result, lesions are more often removed than left in place. This is also where the personality of the doctor comes into play. Many doctors prefer to be on the safe side. They are also not keen to save data, such as pictures of the lesions, which can later be used against them in lawsuits. Legal questions regarding the use of diagnostic tools must therefore be clarified for the acceptance. It should also not be neglected that economic aspects play an important role. Sometimes technical devices are only used for billing reasons, and the actual output of the device is not taken into account in the decision process.

All these factors should be considered when designing a medical device. The use and acceptance of an AI device is therefore highly dependent on the end user and their training including personality and attitude towards patient care.

The Evidence-Based Medicine View

We now present the Evidence-Based Medicine View as defined by two of our domain experts in our team. This complementary viewpoint helps clarify how the AI tool could help.

In addition to the health threat of melanoma, the phenomenon of overdiagnosis is important to understand given that overdiagnosis affects the interpretation of melanoma detection and treatment effects (Johansson et al., 2019).

Overdiagnosis is about correctly diagnosing a condition that would never harm the patient (Brodersen et al., 2018). The impact of overdiagnosis can be of such magnitude that it increases the incidence, prevalence, and survival rates of the condition. This has been estimated also to be the case with melanoma (Vaccarella et al., 2019). An Australian study has estimated the proportion of overdiagnosis of melanoma in women and men to be 54 and 58% respectively in 2012 (Glasziou et al., 2020).

At the moment, there is no means to establish the potential risk of metastazation in localized stage melanomas. Because of this, most melanomas are excised, and patients are therefore labelled with a diagnosis of having cancer–a malignant disease. Thus, those melanomas that would never turn harmful, therefore counts in cancer statistics as successful (early) detection and 100% treatment effects although they would be as successful if they were not diagnosed in the first place. Therefore, mortality rates are the most valid outcome to evaluate the effect of detection and treatment, while survival rates are invalid.

Overdiagnosis also influences the interpretation of test accuracy. In evidence-based medicine (EBM), test accuracy is traditionally described within the Bayesian diagnostic paradigm as a 2×2 table: the result can be either positive or negative and true or false. Due to overdiagnosis, test accuracy requires a 3×2 table: either positive or negative harmful condition, positive or negative harmless condition, as well as true or false (Brodersen et al., 2014).

Compared to overdiagnosis, the phenomenon of underdiagnosis is less investigated and less clearly defined in EBM. In the research field of melanoma, underdiagnosis has among others been described as a wrong diagnosis/misdiagnosis (Kutzner et al., 2020) and defined as "melanoma that was initially diagnosed as a naevus or melanoma *in situ*" (Van Dijk et al., 2007).

Based on the above, the following questions were posed at co-creation: diagnostic precision of *what*? Is it precision in detecting any melanoma, those melanoma in a certain localized stage or tumor thickness, or only the melanoma that would develop into a metastasizing tumor? Or is it precision in detecting non-melanoma? or potential (non-)melanoma that should be assessed by a dermatologist? Because of overdiagnosis, it is imperative to clarify the precision in detail as it influences the interpretation of the results that follow. If not, sophisticated use of AI may detect more cases of melanoma without benefit to patients, public health and society because these additional detected cases could be due to overdiagnosis and thereby only harmful.

The Population Health View

The Population Health Perspective is provided by three of our domain experts. This viewpoint reframes the questions and goals to highlight how the AI tool can promote population health.

From the population health perspective, it is vital to create an instrument that works towards equity and does not further

exacerbate the inequities already present in melanoma diagnosis and treatment (Rutherford et al., 2015). To achieve this, the criteria of inclusivity must be met, which for this use case entails prioritizing the inclusion of the full spectrum of skin tones, as well as age, and socio-economic status, in the training of the AI model. Additionally, while the clinical problem presents as one of overdiagnosis of disease, it is vital to recognize what population this issue pertains to. While melanoma may be most common among white individuals, a diagnosis is far more likely to lead to fatality for people of color. This is driven by the latency of diagnosis for the latter group (Gupta et al., 2016). Therefore, while the aim for one part of the population might be to focus on overdiagnosis, there should be a distinct aim towards addressing underdiagnosis among populations that are non-white.

The Patient View

We now present the patient view as defined by three of our domain experts in our team. This complementary viewpoint helps clarifying how the AI tool could help the patient. Patient expectations and concerns are of main interest during the whole process from design to application of the AI system. For this use case, we included the patient's view and the general citizen perspective during the various workshops, represented by three of our domain experts in our team. In particular, one of the three team experts, works as a journalist and author on AI acceptance among citizens and had undergone dermatologist checks himself for his own research. No additional patient peer review groups were involved at this stage.

Given that the target audience and purpose of the AI system is not yet clearly defined, it is important to explore several directions and how each of them might impact patients' health and health service experience. At this stage of the process when we refer to a *patient*, we did not distinguish between patients at different stages of interactions with clinicians and their relative percentage (which differs country by country): e.g. patients who are worried presenting to a general practitioner (GP) with a skin lesion and maybe being afraid they have malignant melanoma; patients who visits directly a dermatologists or who are referred to a dermatologist by a GP with a mole that is most frequently diagnosed as a benign lesion or in more rare cases as a malignant melanoma—of which many are overdiagnosed.

It is reasonable to assume that the patient's primary goals are to have their malignant melanoma detected early, and to avoid unnecessary treatment for harmless abnormalities. Any AI system intended as a prevention tool to support lay people in self-screening could represent a significant step towards fostering patients' active involvement in prevention efforts, especially given the shift towards telemedicine since the beginning of the COVID-19 pandemic. Research also suggests that patients are open to using new digital solutions for performing skin self-examinations, as long as they receive adequate technical support for using these tools (Dieng et al., 2019) and don't feel that they compromise the doctor-patient relationship (Nelson et al., 2020). However, there are also certain challenges to the introduction of these new tools from the patient's point of view.

Firstly, it would shift the diagnostic process and responsibility from the healthcare provider to patient, without additional clinical knowledge to confirm the diagnosis. Despite being able to self-screen, the patient would, however, remain dependent on the clinician's decision to act upon the AI system's recommendations. In other words, the system might give patients a false sense of autonomy. In cases where patient and clinician have conflicting views about diagnosis, this could strain the doctor-patient relationship. Moreover, it is important to consider that failing to detect a skin lesion, be it due to technical or human error from the patient's side, may lead to feelings of regret and self-blame (Banerjee et al., 2018; Eways et al., 2020).

Secondly, while some patients may be capable and willing to engage in self-screening practices, others may feel overwhelmed and reluctant to do so, or may simply not be willing to engage in preventive activities (Lau et al., 2014).

Thirdly, from the patient's point of view cost and reimbursement are important considerations. Specifically, state healthcare systems need to clarify who will bear the costs of such a software, clarify medical legal liability (as normally this would be squarely in the hands of the product producer if being advertised as a medical tool) and subsequent treatments, and whether a self-examination can replace a regular check-up by a doctor or increase its frequency.

Despite these ethical issues, the benefits of this technology are numerous. For many patients, the option of primary self-examination can reduce the effort and cost of visiting a doctor. This is particularly true in countries in which preventive examinations are not paid for by a state healthcare system or when visiting doctors, such as due to COVID-19 limitations, is more difficult. Another benefit has to do with the AI providing a second "opinion". A well-designed system can offer detailed explanations of its reasoning underlying a diagnosis, which the clinician can communicate to the patient in layman's language. These explanations can include a statistical probability of the presence of malignant melanoma, and offer a visual comparison of the irregularity under investigation with other cases. This makes it easier for patients to understand why a diagnosis is made.

The patient representative expects that the AI system is designed in a way that its outputs can be easily understood by non-experts, especially by using common every-day language rather than medical terms. Especially if the system is used as a standalone version outside the clinical context as an app, its user interface design must also be easy to understand with clear guidelines on what to do and how to avoid mistakes while operating the software. Medical examinations can be very stressful, even more if the examination detects a potential threat to the patient's health. At all steps access to medical support, for example in the form of a call center that can explain findings and necessary steps as well as help to cope with the stress, is therefore mandatory.

At the same time, if the above is implemented, we need to consider what are the risks on the scope and limitation of the model. How do we ensure that all users are educated if the system is open-sourced? Who is accountable if the model is misused?

The AI system must also provide clear explanations for statistical probabilities, that put numbers into perspective for the specific patient rather than for an abstract group. If the AI system suggests an action, like extraction of a skin sample, the patient wants to know, why exactly the algorithms came to that suggestion. This can help to minimize patient's concerns about a "black box" decision from a machine. On a very practical level, patients want to receive all available information as a printed or digital copy in order to seek a second opinion and have evidence for possible discussions with insurance companies. This is an important point supporting explainability.

The patient representative suggests that user-focus-groups with real patients be established during the design of the system in order to gain maximum acceptance of a later use.

While the patient wants to have clear and explained information another addressed potential tension of self-diagnosis is the lack of the human emotional aspect. The patient does not want to be treated as a mere number on which to collect samples from the lab and spit out a 1, 0 diagnosis but wants, also, to receive understanding and rational human support. We, therefore, expect that such self-diagnosis will be complemented with human support. Hence, in the user-focus-groups various designs can be proposed to hamper this lack such as complementing the tool with a support expert team ready to address the various emotional needs arising from its use.

The AI tool would need to be assessed from the patient perspective. Patients may have different viewpoints regarding the tool's use for self-diagnosis and for use by doctors. Some patients may wish to have a more human connection, while others may appreciate remote, self-led, medical diagnosis. Telemedicine has grown during the COVID-19 epidemic and patient perspectives on the use of self-diagnosis and remote tools may also be evolving during this time. The tool would still require the involvement of the clinician and confirmation by a clinician for future treatment, if needed.

Shifting the diagnostic process from doctor to patient can make the process more accessible (home based care, especially as telemedicine has increased during the COVID-19 epidemic) and also add new complexities, affecting cost and liability and especially as the doctor would need to confirm the diagnosis for further treatment.

The pitfalls in shifting from a doctor-based diagnostic process to a self-diagnosing process are many. First of all, if the pre-test probability of disease is decreasing, then the positive predictive value of a positive test result will decrease, resulting in more false positives. Another problem is that spectrum bias will to a large degree affect the diagnostic process so the diagnostic precision will be lower. Finally, empirical evidence strongly supports that screening for malignant melanoma increases overdiagnosis and there is a danger that self-diagnosing will lead to substantial screening, which again will lead to even more overdiagnosis.

According to recent research related to the use of AI in the diagnostics of skin cancer (Nelson et al., 2020), the majority of patients support the use of AI and have high confidence in its use. This might be due to unnecessary and overseen malignancies being expected. It should be noted that in this context, more overidagnosis could be regarded as beneficial, which is against the view point from the experts in evidence based medicine in our team. Patients would prefer a separate assessment by the physician and the AI system and be informed about the respective results.

According to (Jutzi et al., 2020) 75% of patients would recommend an AI system to their families and friends. and 94% wish symbiosis of the physician with AI. The integrity of the physician-patient-relation must be maintained. There is confidence in the accuracy of AI on the one hand, but concerns about the accuracy on the other hand. A marked heterogeneity of patient's perspectives on the use of AI is noted.

The patient perspective would clearly encompass more than just the self-diagnosis and can be taken in many directions.

It was also noted during the analysis, that in different medical systems than Europe (e.g. US), if this AI system is being used for medical diagnosis without a doctor, this will end up placing the liability in the hands of the designers of the tool in some medical systems. Therefore, perhaps an AI system should not replace the GP or the dermatologist but be used as a decision support system.

Measure the Risk of Harming

For this use case, this is our proposal of an exposure analysis presenting a metric of end-user *exposure* which can aid in the development of an AI design with reduced end-user *vulnerability*.

Not harming, or minimizing harming a patient is an important, if not the most important, requirement for an AI for asserting acceptable ethical standards within screening assessment in the medical field. This risk requires a closer look, and quantification of the term "harm" needs to be determined. Further—the result of the quantification will differ depending on the field in which the AI is used.

To verify a superior result of an AI over a standard or current screening method, the considered set of variables must be representative of different stakeholder views, such as the healthcare system in respective countries, physician, and the patient.

For this use-case we have asked the following questions, 1) what ratio of False Positives to False Negatives is reasonable, 2) is there a standard way of quantifying the costs given differing stakeholder perspectives, 3) how do we assess if the AI system is harming or not? These are also relevant for the definition of the fairness criteria, which typically concerns inequalities in e.g. the error rates for different salient groups.

While early detection of melanoma is of prime importance, overassessment is linked with additional medical costs and an unnecessary physical and psychological burden on the affected patients. Measures reducing misdiagnosis while maintaining or increasing sensitivity have the potential to reduce psychological burden and potential consequences arising from unneeded treatments. Further, an acceptable ratio of false positives vs. negatives will also differ between cases where an AI is used. For instance, erroneous brain surgery in the case of a false positive will be far more serious than a false positive in the case of a mole removal that might or might not lead to cancer.

The number of errors must be kept to a minimum, as the number of false positives (the claim that a healthy patient has a

specific diagnosis) and false negatives (the claim that a sick patient is healthy) might result in overdetection of patients that are healthy, or *not* treating patients that are, in fact, unhealthy or ill. It is important to keep the balance between overdetection reducing the global medical and labor costs while still maintaining a high True Positive detection rate hence ensuring people with positive cases are treated rapidly and adequately.

Requirements from a statistical point of view regarding false positives and negatives, could be expressed as: The false **positive** rate is calculated as $\frac{FP}{FP+TN}$, where FP is the number of **false positives** and TN is the number of true **negatives** (FP + TN being the total number of **negatives**). It is the probability that a **false** alarm will be raised: that a positive result will be given when the true value is negative. The false negative rate—also called the miss rate—is the probability that a true positive will be missed by the test. It is calculated as $\frac{FN}{FN+TP}$, where FN is the number of false negatives and TP is the number of true positives (FN + TP being the total number of positives). The number of false positives must be balanced against the number of false negatives. This ideally gives that both AI false positives and AI false negatives should be lower than the current screening method.

A verified histological assessment should be used as a reference measure for comparing standard methods with that of the AI. The measure confirms the level of error, which in turn must surpass that of the current best practice. The test has four cases with outcomes where only two give an unambiguous answer:

I) $FN^{AI} < FN^{cur}$ and $FP^{AI} < FP^{cur}$
II) $FN^{AI} > FN^{cur}$ and $FP^{AI} < FP^{cur}$
III) $FN^{AI} < FN^{cur}$ and $FP^{AI} > FP^{cur}$
IV) $FN^{AI} > FN^{cur}$ and $FP^{AI} > FP^{cur}$

Considering the implications of the outcome groups:

- In case I) the AI has a better outcome for both variables than the current method and the AI would be considered to comply with the "Do no harm" requirement.
- In the case of outcome IV) the AI does not pass the test as both variables have a less favorable outcome than the current method.
- Whether outcome II) and III) falls within an acceptable ethical range will depend on the gravity of the harm in which a false claim will result.

Further, testing AI-devices solely on historical–and potentially outdated–data is, in itself, a liability and warrants testing in an actual clinical setting. Thus, doctors can compare their own assessment with the AI-recommendations based on data generated from clinical studies and not limit the data to pools from historical data or data from one or two sites. Restricting data to a few sites may limit the racial and demographic diversity of patients and create unintended bias (Wu et al., 2021). However, combining real world evidence data with clinical studies together with clinical experience might help GPs and dermatologists reduce the chance of harming the patient to an acceptable level and in addition avoid costs for unnecessary surgery.

From an ethical point of view, both clinical and big data statistical population evidence must provide the direction, although, each specific AI case under consideration must also strive to consider other factors, like indication, patient prognosis and potential treatment implications as mentioned above.

Look for Similarities

During co-creation, it is important to look for similarities. This helps when defining the aim of the AI and allows one to see what challenges might exist for similar technologies.

Similarities With Genetic Testing and Other Forms of Clinical Diagnostics.

As with most clinical diagnostics, in this use case it is crucial to understand what a test can and cannot tell. When trying to predict future outcomes, no test has 100% sensitivity and specificity and there can be many other variables and intervening events that can lead to a different outcome.

This test carries many of the issues of other cancer screenings. From pap smears to mammograms to prostate-specific antigen test (PSA test) to screen for prostate cancer, tests can point to increased risk or may overlook risk, but are not crystal balls. Treatments that then follow may not be necessary and may lead to unnecessary interventions, or the test results may give a false sense of security and lead to worse outcomes, as the diagnosis is then overlooked.

In view of the differences between the Dermatologist's View, the Evidence-based Medicine View and the Patient View and the resulting tensions with regard to overdiagnosis, risk assessment, risk of harming related to early diagnosis, risk communication, and patient understanding of the test result and its implications, one of the ethics team members of the group felt that this use case could benefit from results of a long-standing interdisciplinary debate on the ethical aspects of predictive genetic diagnosis. Aspects addressed in this debate include patient autonomy, the right to know and the right not to know, psychosocial implications of receiving test results, the clinical significance of test results, lifestyle related questions, the question of which treatment options to choose based on predictions, the harm resulting from treatments that may prove unnecessary (Geller et al., 1997; Burgess, 2001; Hallowell et al., 2003; Clarke & Wallgren-Pettersson, 2019). In order to discuss the melanoma use case in question, it could be very helpful to build on this debate and ask: What are the similarities, what are the differences between (predictive) genetic testing and early detection of melanoma?

Undoubtedly, there are parallels of this use case in dermatology to genetic testing administered through medical doctors or geneticists. Insofar, it seems advisable to think about how the concept of counselling could be transferred to this case. Besides helping patients cope with the psychosocial implications of melanoma analysis and treatment, a reflection of what counselling would imply in the context of melanoma diagnosis, prediction and prognosis would help improve patient-doctor communication about risk. For counselling and clinical risk communication with the individual patient is a

complex task filled with difficulties and pitfalls, e.g. that lay people might understand the concept of overdiagnosis and that strong pre-assumption among lay people might create a perception gap (Hoffmann and Del Mar, 2015; Moynihan et al., 2015; Byskov Petersen et al., 2020). Counselling could also help increase patient involvement and give patients the opportunity to decide on whether they prefer a process involving or not involving a ML system. Similar to this use case, there are two forms of "uses" of (predictive) genetic testing: 1) testing administered by a medical doctor/geneticist, embedded in genetic counselling; and 2) direct-to-consumer genetic testing, the latter coming with additional practical and ethical issues (Caulfield & McGuire, 2012).

There are also relevant differences between this use case and genetic testing. In particular, in predictive genetic testing for serious disorders that may develop in the more distant future, such as in the case of predictive genetic testing for Huntington's disease, genetic testing may trigger complex and adverse psychological outcomes. Predictive genetic testing may burden an individual with information about future serious health deterioration or a distant death, or trigger an irreversible intervention, like removal of breasts and ovaries in a young woman following BRCA mutation testing when other life events, including even future treatments, may intervene (Broadstock et al., 2000; Hawkins et al., 2011; Eccles et al., 2015). Moreover, genetics gives us a sense of self and a sense of who we are. In contrast, the melanoma use case discussed here does not have this level of impact. Also, the time scale with melanoma analysis is much shorter and so is not as prone to the problems of predicting distant outcomes. Furthermore, unlike genetic testing for familial disorders, melanoma testing is only about the individual person undergoing diagnosis.

Consider the Aim of the Future Artificial Intelligence System as a Claim

One of the key lessons learned at this point is that there may be tensions when considering what the relevant existing evidence to support a claim is, or, as in this case, to support the choice of a design decision when considering different viewpoints.

At this stage of early design, we suggest to consider the *aim* of the future AI system as a *claim* that needs to be validated before the AI system is deployed. It is known from the literature (Brundage et al., 2020) that "Verifiable claims are statements for which evidence and arguments can be brought to bear on the likelihood of those claims being true". As mentioned by (Brundage et al., 2020) if the AI system is already deployed, "claims about AI development are often too vague to be assessed with the limited information publicly made available".

There is an opportunity to apply claim-oriented approaches in the early design phase of the AI system. We can use in the design of the AI system the Claims, Arguments, and Evidence (CAE) framework–not as an audit process, but rather as a co-design framework.

If we consider the "aim" of the tool we are co-creating as a *claim*, then we consider the aim as an assertion that needs to be evaluated somehow. Beyond verification, we define the validation of claims as to the use of appropriate forms of evidence and argument to interpret claims as true or false with respect to the original problem statement. Adapting the framework, we then consider *arguments* as linking evidence to the aim of the AI system, and *evidence* as to the basis for justification of the aim. Sources of evidence for our use case include the medical research related to the AI system under design and the various viewpoints.

As is the case for the design of this specific AI component, we may discover a *tension* between the various arguments linking evidence to the aim of the system. At this stage it is important to note this tension and document it, so that it can be taken into account during the later stages of the AI design, and if possible resolve the tension with a trade off.

We list some of the arguments linking evidence to the aim of the system. For some of the Arguments, i.e. Arguments 2 till arguments 5, tensions between different expert view points were also identified.

Argument 1: Malignant melanoma is a very heterogeneous tumor with a clinical course that is very difficult to predict. To date, there are no reliable biomarkers that predict prognosis with certainty. Therefore, there exist subgroups of melanoma patients with different risk for metastasization, some might never metastasize and diagnosing them would be overdiagnosis.

Tensions

For this use case, there are *tensions* between the various arguments linking evidence to the aim of the system, derived from the different viewpoints expressed by domain experts.

Argument 2: View Point: The Dermatologist.

The dermatologists in our team consider harms from treatment as rather small because of a generally small exzision with no need for general anaesthesia.

Counter Argument: View Point: The Evidence Based Medicine.

The evidence based experts in our team do consider harms from treatments.

Argument 3: View Point: Patient representative.

Patients should be informed about the consequences of screening and therapy before they opt to do it. And, of course, about the prognosis after a diagnosis has been made.

Counter Argument: *View Point: The Evidence Based Medicine; The population health view.*

According to the WHO screening principles and many countries' screening criteria, screening should not be offered before robust evidence from high quality evidence (RCTs) shows that the benefits outweigh the harms of screening for malignant melanomas. One of our experts has co-authored a Cochrane review where they did not find this kind of evidence and until then no screening should be performed. Moreover, making a free evidence-based informed choice of whether to be screened or not is not "free", is not "informed" and is "framed" (Henriksen et al., 2015; Johansson et al., 2019; Byskov Petersen et al., 2020; Rahbek et al., 2021).

Argument 4: View Point: The Dermatologist.

Early detection of malignant melanoma is critical, as the risk of metastasis with worse prognosis increases the longer melanoma remains untreated.

Counter Argument: *View Point: The Evidence Based Medicine*.

There are no reliable biomarkers that can predict the prognosis of melanoma before excision. There are patients who survive their localized melanoma without therapy. Therefore, the early diagnosis does not necessarily mean a better prognosis; on the contrary, there is a risk of poor patient care due to overdiagnosis.

Argument 5: *View Point: The Dermatologist and AI Engineer.*

Screening (in the future with AI-devices with even a higher sensitivity) will detect more early, localized melanomas that would have metastasized in a proportion of patients. The benefit of preventing this in this subgroup of patients outweighs the risk of overtreatment of other patients by a small harm - a small excision.

Counter Argument: View Point: The Evidence Based Medicine.

Pivotal ethical value as a physician when conducting clinical work as a GP is "primum non nuocere"—as stated in the original Hippocratic oath–first, do no harm. Therefore, as long as there is lack of robust evidence of high quality that early diagnosis of, or screening for, the melanoma result in reduced morbidity and mortality it cannot support such approaches. At the same time, screening for melanoma will inevitably result in substantial overdiagnosis. Therefore, there is the tendency to plead against screening for a melanoma, and early diagnosis of melanoma (and plead for timely diagnosis of clinical relevant melanoma) until it is provided with robust evidence of high quality that early diagnosis of, or screening for, a melanoma actually result in reduced morbidity and/or mortality of the disease.

Counter Argument: View Point: Patient representative.

We cannot judge as a clinician what is a "small harm". This can only be judged by the patient.

Is Bias justifiable?

Observing the current literature on AI fairness, there is a tendency to assume that any presence of bias automatically renders the tool ethically unjustifiable. This is an imperfect assumption, however. From a consequentialist perspective, the presence of bias becomes irremediably objectionable only at the moment the harm of bias outweighs any potential good that the tool might bring. While major bias in gender, race and other sensitive areas may often prove ethically challenging, coming to a final conclusion in any particular case will entail argumentation and potential disagreements. Regardless, conclusions are not automatic, and in this regard the current use case can serve as an interesting example.

At first glance, the fact that the tool was predominantly developed for skin types typically found in Caucasians, and that it exhibited considerable bias against darker skin types might lead to a criticism similar to that levelled by Obermeyer et al. (2019) against a different tool. The authors showed that a commercial tool predicting complex health needs exhibited considerable bias against black patients, i.e. getting the same score, black patients were considerably sicker. From the perspective of classical fairness, black and white patients should typically be treated in the same way with regards to access to healthcare resources, and therefore such a tool raises ethical objections.

The argument could be made, however, that the situation in the skin cancer case is different. In contrast to the Obermeyer example, white and black patients do *not* have the same resource need when it comes to melanoma. The incidence of melanoma in the black population is for example 20–30 times lower than in the white population (Culp & Lunsford, 2019). This makes melanoma in the black population a rare disease, whereas in the white population it is relatively common and a major public health challenge.

Under such circumstances, with differing needs with regards to access to healthcare resources, bias in the given tool, i.e. a development targeted at lighter skin types, could be justifiable in accordance with classical fairness as unequal patients are treated with proportionally unequal resources. It follows that the tool is fair and ethically justified not despite the bias, but because of it (Brusseau, 2021). This paradox is one of the case's more remarkable features.

It is outside of the scope of this work to reach a final conclusion with regards to this ethical challenge since further arguments need to be considered. For example the fact that the incidence of melanomas in black people is lower but the mortality is higher (Chao et al., 2017) must be weighed as well. Additionally, some of this difference in incidence might be the cause of different degrees of overdiagnosis–again caused by social inequality in access to healthcare. Overdiagnosis in malignant melanoma has the opposite social inequality: those who are highly educated and the richest are those with the highest degree of overdiagnosis (Welch and Fisher, 2017).

Still, the argument remains that the presence of statistical bias should not trigger the automatic response that a tool must be rejected on ethical grounds.

Even if bias in some AI tools could be ethically justifiable, challenges remain. From the perspective of the individual patient, there is still a bias in the system which needs to be clearly and firmly communicated to patients or recalibrated for specific populations before distribution or otherwise it could lead to considerable harm for some patients, though this will be for a minority.

If we consider a health system like in the US, such an AI tool could come with a considerable risk to its producers, given the known bias and the medical liability that could accompany this if not fully communicated, if marketed as a medical diagnostic tool replacing a doctor's diagnosis.

We do stress the importance of communicating the risks and benefits of the AI tool in different populations.

Verify if Transparency is a prerequisite for Explanation

Explanation and explainability concerns knowing the rationale on the basis of which an AI produces an output. To explain an algorithmic output, it is not sufficient simply to describe how it produces it. To explain the output requires some account of *why* the output is produced. It is worth noting that transparency is not well defined in the principles of Trustworthy AI, making the

relation to explainability very difficult to establish. While there is still no general consensus on what constitutes explainability, we find that three different notions can usefully be distinguished for the purpose of determining whether transparency is a prerequisite for explanation.

On the basis of an analysis of notions of explainability in AI-related research communities Doran et al. (2017, p. 4) distinguish between three ways in which a user may relate to a system. In opaque systems "the mechanisms mapping inputs to outputs are invisible to the user" (Doran et al., 2017, p. 4). In interpretable systems a user can see, study, and understand "how inputs are mathematically mapped to outputs" (Doran et al., 2017, p. 4). Thus, a necessary condition for interpretability is that the system is transparent as opposed to black-boxed. Finally, Doran et al. characterize a system as comprehensible if it "emits symbols along with its output (. . .)." These symbols "allow the user to relate properties of the inputs to their output" (Doran et al., 2017, p. 4).

While both comprehensible and interpretable systems may be considered improvements as compared to opaque systems, in that they enable explanations of why features of the input led to the output, the explanations given will still depend on human analysis. Transparency should thus be considered a necessary but not a sufficient condition in these cases. In other words, transparency allows for but does not imply explainability. Only a truly explanatory AI system designed as an autonomous system will produce an explanation by itself independent of a human analyst and the contingencies of their context and background knowledge. Given the autonomous nature of these type of systems, explanations will not require transparency about the "inner mechanisms" of the model (Doran et al., 2017, p. 7).

In this case transparency does not mean understanding the mathematical mapping process but identifying/reconstructing "important" drivers that led the model to make a given prediction, and make this understandable from clinicians. Several state-of-the-art approaches exist to compute local explanations of the predictions and to reconstruct these drivers. One introduced by Simonyan et al., is the saliency map method (Simonyan et al., 2014), that assigns a level of importance to each pixel in the input image, using the gradient of prediction with respect to each pixel. This allows one to localize the region of interest in the image, for this use-case, to localize the melanoma.

Another approach to computing local explanations is to perturb the input image. The idea is to modify part of the image, by replacing some pixels, and observing changes in the prediction. When the parts of the image that are important to the prediction are disturbed, the output is changed, while when they are unimportant, the output does not change much. Deep SHAP (Lundberg and Lee, 2017) was introduced as a combination of Deep Lift (Shrikumar et al., 2017) and Shapley Additive Explanation to leverage the explanation capabilities of SHAP values with deep networks. Using concepts from game theory to evaluate different perturbations of the input, it shows the positive and negative contribution of each pixel to the final prediction.

More recently, Lucieri et al. proposed the concept-based Concept Localization Map (CLM) explanation technique (Lucieri, et al., 2020b) as an improvement of previous saliency-based methods, by allowing to highlights groups of pixels representing individual concepts learned by the AI. This is the approach considered in the use-case. Other notable methodologies used in similar cases include adding an attention module that highlight salient features, as was done by Schlemper et al. who used attention for segmentation in abdominal CT scans. (Schlemper et al., 2019).

Given that the intended purpose of the system at hand is to support clinical decision-making, we argue that transparency is a necessary condition for explanations that are dependent on human analysis.

Involve Patients

exAID currently serves as a "trust-component" for existing AI systems. It provides dermatologists with an easy-to-understand explanation that can help guide the diagnostic process. But how can this information be translated and presented to patients, so as to engage them in the decision-making process? What might a discussion aid for the clinical encounter look like?

From a patient-centric view, we need the input of patients to answer these questions and involve them at every stage of the design process. There is indeed a growing body of evidence, indicating that involving patients in healthcare service design can improve patients' experiences (Tsianakas et al., 2012; Reay et al., 2017). Here it is particularly important to ensure that the views, needs, and preferences of vulnerable and disadvantaged patient groups are taken into account to avoid exacerbating existing inequalities (Amann and Sleigh, 2021).

When faced with decisions about their health, patients should be provided with all available and necessary information that is relevant for making an informed decision. The effective design of explanations of the AI system intended for the patient must still be the task of future investigations: If the system is communicating statistically correct results and therefore also shows very low and low probabilities for the presence of malignant melanoma, which would unlikely lead to the physician taking a sample, this might cause irritation and leave the patient with the feeling of "not having done everything possible". If, on the other hand, the system would communicate more clearly and was designed to give an own assessment for a "yes" or "no" sample collection, the message would be more clear and understandable for patients, but shift the decision from the physician to the AI system, which is not intentional. AI diagnostic systems, as they are currently designed, do not generally garner the trust of patients, even when they perform much better than human physicians (Longoni et al., 2019). A patient-centered co-design approach could end up reversing these preferences.

Consider the Legal and Ethical Perspectives: Mapping to the Trustworthy Artificial Intelligence Requirements

This AI tool is being developed for use by clinicians to support their decision making about the necessity of next steps for skin lesions and thereby potentially improving diagnostic

performance. There are a variety of legal factors that should be considered at this stage. A non-exhaustive discussion of some of these issues is set out below, under the headings defined by the AI HLEG. Not considered here are the licencing and other regulatory requirements of the jurisdiction in which the ML is to be used.

Transparency

Grote and Berens (2020) note that the deployment of machine learning algorithms might shift the evidentiary norms of medical diagnosis. They note; "as the patient is not provided with sufficient information concerning the confidence of a given diagnosis or the rationale of a treatment prediction, she might not be well equipped to give her consent to treatment decisions". In other words, when a patient may be harmed by an inaccurate prediction, if no explanation for the resulting decision is possible, their truly informed consent cannot be given. This threatens transparency and thereby evidence-based clinical practice, further research and academic appraisal.

Diversity, Non-discrimination, and Fairness

While the notions of bias and fairness are mentioned as issues relating to epistemological risk in section (2.1), there is also a genuine ethical concern about bias, fairness, and equality with respect to the development and use of ML in healthcare (Larrazabal et al., 2020). In general, AI encodes the same biases present in society (Owens & Walker, 2020). This is true when the data is used as is, but if an engineering team works on transforming the data to remove biases, then AI will encode a subset or even a distorted version of these biases.

It will be necessary to identify the ways this ML responds to different races and genders and how any discriminatory effects can be mitigated.

Issues of bias, fairness and equality relate to the issue of trust. Both clinicians and the public may become skeptical about ML systems in diagnostics as a result of problematic cases of inequality in performance across socially salient groups. The opposite could also occur: the public and clinicians trust the AI despite the lack of evidence of the effect of AI on patients' prognosis–or even evidence showing that the AI is creating more harm than good.

Human Agency and Oversight

The AI HLEG specifically recognizes a "right not to be subject to a decision based solely on automated processing when this (. . .) significantly affects them" (AI HLEG, 2019, p. 16). For this case, the ML is used as a support mechanism for the decision making of the clinician. On the face of it that seems unproblematic from a human agency and oversight perspective. However, the design process should ensure that it is possible for those who are impacted by the decisions made by AI to challenge them and that the level of human oversight is sufficient (Hickman and Petrin, 2020).

The design process should consider the extent to which a clinician's agency and autonomous decision-making are or could be reduced by the AI system. The assumption is that the clinicians would only use the AI system to support their classification but any requirement to use it as support may be eroded over time if clinicians were to consider the AI system highly accurate.

Privacy and Data Governance

According to the General Data Protection Regulation (GDPR), an informed and explicit consent is required for the processing of sensitive data, such as health data. A data protection impact assessment in accordance with Article 35 GDPR will need to be carried out. GDPR requirements are extensive and evolving (European Parliament and Council of European Union, 2016). Specific concerns for this use case include the patient's need to consent that an AI system is included in the process and that personal data, in the form of images with related information, may need to be stored for the development of the ML. The possibility of a right to explanation under the GDPR may also pose difficulties to the extent that human understanding of the AI process is limited. The exAID is trying to mitigate this.

A full legal review will be needed to assess the compliance of the system with the GDPR's requirements. The goal of the GDPR is the protection of fundamental rights and freedoms of natural persons (Art. 1). These are determined in accordance with the Charter of Fundamental Rights of the European Union and the European Convention on Human Rights. This also includes the right to non-discrimination pursuant to Article 21 Charter of Fundamental Rights. This is relevant for this use case in light of the issues, discussed above, relating to accuracy for different skin colours.

Possible Accountability Issues

The AI HLEG trustworthy AI guidelines require "that mechanisms be put in place to ensure responsibility and accountability for AI systems" and emphasizes the importance of redress when unjust adverse impact occurs (AI HLEG, 2019, p. 19f). In matters of human health, the potential harm can be substantial both in nonmonetary and in monetary terms. Mechanisms that allow for redress in the event of harm or adverse impact are therefore particularly important.

In the application of this ML tool, different actors (such as the institution using the AI, the manufacturers of the AI, or those in charge of oversight of the AI) could potentially be responsible for any harm caused. This would create difficulties for any injured person to prove specific causation or to show that an AI system was "defective". In the design phase, thought should be given to putting mechanisms in place to provide information to (end-) users and third parties about opportunities for redress, as required by the AI HLEG trustworthy AI guidelines (AI HLEG, 2019, p. 31). This may be challenging given the transparency issues discussed above.

As long as the algorithm is unknown, nothing more than general guidelines can be disclosed to the actors involved in the application. Still, the parties involved in designing, developing, deploying, implementing, and using the AI system should consider how–in line with the AI HLEG trustworthy AI guidelines–they can enhance the accountability factors mentioned above. This could include facilitating audit processes, if appropriate via evaluation by internal and external auditors, and creating avenues for redress apart from

the pre-existing legal avenues available to those negatively affected by AI. This could also involve using more explainable AI models (e.g., bayesian network), or using tools that can create whitebox models from blackbox models to understand the feature importance.

Limitations

Findings of our assessment should be interpreted in light of some limitations. First, it must be recognized that identifying a list of specific evaluation criteria that is complete and as exhaustive as possible in the midst of the evaluation workshop discussions is a significant challenge, even if the requirements identified in the AI HLEG trustworthy AI guidelines can be used to frame discussions on this matter. We aimed to mitigate these challenges by closely following the Z-Inspection® process, which ensures constant exchange and reflection within the research team.

The interpretation and potential misinterpretation of the results of the exchanges and of the ethical evaluation established by the Z-Inspection®, as the process does not aim at granting an ethical "approval" to the AI system under review, but rather at setting up an open discussion process which, although it explicitly aims at compensating for the shortcomings of ethical evaluations that are limited to the compliance to certain pre-established rules, may not fit easily into the practices of certain stakeholders.

In this phase of the co-design, we did not include a patient peer group representative in our assessment process, besides the expert in our team. This is a limitation. In the next phase of the co-design for this use case, as suggested by the patient representative in our team, we plan to involve patient peer groups for further depth and richness of the assessment.

RELATED WORK

Our research work is addressing the need for co-design of trustworthy AI using a holistic approach, rather than static checklists. There are a number of AI ethics checklists being produced for AI systems. Madaio et al. (2020) mention that unless such checklists are grounded in practitioners' needs, they may be misused.

Co-design approaches have been shown paramount to improve the adoption of AI within the healthcare field. Kocaballi et al. (2020) sought to understand the potential role of future AI documentation assistant in primary care consultation by carrying out co-design workshops with general practitioners. The results of such activities raised concerns about different topics like medico-legal aspects on processing patient data continuously; possible deviations in treatments due to focus of AI algorithms on improving efficiency as opposed to patient care; and human conversation and empathy remaining the core tasks of doctors despite AI advances. Therefore, the study demonstrates that human-AI collaboration models within the healthcare field need to be designed by involving an interdisciplinary team that assesses the AI system in several spheres of patient care (i.e. medical ethical, legal, technological).

Human-AI collaboration in healthcare demands rigor in evaluation. Studying the use of AI for decision support in healthcare settings, Lai et al. (2021) found that some of the reviewed cases still omit collaborative system evaluation, and call for more field studies to obtain a deeper understanding of the practical setting. Beede et al. (2020) introduce such an approach for evaluating the diagnostic use of AI for diabetic retinopathy. The findings show that several socio-environmental factors, which impact model performance, could not be foreseen during development. Our previous work within Z-inspection® (Zicari, et al., 2021b), makes a similar finding when assessing a case where AI is used to detect sound patterns of callers to an emergency line, to indicate the probability of a cardiac arrest situation to an operator. The descriptive statistics of the modeling solution show improved ability for detection; however, a further examination of the collaborative system shows that operator trust toward the system is low (Blomberg et al., 2021).

Ontological concerns include the fluidity of concepts such as disease, and diagnosis and even the roles of patients, clinicians, and healthcare authorities. Research into clinical overdiagnosis has shown that increasingly more harmless abnormalities are identified and diagnosed as disease despite being asymptomatic for the "patient" (Brodersen et al., 2018). Thus, the concepts of being apparently healthy and being a patient can change as well. The role of the patients also changes when they turn into a position where they can start self-diagnosing via publicly available AI (Lupton, 2013).

Other concerns that have been raised include the extent to which an AI system could replace clinicians, if decisions made based on the result of an AI system could weaken the authority of clinicians, introduce paternalism, threaten patients` autonomy, disrupt the interaction between medical doctors and patients (Gerke, et al., 2020a; Gerke, et al., 2020b). However, instead of replacing clinicians, the line of action should promote collaboration between professionals and technologies, for instance, incorporating human expert knowledge in order to enhance the AI algorithms outcomes by improving its accuracy and trustworthiness.

Specifically for this use case, skin is the largest organ of the body and the first barrier in protecting us from environmental stressors such as ultraviolet radiation, pollution, and abrasion. UV light, particularly, contributes significantly in affecting the molecular composition of skin and skin cells, which sometimes may result in skin cancer—melanoma (Laikova et al., 2019). Melanomas account for more than 90% of all deaths caused by skin tumors (Leitlinienprogramm Onkologie et al., 2020, p. 25). Melanomas originate when skin cells (melanocytes) accumulate damage and undergo uncontrolled, abnormal and very rapid division which invades healthy surrounding skill cells and is not controlled by normal cell death. Early detection of melanoma is very important to minimize the spread of these abnormal cells hence improving the chances of survival. The five-year survival rate varies from 27 to 99% from distant metastasized to localized proliferation of the disease during diagnosis (American Cancer Society, 2021, p. 21) A decrease in delay between the diagnostic biopsy and surgical excision has been

associated with a significant increase in survival rate (Adamson et al., 2020).

A visit to a dermatologist can be triggered by many reasons including conspicuous alterations of the skin. In addition, certain risk factors foster the development of malignant melanoma and therefore warrant regular examination. The most common risk factors include light skin color, personal or family history of melanoma, presence of atypical, large or numerous moles and (history of) excessive exposure to UV-radiation and immunosuppression (American Cancer Society, 2021, p. 24).

The visual resemblance of benign skin lesions like nevi to malignant melanomas poses a major difficulty in the early detection of skin cancer (Grant-Kels et al., 1999). The desire for low false negative rates, for the sake of patient survival, typically results in high false positives, which is revealed after excision and histological examination of the tissue. However, misdiagnosis leads to the unnecessary physical and psychological burden of affected patients. Average dermatologists excise 20–30 benign lesions to find a single malignant melanoma (Johnson et al., 2017; Kutzner et al., 2020).

Visual examination of skin lesions through dermoscopy is a strenuous task demanding tedious training and experience of medical specialists (Kittler et al., 2002). The deployment of a digital assistant, augmenting the recognition capabilities of human experts could potentially result in more consistency and an overall increase in diagnostic performance, as indicated by previous work (Brinker T. J. et al., 2019; Brinker et al., 2019 TJ.).

Currently, most dermatologists or GPs will not hesitate to remove a suspicious mole as early detection is a matter of life and death even though it later should turn out to be a false positive by closer inspection at the lab (Huff et al., 2012). However–this may in turn lead to overdiagnosis. A "defensive overdiagnosis" automatically translates into a "defensive overtreatment". On the other hand, not performing surgery of a true positive might result in death or serious illness and costly and extensive cancer treatment (Troxel, 2003).

Welch et al. (2021) point out that with "absent metastasis, no definitive diagnostic criteria for the pathological diagnosis of melanoma exist. Because the diagnosis is subjective, pathologists disagree about whether melanoma is present, particularly when faced with lesions in the diagnostic gray zone". They also mention that in their opinion the most important step to reduce melanoma overdiagnosis is to stop population-wide screening for skin cancer.

CONCLUSION

Artificial intelligence systems can raise ethical and societal concerns from direct stakeholders such as patients in Healthcare environments and from indirect stakeholders such as politicians or general media. The nature of these concerns can vary and include a vast array of topics like data security, biases, cost-benefit-debates, technical dependencies, or technical supremacy. The multidisciplinary approach of the evaluation can help to identify these concerns in many different fields, already at very early development stages.

In the public, AI systems are increasingly criticized in their entirety because of their "black box" character. Communicating the co-design process itself can help reinforce trust in such a system by making its exact workings transparent, even to non-specialist project staff. This transparency helps funding agencies, oversight boards, and executive teams explain their decisions about funding and governing decisions as well as the system's operation. In the healthcare domain, lack of explainability limits a wider adoption of AI solutions since healthcare workers often find it challenging to trust complex models since they require high technical and statistics knowledge (Moreno-Sanchez, 2020).

Co-designing trustworthy AI with a holistic approach requires some unique aspects in the structure and design of the process: An interdisciplinary team with experts coming from the domain of the AI application, i.e. healthcare, as well as other fields like technical, legal, social and ethical. For this purpose, Z-Inspection® proposes a co-design methodology where an interdisciplinary team of experts works together with the AI designers and their managers to explore and investigate possible ethical, legal and technical issues that could arise from the future use of the AI system.

Several benefits for improving the ethical quality of the design of the AI system have been shown in this paper. The interdisciplinary nature of co-design can lead to actions that optimize current or future versions of the AI system by detecting unforeseen problems. For example, concerns from the evaluation team regarding potential biases in the training data of a used external machine learning model can lead to the formulation of new and more detailed requirements for the selection of external system components.

The problem of overdiagnosis of melanomas was previously not addressed by the team of engineers who built a first prototype system, with the aim of helping doctors to diagnose melanomas. During the evaluation, it became clear that the project team had to address such an important external issue before deploying the system.

The interdisciplinary approach of the evaluation can also help uncover potential conflicts of interest early on and even in indirect stakeholder groups, since they inevitably become visible during a process that includes neutral actors from many disciplines. During the evaluation, discussions in the workshops are recorded and shared with whatever entities solicited the inspection and therefore provide easy reference to different view points.

Evaluation of an AI development with a holistic approach like Z-Inspection® creates benefits related to general acceptance or concerns inside and outside the institution that applies an AI project, as well as benefits related to the quality of the project's processes, transparency about possible conflicts of interest, and in general comprehensibility of the system, which improves the quality of communication for any kind of stakeholder.

The analysis of this use case leads to challenging tensions from a translational point-of-view. Best practice guidelines for AI in healthcare development stress the continuous alignment of technical development with a clinical use case (Higgins and Madai, 2020).

This is owing to the fact that purely technical teams are less suited to understand the nuances of clinical decision making and the clinical workflow. Close collaboration with medical experts

and medical experts as part of the team are highly encouraged. Analysing this use case led, however, to the identification of different views *within* the medical community on the given dermatological use case. This sheds light on the observation that modern medicine is not only guided by pure science and clinical guidelines, but also clinical practice is sometimes also influenced by traditions, different cultural viewpoints and differing interpretations of the available scientific literature. Thus, different national guidelines and strategies might exist. This poses a serious challenge for teams developing tools for the clinical setting. If they are not aware of these differences, their tool might only be applicable in a certain country or region despite their best efforts to include clinicians and develop a tool fitting to the clinical setting.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

Conception/design—all authors; paper editing—RZ, JA, SB, JoB, JaB, MC, AG, PG, CH, EH (18th author), EH (19th author), SH, PK, UK, AL, VM, MO, ES, AS, JT, DV, HV, MW, RW, FB, BD, EG, and OM; final approval—all authors. Authors are mentioned in alphabetical order (except for 1st author).

FUNDING

DV received funding from the European Union's Horizon 2020 Research and Innovation Program "PERISCOPE: Pan European Response to the ImpactS of COvid-19 and future Pandemics and Epidemics" under grant agreement no. 101016233, H2020-SC1-PHE-CORONAVIRUS-2020-2-RTD. TH was supported by the Cluster of Excellence "Machine Learning—New Perspectives for Science" funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—Reference Number EXC 2064/1—Project ID 390727645

ACKNOWLEDGMENTS

We would like to thank Valentina Beretta, Marianna Ganapini, Sara Gerke, Georgios Kararigas, Leonardo Espinosa-Leal, and Andréane Sabourin Laflamme for their valuable feedback on earlier versions of this paper. Thank you to Timo Eichhorn, Georgios Kararigas, Todor Ivanov, Melissa McCullough, Karsten Tolle, Gemma Roig, Norman Stürtz and Irmhild van Halem for their invaluable contribution to the definition of the Z-Inspection® process.

REFERENCES

Adamson, J., Charles, J., Darden, A., Lee, F., and Lowe, M. (2020). Foresight into AI Ethics in Healthcare (FAIE-H): A Toolkit for Creating an Ethics Roadmap for Your Healthcare AI Project. Open Roboethics Institute. Available at: https://openroboethics.org/wp-content/uploads/2020/02/FAIE-H-Final-to-Upload.pdf.

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

Amann, J., and Sleigh, J. (2021). Too Vulnerable to Involve? Challenges of Engaging Vulnerable Groups in the Co-production of Public Services through Research. *Int. J. Public Adm.* 44 (9), 715–727. doi:10.1080/01900692.2021.1912089

American Cancer Society (2021). Cancer Facts & Figures 2021. American Cancer Society. Available at: https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2021/cancer-facts-and-figures-2021.pdf.

Argenziano, G., Fabbrocini, G., Carli, P., De Giorgi, V., Sammarco, E., and Delfino, M. (1998). Epiluminescence Microscopy for the Diagnosis of Doubtful Melanocytic Skin Lesions. *Arch. Dermatol.* 134 (12), 1563–1570. doi:10.1001/archderm.134.12.1563

Baade, P. D., Youl, P. H., Janda, M., Whiteman, D. C., Del Mar, C. B., and Aitken, J. F. (2008). Factors Associated with the Number of Lesions Excised for Each Skin Cancer. *Arch. Dermatol.* 144 (11), 1468–1476. doi:10.1001/archderm.144.11.1468

Banerjee, S. C., D'Agostino, T. A., Gordon, M. L., and Hay, J. L. (2018). "It's Not JUST Skin Cancer": Understanding Their Cancer Experience from Melanoma Survivor Narratives Shared Online. *Health Commun.* 33 (2), 188–201. doi:10.1080/10410236.2016.1250707

Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., et al. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.*, 1–12. doi:10.1145/3313831.3376718

Bissoto, A., Fornaciali, M., Valle, E., and Avila, S. (2019). (De)Constructing Bias on Skin Lesion Datasets. 0–0. Available at: https://openaccess.thecvf.com/content_CVPRW_2019/html/ISIC/Bissoto_DeConstructing_Bias_on_Skin_Lesion_Datasets_CVPRW_2019_paper.html. doi:10.1109/cvprw.2019.00335

Bissoto, A., Valle, E., and Avila, S. (2020). Debiasing Skin Lesion Datasets and Models? Not So Fast. Available at: https://openaccess.thecvf.com/content_CVPRW_2020/html/w42/Bissoto_Debiasing_Skin_Lesion_Datasets_and_Models_Not_So_Fast_CVPRW_2020_paper.html. 740–741.

Blomberg, S. N., Christensen, H. C., Lippert, F., Ersbøll, A. K., Torp-Petersen, C., Sayre, M. R., et al. (2021). Effect of Machine Learning on Dispatcher Recognition of Out-Of-Hospital Cardiac Arrest during Calls to Emergency Medical Services. *JAMA Netw. Open* 4 (1), e2032320. doi:10.1001/jamanetworkopen.2020.32320

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., et al. (2019b). Deep Learning Outperformed 136 of 157 Dermatologists in a Head-To-Head Dermoscopic Melanoma Image Classification Task. *Eur. J. Cancer* 113, 47–54. doi:10.1016/j.ejca.2019.04.001

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., et al. (2019a). A Convolutional Neural Network Trained with Dermoscopic Images Performed on Par with 145 Dermatologists in a Clinical Melanoma Image Classification Task. *Eur. J. Cancer* 111, 148–154. doi:10.1016/j.ejca.2019.02.005

Broadstock, M., Michie, S., and Marteau, T. (2000). Psychological Consequences of Predictive Genetic Testing: A Systematic Review. *Eur. J. Hum. Genet.* 8 (10), 731–738. doi:10.1038/sj.ejhg.5200532

Brodersen, J., Schwartz, L. M., Heneghan, C., O'Sullivan, J. W., Aronson, J. K., and Woloshin, S. (2018). Overdiagnosis: what it Is and what it Isn't. *Bmj Ebm* 23 (1), 1–3. doi:10.1136/ebmed-2017-110886

Brodersen, J., Schwartz, L. M., and Woloshin, S. (2014). Overdiagnosis: How Cancer Screening Can Turn Indolent Pathology into Illness. *APMIS* 122 (8), 683–689. doi:10.1111/apm.12278

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., et al. (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *ArXiv:2004.07213 [Cs]*. doi:10.5772/intechopen.90859

Brusseau, J. (2021). *Using Edge Cases to Disentangle Fairness and Solidarity in AI Ethics*. Manuscript Submitted for Publication.

Burgess, M. M. (2001). Beyond Consent: Ethical and Social Issues in Genetic Testing. *Nat. Rev. Genet.* 2 (2), 147–151. doi:10.1038/35052579

Byskov Petersen, G., Sadolin Damhus, C., Ryborg Jønsson, A. B., and Brodersen, J. (2020). The Perception gap: How the Benefits and Harms of Cervical Cancer Screening Are Understood in Information Material Focusing on Informed Choice. *Health Risk Soc.* 22 (2), 177–196. doi:10.1080/13698575.2020.1778645

Caulfield, T., and McGuire, A. L. (2012). Direct-to-Consumer Genetic Testing: Perceptions, Problems, and Policy Responses. *Annu. Rev. Med.* 63 (1), 23–33. doi:10.1146/annurev-med-062110-123753

Chao, L. X., Patterson, S. S. L., Rademaker, A. W., Liu, D., and Kundu, R. V. (2017). Melanoma Perception in People of Color: A Targeted Educational Intervention. *Am. J. Clin. Dermatol.* 18 (3), 419–427. doi:10.1007/s40257-016-0244-y

Clarke, A. J., and Wallgren-Pettersson, C. (2019). Ethics in Genetic Counselling. *J. Community Genet.* 10 (1), 3–33. doi:10.1007/s12687-018-0371-7

Codella, N. C. F., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., et al. (2018). Skin Lesion Analysis toward Melanoma Detection: A challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). 2018 IEEE 15th International Symposium on Biomedical Imaging, Washington, April 2018. ISBI 2018, 168–172. doi:10.1109/ISBI.2018.8363547

Culp, M. B., and Lunsford, N. B. (2019). Melanoma Among Non-hispanic Black Americans. *Prev. Chronic Dis.* 16. doi:10.5888/pcd16.180640

Dieng, M., Smit, A. K., Hersch, J., Morton, R. L., Cust, A. E., Irwig, L., et al. (2019). Patients' Views about Skin Self-Examination after Treatment for Localized Melanoma. *JAMA Dermatol.* 155 (8), 914–921. doi:10.1001/jamadermatol.2019.0434

Doran, D., Schulz, S., and Besold, T. R. (2017). What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *ArXiv:1710.00794 [Cs]*.

Eccles, B. K., Copson, E., Maishman, T., Abraham, J. E., and Eccles, D. M. (2015). Understanding of BRCA VUS Genetic Results by Breast Cancer Specialists. *BMC Cancer* 15 (1), 936. doi:10.1186/s12885-015-1934-1

English, D. R., Del Mar, C., and Burton, R. C. (2004). Factors Influencing the Number Needed to Excise: Excision Rates of Pigmented Lesions by General Practitioners. *Med. J. Aust.* 180 (1), 16–19. doi:10.5694/j.1326-5377.2004.tb05766.x

European Parliament & Council of European Union (2016). GDPR) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation). *Official J. Eur. Union, L* 119, 1–88.

Eways, K., Bennett, K., Hamilton, J., Harry, K., Marszalek, J., Marsh, M.-J., et al. (2020). Development and Psychometric Properties of the Self-Blame Attributions for Cancer Scale. *Onf* 47 (1), 79–88. doi:10.1188/20.ONF.79-88

Fitzpatrick, T. B. (1988). The Validity and Practicality of Sun-Reactive Skin Types I through VI. *Arch. Dermatol.* 124 (6), 869–871. doi:10.1001/archderm.1988.01670060015008 10.1001/archderm.124.6.869

Geller, G., Botkin, J. R., Green, M. J., Press, N., Biesecker, B. B., Wilfond, B., et al. (1997). Genetic Testing for Susceptibility to Adult-Onset Cancer. *JAMA* 277 (18), 1467–1474. doi:10.1001/jama.1997.03540420063031

Gerke, S., Babic, B., Evgeniou, T., and Cohen, I. G. (2020a). The Need for a System View to Regulate Artificial Intelligence/machine Learning-Based Software as Medical Device. *Npj Digit. Med.* 3 (1), 1–4. doi:10.1038/s41746-020-0262-2

Gerke, S., Minssen, T., and Cohen, G. (2020b). "Ethical and Legal Challenges of Artificial Intelligence-Driven Healthcare," in *Artificial Intelligence in Healthcare*. Editors A. Bohr and K. Memarzadeh (Academic Press), 295–336. doi:10.1016/B978-0-12-818438-7.00012-5

Glasziou, P. P., Jones, M. A., Pathirana, T., Barratt, A. L., and Bell, K. J. (2020). Estimating the Magnitude of Cancer Overdiagnosis in Australia. *Med. J. Aust.* 212 (4), 163–168. doi:10.5694/mja2.50455

Grant-Kels, J. M., Bason, E. T., and Grin, C. M. (1999). The Misdiagnosis of Malignant Melanoma. *J. Am. Acad. Dermatol.* 40 (4), 539–548. doi:10.1016/S0190-9622(99)70435-4

Grote, T., and Berens, P. (2020). On the Ethics of Algorithmic Decision-Making in Healthcare. *J. Med. Ethics* 46 (3), 205–211. doi:10.1136/medethics-2019-105586

Gupta, A. K., Bharadwaj, M., and Mehrotra, R. (2016). Skin Cancer Concerns in People of Color: Risk Factors and Prevention. *Asian Pac. J. Cancer Prevention : APJCP* 17 (12), 5257–5264. doi:10.22034/APJCP.2016.17.12.525710.7314/apjcp.2016.17.s2.19

Hallowell, N., Foster, C., Eeles, R., Ardern-Jones, A., Murday, V., and Watson, M. (2003). Balancing Autonomy and Responsibility: the Ethics of Generating and Disclosing Genetic Information * Commentary * Author's Reply. *J. Med. Ethics* 29 (2), 74–79. doi:10.1136/jme.29.2.74

Hansen, C., Wilkinson, D., Hansen, M., and Argenziano, G. (2009). How Good Are Skin Cancer Clinics at Melanoma Detection? Number Needed to Treat Variability across a National Clinic Group in Australia. *J. Am. Acad. Dermatol.* 61 (4), 599–604. doi:10.1016/j.jaad.2009.04.021

Hawkins, A. K., Ho, A., and Hayden, M. R. (2011). Lessons from Predictive Testing for Huntington Disease: 25 Years on. *J. Med. Genet.* 48 (10), 649–650. doi:10.1136/jmedgenet-2011-100352

Henriksen, M. J. V., Guassora, A. D., and Brodersen, J. (2015). Preconceptions Influence Women's Perceptions of Information on Breast Cancer Screening: a Qualitative Study. *BMC Res. Notes* 8 (1), 404. doi:10.1186/s13104-015-1327-1

Hickman, E., and Petrin, M. (2020). Trustworthy AI and Corporate Governance - the EU's Ethics Guidelines for Trustworthy Artificial Intelligence from a Company Law Perspective. *SSRN J.* doi:10.2139/ssrn.3607225

Higgins, D., and Madai, V. I. (2020). From Bit to Bedside: A Practical Framework for Artificial Intelligence Product Development in Healthcare. *Adv. Intell. Syst.* 2 (10), 2000052. doi:10.1002/aisy.202000052

Hoffmann, T. C., and Del Mar, C. (2015). Patients' Expectations of the Benefits and Harms of Treatments, Screening, and Tests. *JAMA Intern. Med.* 175 (2), 274–286. doi:10.1001/jamainternmed.2014.6016

Huff, L. S., Chang, C. A., Thomas, J. F., Cook-Shimanek, M. K., Blomquist, P., Konnikov, N., et al. (2012). Defining an Acceptable Period of Time from Melanoma Biopsy to Excision. *Dermatol. Rep.* 4 (1), 2. doi:10.4081/dr.2012.e2

J. Keulartz, M. Korthals, M. Schermer, and T. Swierstra (2002). *Pragmatist Ethics for a Technological Culture* (Springer Netherlands), Vol. 3. doi:10.1007/978-94-010-0301-8

Johansson, M., Brodersen, J., Gøtzsche, P. C., and Jørgensen, K. J. (2019). Screening for Reducing Morbidity and Mortality in Malignant Melanoma. *Cochrane Database Syst. Rev.* 6. doi:10.1002/14651858.CD012352.pub2

Johnson, M. M., Leachman, S. A., Aspinwall, L. G., Cranmer, L. D., Curiel-Lewandrowski, C., Sondak, V. K., et al. (2017). Skin Cancer Screening: Recommendations for Data-Driven Screening Guidelines and a Review of the US Preventive Services Task Force Controversy. *Melanoma Management* 4 (1), 13–37. doi:10.2217/mmt-2016-0022

Jutzi, T. B., Krieghoff-Henning, E. I., Holland-Letz, T., Utikal, J. S., Hauschild, A., Schadendorf, D., et al. (2020). Artificial Intelligence in Skin Cancer Diagnostics: The Patients' Perspective. *Front. Med.* 7, 233. doi:10.3389/fmed.2020.00233

Kawahara, J., Daneshvar, S., Argenziano, G., and Hamarneh, G. (2019). Seven-Point Checklist and Skin Lesion Classification Using Multitask Multimodal Neural Nets. *IEEE J. Biomed. Health Inform.* 23 (2), 538–546. doi:10.1109/JBHI.2018.2824327

Kittler, H., Pehamberger, H., Wolff, K., and Binder, M. (2002). Diagnostic Accuracy of Dermoscopy. *Lancet Oncol.* 3 (3), 159–165. doi:10.1016/S1470-2045(02)00679-4

Kocaballi, A. B., Ijaz, K., Laranjo, L., Quiroz, J. C., Rezazadegan, D., Tong, H. L., et al. (2020). Envisioning an Artificial Intelligence Documentation Assistant for Future Primary Care Consultations: A Co-design Study with General Practitioners. *J. Am. Med. Inform. Assoc.* 27 (11), 1695–1704. doi:10.1093/jamia/ocaa131

Kutzner, H., Jutzi, T. B., Krahl, D., Krieghoff-Henning, E. I., Heppt, M. V., Hekler, A., et al. (2020). Overdiagnosis of Melanoma - Causes, Consequences and Solutions. *JDDG: J. Der Deutschen Dermatologischen Gesellschaft* 18 (11), 1236–1243. doi:10.1111/ddg.14233

Lai, Y., Kankanhalli, A., and Ong, D. (2021). Human-AI Collaboration in Healthcare: A Review and Research Agenda. *Hawaii Int. Conf. Syst. Sci.* doi:10.24251/HICSS.2021.046

Laikova, K. V., Oberemok, V. V., Krasnodubets, A. M., Gal'chinsky, N. V., Useinov, R. Z., Novikov, I. A., et al. (2019). Advances in the Understanding of Skin Cancer: Ultraviolet Radiation, Mutations, and Antisense Oligonucleotides as Anticancer Drugs. *Molecules* 24 (8), 1516. doi:10.3390/molecules24081516

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender Imbalance in Medical Imaging Datasets Produces Biased Classifiers for Computer-Aided Diagnosis. *Proc. Natl. Acad. Sci. USA* 117 (23), 12592–12594. doi:10.1073/pnas.1919012117

Lau, S. C. M., Chen, L., and Cheung, W. Y. (2014). Protective Skin Care Behaviors in Cancer Survivors. *Curr. Oncol.* 21 (4), 531–540. doi:10.3747/co.21.1893

Leikas, J., Koivisto, R., and Gotcheva, N. (2019). Ethical Framework for Designing Autonomous Intelligent Systems. *JOItmC* 5 (1), 18. doi:10.3390/joitmc5010018

Onkologie, Leitlinienprogramm., Krebsgesellschaft, Deutsche., and Deutsche Krebshilfe, A. W. M. F). (2020). Diagnostik, Therapie und Nachsorge des Melanoms, Langversion 3.3 (032/024OL). Available at: http://www.leitlinienprogramm-onkologie.de/leitlinien/melanom/

Longoni, C., Bonezzi, A., and Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *J. Consumer Res.* 46 (4), 629–650. doi:10.1093/jcr/ucz013

Lucieri, A., Bajwa, M. N., Alexander Braun, S., Malik, M. I., Dengel, A., and Ahmed, S. (2020a). On Interpretability of Deep Learning Based Skin Lesion Classifiers Using Concept Activation Vectors. *Int. Jt. Conf. Neural Networks (Ijcnn)*, 2020, 1–10. doi:10.1109/IJCNN48605.2020.9206946

Lucieri, A., Bajwa, M. N., Dengel, A., and Ahmed, S. (2020b). "Explaining AI-Based Decision Support Systems Using Concept Localization Maps," in *Neural Information Processing*. Editors H. Yang, K. Pasupa, A. C.-S. Leung, J. T. Kwok, J. H. Chan, and I. King (Springer International Publishing), 185–193. doi:10.1007/978-3-030-63820-7_21–193)

Lucivero, F. (2016). *Ethical Assessments of Emerging Technologies: Appraising the Moral Plausibility of Technological Visions*. 1st ed. Imprint: Springer International Publishing Springer. doi:10.1007/978-3-319-23282-92016

Lundberg, S. M., and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Adv. Neural Inf. Process. Syst.* 30, 4765–4774.

Lupton, D. (2013). Quantifying the Body: Monitoring and Measuring Health in the Age of mHealth Technologies. *Crit. Public Health* 23 (4), 393–403. doi:10.1080/09581596.2013.794931

Madaio, M. A., Stark, L., Wortman Vaughan, J., and Wallach, H. (2020). Co-Designing Checklists to Understand Organizational Challenges and Opportunities Around Fairness in AI. *Proc. 2020 CHI Conf. Hum. Factors Comput. Syst.*, 1–14. doi:10.1145/3313831.3376445

Malvehy, J., Hauschild, A., Curiel-Lewandrowski, C., Mohr, P., Hofmann-Wellenhof, R., Motley, R., et al. (2014). Clinical Performance of the Nevisense System in Cutaneous Melanoma Detection: An International, Multicentre, Prospective and Blinded Clinical Trial on Efficacy and Safety. *Br. J. Dermatol.* 171 (5), 1099–1107. doi:10.1111/bjd.13121

Mendonça, T., Ferreira, P. M., Marques, J. S., Marcal, A. R. S., and Rozeira, J. (2013). PH2 - A Dermoscopic Image Database for Research and Benchmarking. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Osaka, Japan, July 3–7, 2013. EMBC, 5437–5440. doi:10.1109/EMBC.2013.6610779

Moreno-Sanchez, P. A. (2020). Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees. 2020 IEEE International Conference on Big Data, December 13–20, 2020. Big Data, 4902–4910. doi:10.1109/BigData50022.2020.9378460

Moynihan, R., Nickel, B., Hersch, J., Doust, J., Barratt, A., Beller, E., et al. (2015). What Do You Think Overdiagnosis Means? A Qualitative Analysis of Responses from a National Community Survey of Australians. *BMJ Open* 5 (5), e007436. doi:10.1136/bmjopen-2014-007436

Nelson, C. A., Pérez-Chada, L. M., Creadore, A., Li, S. J., Lo, K., Manjaly, P., et al. (2020). Patient Perspectives on the Use of Artificial Intelligence for Skin Cancer Screening. *JAMA Dermatol.* 156 (5), 501. doi:10.1001/jamadermatol.2019.5014

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. *Science* 366 (6464), 447–453. doi:10.1126/science.aax2342

Owens, K., and Walker, A. (2020). Those Designing Healthcare Algorithms Must Become Actively Anti-racist. *Nat. Med.* 26 (9), 1327–1328. doi:10.1038/s41591-020-1020-3

Petty, A. J., Ackerson, B., Garza, R., Peterson, M., Liu, B., Green, C., et al. (2020). Meta-analysis of Number Needed to Treat for Diagnosis of Melanoma by Clinical Setting. *J. Am. Acad. Dermatol.* 82 (5), 1158–1165. doi:10.1016/j.jaad.2019.12.063

Rahbek, O. J., Jauernik, C. P., Ploug, T., and Brodersen, J. (2021). Categories of Systematic Influences Applied to Increase Cancer Screening Participation: A Literature Review and Analysis. *Eur. J. Public Health* 31 (1), 200–206. doi:10.1093/eurpub/ckaa158

Reay, S., Collier, G., Kennedy-Good, J., Old, A., Douglas, R., and Bill, A. (2017). Designing the Future of Healthcare Together: Prototyping a Hospital Co-design Space. *CoDesign* 13 (4), 227–244. doi:10.1080/15710882.2016.1160127

Robertson, L. J., Abbas, R., Alici, G., Munoz, A., and Michael, K. (2019). Engineering-Based Design Methodology for Embedding Ethics in Autonomous Robots. *Proc. IEEE* 107 (3), 582–599. doi:10.1109/JPROC.2018.2889678

Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., et al. (2021). A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context. *Sci. Data* 8 (1), 34. doi:10.1038/s41597-021-00815-z

Rutherford, M. J., Ironmonger, L., Ormiston-Smith, N., Abel, G. A., Greenberg, D. C., Lyratzopoulos, G., et al. (2015). Estimating the Potential Survival Gains by Eliminating Socioeconomic and Sex Inequalities in Stage at Diagnosis of Melanoma. *Br. J. Cancer* 112 (1), S116–S123. doi:10.1038/bjc.2015.50

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., et al. (2019). Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *Med. Image Anal.* 53, 197–207. doi:10.1016/j.media.2019.01.012

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning Important Features through Propagating Activation Differences. Proceedings of the 34th International Conference on Machine Learning, PMLR 70: 3145–3153. Available at: http://proceedings.mlr.press/v70/shrikumar17a.html

Sidhu, S., Bodger, O., Williams, N., and Roberts, D. L. (2012). The Number of Benign Moles Excised for Each Malignant Melanoma: The Number Needed to Treat. *Clin. Exp. Dermatol.* 37 (1), 6–9. doi:10.1111/j.1365-2230.2011.04148.x

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *ArXiv: 1312.6034 [Cs]*.

S. Vaccarella, J. Lortet-Tieulent, R. Saracci, D. I. Conway, K. Straif, and C. P. Wild (2019). in *Reducing Social Inequalities in Cancer: Evidence and Priorities for Research* (International Agency for Research on Cancer.). Available at: http://search.ebscohost.com/login.aspx?direct=true&site=edspub-live&scope=site&type=44&db=edspub&authtype=ip, guest&custid=ns011247&groupid=main&profile=eds&bquery=AN%2020337414

Troxel, D. B. (2003). Pitfalls in the Diagnosis of Malignant Melanoma. *Am. J. Surg. Pathol.* 27 (9), 1278–1283. doi:10.1097/00000478-200309000-00012

Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., et al. (2020). Human-computer Collaboration for Skin Cancer Recognition. *Nat. Med.* 26 (8), 1229–1234. doi:10.1038/s41591-020-0942-0

Tsianakas, V., Robert, G., Maben, J., Richardson, A., Dale, C., and Wiseman, T. (2012). Implementing Patient-Centred Cancer Care: Using Experience-Based Co-design to Improve Patient Experience in Breast and Lung Cancer Services. *Support Care Cancer* 20 (11), 2639–2647. doi:10.1007/s00520-012-1470-3

Van Dijk, M. C. R. F., Aben, K. K. H., Van Hees, F., Klaasen, A., Blokx, W. A. M., Kiemeney, L. A. L. M., et al. (2007). Expert Review Remains Important in the Histopathological Diagnosis of Cutaneous Melanocytic Lesions. *Histopathology* 52 (2), 139–146. doi:10.1111/j.1365-2559.2007.02928.x

Welch, H. G., and Fisher, E. S. (2017). Income and Cancer Overdiagnosis - when Too Much Care Is Harmful. *N. Engl. J. Med.* 376 (23), 2208–2209. doi:10.1056/NEJMp1615069

Welch, H. G., Mazer, B. L., and Adamson, A. S. (2021). The Rapid Rise in Cutaneous Melanoma Diagnoses. *N. Engl. J. Med.* 384 (1), 72–79. doi:10.1056/NEJMsb2019760

Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D. E., and Zou, J. (2021). How Medical AI Devices Are Evaluated: Limitations and Recommendations from an Analysis of FDA Approvals. *Nat. Med.* 27, 582–584. doi:10.1038/s41591-021-01312-x

Zicari, R. V., Brodersen, J., Brusseau, J., Dudder, B., Eichhorn, T., Ivanov, T., et al. (2021a). Z-inspection: A Process to Assess Trustworthy AI. *IEEE Trans. Technol. Soc.* 2 (2), 1. doi:10.1109/TTS.2021.3066209

Zicari, R. V., Brusseau, J., Blomberg, S. N., Collatz Christensen, H., Coffee, M., Ganapini, M. B., et al. (2021b). On Assessing Trustworthy AI in Healthcare Best Practice for Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Front. Hum. Dyn.* doi:10.3389/fhumd.2021.673104

Conflict of Interest: JT was employed by the company Intel Labs. Also for JT the research was conducted in the absence of potential conflict of interest.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.