Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients

Himanshi Allahabadi, Julia Amann, Isabelle Balot, Andrea Beretta, Charles Binkley, Jonas Bozenhard, Frédérick Bruneault, James Brusseau, Sema Candemir, Luca Alessandro Cappellini, Subrata Chakraborty, *Senior Member, IEEE*, Nicoleta Cherciu, Christina Cociancig, Megan Coffee, Irene Ek, Leonardo Espinosa-Leal, Davide Farina, Geneviève Fieux-Castagnet, Thomas Frauenfelder, Alessio Gallucci, Guya Giuliani, Adam Golda, Irmhild van Halem, Elisabeth Hildt, Sune Holm, Georgios Kararigas, Sebastien A. Krier, Ulrich Kühne, Francesca Lizzi, Vince I. Madai, Aniek F. Markus, Serg Masis, Emilie Wiinblad Mathez, Francesco Mureddu, Emanuele Neri, Walter Osika, Matiss Ozols, Cecilia Panigutti, Brendan Parent, Francesca Pratesi, Pedro A. Moreno-Sánchez, Giovanni Sartor, Mattia Savardi, Alberto Signoroni, Hanna-Maria Sormunen, Andy Spezzatti, Adarsh Srivastava, Annette F. Stephansen, Lau Bee Theng, *Senior Member, IEEE*, Jesmin Jahan Tithi, Jarno Tuominen, Steven Umbrello, Filippo Vaccher, Dennis Vetter, Magnus Westerlund, Renee Wurth, and Roberto V. Zicari

***Abstract*—This article's main contributions are twofold: 1) to demonstrate how to apply the general European Union's High-Level Expert Group's (EU HLEG) guidelines for trustworthy AI in practice for the domain of healthcare and 2) to investigate the research question of what does "trustworthy AI" mean at the time of the COVID-19 pandemic. To this end, we present the results of a *post-hoc* self-assessment to evaluate the trustworthiness of an AI system for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients, developed and verified by an interdisciplinary team with members from academia, public hospitals, and industry in time of pandemic. The AI system aims to help radiologists to estimate and communicate the severity of damage in a patient's lung from Chest X-rays. It has been experimentally deployed in the radiology department of the *ASST Spedali Civili clinic* in Brescia, Italy, since December 2020 during pandemic time. The methodology we have applied for our *post-hoc* assessment, called Z-Inspection®, uses sociotechnical scenarios to identify ethical, technical, and domain-specific issues in the use of the AI system in the context of the pandemic.**

***Index Terms*—Artificial intelligence, case study, COVID-19, ethical tradeoff, ethics, explainable AI, healthcare, pandemic, radiology, trust, trustworthy AI, Z-Inspection®.**

Manuscript received 8 December 2021; revised 13 July 2022; accepted 18 July 2022. Date of publication 29 July 2022; date of current version 15 December 2022. The work of Julia Amann was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 777107 (PRECISE 4Q). The work of Andrea Beretta, Cecilia Panigutti, and Francesca Pratesi was supported by the ERC Advanced through XAI Science and Technology for the Explanation of AI Decision Making under Grant 2018-834756. The work of Walter Osika was supported by the European Union's Horizon 2020 Research and Innovation Program under Grant 101016233 (PERISCOPE). The work of Matiss Ozols was supported by the Wellcome Trust under Grant 206194. The work of Giovanni Sartor was supported by the European Union's Justice Programme (2014–2020) through the H2020 ERC Project "CompuLaw" under Grant 833647. The work of Mattia Savardi, Alberto Signoroni, Filippo Vaccher, and Davide Farina was supported by the Italian Ministry of University and Research ("ResponsiX: Responsible and Deployable AI-Driven Evaluation of COVID-19 Disease Severity on Chest X-Rays") under Grant FISR2020IP_02278. The work of Dennis Vetter was supported in part by the European Union's Horizon 2020 Research and Innovation Program under Grant 101016233 (PERISCOPE), and in part by the Connecting Europe Facility of the European Union under Grant INEA/CEF/ICT/A2020/2276680 (xAIM). *(Corresponding author: Roberto V. Zicari.)*

Please see the Acknowledgment section of this article for the author affiliations.

This article has supplementary downloadable material available at https://doi.org/10.1109/TTS.2022.3195114, provided by the authors.

Digital Object Identifier 10.1109/TTS.2022.3195114

I. Introduction

THE COVID-19 pandemic led to a high saturation of healthcare facilities and a significant rate of respiratory complications. In this context, quick assessment of the severity of a patient's condition played an essential role in the management of patients, clinicians, and medical resources. Most decisions were made clinically, but the primary radiologic tools for facilitating these fast paced decisions were chest X-ray (CXR) and computed tomography (CT) imaging. Among those two, CT images convey more information. However, CT scan exposes patients to more radiation than CXR and, thus, as a more frequent testing tool, CXR is preferred.

The main reason for CXR over CT is to avoid spreading COVID-19, reducing the exposure time to healthcare workers and all other people in a hospital, as CXR could be brought to the bedside, with no need for a patient to be moved in the hospital and to go inside a machine. Second, costs and time (CT scans will have a limited number of slots for patients in

a day and each scan takes longer) also matter in a pandemic which limits resources. Third, a critically ill patient needs to be monitored in a CT scanner. As decisions in COVID-19 clinical care are not usually based on imaging, the risk to a patient of being untended and less monitored in a CT scan machine (while needed personnel is pulled away from other activities) if critically ill and of increased risks of COVID-19 exposures, would be unsupportable if unlikely to change management.

This has made CXR the first diagnostic imaging option for COVID-19 severity assessment and monitoring, despite its reduced sensitivity.

The estimation of the severity of a patient's lung condition, however, may be hampered on CXR by the projective nature of the image generation process. In addition, when serial CXR are performed during hospitalization, descriptive reports may fail to communicate directly and clearly the evolution of the disease to the referring clinicians. In order to provide an unambiguous description of the extent of COVID-19 pneumonia, in March 2020, Borghesi and Maroldi introduced the Brixia score, a semiquantitative multivalued scoring system, which translates radiologists' judgements onto numerical scales, thus providing a supplementary diagnostic tool to improve communication among specialists [1]. During multidisciplinary meetings, the scoring system was shared and discussed by clinicians and radiologists of the *ASST Spedali Civili di Brescia,* and, from March 2020, integrated in the daily routine. This requires that, for every CXR acquired from COVID-19 subjects, the radiologist on duty determines the Brixia score and integrates it in the standard descriptive report. This was made compulsory during the period of the highest emergency and hospital saturation (first Italian pandemic peak) and then continued on a voluntary basis.

In this scenario, it was hypothesized that an AI system could be trained to support the radiologist in estimating the score. The collaboration between a group of engineers and radiologists of the University of Brescia allowed to design and develop such a system, i.e., BS-Net.

A. AI Solution

The BS-Net system [2] is an end-to-end AI system that is able to estimate the severity of damage in a COVID-19 patient's lung by assigning the corresponding Brixia score to a CXR image. The system is composed of multiple task-driven deep neural networks working together and was developed during the first pandemic wave. After the Institutional Board (*Comitato Etico di Brescia*) clearance in mid-May 2020, the system was trained and its performance was verified on a large portion of all CXRs acquired during the first pandemic peak from COVID-19 patients within ASST Spedali Civili of Brescia, Italy. The results of the internal validation, as well as those related to an external public dataset, were first published as a preprint in June 2020. The system not only assesses lung damage in CXR images, it also generates confidence values and creates explainability maps that highlight which sections of the image are most influential in generating the severity score, hence making the AI decision process more transparent to the radiologists.

Aiming at facilitating clinical analyses and considerations, the system has been also experimentally deployed in the radiology department of ASST Spedali Civili of Brescia, Italy, since December 2020. A team of radiologists working at the hospital assisted the engineers in the design of the implemented solutions described in [2].

The AI systems are at the time of writing in an experimental stage, but the severity score estimation and explainability maps computed on CXRs of all incoming COVID-19 patients are available for radiologists that take part in the current and future test activities in a fully integrated way with respect to the standard CXR reading workflow. Through a noncommercial collaboration with the provider of the radiology information system (RIS)—which was already in use at the hospital—the integration of the BS-Net system within the radiological workflow was carried out. All CXRs of COVID-19 patients are processed by BS-Net right after the acquisition and the radiologist has the option of obtaining support for AI during the definition of the Brixia score by opening a dedicated panel from the RIS interface.

The AI system and its explanations received positive feedback from the radiologists working at the hospital [2].

Currently, at the time of writing of this article, the conditions of extreme overload that characterized the first wave of COVID-19 in Brescia, and that gave rise to the need for the systematic evaluation of the Brixia score, did not reappear in the following waves. The use of the score is no longer critical nor mandatory within the hospital. However, despite not being routinely used, the integrated system continues to work in background (thus allowing performance monitoring) and is being used for ongoing clinical studies about the impact of AI on radiologists' work.

B. Research Questions

We conducted a *post-hoc* self-assessment focused on answering the following two questions.

1) What are the technical, medical, and ethical considerations determining whether or not the system in question can be considered trustworthy?
2) How may the unique context of the COVID-19 pandemic change our understanding of what trustworthy AI means in a pandemic?

We expand on these questions in the following.

Is the AI system trustworthy? The AI system is used to support high-stakes decisions. Wrong or systematically biased decisions can result in adverse effects for individuals or whole population subgroups.

Is the use of this AI system trustworthy? AI systems are never used in isolation, but always as part of complex sociotechnical systems. For trustworthy use of an AI system, it needs to be ensured that the users know about the system's intended purpose, abilities, and limitations, and are able to ensure respect for human autonomy, prevention of harm and fairness in its application.

What does "trustworthy AI" in time of a pandemic mean? The current pandemic is an extreme situation, in which the healthcare system is frequently brought to its limits. How does

this pressure influence the need for trustworthiness? Are less trustworthy systems acceptable as long as they help reduce the load on the overwhelmed systems or is trustworthiness even more important, as it is likely that a larger part of the decision will be made by the AI system?

C. Standards of Care and Ethics in Times of COVID-19

The European Commission has proposed a general framework for *Trustworthy AI* (not specific to healthcare) based on four ethical principles, rooted in fundamental rights [3]:

1) respect for human autonomy;
2) prevention of harm;
3) fairness;
4) explicability

and proposed seven requirements for their operationalization, namely:

1) human agency and oversight;
2) technical robustness and safety;
3) privacy and data governance;
4) transparency;
5) diversity, nondiscrimination, and fairness;
6) societal and environmental wellbeing;
7) accountability.

The COVID-19 pandemic is an example of modern unpredictable scenarios which have challenged the traditional way of operating, particularly in the healthcare field, where the high saturation of healthcare facilities was an almost unprecedented event in recent history. The introduction of novel technological devices in the clinical setting is usually a long-winded process. Regulatory and ethical requirements aim to ensure that those technologies meet the highest standards of care to ensure patient safety.

Are the four ethical principles defined above only for "normal, business-as-usual times"? Is it acceptable to modify or weaken them because of a pandemic that causes a state of emergency? It seems to be more acceptable to shift (lower) standards of care in the case of a "fast" pandemic where health services are overwhelmed, such as the COVID-19 pandemic; whereas for "slow" pandemics where we see a steady rise in chronic health conditions, lowering standards might not be appropriate.

It is also important to consider which standards may be subject to change and whether or why a change in circumstances may allow that. For example, could it be adequate to not require consent from patients in a situation where a pandemic is overwhelming, and we want to develop an AI tool to assist doctors, while we require consent in less overwhelming but no less fatal contexts.

More generally, the justification for lowering standards in the case of a pandemic tends to revolve around a lack of resources, time, and counterfactual risk. How and when is such lowering of standards legitimate? Who should make these decisions? A key starting point in cases where standards of care are altered is to do so in a transparent manner and not in secret through backdoors. Adapting standards of care may be called for or even inevitable in certain high-risk situations that require immediate action (the alternative might be more unethical); however, there should be clear procedures and governance structures to monitor and document these adaptations.

We suggest that there might be important lessons to be learned about these and related questions. We will present some of the lessons learned in assessing this use case in Section V.

In this article, we present and evaluate an AI system experimentally deployed in a pandemic context in a public hospital in Italy [2]. The system predicts a multiregional score conveying the degree of lung compromise in COVID-19 patients.

What exactly was the goal of the system, and how were standards of care altered and justified? Why should such standards not be weakened when looking forward and outward with respect to nonpandemic contexts? In our *post-hoc* assessment, and in line with the recent legislative proposals [4], we consider amongst others the notions of transparency and trustworthiness.

The context of the pandemic also gives rise to considerations about how best to assess the performance of an AI. As it will be clear from the use case, the AI system is supposed to help tired and exhausted radiologists and doctors. If this is the reality in which the system is to be used, then the reference standard against which it should be evaluated might not be 'rested' medical doctors in nonemergency contexts. In other words, the comparison needs to be adjusted.

Similar considerations can apply to a use context where the system assists junior doctors. In emergency/overload contexts, the AI system performance, if maybe not proven to be at the level of leading experts, should be at least at sufficient level of accuracy to be helpful. Thus, a lesson that might be learned from the pandemic case is that the test of AI performance must be matched to the clinical reality in which it is supposed to add value, and that might not require top performance; in some cases, such as the present case study, "better than average" would already present a significant improvement. Of central importance is of course the question of what is ultimately in the best interest of the patient.

D. Contribution and Paper Structure

In this article, we present the results of a *post-hoc* self-assessment to evaluate the trustworthiness of an AI system for predicting a multiregional score conveying the degree of lung compromise in COVID-19 patients, experimentally deployed in a public hospital in the time of Covid-19 pandemic.

This article is structured as follows: Section II introduces the methodology we have used to assess the AI system; Section III presents the use case, the analysis of sociotechnical scenarios, how to define the mappings to the trustworthy AI framework, the key issues we have identified and some recommendations; and Section IV presents some considerations on the trustworthiness of AI in times of pandemic. In Section V, we present some reflections on what we have learned from this *post-hoc* assessment that can be useful for similar cases in the future, together with an evaluation of our methodology compared with the related work.

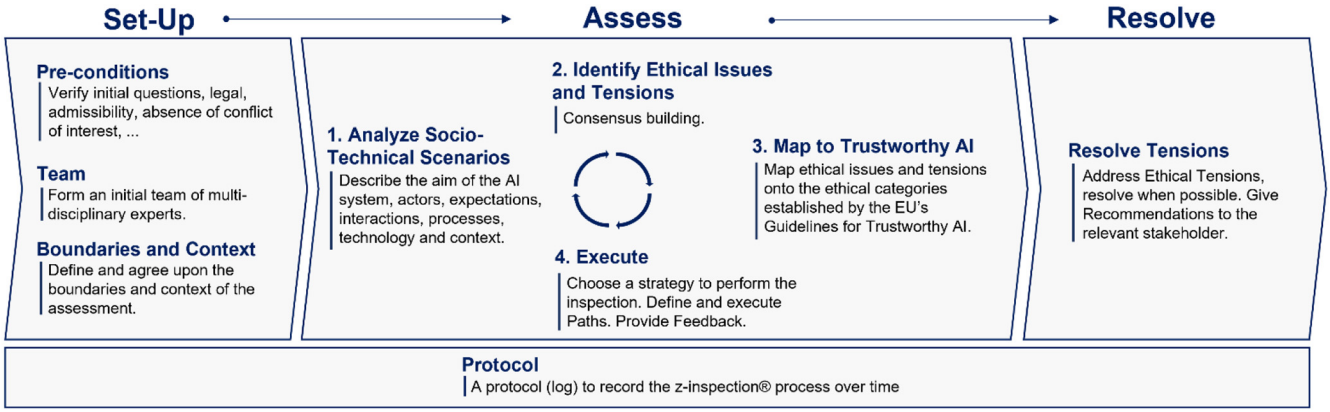

Fig. 1. Z-Inspection® process in a nutshell (adapted from [5]).

II. Methodology

In this section, we give a high-level overview of the methodology we have used for this *post-hoc* self-assessment. The process is described in detail in Section III.

A. Z-Inspection® Process

We used a process to assess trustworthy AI in practice, called Z-Inspection® [5], which expands upon the "Framework for Trustworthy AI" as defined by the High Level Experts Groups set up by the European Commission [3]. The Z-Inspection® is a holistic process based on the method of evaluating new technologies according to which ethical issues must be discussed through the elaboration of sociotechnical scenarios. The Z-Inspection® process is depicted in Fig. 1, and it is composed of three main phases: 1) the Set Up Phase; 2) the Assess Phase; and 3) the Resolve Phase. The process has been successfully applied to both assess *post-hoc* [6] and *ex-ante* [7] trustworthiness of AI systems used in healthcare.

B. Creation of Interdisciplinary Team

In the Set Up phase, we created an interdisciplinary assessment team composed of a diverse range of experts. For this use case, the team included: philosophers, healthcare ethicists, healthcare domain experts (such as radiologists, and other clinicians, and public health researchers), legal researchers, ethics advisory, social scientists, computer scientists, and patient representatives.

The choice of experts required for this use case had an ethical dimension since the quality of the analysis and the results depended on the diligent selection and quality of experts including them not being biased or in a position of conflict of interest. Domain experts may need to include several classes of expertise and practice, especially as a tool may impact the workflow of different categories of professionals. Since this was a self-assessment of an AI system, special considerations have been taken into account of the potential behavioral bias of the stakeholders *owing to the* use case in the process of the evaluation.

Team members were selected based primarily on required skills and expertise. To ensure the quality of the inspection process, it was important that all team members respect specific areas of competency of each other. Later additions of experts to the team were limited. It is preferable that later additions are avoided to keep the team's viewpoints balanced and the workflow of the team stable.

The team composition was as follows (all team members are coauthors of this article).

Lead: Coordinated the process and the finalization of the interim issues report.

Rapporteur: Wrote minutes of all Zoom-meetings in a shared google doc.

Ethicist(s): Helped the other experts identify ethical tensions and dilemmas and how to solve these.

Domain Expert(s): We had more than one to bring different viewpoints (specialized radiologists and generalistic medical doctors), assisted inter alia in establishing whether there was a ground truth regarding the problem domain, and what this was.

Legal Expert(s) Specialized for the Specific Domain: due to being highly specialized in the field legal experts had to be familiar with the problem domain area and/or have some understanding of the legal aspects of data protection and human rights.

Technical Expert(s): With specialty in Machine Learning, Deep Learning, Imaging and data science.

The team included also *Social Scientists, Policy Makers,* and *Communication specialists.*

The Role of Philosophers/Ethicists: Philosophers/Ethicists acted as "advisors" to the rest of the team in order to assist team members with little ethics background in the interpretation of the four ethical principles and the seven requirements identified in the EU guidelines for Trustworthy AI.

This interdisciplinarity is one of the most important aspects of our approach to ensure that a variety of viewpoints are expressed when assessing the trustworthiness of an AI system.

The set-up phase also includes the definition of the boundaries of the *post-hoc* assessment, taking into account that we do not assess the AI system in isolation but rather consider the social-technical interconnection with the ecosystem(s) where the AI is developed and/or deployed. This case is a special one, since we considered the context of the pandemic.

C. Split the Work in Working Groups

Initially, the experts team met together with the stakeholders owning the use case in a number of workshops (via video conference) to define sociotechnical scenarios of use of the AI systems. We use the term *stakeholders* in the rest of this article to denote the actors who have direct ownership on the development and deployment of the AI system.

Later, the team was split in a number of working groups (WGs), grouped together by homogeneous expertise, namely, eight WGs.

WG Technical: It is composed of 21 experts in Deep Learning/and Medical Image recognition.

WG Ethics: It is composed of four experts in ethics.

WG Ethics/Healthcare: It is composed of four experts in healthcare ethics.

WG Healthcare Radiologists: It is composed of three experts in radiology (independent from the radiologist of the hospital).

WG Healthcare Medical Doctors/Others: It is composed of 15 experts in various areas of medicine.

WG Law/Healthcare; Law; Data Privacy, and Data Protection: It is composed of four experts in law, data protection, and data privacy.

WG Social Science/Ethics/A.I./Policy Makers: It is composed of five experts in Social Science, Policy makers, and representatives of patients.

WG Lead: It is composed of two experts who coordinated the assessment.

D. Creation of Reports

Each WG analyzed the sociotechnical scenarios and produced preliminary reports—working independently and in parallel to avoid cognitive biases and take advantage of their unique perspective and expertise. Such preliminary reports were then shared with the entire team for feedback and comments. These interdisciplinary interactions among experts with different backgrounds allowed each WG to consider the viewpoints of other experts when delivering their final reports. Each final report was written using free text and open vocabulary to describe the possible risks and issues found when analyzing the AI system.

Specifically, each WG report listed the identified ethical, technical, domain specific (i.e., medical) described using an open vocabulary. In this article, we will not consider legal issues.

E. Mappings to the Framework of Trustworthy AI

The *issues* described in free text were then mapped by each WG using templates (called rubrics) [5], [8] to some of the four ethical principles and the seven requirements defined in the EU framework for trustworthy AI [3]. With this mapping, the reports developed from an open vocabulary to a closed vocabulary (i.e., the templates). We call these *mappings*. Each WG worked independently from each other, and adopted different/similar strategies to perform such mappings. We will present in Section III an example of how a WG performed such a mapping strategy.

F. Consolidation Process of Mapping

At this point, we consolidated the *mappings* produced by the various WGs into a consistent list. This was done by creating a dedicated WG who grouped the issues that had been mapped to the same requirements of the EU framework for trustworthy AI. The consolidated lists of WG issues for each of the seven requirements were reviewed so commonalities and differences could be identified and discussed before the final consolidation. The method highlighted how different perspectives could lead to similar issues being mapped to different requirements. We will show in Section III the results of such consolidated mappings for this use case.

G. Give Recommendations

The resolve phase completes the process by addressing ethical tensions and by giving recommendations to the key stakeholders. It is crucial to monitor that the AI system that fulfilled the trustworthy AI requirement at launch continues to do so over time. Therefore, when required, the resolve phase includes conducting a trustworthy monitoring over time of the AI system (we call it "ethical maintenance"). In [9], we have defined an AI ethical maintenance process based on an adapted version of the reliability-centered maintenance (RCM) model [10]. This is not part of this initial *post-hoc* assessment and it will be performed in the second stage.

III. Assessing Trustworthy AI in Times of COVID-19: Deep Learning for Predicting a Multiregional Score Conveying the Degree of Lung Compromise in COVID-19 Patients

The Assess Phase of the process begins with the creation of sociotechnical scenarios.

A. Phase I: Sociotechnical Scenarios

We considered three possible scenarios in which the AI system could be used.

1) *The current scenario* is a single-site deployment in a radiology department at the hospital, where the system supports radiologists in their daily workflow by providing a second expert opinion to reduce oversights and fatigue-related mistakes.
2) *Possible future applications* of the system include access via a Web-interface where users can upload CXR images and the system then provides them with a severity estimation and an explanation map. In this scenario, the system can serve as a readily available expert opinion in areas where access to qualified radiologists is limited. An initial prototype for this process was already developed.
3) *Another possible future application* is large-scale image analysis where the system can rate large archives of historical data, e.g., for use in retrospective studies. Here, the system can be used to annotate large datasets and lighten the workload of radiologists labeling historical data.

1) Aim of the AI System: The main goal of the system is to alleviate the load on overwhelmed radiologists, and improve quality and timeliness of patient care and management. It should act as a support system that assists the radiologists in pneumonia assessments for COVID-19 patients in all hospitalization phases and improve the radiologists' performance, especially by acting as a safety net to catch avoidable errors related to fatigue, misinterpretations or similar causes.

Furthermore, in the intention of the engineers and the radiologists who implemented the system, the system should act as a stable reference for different kinds of clinical studies, where it could fulfill the role of a radiologist capable of annotating large amounts of images in a short time. The stable reference is important, as CXR images allow for some degree of subjectivity in their interpretation [1], [2] and the system could therefore abstract away from the different levels of experience among the 50 radiologists in the hospital and it could support junior radiologists during training with fast access to an expert opinion and explanations.

While, thanks to the verified high-performance, the system could in theory be used without a radiologist to combat temporary shortages of personnel, this was not the case in the hospital where it was developed and verified. Autonomous system functioning is to be discouraged at this stage since: 1) it would require external validation and possible fine-tuning if used outside the native context and 2) machine and human errors only statistically compensate (in favor of the machine) but remain different in nature, therefore discouraging fully autonomous working and leading to the need of deeper evaluations about the deployment modalities [11].

2) Identification of Actors: The system is directly and indirectly in contact with a multitude of actors. Depending on the type of contact, we grouped the actors into primary, secondary, and tertiary actors.

Primary actors are in direct contact with the system during day-to-day business or directly affected by the system. This includes patients, reporting radiologists and other clinicians, as well as the clinical and technical staff that handle and assist system development.

Secondary actors are in contact with the system but do not use it in their workflow or are directly affected by its decisions. This includes the supporting RIS and picture archiving and communication system (PACS) vendors which worked to facilitate data collection and to integrate the system in existing radiology devices and workflows and provide assistance in data management (anonymization), as well as the hospital IT services that support the research team to ensure smooth operations.

Tertiary actors potentially benefit from the system, even though they are neither working with the system nor are they directly affected by its decisions. We identified the University of Brescia and the hospital where the system is deployed as tertiary actors. Other tertiary actors include unrelated researchers that use the publicly available dataset and/or the BS-Net for their own research.

Actors Expectations and Motivations: Depending on their levels of contact and involvement the actors have different motivations for working on/with the system and different expectations toward it. None of the actors has a commercial interest regarding the application of the system. During workshops, we identified the following main expectations.

Primary actors want help during the ongoing pandemic. Their motivation is a reduction of overload on personnel, better communication, delivery of care, response to clinical needs, and overall better treatment. Therefore, they expect the system to produce quick, stable and reliable severity scores.

Secondary actors want to help in a complex situation and collect experience in provisioning a new kind of medical services.

Tertiary actors expect improved treatment of patients and increased visibility/reputation from successfully employing a complex AI system in clinical contexts during an emergency situation.

Potential tension in the interests between actors was flagged in the work of one of the WGs: patients and developers have different interests regarding the collection, control, and use of personal and sensitive data.

3) Context and Processes, Where the AI System Is Used: Currently, the system is experimentally deployed in the radiology department of the Brescia Public Hospital, where it is tightly integrated in the radiologists' workflow. The system handles CXR images of all incoming COVID-19 patients. For these images, it provides a severity estimation to the radiologist who requests it, along with an explainability map and the system's confidence in its prediction. From inputting an image into the system to outputting scores and explainability maps, the whole process takes less than 1 min. After seeing the CXR images and the system's prediction and reasoning, the radiologists can freely adapt the scores according to their judgement and they can use the explainability map and the provided confidence scores to resolve disagreements with the system. In an earlier deployment, the radiologists could only access the system's output after making their own decision, but they requested earlier access to the system's output to better integrate the use of the system with their workflow. From the system's initial deployment in December 2020 until mid-April 2021, more than 19 000 images were assessed by the system. However, only a fraction of these was handled through the dedicated interface since the hospital conditions in that period were far from the overwhelming ones that urged the design and development of this solution. We note that the system operativeness remains highly relevant in terms of stability and robustness, allowing and streamlining realistic dedicated clinical evaluations and continuous performance monitoring

4) Technology Used: The system is an end-to-end system, where the input is a CXR image, and the output is the same image, annotated with the severity scores, confidence scores, and an explainability map. It consists of multiple specialized networks for solving the subtasks of segmentation, alignment, feature extraction, and scoring. All of the components are trained in isolation to satisfy performance on their respective tasks, after which the complete system is trained end-to-end for further performance improvements [2].

The segmentation subtask is performed by a U-net++ network [12], a specialized architecture for medical image

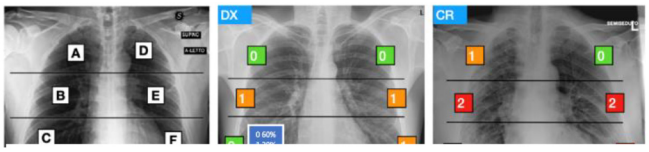

Fig. 2. Brixia score. (a) six zones definition and (b) and (c) examples of scores (either defined by the radiologist or estimated from the AI). In (b), confidence values generated by the AI prediction are shown (modified version of the figure in [1]).

segmentation. The goal is to output a probability mask of the lung's location in the image, so that following steps can focus their efforts on this region.

The alignment subtask is performed with the help of a spatial transformer network [13]. Input for this network is the segmentation mask from the previous image, output are the coefficients of an affine transformation that is then used for resampling and aligning the original image and features detected by later steps. Alignment is performed to center, rotate and zoom the lungs, so that their final position is approximately the same for every image. This makes the system more robust against the different perspectives from which the CXR images are taken.

Feature extraction is performed by the so-called "backbone," a pretrained state-of-the-art convolutional network. The default backbone is a ResNet18 [14], but other backbones, such as Inception Net [15] or DenseNet [16] can be used as plug-and-play replacements. The backbone outputs feature maps at different resolutions, which are then used in later steps.

In the last step, the aligned features at different resolutions are pooled based on their corresponding lung region and then used to estimate the Brixia Score for each of the six lung regions.

In the multiregion 6-valued Brixia-score [1], lungs in anteroposterior (AP) or posteroanterior (PA) views are subdivided into six zones, three for each lung, almost equal in height [Fig. 2(a)], and the referring radiologist assigns to each region an integer rating from 0 to 3 [Fig. 2(b) and (c)], based on the local assessed severity of lung compromise: 0—no lung abnormalities, 1—interstitial infiltrates, 2—interstitial and alveolar infiltrates (interstitial predominance), and 3—interstitial and alveolar infiltrates (alveolar predominance).

Explanations are generated via a LIME [17]-inspired approach. First, the image is divided into superpixels, regions of similar intensity and pattern. The importance of each of these superpixels is then estimated by masking the superpixel (i.e., setting all pixel values to the background value of 0) and then checking how the prediction changes if the information in this superpixel is not used.

The collected image database corresponds to the whole flow of CXR produced in one month during the main pandemic peak in north-Italy from all the COVID-19 patients admitted to the hospital from the end of March 2020 to the end of April 2020. Annotations were performed by the different radiologists employed and on duty in the hospital, thus corresponding to the real clinical activity of two radiology wards counting about 50 radiologists. In total, the dataset comprises 4703 CXRs. Since more than one image can be associated with the same patient (especially more compromised patients who underwent CXR exams even on a daily basis), the training/validation/test splitting has been given on a patient basis. In particular, the test set comprised about 450 images and 150 of them have been further annotated with the agreed score of five different radiologists. During the training process, training images were augmented by applying geometric transformations, random changes in brightness and contrast, as well as flipping of images and labels.

A detailed report on the implementation details is available in [2].

5) AI Design Decisions and Tradeoffs: During development, multiple different backbone networks were tested. The final network, ResNet18, was selected as it provides the best tradeoff between quality of extracted features and resources required for inference.

The custom explainability method was developed, as existing methods such as GradCAM [18] did not create explanations of the desired spatial localization and precision. Furthermore, the output of GradCAM and related methods was found to be more difficult to understand by the radiologists.

There was also the conscious decision against continuous learning as a more stable system is preferred. For increased stability, the system is also monitored for concept drift—a concept drift, for example, happens when the data statistics to predict change over time and the training set is not representative anymore—and if the performance deviates from the expected behavior, the system will be retrained. In particular, every prediction is tracked, and results feed a back-end dashboard where statistics are constantly monitored and alarms are generated in case of malfunctions, out-of-service occurrences, or score statistics abnormalities. The system went down very few times only due to external reasons, while at the time of writing, no alterations occurred which would have made necessary a model tuning.

6) Process Workflow: An important decision made by the key stakeholders of this use case—namely, the radiologists at the hospital—was to set as default that the results of the AI score prediction are immediately visible to the radiologists when they report on the COVID-19 form integrated in the reporting workflow (see Fig. 3). Specifically, in the RIS interface, there is a button to access a so-called COVID-19 form that can be opened, allowing the radiologist (who is contextually viewing the patient CXR on a separate diagnostic monitor) to confirm or freely modify the predicted score. Confidence values for each regional score and explainability maps are also available on the form. The AI system is active within this form. Radiologists who are not willing to use the AI system can write directly her/his report without entering into the COVID-19 form. This specific opt-in policy is justified by the experimental nature of the deployment and from the fact that, given that the whole radiology team at Brescia's hospital counts more than 50 specialists and is subdivided into two departments, it was not considered necessary to force the whole staff to always estimate the score, or to use the specific interface, in periods where the hospital is not in presaturation conditions and the score estimation is not

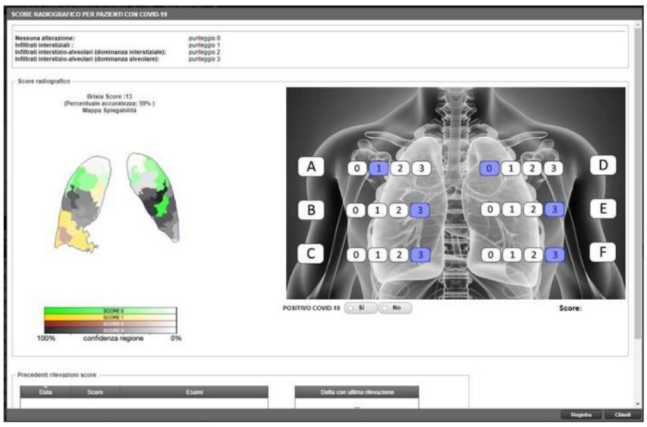

Fig. 3. COVID-19 reporting form.

made mandatory anymore. However, the fact that the system is continuously working and promptly answers on-demand, *de-facto* enables further experimentations, e.g., the ones that are currently involving radiology residents and the role of the AI system in their training (this is an ongoing work and is directly linked to the concerns coming from Radiologists in Section III-A1).

7) Intellectual Property: The dataset, model architecture, and the weights of the trained model are publicly available under an opensource license on the project website https://brixia.github.io.

8) Legal Framework: Deployment in the radiology department was made possible through the integration of the system as an experimental add-on by the RIS vendor (see Fig. 3). The user manual of the system has limited-liability disclaimers and explicitly informs users about the importance of oversight since, despite statistically compensating in favor of the AI, radiologists and AI make different errors in nature.

Images used for training of the system were anonymized to comply with data protection regulations and a safe anonymization was guaranteed by hospital IT. During the first months of the pandemic, the team also received a special waiver from regulatory bodies and ethics committees. To ensure compliance with GDPR and patient rights, we were told that the development team consulted lawyers during system design and received help with drafting the license agreement for the dataset. Full ownership of the data remains with the hospital.

9) Protocol: The protocol of the assessment is a shared google doc that kept being updated/commented during the all process.

B. Phase II: Analyzing of the Sociotechnical Scenarios From Different Viewpoints

We present a summary of the analysis of selected WGs. The analysis was conducted in parallel by the various WGs and, intentionally, we allowed that the results had possible duplications, and overlapping of content. In the consolidation phase later on, we addressed these overlapping and duplications.

1) View of WG Healthcare Radiologists: The team of independent radiologists consider the present AI algorithm as a robust method for the semiquantitative assessment of COVID-19 disease.

In their opinion, the present data show that the algorithm can segment the lung very accurately. The user interface is very well set up and clear. The assessment of conventional images is a routine task, which takes place in 2–3 min. Additional time lost by incorporating or using additional components would result in it not being used by radiologists.

They concluded that from the radiological and technical point of view, the system can be easily integrated into a PACS system. The fact that the radiologist does not wait for the score has two effects: on the one hand, the radiologist does not lose time for reading and reporting the images, on the other hand, they may get biased by the presented score. This might be a problem, when young radiologists are reading the images.

It should be noted that the algorithm only evaluates a momentary status according to the image present. This momentary status includes the general health of the patient as well as his/her actual status during CXR, which may influence inspiration depth. Furthermore, the technical skills of the technician may influence the image quality and, thus, the calculated score.

The algorithm does not allow for reliable longitudinal observation because, in particular, changes in respiratory position (after intubation, for example) are not included. Although the algorithm and the Brixia Score were developed for COVID-19 related evaluation of CXRs, it is very unspecific to that disease. The score can be applied equally well to any other disease, meaning that it has not been developed specifically for a certain pattern. Therefore, it is mandatory that the physician is informed about the diagnosis and the clinical status of the patient concerned. Regarding the COVID-19 disease itself, the score does not allow differentiation between diseases and between different stages such as the transition from infiltrates to the Acute Respiratory Distress Syndrome.

Nevertheless, the score provides a certain standardization in itself, but in the opinion of the independent radiologists who assessed the system, it is not robust enough with respect to the variations in imaging acquisition, which as such is not standardized enough. The severity score correlates with the patient outcome; this is rather because the severity score correlates with high opacity in case of severe disease and relatively low opacity in case of mild disease. As mentioned previously from a radiological point of view, the score is easy to implement and use. It provides guidance and does not disempower the radiologist who still needs to be aware of the type of patient and disease present. Thus, this score does not help to differentiate atelectasis versus consolidations. Both atelectasis and consolidation lead to an increase in density and therefore to a higher score.

In case of, e.g., poor inspiration, the score cannot replace the radiologist, who primarily has to check the quality of the image with regard to inspiration depth, exposure, and superimposition.

The radiologist must include information about any underlying lung diseases (e.g., UIP, Emphysema), which are not captured by the algorithm. The score only assesses the pattern density and leaves the interpretation of the findings unchanged

to the radiologist. Therefore, in the opinion of the team of independent radiologists, it does not disempower the radiologist but supports them in interpreting the severity of a disease.

The algorithm has the following limitations that have to be taken into account: (1) it was not tested for a pediatric population and (2) the training/test data are curated from a specific country and the same hospital. The system's generalizability has not been appropriately tested on external and more diverse data. It would need a large dataset with diverse, high-quality images curated from multiple institutions and different geographic areas in order to claim and ensure the generalizability of an AI system intended for clinical usage [19], [20].

The specific application of the Brixia score to the CXRs is another key point. Most of the published papers have primarily focused on the use of AI in CT to diagnose pulmonary COVID-19, later, the focus shifted toward quantification [21]. In the case of CXRs, it is indeed useful to have a quantification system, as the visual assessment is subject to wide variability among radiologists. A deep-learning-based system that can provide quantification of the severity of lung engagement is welcome because it helps in clinical practice. The team of independent radiologists believe that this work, similar to others like it, stimulates the concept of "AI to support the diagnosis," as opposed to the concept of "diagnostic AI." In fact, we must think of an AI that assists the doctor and not an AI that replaces the doctor. Therefore, it is important that the radiologist reviews it after reporting in order to not be biased by the results.

2) View of WG Healthcare Medical Doctors:

a) Dataset and population data selection: A major issue with the algorithm design lies in the dataset, primarily how representative the training data are for prospective populations. The performance of the tool would need to be evaluated with diverse demographic features, as it presently skews toward the relatively homogeneous patients living in Northern-Italy. Image quality may vary between different geographic locations and different datasets may have varying degrees of quality. Beyond generalizability issues, the overall absence of demographic and relevant metadata could lead to other biases. For instance, knowing medical history would provide information on pre-existing lung issues that might influence the Brixia score independent of COVID-19 severity.

Questions remain as to how well the algorithm is able to accommodate potential heterogeneity of image quality. If data collection does not include low resource regions of the world, where image quality and different underlying diseases play a role, but the tool is used in such settings, as suggested by the creators, systematic error would bias the results.

b) Clinical usage: This tool, though directed at radiologists, is intended to support clinical care and hence primarily impacts three groups: 1) radiologists; 2) patient-facing clinicians; and 3) patients.

The information communicated by this tool is very different from what a radiologist would normally communicate or how a clinician would independently read a film. The tool utilizes the Brixia score [1], which divides the X-Ray lung images into six fields, summing a quantification of the opacification in each field to create a total score. This is different from how clinicians intuitively divide lungs, as with the heart on the left, there are three lobes on the right, and two on the left. This is also different from how radiologists and patient-facing clinicians normally would examine or communicate about an X-Ray. Radiology reads provide narratives to supplement a clinician's independent read and describe visual findings. A narrative will include a differential diagnosis, as well as a description of pertinent attributes. This conveys a more textured description of lung fields and extra-pulmonary findings (heart, trachea, and bones). The read will also identify if one side or lobe is heavily affected or if the lungs are diffusely affected. These findings will contribute to a clinician's understanding of the current patient status, past history, and expected outcomes, as well as expected lung functioning. A single score as well as the values of the six pulmonary regions, which, in effect, averages the opacification, will not capture this more complex information. As such, this tool is not intended to recreate a standard radiology read, but instead to provide a different metric and change the radiologist workflow.

The COVID-19 epidemic pulled clinicians away from the bedside and reduced clinical exams, leading to great reliance on computer-based information so it is more important than ever to understand how real life clinical practice incorporates the tool and whether this improves workload or outcomes. New tools, even if accurate, can create unexpected impacts or distractions and come with a cost of adaptation, which may be difficult during a surging epidemic.

Radiologists working at the Brescia hospital report that the metric (Brixia score) would, in their system, only be shared with clinicians at their institution with a descriptive report by the radiologist, who may choose not to include it.

Although the radiology read will be included with the score, it is important to understand how this affects clinician understanding and decision making. The score focuses on findings which may not be crucial in clinical care. Degree of opacification is certainly a factor in evaluating COVID-19 films, but it often does not change management. This is also not a difficult feature for clinicians or radiologists to interpret, but in a busy ward, this metric may be over-relied on. This matters because identification of other findings may necessitate immediate or specific clinical interventions. Some specific patterns, beyond simply opacification, may convey further information about potential superinfection or other interstitial lung processes, such as effusion and edema (diffuse fluid accumulation in the lungs) which increase the score but require different management, not directly related to COVID-19. Some radiologic findings related to COVID-19 which could substantially impact a clinical plan, such as a pneumothorax (collapsed lung) would not increase the score and it is important clinicians understand this. These specific findings may require more immediate or very different clinical intervention (such as a placement of a chest tube or diuresis) and it is important to determine whether a low score could falsely lull clinicians into slower response.

Moreover, as we have learned more about COVID-19, we understand its impact extends well beyond the lungs. It is essentially a multiorgan disease [22], [23] and clinical progression is tied to much more than pulmonary findings. Other

substantive pulmonary findings due to COVID-19 may not be evident at all on CXR, such as a pulmonary embolism.

Many patients may also have pre-existing lung issues seen on Xray, such as old tuberculosis, which may affect this metric's scoring and may or may not be associated with worsened outcomes. Those with cardiomegaly, an enlarged heart, or prior structural lung disease may have even less visible lung volumes and this may further disproportionately affect the six lung fields of this metric.

Less is often more in medicine. The introduction of any new tool can bring costs and risk automation bias. The cost may simply be the time for learning and implementation affecting productivity. More tools can create more distractions for clinicians trying to streamline clinical decision making when patient numbers are rising and time is limited in a pandemic. The time needed to click on a computer for the score may be extra time a clinician does not have in a surging pandemic.

Overall, as the workload snowballs in a surge and more cursory reads may be relied on, it is important that the tool is studied and validated with an eye to clinical impact. It would be best if the tool was included in a clinical study, particularly a clinical trial, to determine whether its addition benefitted clinical care. It could be determined whether it quickened the workflow, added to decision making or clinical outcomes. It will also be important to assess the tool based on different stages of disease development, as COVID-19 clinical presentation rapidly evolves and patients in a surge may present at different points depending on hospital bed availability, testing rates, and denial.

Clinicians often look for clinical severity tools to guide management. At this time, this tool has not been validated as a predictive clinical tool, but is used to describe some elements of the severity of the radiologic findings. Further studies can look to see if the tool improves decision making at key decision points (such as hospital admission and allocation of intensive care beds).

At this time, the tool may be useful in streamlining clinical trial population evaluation, which is important for clinicians involved in research during the outbreak.

Ethically, given there are hard choices to be made when resources are insufficient in the face of surging cases, it is important that an unvalidated tool should not be used to make life altering decisions, such as to limit care. Radiographic findings do not always correlate with current clinical status or outcomes, especially given the complex nature of COVID-19, and it would be important not to have an unvalidated metric guide care. The tool would benefit from more study in a clinical setting to determine and validate its ability to predict clinical severity and whether it assists clinicians in their care of patients.

c) Autonomy/human oversight of AI: The main goal of the system was to support (not replace) radiologists in assessing pneumonia severity for COVID-19 patients in all hospitalization phases.

At first, the score given by the AI system could only be seen after the radiologist saved their report. It was then requested by radiologists to be able to access the AI system's output earlier, as having the score available integrates better into the workflow of clinicians.

Brescia researchers are investigating whether radiologists blindly confirm the tool's suggestions, or use it as a helpful second opinion. The issue here is whether the Brescia radiologists are being influenced or biased in their decision by the score, if they look at it before they analyze the CRX themselves.

d) Post-pandemic use: Issues around generalizability of the algorithm and clinical utility bring to question how this tool might be used in non clinical settings and even outside the current pandemic. This work may be a guide as to how to create a workflow to develop such tools. Future epidemics may require very different tools and even in radiology may require more nuanced reads of films. The tool would also need to be separately evaluated on any different disease processes as its utility may vary with different diseases. As there begin to be viral co-diagnoses with COVID-19 as other respiratory viruses bounce back, this may make it more difficult to interpret these findings on COVID+ but now, RSV+ images, when in the initial outbreak COVID-19 alone, caused viral respiratory infections. The tool may be an excellent starting point for research stratification of severity of COVID-19 as we learn more about this disease in the coming years.

e) Effect on healthcare: Foremost for clinicians will be whether this tool benefits patient care. This could be through simplified triage, streamlining case management, predicting clinical progression, or communicating findings to family members and patients, but it would require further evaluation and testing for such uses. If it is incorporated into use, such a tool may impact healthcare workers and their workload. Increased automation can lead to deskilling workers. The tool might also, if further iterations of the tool can accomplish more, be able to reduce the number of healthcare workers needed, which can be disruptive or beneficial depending on the context. It could also, if this were ever used in lieu of a radiologist read if there were ever too few radiologists and too many films, upskill end-user clinicians who would need to rely more on their own radiologic reads for the finer points of X-Ray interpretation. Likewise, just as the tool might be used to evaluate inter-rater variability between radiologist reads, it might also be used to standardize clinical inputs in the evaluation of clinical decision making among patient-facing clinicians and clinical centers. This tool is likely best suited as a metric for research, which can facilitate clinicians being involved in real time research, which is very much needed in a pandemic in order to determine best means of clinical care. It also can involve clinicians in the iterative of involving new tools, as AI expands its role in medicine and how these can be used more easily and safely in emergency conditions.

f) Liability: Liability remains a concern with this tool. This tool would need a clinical evaluation with an ethical committee involvement. This would affect both radiologists and clinicians and without clear communication, it is possible that clinicians might believe this tool had more validation within the field of radiology than it has had—or possibly, to the contrary, distrust it more than it is justified.

The tool also avoids some of the blackbox concerns other AI tools may have, which may make it easier for the tool to be understood and adopted by clinicians. However, a tool that does not incorporate the full range of clinical findings may cause questions of liability. Some clinicians may assume that a low score means a deprioritization of the evaluation of the film or patient, which could lead to delaying a response to a pneumothorax or other findings requiring emergency intervention.

The designers of the tool do show caution in its use [2], as the tool would have limited decision making capacity, instead decision making would be left to the humans (radiologists, clinicians). Any use for triage or clinical guidance would require further clinical study to determine its use and benefit. Given the constant evolution of COVID-19 surges, there may be many different clinical contexts in which it could be. Otherwise, it is unclear how liability would be resolved, especially if clinicians had faith in the tool, without adequate preclinical usage testing and evaluation.

There are also always concerns regarding data protection and cybersecurity with any clinical tool. The tool does not use metadata and fewer protections are required for CXR imaging, but it will continue to be important to see how it is implemented in different medical systems. The system does not appear to place the individual or patient in control of their own data.

A more detailed risk management plan and governance structure would need to be in place if it were to be expanded or scaled up. It is unclear who is accountable for the system making mistakes, or how liability would be resolved, even if in the end, the radiologists and clinicians are the final decision makers. There would need to be a clear process for complaints.

3) View of WG Technical: In this section, we summarize the technical issues in the system that can potentially give rise to ethical, legal, or even performance issues and limit the applicability of the system. The issues are divided in three categories: 1) training data; 2) data labeling; and 3) model definition and maintenance.

a) Data distribution: The model was trained on data collected from COVID-19 patients over the course of one month during the first wave of COVID-19 in one of the largest hospitals in Italy. The Brixia dataset contains almost 5000 CXR images for training the classifier and 1000 CXR images for training the segmentation and alignment. We consider the following the most pressing issues with this training dataset.

Small Size: While the model employs transfer learning to reduce the number of images required, we are not sure if 5000 images are enough to capture this complex problem's variance. Even though this is large for a medical dataset and first evaluations against the publicly available datasets suggest that the model generalizes well, additional future evaluation is needed to ensure that the dataset stays representative of the cases seen in the hospital.

Representational Fairness: At the time of collection, age is skewed toward older patients and mostly excludes patients less than 18 years olds (Fig. 4). The patients' gender is also biased toward male. Based on the evaluation so far, there appears to be no statistically significant difference in the performance between age groups or sexes, but this might change if the AI is deployed in areas with different demographic distribution. Ethnicity is naturally dominated by the Italian demographic (∼80%), given the location of data collection and model deployment. Since further ethnic information was not collected from patients, ethnic representation could not be verified.

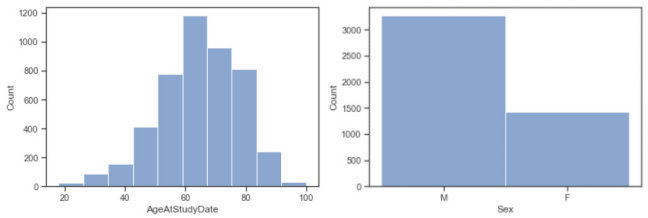

Fig. 4. Distribution of patients' age (left) and sex (right) in the training dataset.

Limited Set of Devices: In addition to the limited demographic diversity, over 90% of CXR images in the dataset were taken with devices from only three manufacturers. Changes or upgrades to the existing devices (e.g., new software for pre/postprocessing, new denoising algorithms, firmware/functionalities updates…etc.) demand additional validation efforts to ensure the changes do not affect the prediction power of the trained models, or even invalidate the whole model. This issue is especially sensitive, as updates to the X-ray machines' software are rolled out often without the hospital's control.

b) Data labeling: The label for each image is its Brixia Score, a method developed by two of the authors of the algorithm under assessment [1]. The Brixia Score is the total sum of the assigned discrete values of (0, 1, 2, 3) for each of the six predefined regions of the lung, whereby, a score of 0 signifies no lung abnormalities, and higher scores indicate more abnormalities in the corresponding lung region.

No "Hard" Ground Truth: With the semiquantitative Brixia Score, there is no hard ground truth and two different radiologists' scores can differ a lot, without one being more correct than the other. The majority of the dataset consists of images annotated by one radiologist. This is also the part that is used for training. Only for a small part (approx. 150 of 5000 images) the images are annotated with the consensus of a group of radiologists, and these images are only used for evaluation.

Score Does Not Describe COVID-19 Specifically: The score is only information on the damage of the lung section and not on what caused the damage. As all images come from patients highly suspected of suffering from COVID-19, subjects with a score of 0 are therefore assumed to be suffering from COVID-19 but not (yet) from related pneumonia. The developers are aware of this limitation and suggest that the system should not be interpreted as a disease detection system.

Potentially Biased: The radiologists employed for labeling all come from the same hospital which might lead to certain unconscious biases in the labeling process. Furthermore, some of the coauthors of the publication describing the AI system were also involved in the development of the Brixia Score, and might therefore be biased.

c) Model definition and maintenance: The system consists of multiple components, one for each of the following subtasks: 1) image segmentation into lung and background; 2) registration of the lung with geometric transformations; 3) separation of the lung into regions; 4) feature extraction; 5) scoring of the regions; and 6) generation of explanations for the radiologists.

Subtasks Might Not Need AI: When creating a stable and trustworthy production environment it might be more beneficial to use deterministic computer vision approaches for image registration and segmentation over state-of-the-art neural networks to increase transparency and reliability of these steps, and make the link between input image and output severity score clearer.

No Detailed Evaluation of Existing Techniques: The system uses a custom LIME-like explanatory technique based on superpixels that highlight which areas contribute significantly to the final score. This allows the radiologists to view parts of the image which are especially important for the classification in a very localized way. While the developers claim that the explanations generated by existing explanation techniques like GradCAM are not localized for the needs of the radiologists, the limitations of the current method are not discussed. As a deviation of LIME, it likely suffers from similar issues such as confusion from high variability of explanations [24], [25] and the superpixel-related tradeoff between fidelity and consistency on the one hand and comprehensibility on the other [26].

In conclusion, we have listed what the WG considers to be the major relevant technical issues, representing the consensus reached between the members of the technical WG. Many more technical issues were identified but not listed for being irrelevant, not probable or not fitting the scope of the analysis. As an example, they consider various missed UI/UX features which can improve the readiness and helpfulness of the system, like displaying confidence values or interaction with the explanations. Another potential tension is linked with the data privacy and proper anonymization of the data to obey GDPR, which is a mostly legal issue.

4) Some Thoughts on the Influence of Pandemic-Related Factors: Specific to this case study is that the AI system was developed during a pandemic. This comes with certain unique characteristics and implications. The system is about one disease, it is specific, relates to the immediate situation, comprises experimental features and is tailored to a particular local context. None of this can be avoided during (the early phase of) a pandemic.

Building such a system requires good and effective collaboration between AI investigators that develop the system, clinical investigators that assist in the development, supporting companies, and hospital IT services. It seems like the stronger interdisciplinary collaboration and understanding between different players and groups during "normal" times, the earlier the hospital/healthcare unit is able to react and develop an AI system applicable to the pandemic situation. Patients should also be willing to consent to their data being used, and radiologists and clinicians must be willing to actually use the system.

It is also important to streamline radiologists' activities when hospitals are overloaded and overwhelmed, to gain time and reduce stress. Supporting the radiologists is crucial in the acute phase of the pandemic; such AI systems are developed to provide a second opinion and cement radiologists' assessments. When the radiologist uses the software, they see the assessment provided by the system. Thus, the radiologist in practice primarily confirms the system's assessment/scoring, or in case of deviations adds comments or modifies it.

In principle, the system could run even without a radiologist, at least for a limited period of time. In an overstrained hospital situation, this may seem more acceptable than during normal times but is certainly not unproblematic and would still require a certain level of oversight. During nonpandemic times, there are equally legitimate concerns over potentially replacing medical professionals/radiologists and deskilling risks that may in the long run require policy interventions.

Ultimately, developers should also seek to develop a less subjective and more homogenous assessment algorithm, so that the quality of the diagnoses depends less on the experience, fatigue and stress-related situation of the radiologists on duty. This may prevent that major responsibilities lie on a few experienced professionals already overwhelmed by fatigue and overwork; however, in the long run, training would be expected so that all radiologists reach a similar level of performance. The system could potentially be a valuable tool used in: 1) retrospective studies or 2) after verifications, in post-COVID-19 patient follow-up.

Pandemics present a state of emergency where usual considerations, timelines, and resources will not necessarily be adequate. Ethics committees, developers and users of AI systems will therefore inevitably face tradeoffs and will need to ask themselves certain questions. For example:

1) in light of limited time, what is an adequate procedure for ethics committee approval in a pandemic? Could it be justified to temporarily lower standards in order to speed up the process?
2) how can patient rights be adequately secured in this situation? There may be situations where a system's attributes makes it difficult to transfer to other contexts, for example, if a system is mostly applicable to a local population (given age, ethnicity, etc.); yet the system may still prove necessary and useful locally despite such shortcomings. Having said that, what can be done to increase the diversity of the dataset, and the accuracy of the system as a whole? What can be done to make the system more applicable to other hospitals and other regions? Failing to address this might mean that several similar systems may be developed in parallel, leading to redundancies and a suboptimal use of resources;
3) data protection regulation (e.g., GDPR) is complex, and legal regulations may differ in different countries. How can one obtain ethically and legally valid informed consent during the pandemic? What is the most effective way to approach patients? What information can and cannot be conveyed? Are the datasets used anonymized or only pseudonymized?

To answer these questions usefully, it is crucial to consider burdens, costs, and (potential) benefits. Again, this highlights the importance of clear communication lines and cross-disciplinary teams. The role of ethics committees and the scope of their work needs to be reviewed to ensure an adequate balance between the provision of care and the consideration of potential risks. For example, ethics committees might consider an accelerated process to issue waivers when urgency dictates so.

C. Phase III: Mapping to the Framework for Trustworthy AI

After each WG completed their reports, we started with the mapping phase. The goal of the mapping is to identify the issues identified in each WG report and map them to the EU's requirements for trustworthy AI.

Following the EU guidelines, we mapped to three levels: 1) four *ethical pillars*; 2) seven *key requirements*; and 3) multiple *sub requirements*. The result is a focused list of issues, with each issue referring to one problem identified in the report. Using so-called rubrics [5], [8] each issue was then mapped to the corresponding pillars, requirement(s) and sub requirement(s).

This helped nonethicists as part of the team to understand how problems can impact the trustworthiness of the AI system by providing them with a list of identified issues. It also helps to highlight different perspectives and implications for each of the problems.

1) Example of Mapping Strategy: In one working group (WG Ethics and Healthcare), the mapping of issues identified in the WG report was organized using the following process: at the initial meeting, they made a list of the key issues that they found to be present in the WG report. The list merely stated key words, no description of the issues. They then divided the issues between them and each member of the group made a description of her selection of the issues. The descriptions formed the basis of another meeting at which they initiated the mapping of issues to ethical pillars, requirements and subrequirements. At the second meeting, they discussed the mapping of a couple of the issues identified. This involved quite a bit of clarification and discussion of their understanding of the pillars and requirements. Moreover, the discussion of what was covered by the pillars and requirements shaped and structured the way they understood the issues. At the meeting, they did not get around to mapping all the issues to the pillars, requirements, and subrequirements. Instead, they decided that they would each map the issues they had described and then meet and discuss these suggested mappings. They did this at the third meeting. At this point, they seemed to have reached a common understanding of the pillars and requirements as well as of the issues described.

2) Challenges of Mapping: The difficulty is that it is often not obvious which of the pillars or requirements applies, in many cases, multiple pillars or requirements can apply or a decision is made which one is the most applicable. The WG team found that the mapping of an issue is often debatable and strongly depends on the background of the person performing the mapping. Disagreements regarding the mappings within the groups were resolved by group consensus.

Across the different WGs, the whole team identified a large number of issues (over 50) which need further consolidation.

D. Phase IV: Consolidation of the Mappings

At this point, we created a special team of so-called "mappers" (i.e., seven experts from the various WGs), whose task was to consolidate the various mappings produced by the WGs into a consistent list.

1) Strategies: Due to the large number of identified issues, the consolidation was performed in two steps. First, issues mapped to the same key requirement of the EU framework were grouped together to identify and combine related issues from similar groups. Then, the consolidated lists of WG issues for each of the seven requirements were reviewed so commonalities and differences could be identified and discussed before final consolidation. This helped us find and combine similar issues mapped to different key requirements, which is possible due to the subjective mapping performed by the groups. We found key requirements to be the right level of granularity for the mapping process, with a focus on ethical principles the mapping is too coarse, when focusing on subrequirements, the multitude of options makes the mapping too difficult.

2) Challenges: A central problem was how to handle the ambiguity of the mapping from issue to key requirement. We observed that the different groups frequently mapped issues to different key requirements which made the first step of our mapping less effective as planned. In the second step, however, we found similar issues identified by different groups and mapped to different key requirements. To us, this showed that while we agreed on the issues, the different backgrounds provided different perspectives on the underlying problem and its implications. Similar to the previous step, if an issue was found to be mapped to different requirements, we tried to find a consensus within the group which of them were most applicable, while also accepting that different points of view could lead to different mappings (i.e., an issue being mapped to more than one requirement).

3) Findings: In the following, we present a selection of five key issues that were identified across different groups and perceived to be the most important issues for the system at hand, along with their mapping.

Issue 1: Clinical benefit of the system is not sufficiently proven.

Description: The AI system's clinical benefit and absence of clinical harm have not been proven since a clinical trial is missing. A clinical trial by comparing the performances of the unsupervised system, a resident radiologist and the two combined would settle the issue. The clinical benefits are proven at least when the radiologist and the AI combined outperform the AI alone. Similarly, the absence of harm is easily proven showing that the radiologist underperforms compared to the other two scenarios (system and system plus radiologist).

Consolidated Rubric
Identified by two WGs: healthcare, and healthcare & ethics
Ethical Principles: *Prevention of harm*
Trustworthy AI Key requirements: *Technical robustness and safety*.

Issue 2: Concerns about protection of patients' data.

Description: Patients and developers have different interests regarding the collection, control, and use of personal and sensitive data. Getting informed consent from hospitalized COVID-19 patients proved difficult, therefore, the ethics committee and regional government passed a waiver softening data protection requirements for the developers. As a data management plan is missing, it is not clear if/when this waiver was retracted. Furthermore, it is not clear if the patient's data are anonymized or only pseudonymized. This has implications, as per GDPR, anonymized data can be used without explicit patient consent, but pseudonymized data cannot.

Consolidated Rubric
Identified by four WGs: social, healthcare & ethics, ethics, technical, legal
Ethical Principles: *Prevention of harm*, *Explicability*
Trustworthy AI Key requirements: *Privacy and data governance, Transparency*.

Issue 3: System lacks transparency.

Description: The scoring function represents a momentary situation of the patient's lung condition influenced by medical or technical conditions, and varying image quality. Importantly, the score does not describe COVID-19 specifically, but only the degree of lung damage, without considering the patient history. It is not clear if the patient is informed that an AI system provides decision support for the diagnostic process and whether doctors and patients are informed if out-of-distribution patient data has been used for inference.

Consolidated Rubric
Identified by three WGs: radiologists, healthcare, technical
Ethical Principles: *Prevention of harm, Explicability*
Trustworthy AI Key requirements: *Technical robustness and safety, Transparency*.

Issue 4: The AI system might bias the radiologists.

Description: In the current workflow, the radiologists see the score and explainability without analyzing the CXR image alone first. Hence, the radiologist may fall victim to a "priming" or "anchoring" effect of the suggested scores. Such an effect has been proven to influence human behavior and numeric judgement [27].

The developers are currently investigating whether the radiologists blindly confirm with the tool or if they use it as a helpful second opinion.

Consolidated Rubric
Identified by five WGs: radiologists, technical, ethics, social, legal
Ethical Principles: *Respect for human autonomy, Fairness*
Trustworthy AI Key requirements: *Human agency and oversight, Accountability*.

Issue 5: Dataset small and not representative.

Description: The dataset used for training is likely not representative of the general population it is currently used on. Limited geographic origins, past medical history, gender, and age also limit the system's applicability in other regions/hospitals. Furthermore, the dataset contains only 5000 images, which is likely not enough to cover a wide enough range of possible lung damages. In addition, the vast majority of the images in the dataset is based on only three of the hospital's nine types of X-ray machines.

Consolidated Rubric
Identified by three WGs: healthcare, technical, ethics
Ethical Principles: *Fairness, Prevention of harm*
Trustworthy AI Key requirements: *Diversity, nondiscrimination and fairness, Technical robustness and safety*.

4) Recommendations: The following are some of the key recommendations we offer to the main stakeholders of this use case.

1) There is a need of a large dataset with diverse, high-quality images curated from multiple institutions and different geographic areas in order to claim and ensure the generalizability of the AI system intended for clinical usage.
2) A feedback mechanism should be put in place so that the radiologist reviews the system's output after reporting in order to not be biased by the results.
3) It will be important to form or contact a panel of patients' representatives, in order to collect, identify, register, understand—and hopefully respond in a satisfactory fashion to—their views, requirements, expectations, and concerns.
4) A study on how the AI tool is incorporated into clinical decision making should be conducted and results of such study should be shared to all involved stakeholders and patient representatives.
5) A detailed risk management plan and governance structure would need to be in place if the AI system were to be expanded or scaled up.
6) Policies on how to secure informed consent and to protect patient rights should be put in place prior to developing systems for collecting data early on and building a database to be used later.
7) Provide a test branch and service with the public repository that allows external parties to test the model directly with test data.
8) Make sure that an external audit tests the model publicly available. The auditor will need to certify certain ethical and healthcare standards.
9) It cannot be assumed that the system and the radiologist are more beneficial and less harmful than the radiologist alone. Subject the system to a trial comparing the system alone, a radiologist alone, and the system and the radiologist in order to show that, as a minimum, the system and the radiologist together perform better than the radiologist alone. (This is especially important when it is proposed that one benefit of the system is for overworked and tired radiologists.)

10) We recommend to conclude and publish the results of the clinical trials currently on going.

5) Small Versus Large Team of Experts: The Z-Inspection® process has been developed (starting January 2019 and published in June 2021) by a core team of experts working closely together with interdisciplinary experts. During the development of Z-Inspection® and working on different use cases, we found, on the one hand, that each use case required different expertise and, on the other hand, we experienced that because of the novelty of the research area that a larger group of experts added new and important perspectives to the development and especially the testing and refinement of the Z-Inspection® process itself.

The size of the team is correlated to the complexity of the AI assessment. We have been working on different use cases with different team sizes. In this use case, we had a large team including over 50 interdisciplinary experts and we had to split the work in parallel WGs. In other use cases, we did not have to split the work in parallel WGs since we had a midsize team including around 20 interdisciplinary experts. There are pros and cons for this decision. If the team is too small and it does not reflect the true interdisciplinary nature of the assessment work, the assessment work might be incomplete. If the team is too big with too much overlap of knowledge and expertise, the assessment process may become cumbersome and delayed.

The current use case drew significant interest in the Z-Inspection® initiative and benefitted from a wide group of experts and expertise. This allowed rich exchanges and a truly interdisciplinary approach. In a research context, nothing speaks against such a large group of experts, however, Z-Inspection® does not require—by design—such a large group of experts. Thus, when considering commercial real-life deployments, even under crisis pressures, a much leaner consortium could be assembled that could produce results quickly.

During the development process of Z-Inspection®, the need for trained and qualified interdisciplinary Z-Inspection® experts became obvious. We therefore created a number of so-called affiliated Trustworthy AI Labs based on the Z-Inspection® process, and we have implemented since 2021 a qualification training. The vision is that in the future this will increase the efficiency of the assessment process, improve the overall quality of the assessment and create a network of qualified experts.

6) Shortcomings: We recognize that our *post-hoc* self-assessment for this use case has some limitations, namely:

1) the assessment team of experts has only limited knowledge of the AI system, of how the situation developed at the hospital, of the ethics committee's decision-making and decision-making process;
2) it has not been investigated whether and how the AI system actually influenced the radiologists routine and decision making;
3) both the mappings and the consolidation of the mappings involve subjective decision-making components;
4) the mapping process relies on the European guidelines. On the one hand, this clearly shapes the process, as it provides a framework for assessment. On the other hand, as the guidelines stress certain ethical concepts and principles in their pillars and requirements, the whole assessment tends to stress those ethical concepts and principles. This may bear the risk of disregarding other ethical concepts and principles relevant in the context of the use case;
5) overall, the *post-hoc* assessment is shaped and also limited by the team members' focus of work and expertise. While a very large interdisciplinary team worked on the use case, it is not possible to ensure that every perspective was covered or was equally covered;
6) the patient perspective is missing, especially the perspective of those patients who underwent COVID-19 treatment while the AI system was used.
7) this version of the assessment does not address legal aspects.

IV. Related Work

A. Ethics-Based Auditing

The Z-Inspection® process can itself be assessed according to what has been called an ethics-based auditing (EBA) [35], which is not "a kind of auditing conducted ethically, nor [...] the ethical use of ADMS [Automated Decision-Making Systems] in auditing, but [...] an auditing process that assesses ADMS based on their adherence to predefined ethics principles [...]. EBA shifts the focus of the discussion from the abstract to the operational, and from guiding principles to managerial intervention throughout the product life cycle, thereby permeating the conceptualization, design, deployment and use of ADMS" [36].

Although there are a large variety of tools that were designed in order to help the governance mechanisms of AI systems, we believe that Z-Inspection® is particularly helpful since it is able to combine the three main components of auditing processes: 1) functionality auditing; 2) code auditing; and 3) impact auditing [36].

The Z-Inspection® is designed to allow us: 1) to assess the design process that led to the conception of the AI system itself (usually with some representatives of the organization that created the system); 2) to address issues related to the AI system's source code as well as issues related to the training of the algorithm, especially the dataset that was used for the training; and 3) to consider the different impacts the AI system might have on users, patients, and society, in general.

The interdisciplinary approach and the specific procedure of the Z-Inspection® make it possible to raise all relevant questions (related to the three components) in a coherent and unified process.

We must understand the Z-Inspection® as an on-going process to highlight potential ethical issues, rather than a procedure designed to provide a final answer regarding the ethical worth of an AI system. This is the first requirement of EBA identified in [35]: the Z-Inspection® is: 1) continuous.

The four other requirements are also met by Z-Inspection®, it is: 2) holistic since the AI system is understood in its connection with other tools but also with institutions and social processes; 3) dialectic, because it is based on a dialogue between a wide range of interdisciplinary stakeholders; 4) strategic, the focus of the discussion being aimed at action regarding the use of the AI system; and 5) design driven, since

it is designed to provide insights to the developers themselves, in order to improve the AI system.

We strongly believe that the Z-Inspection® can thus be considered as an ≪EBA [...] a governance mechanism that helps organizations not only to ensure but also demonstrate that their ADMS adhere to specific ethics principles≫ [36], more precisely the ones provided by the EU Ethics Guidelines for Trustworthy AI.

B. Standardization of Trustworthy AI

The Z-Inspection® process is a formalized and principled approach for evaluating the design, deployment, and use of AI-based systems toward, aimed at ensuring that the final system iteration is both trustworthy and trusted. It is positioned within the broader trend to design and assure trustworthy AI systems. It can be used at various stages of the AI development and maintenance process. First, in the design phase, the Z-Inspection® methodology can be utilized as a co-creation process to ensure an AI system meets the trustworthy AI criteria. Both before and after AI deployment, Z-Inspection® can be used as a validation process to assess the trustworthiness of the AI system being developed. Additionally, it can form part of an AI certification, audit or monitoring process. The latter can be considered a part of "ethical maintenance" for trustworthy AI.

Among recent attempts to devise AI auditing frameworks, the IEEE Standards Association (IEEE SA) has developed and released the IEEE CertifAIEd Mark [37], [38]. The mark aims to be a recognizable signal for the trustworthiness of an AI system. The IEEE SA describes the CertifAIEd mark as follows:

> The IEEE CertifAIEd mark recognizes that your product, service, or system has been verified to meet relevant ethical criteria, contributing toward a greater level of confidence and demonstrating a proactive approach to building public trust in your AI system. It sets the standard that AI products, services and systems should meet in order to deliver authentic and practical value and trust [37].

For example, the Z-Inspection® process could be used to verify if a system is compliant with the IEEE CertifAIEd mark. As this article has demonstrated with the particular use case, the Z-Inspection® process helps organizations designing AI-based systems to operationalize and implement the design values specified in the European Union's High-Level Expert Group (EU HLEG) guidelines. Among such values are transparency, accountability, reduction of algorithmic bias, and preservation of privacy and data protection. These values are all reflected in the Z-Inspection® process, meaning that compliance and adherence would also satisfy the requirements for receiving the IEEE CertifAIEd mark.

V. Conclusion

In this article, we showed how to assess Trustworthy AI in practice in times of pandemic.

In particular, we have assessed a deep-learning-based solution deployed at the public hospital in the city of Brescia.

This work is, to the best of our knowledge, one of the most comprehensive Trustworthy AI assessments in times of pandemic based on the EU guidelines.

In our assessment, we did not have direct involvement of affected people. This is a limitation. In the future, we plan to amend this by considering these two viewpoints: 1) evaluating when to involve patient representatives and family members/informal caregivers in the assessment to enhance the stakeholders view to be considered versus 2) evaluating when involving more perspectives makes the process more cumbersome. This depends case by case on the definition of affected people for the specific use case.

Disclaimer

The views expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations.

Author Contribution

Authors are listed alphabetically. Conception/design–all authors; Coordinators of the Working Groups: WG technical: AFM, PAMS, DV, AFS, HA, LEL, SM, AG, JJT, MW, AS; WG ethics: EG, FB; WG healthcare ethics: EH, SH; WG radiologists: EN, TFG; healthcare: MC, RW; WG law: IB, NC; WG social: EWM, IB; WG mapping: AFM, EH, DV, EH, EWM, GFC, IB, IVH, JJT, LEL, OM, SH, S, RVZ (who were responsible for producing the consolidated mappings of Ethical Issues and Tensions); WG Lead: DV, RVZ; JB co-conceived ethics template and contributed the method of combining diverse ethical viewpoints; final approval–all authors.

Acknowledgment

The authors would like to thank Elia Belussi, Matthias Braun, Helga Brøgger, Andrew Bushell, Marcelo Corrales Compagnucci, Boris Düdder, Mads Kjolby, Maria Forss, Fosca Giannotti, Thomas Gilbert, David Higgins, Asiatu Agnes Jalloh, Ahmed Khali, Federica Lucivero, Oriana Medlicott, Timo Minssen, Belona Sonna, Leonoor Tideman, and Karsten Tolle who actively participated in Zoom meetings and/or offered valuable comments. Brendan Parent is the principal investigator on a Robert Wood Johnson Foundation grant to study the ethics of big data for AI applications in health care.

Himanshi Allahabadi is with the Enterprise Intelligence Department, EY Netherlands, 1083 HP Amsterdam, The Netherlands.

Julia Amann is with the Health Ethics and Policy Lab, Department of Health Sciences and Technology, ETH Zurich, 8092 Zürich, Switzerland.

Isabelle Balot is with the Postgraduate Studies in Diplomacy and International Relations, Center for Diplomatic & Strategic Studies, 75015 Paris, France.

Andrea Beretta and Francesca Pratesi are with the Institute of Information Science and Technologies, National Research Council of Italy (CNR), 56124 Pisa, Italy.

Charles Binkley is with the Bioethics Center, Hackensack Meridian Health, Edison, NJ 08820 USA.

Jonas Bozenhard is with the Faculty of Philosophy, University of Oxford, Oxford OX2 6GG, U.K.

Frédérick Bruneault is with the Philosophie Department, Collège André-Laurendeau, Montreal, QC H8N 2J4, Canada, and also with the École des médias, Université du Québec à Montréal, Montreal, QC H2L 2C4, Canada.

James Brusseau is with the Philosophy Department, Pace University, New York, NY 10038 USA.

Sema Candemir is with the Wexner Medical Center, Department of Radiology, The Ohio State University, Columbus, OH 43210 USA.

Luca Alessandro Cappellini is with the Department of Radiology, Humanitas Research Hospital, 20089 Milan, Italy, and also with the Department of Biomedical Sciences, Humanitas University, 20089 Milan, Italy.

Subrata Chakraborty is with the Faculty of Science, Agriculture, Business and Law, University of New England, Armidale, NSW 2351, Australia, and also with the Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, NSW 2007, Australia.

Nicoleta Cherciu is with the European Centre of Excellence on the Regulation of Robotics & AI, Scuola Superiore Sant'Anna, 56127 Pisa, Italy.

Christina Cociancig is with the Group of Computer Architecture, University of Bremen, 28359 Bremen, Germany.

Megan Coffee is with the Department of Medicine, Division of Infectious Diseases and Immunology, New York University Grossman School of Medicine, New York, NY 10016 USA.

Irene Ek is with the AI Research Section, Digital Institute, 16731 Stockholm, Sweden.

Leonardo Espinosa-Leal is with the Department of Business Management and Analytics, Arcada University of Applied Sciences, 00550 Helsinki, Finland.

Davide Farina and Filippo Vaccher are with the Department of Medical and Surgical Specialties, Radiological Sciences, and Public Health, University of Brescia, 25121 Brescia, Italy.

Geneviève Fieux-Castagnet is with the Ethique Groupe, SNCF Reseau SA, 93418 La Plaine, France.

Thomas Frauenfelder is with the Institute of Diagnostic and Interventional Radiology, University Hospital Zurich, 8091 Zürich, Switzerland.

Alessio Gallucci is with the Department of Mathematics and Computer Science, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands.

Guya Giuliani is with the Data Privacy Advisor, Ericsson, 11331 Stockholm, Sweden.

Adam Golda is with the Department of Cardiology, 4th Gliwice Municipal Hospital, 44-100 Gliwice, Poland.

Irmhild van Halem and Emilie Wiinblad Mathez are with the Z-Inspection® Initiative, Frankfurt, Germany.

Elisabeth Hildt is with the Center for the Study of Ethics in the Professions, Illinois Institute of Technology, Chicago, IL 60616 USA.

Sune Holm is with the Department of Food and Resource Economics, University of Copenhagen, 1165 Copenhagen, Denmark.

Georgios Kararigas is with the Department of Physiology, Faculty of Medicine, University of Iceland, 101 Reykjavik, Iceland.

Sébastien A. Krier is with the Cyber Policy Center, Stanford University, Stanford, CA 94305 USA.

Ulrich Kühne is with the Department for Dermatology, Hautmedizin Bad Soden, 65812 Bad Soden, Germany.

Francesca Lizzi is with the Data Science Department, Scuola Normale Superiore, 56126 Pisa, Italy.

Vince I. Madai is with the Charité Lab for Artificial Intelligence in Medicine, Department of Neurosurgery, the QUEST Centre for Responsible Research, Berlin Institute of Health at Charité, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany, and also with the Faculty of Computing, Engineering and the Built Environment, School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, U.K.

Aniek F. Markus is with the Department of Medical Informatics, Erasmus University Medical Center, 3015 GD Rotterdam, The Netherlands.

Serg Masis is with the CADS Department, Syngenta, Research Triangle Park, NC 27709 USA.

Francesco Mureddu is with the Policy Research, The Lisbon Council, 1040 Brussels, Belgium.

Emanuele Neri is with the Department of Translational Research, Academic Radiology, University of Pisa, 56126 Pisa, Italy.

Walter Osika is with the Center for Psychiatry Research, Department of Clinical Neuroscience, Karolinska Institutet, 171 77 Stockholm, Sweden.

Matiss Ozols is with the Department of Medicine, The University of Cambridge, Addenbrooke's Hospital, Cambridge CB2 0QQ, U.K., also with the School of Cell Matrix and Regenerative Medicine, The University of Manchester, Manchester M13 9PG, U.K., and also with the Department of Human Genetics, The Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge CB10 1RQ, U.K.

Cecilia Panigutti is with the Department of Computer Science, University of Pisa, 56127 Pisa, Italy.

Brendan Parent is with the Division of Medical Ethics, Department of Population Health, NYU Grossman School of Medicine, New York, NY 10016 USA.

Pedro A. Moreno-Sánchez is with the School of Healthcare and Social Work, Seinäjoki University of Applied Sciences, 60100 Seinäjoki, Finland.

Giovanni Sartor is with the Law Department, European University Institute, 50139 Firenze, Italy, and also with the CIRSFID-Alma AI and Law Department, University of Bologna, 40121 Bologna, Italy.

Mattia Savardi and Alberto Signoroni are with the Department of Information Engineering, University of Brescia, 25121 Brescia, Italy.

Hanna-Maria Sormunen is with the Advanced Analytics, Finnish Tax Administration, 00510 Helsinki, Finland.

Andy Spezzatti is with the Artificial Intelligence Research, AI for Good Foundation, El Cerrito, CA 94530 USA, and also with the Industrial Engineering and Operation Research Department, University of California at Berkeley, Berkeley, CA 94720 USA.

Adarsh Srivastava is with the Data Services Function in Roche Diagnostics, Roche, Pune 411005, India.

Annette F. Stephansen is with the Digital Systems, NORCE Norwegian Research Centre AS, 5008 Bergen, Norway.

Lau Bee Theng is with the School of Research, Swinburne University of Technology (Sarawak), Kuching 93350, Malaysia.

Jesmin Jahan Tithi is with the Parallel Computing Labs, Intel, Santa Clara, CA 95054 USA, and also with the Department of Computer Science, Stony Brook University, Stony Brook, NY 11794 USA.

Jarno Tuominen is with the Department of Psychology and Speech-Language Pathology, University of Turku, 20500 Turku, Finland.

Steven Umbrello is with the Department of Values, Technology and Innovation, Delft University of Technology, 2628 BX Delft, The Netherlands.

Dennis Vetter is with the Computational Vision and Artificial Intelligence Lab, Goethe University Frankfurt, 60325 Frankfurt, Germany.

Magnus Westerlund is with the Department of Business Management and Analytics, Arcada University of Applied Sciences, 00550 Helsinki, Finland, and also with the School of Economics, Innovation and Technology, Kristiania University College, 0107 Olso, Norway.

Renee Wurth is with the T. H. Chan School of Public Health, Harvard University, Boston, MA 02115 USA.

Roberto V. Zicari is with the Department of Business Management and Analytics, Arcada University of Applied Sciences, 00550 Helsinki, Finland, and also with the Data Science Graduate School, Seoul National University, Seoul 08826, South Korea (e-mail: roberto@zicari.de).

REFERENCES

[1] A. Borghesi *et al.*, "Chest X-ray severity index as a predictor of in-hospital mortality in coronavirus disease 2019: A study of 302 patients from Italy," *Int. J. Infect. Dis.*, vol. 96, pp. 291–293, Jul. 2020, doi: 10.1016/j.ijid.2020.05.021.

[2] A. Signoroni *et al.*, "BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset," *Med. Image Anal.*, vol. 71, Jul. 2021, Art. no. 102046, doi: 10.1016/j.media.2021.102046.

[3] (AI HLEG) High-Level Expert Group on Artificial Intelligence. "Ethics Guidelines for Trustworthy AI." European Commission. Apr. 2019. [Online]. Available: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1 (Accessed: Oct. 26, 2020).

[4] (Eur. Commission, Brussels, Belgium). *Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. (Apr. 2021). Accessed: Nov. 30, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206

[5] R. V. Zicari *et al.*, "Z-inspection®: A process to assess trustworthy AI," *IEEE Trans. Technol. Soc.*, vol. 2, no. 2, pp. 83–97, Jun. 2021, doi: 10.1109/TTS.2021.3066209.

[6] R. V. Zicari *et al.*, "On assessing trustworthy AI in healthcare. Machine learning as a supportive tool to recognize cardiac arrest in emergency calls," *Front. Human Dyn.*, vol. 3, p. 30, Jul. 2021, doi: 10.3389/fhumd.2021.673104.

[7] R. V. Zicari *et al.*, "Co-design of a trustworthy AI system in healthcare: Deep learning based skin lesion classifier," *Front. Human Dyn.*, vol. 3, p. 40, Jul. 2021, doi: 10.3389/fhumd.2021.688152.

[8] J. Brusseau, "What a philosopher learned at an AI ethics evaluation," *AI Ethics J.*, vol. 1, no. 1, p. 4, Dec. 2020, doi: 10.47289/AIEJ20201214.

[9] B. Düdder, F. Möslein, N. Stürtz, M. Westerlund, and R. V. Zicari, "Ethical maintenance of artificial intelligence systems," in *Artificial Intelligence for Sustainable Value Creation*, M. Pagani and R. Champion, Eds. Cheltenham, U.K.: Edward Elgar Publ., 2020.

[10] J. Moubray, *Reliability-Centered Maintenance*, 2nd ed. New York, NY, USA: Ind. Press, 2000.
[11] D. Sayers *et al.* "The Dawn of the Human-Machine Era: A Forecast of New and Emerging Language Technologies." 2021. doi: 10.17011/jyx/reports/20210518/1.
[12] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11, doi: 10.1007/978-3-030-00889-5_1.
[13] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," Dec. 2015, *arXiv:1512.00567*.
[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.
[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, New York, NY, USA, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.
[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017. pp. 618–626. Accessed: Nov. 18, 2021. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html
[19] S. H. Park and K. Han, "Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction," *Radiology*, vol. 286, no. 3, pp. 800–809, Mar. 2018, doi: 10.1148/radiol.2017171920.
[20] J. R. Geis *et al.*, "Ethics of artificial intelligence in radiology: Summary of the joint European and North American multisociety statement," *Radiology*, vol. 293, no. 2, pp. 436–440, Nov. 2019, doi: 10.1148/radiol.2019191586.
[21] R. M. Summers, "Artificial intelligence of COVID-19 imaging: A hammer in search of a nail," *Radiology*, vol. 298, no. 3, pp. E162–E164, Mar. 2021, doi: 10.1148/radiol.2020204226.
[22] M. A. Lim, R. Pranata, I. Huang, E. Yonas, A. Y. Soeroto, and R. Supriyadi, "Multiorgan failure with emphasis on acute kidney injury and severity of COVID-19: Systematic review and meta-analysis," *Can. J. Kidney Health Dis.*, vol. 7, Jan. 2020, Art. no. 2054358120938573, doi: 10.1177/2054358120938573.
[23] V. G. Puelles *et al.*, "Multiorgan and renal tropism of SARS-CoV-2," *New Engl. J. Med.*, vol. 383, no. 6, pp. 590–592, Aug. 2020, doi: 10.1056/NEJMc2011400.
[24] D. Alvarez-Melis and T. S. Jaakkola, "On the robustness of interpretability methods," Jun. 2018, *arXiv:1806.08049*.
[25] M. R. Zafar and N. M. Khan, "DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems," Jun. 2019, *arXiv:1906.10263*.
[26] M. Robnik-Šikonja and M. Bohanec, "Perturbation-based explanations of prediction models," in *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, J. Zhou and F. Chen, Eds. Cham, Switzerland: Springer Int., 2018, pp. 159–175, doi: 10.1007/978-3-319-90403-0_9.
[27] B. R. Newell and D. R. Shanks, "Prime numbers: Anchoring and its implications for theories of behavior priming," *Soc. Cogn.*, vol. 32, pp. 88–108, Jun. 2014, doi: 10.1521/soco.2014.32.supp.88.
[28] H. Else. "COVID 'Fast Grants' Sped Up Pandemic Science." Nature. Aug. 2021. doi: 10.1038/d41586-021-02111-7.
[29] "The Territorial Impact of COVID-19: Managing the Crisis across Levels of Government—OECD." OECD. Nov. 2020. [Online]. Available: https://read.oecd-ilibrary.org/view/?ref=128_128287-5agkkojaaa&title=The-territorial-impact-of-covid-19-managing-the-crisis-across-levels-of-government (Accessed: Dec. 6, 2021).
[30] G. P. Pisano, R. Sadun, and M. Zanini. "Lessons From Italy's Response to Coronavirus." Harvard Business Review. Mar. 2020. [Online]. Available: https://hbr.org/2020/03/lessons-from-italys-response-to-coronavirus (Accessed: Dec. 6, 2021).
[31] A. Lal, H. C. Ashworth, S. Dada, L. Hoemeke, and E. Tambo, "Optimizing pandemic preparedness and response through health information systems: Lessons learned from Ebola to COVID-19," *Dis. Med. Public Health Preparedness*, vol. 16, no. 1, pp. 333–340, Feb. 2022, doi: 10.1017/dmp.2020.361.
[32] W. Naudé (Soc. Sci. Res. Netw., Rochester, NY, USA). *Artificial Intelligence Against Covid-19: An Early Review*. (Apr. 2020). doi: 10.2139/ssrn.3568314.
[33] W. Naudé, "Artificial intelligence vs COVID-19: Limitations, constraints and pitfalls," *AI Soc.*, vol. 35, pp. 761–765, Apr. 2020, doi: 10.1007/s00146-020-00978-0.
[34] E. Mbunge, B. Akinnuwesi, S. G. Fashoto, A. S. Metfula, and P. Mashwama, "A critical review of emerging technologies for tackling COVID-19 pandemic," *Hum. Behav. Emerg. Technol.*, vol. 3, no. 1, pp. 25–39, 2021, doi: 10.1002/hbe2.237.
[35] J. Mökander and L. Floridi, "Ethics-based auditing to develop trustworthy AI," *Minds Mach.*, vol. 31, no. 2, pp. 323–327, Jun. 2021, doi: 10.1007/s11023-021-09557-8.
[36] J. Mökander, J. Morley, M. Taddeo, and L. Floridi, "Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations," *Sci. Eng. Ethics*, vol. 27, no. 4, p. 44, Aug. 2021, doi: 10.1007/s11948-021-00319-4.
[37] "IEEE CertifAIEd—The Mark of AI Ethics," IEEE SA—The IEEE Standards Association. [Online]. Available: https://engagestandards.ieee.org/ieeecertifaied.html (Accessed: Nov. 23, 2021).
[38] "The Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS)." IEEE SA—The IEEE Standards Association. [Online]. Available: https://standards.ieee.org/industry-connections/ecpais.html (Accessed: Nov. 23, 2021).