Accountable AI for Healthcare IoT Systems

Prachi Bagave, Marcus Westberg, Roel Dobbe, Marijn Janssen, and Aaron Yi Ding*
Department of Engineering Systems and Services
Delft University of Technology

***Abstract*—Various AI systems have taken a unique space in our daily lives, helping us in decision-making in critical as well as non-critical scenarios. Although these systems are widely adopted across different sectors, they have not been used to their full potential in critical domains such as the healthcare sector enabled by the Internet of Things (IoT). One of the important hindering factors for adoption is the implication for accountability of decisions and outcomes affected by an AI system, where the term accountability is understood as a means to ensure the performance of a system. However, this term is often interpreted differently in various sectors. Since the EU GDPR regulations and the US congress have emphasised the importance of enabling accountability in AI systems, there is a strong demand to understand and conceptualise this term. It is crucial to address various aspects integrated in accountability and understand how it affects the adoption of the AI systems. In this paper, we conceptualise these factors affecting accountability and how it contributes to a trustworthy healthcare AI system. By focusing on healthcare IoT systems, our conceptual mapping will help the readers understand what system aspects those factors are contributing to and how they affect the system trustworthiness. Besides illustrating accountability in detail, we also share our vision towards causal interpretability as a means to enhance accountability for healthcare AI systems. The insights of this paper shall contribute to the knowledge of academic research on accountability, and benefit AI developers and practitioners in the healthcare sector.**

***Index Terms*—accountability, trustworthiness, healthcare AI, healthcare internet of things (IoT)**

I. Introduction

With the advent of AI, many industries are getting smarter by monitoring the important factors and predicting events before they may arise. This has been made possible by the complex data modelling techniques like machine learning and deep learning which produce precise parameterized probabilistic models based on the historic data [1]. AI has enabled sensing of various activities, collecting their data, making complex data models and using them for predicting events for better management of the system. Such monitoring systems have been implemented in various sectors of our society like autonomous driving [2], judicial systems [3], recommendation systems [4], and finance [5]. Thus, AI is now impacting various aspects of our lives and has become an inseparable element of our society. This has also called for the development of these systems from a more socio-technical approach.

Healthcare, one of society's critical industries, has also been transformed by AI. This transformation has led to the creation of huge electronic medical data-sets at hospitals to enable data modelling [6]. Some of the application areas in this industry are assistance in diagnosis and treatment, disease prevention, drug research and healthcare management [7]. In addition to hospital care, remote healthcare monitoring at home and social welfare centres has also enabled monitoring of patients post hospitalization. This has proven helpful especially for chronic patients. Specifically, IoT-based AI systems may now be used for disease prevention like physiological conditions, epidemic spread prevention systems, treatment of Parkinson's disease and diabetes. These are done with the help of sensors like ECG monitors, accelerometers, gyroscopes, microphones, heart rate monitors, blood pressure monitors, or blood glucose monitors [8]–[10]. The data collected by these sensors are used to create models which are either located in a centralized fashion at the cloud or distributed within the network by performing edge or fog computing [11]–[13]. The decision made by this model are later be used to communicate about a patients' whereabouts to various healthcare applications like, emergency calls, hospitals, or online help [14]. Figure 1 shows the high-level decision making workflow of such a system.

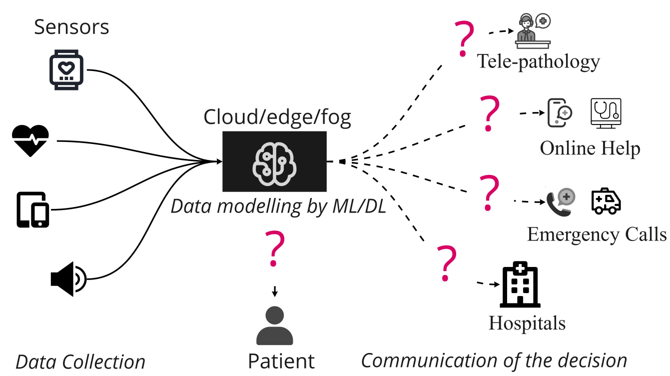

Fig. 1. Accountability issues in healthcare IoT systems

In spite of such development, the adoption of AI healthcare assistance is limited due to a lack of trustworthiness in these model based systems. Even though AI in healthcare has an optimistic view in media, and people believe it is efficient, but they do not trust it due to accountability reasons. The most common reason for distrust in the adoption of such systems is due to the lack of the scrutiny it undergoes [15]–[17]. One of the major reasons behind such problems is the opaque nature of the deep learning systems. In the process of building a precise data model, the deep learning models undergo multiple iterations and generate a complex model [1]. These models are so complex that it is difficult even for the data engineers to comprehend them. Thus, when a decision is made by the AI

* Corresponding Author: Aaron Ding (aaron.ding@tudelft.nl)

system, it is very difficult to understand the reasoning behind them and therefore realise if the decision was right or wrong. Moreover, these models are inherently based on the data used for training the model [18]. Thus, they may be prone to data biases, unjustified reasoning and unethical decision making [19]. So, it is important that the model is able to justify its decisions.

Currently, Explainable AI (XAI from here on), one of the emerging fields of AI has enabled to explain the opaque nature of deep learning algorithms. It is also assumed that XAI can inherently improve the transparency and hence the accountability of the algorithms [20], [21]. However, current studies do not describe how this can be achieved. Moreover, explaining does not necessarily mean that the explanations are interpretable by the people. Thus, using XAI methods to explain does not inherently make them accountable. Nonetheless, having clear objectives may help us create meaningful explanations. Since the EU ethics guidelines [22] and the US commission [23] have also promoted the development of accountability in algorithms, there is a need to understand what accountability means as a socio-technical element. Thus, in this paper we specifically reflect on accountability for healthcare IoT systems by conceptualizing the different elements, present how it can enhance trustworthiness and our vision to enable it using causal interpretability as an AI design element. This paper provides a good blend of social as well as technical factors for understanding these terms. Accountability may mean differently from a social, legal, regulatory, or a system point of view. In this paper accountability is the ability of an AI to justify its algorithmic decision making. Moreover, since causality has been in the discussions for providing interpretable explanations, in this paper we explain how it might also help make more accountable explanations. Thus, Section II explains the importance of different stakeholders for a healthcare IoT system, section III describes how accountability contributes to the trustworthiness of the system, Section IV conceptualizes the factors of accountability, Section V outlines the different trade-offs, and Section VI provides final conclusion remarks and our outlook.

II. Stake-holder positioning for accountability in healthcare IoT Systems

In this section we analyse the various users and stakeholders of the healthcare IoT system and map them on the basis of their interest in the development of accountability aspects of the system and the power they hold in the decision making process. Figure 2 shows various stakeholders on the power interest matrix. Naturally, the healthcare professionals are the people making critical decisions in a healthcare scenario. This group of people consists of doctors, surgeons, and other professional directly involved in the care-taking of the patients. Thus these are the experts who understand the system well, how an AI assistance may help as well as how it may case troubles when deployed in the healthcare systems. So, these people have the highest level of interested and may also take decision whether to use an AI assistance system or not. Healthcare assistants are people like nurses, who may not have as much power as the doctors, but still assist them in decision making. AI developers are the people who develop the assistance systems. They have the power to design, implement and evaluate various elements of the system. As a result, they have direct control over the AI development process and also might be the people to face questions when AI faces accountability issues. System administrators and auditors are the people making high stake decisions about whether a particular AI system should be deployed in the healthcare domain. They might not have much interest in the technical details of how the accountability is being designed in the system, but still hold power to pass or deny it from use in professional environments based on their quality checks. Academic researchers studying accountability, other socio-technical aspects of AI, as well as those working on AI systems other than healthcare may also be interested in this study as they could use the knowledge created in this research to reflect on their own systems. Vulnerable patients would also have a lot at stake when using such an AI assistance tool. They may also be highly interested in knowing if this system is accountable in its decision making process or not. However, such users only have a limited influence on the acceptability and usage of the tools, as their decision primarily comes down to whether they want to use the assistance tools or not. Non-vulnerable users may be able to try on new types of monitoring devices without increasing the stakes involved. Thus they would have relatively lower interest and less power in the accountability aspects of AI assistance tools in healthcare systems.

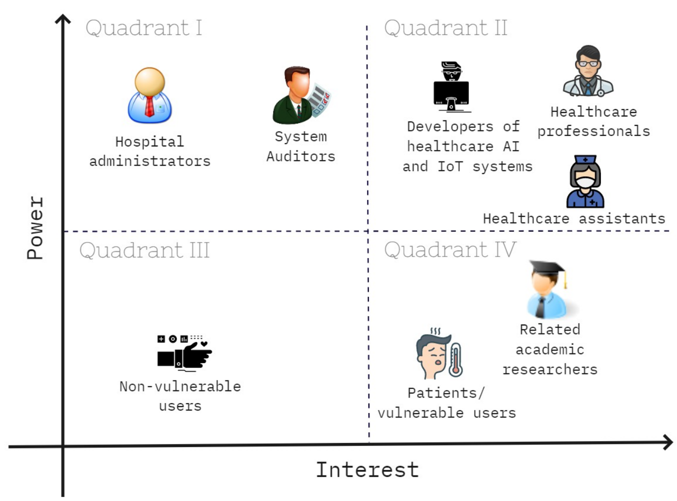

Fig. 2. Positioning stake-holders of healthcare IoT systems based on their power and interest in the accountability of the system

III. Accountability as a Trustworthiness factor for AI in Healthcare IoT Systems

In this section we discuss how accountability contributes to trustworthiness in IoT based systems. This study was specifically performed for healthcare systems. Thus it is specifically oriented towards this field. But most of the elements of this conceptual mapping are also true for other IoT based AI systems and can be expanded to different fields. Here we discuss various key-terms that we came across while studying trustworthiness in AI systems. Thus, we discuss the different terms under this broad trustworthiness umbrella. However, this is a very broad topic, and so the terms discussed here are non-exhaustive.

A. *Conceptualizing Trust and Trustworthiness*

The most common definition of trust from interpersonal studies is: *the anticipation of someone's behaviour in vulnerable situations* [24], [25]. It was observed that trust was only achieved in situations of vulnerability, where a person A anticipates if the person B will act in their favour. Thus trusting someone requires a sense of reliability on how well can the task be performed based on one's prior knowledge. This kind of intuition used for trusting someone is known as intrinsic trust [24]. However, for achieving trust in AI systems, some form of performance evaluation and verification can be employed in order to increase trust levels among users. Using these kind of external means for gaining trust is known as extrinsic trust [24]. Along similar lines, the term *trustworthiness* is used as a means of warranting trust by providing formal statements on the quality of the systems [26]. Thus the two major factors influencing trustworthiness of an AI system are the *performance* of the system and the verification mechanisms referred to as *accountability* in our conceptual mapping depicted in Figure 3 [27].

Since we define trust as a relationship between a system and its users, we must consider the aspects that we can design from a system perspective as well as the factors influencing from the users. Thus, in the Figure 3 we categorize the factors of trustworthiness into the system aspects, user aspects, and the interface properties used to communicate information between the two.

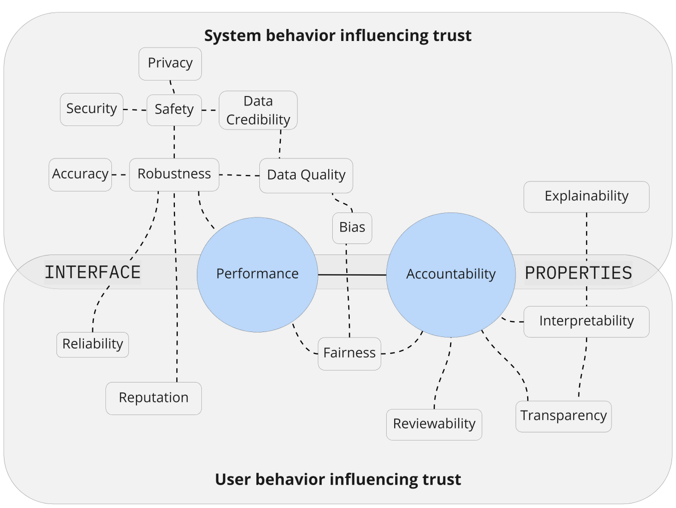

Fig. 3. Trustworthiness factors for AI in healthcare IoT systems

B. *Performance Aspects*

Robustness: Robustness is commonly used as an attribute to measure the performance of the system not just based on how precisely it performs but also on how well it can handle unforeseen adverse conditions. However, there is no single parameter to measure this. Under this large concept, various other aspects of the system are measured like accuracy, safety, and reliability. The EU ethics guidelines describes robust AI as a system that is safe, secure and reliable, and safeguards against any unintended adverse impacts from technical as well as social perspective [22].

Accuracy: Performance is very commonly measured by how accurately a system performs. In AI systems *accuracy* is most commonly used as a measure of how accurately the machine learning model predicts the value of test data sets. For classification models, this is the mean of true predictions over different classes. Additionally the other commonly used measures of performances for machine learning models are also precision, recall, and F1 score [28].

Security: Security attributes of the system protect them from external factors. These could be to develop mechanisms to protect them from various attacks, ensuring the data sources used for building the systems or mechanisms to validate new members in an IoT network. The term security is usually used to detect threats, attacks and malicious components in the IoT network [29]–[31].

Privacy: Since healthcare data contains a lot of personal information of an individual, this data may be considered as privacy sensitive and needs to be encrypted in such a way that the privacy of the individual is safeguarded when sent over untrusted servers [32]. Thus, in addition to provide secure data storage and data sharing, preserving the private information in the data is also a key performance factor for healthcare systems [29].

Data Credibility and Data Quality: Another method to ensure safety of the models is to use only the credible data. The existing methods use filtering methods to remove the outlier data or use signal to noise ratio to remove the noisy data [33], [34]. Although ensuring data quality in this way may help to make stable and generalizable models, the data models may still produce unaccurate results. Stacke, K. et al. [35] uses representation shift as a method to quantify how an actual data in real time is shifted from the training data. The paper shows how large shifts in data may cause the model to perform unreliably and discuss methods to tackle this issue. Wickstrom, K. et al. [36] uses the relevance scores produced by the XAI models to determine the certainty of the explanations. This method is used to provide stabilization for the generation of explanations. It uses ensemble method to choose between the most certain XAI method.

Reputation: Humans quite often base their trust in organisations on reputation and personal experiences [37]. Reputation, here refers to how well something is known in a community of users. Similar methods are used to find the credibility of the source in crowd-sourced data [38]. Blockchain, one of the highly successful methods in securing data for financial as well as many other domains runs along the similar line. It has also been extended to provide security measures in the IoT network for healthcare systems [39], [40].

Biases and Fairness: Data biases may be infused in the models during the measurement or collection of the data. These could be in the form of measurement bias, representation bias, aggregation bias, omitted variable bias, or sampling bias. These biases in the data may cause the model to cause unfair advantages to certain groups of people causing discrimination [41]. Thus ensuring fairness has been used as a means of ensuring that people are not mistreated by the algorithmic decision making due to their color, ethnic background, gender or other racial identities.

C. Accountability Aspects

Reviewability : Reviewability is the process which enables the system to be auditable. Auditability refers to the process of record keeping, logging and documenting of processes in a system, which can be used for reviewing the system behaviour for desirable attributes and legal compliance. Reviewability, in a similar vein, can be defined as a holistic and systemic approach to accountability via an iterative process of review, feedback, and revision [42].

Explainability and Interpretability : There is a growing discussion on explainability and interpretability. Some papers use these terms interchangeably, but a few distinguish between them. According to R. Nassih et al. [43] explanations are expressed as a collection of interpretations and contextual information, used to understand decision making, whereas R. Calegari et al. and M. Clinciu [44], [45] refer to explanations as a tool which essentially helps user interpret the decisions made by machine learning models.

For interpretability, we use the definitions provided in [44], [45] as a cognitive effort required by humans to provide meaning to the way they understand the working of the algorithm. In addition, these sources define explanations as a set of statements used to make something clear or to provide justifications to the actions taken by the ML algorithms. In context of XAI, explainability is the ability of the methods to provide the explanations. Thus in our conceptual map in the Figure 3, explainability in itself cannot provide accountability, unless it is understood by the human observer trying to assign meaning to the explanations.

Transparency and Interpretability : Since we accept the definition of interpretability as the effort required to understand the explanations [44], [45], in the context of opaque models, this helps bring transparency to the working of the models. Transparency as a term is often found accompanying interpretability in XAI literature, but technical definitions are less frequent. In Clinciu & Hastie [46], transparency is described as a blanket concept to which intelligibility, interpretability and explainability are facets. One should also bear in mind that transparency is only relevant when put in the context of the audience [47]. In addition, transparency is also identified as a key component in improving user trust in a system [48].

D. Interface Properties

The field of Human-Computer Interaction (HCI) studies a number of factors influencing human to AI trust, and how the interface with the AI system plays a role in it. Along with a number of visualization tools used in the HCI study, we broadly classify three important factors responsible for developing the trust .

Interaction: Representation of information plays an important role in the cognition effort needed for people to understand it. This could be in the form of static representation like verbal, textual, or graphical, or even in the form of dynamic interactions like in virtual and augmented reality [14]. Although interactive systems are able to improve the comprehension of a system, they come with a trade-off of more time consumption [49]. Thus, in the context of explaining AI decisions, where comprehension of the explanations is a key element in developing trustworthiness, the time investment needed by the healthcare professionals might be a critical constraint in designing the interface for healthcare systems.

Heuristics: While interpersonal trust (human-to-human) has been observed as a complex term, it was observed that humans trust machines much more easily than other humans. This is because people think that the machines do not judge them, thus they are more open to share their personal information with them. For example, people are more comfortable sharing their credit card details for online shopping, or share their personal stories in online therapy sessions with chat-bots. This information might be at a risk of being used, tracked, traced, sold or stored by the machines in some other ways. But people usually ignore such possibilities. This is due to the nature of humans to use the heuristics (mental shortcuts) where they avoid going into the details. The heuristic belief that machines are objective and incapable of biases is also called as machine heuristics [50]. In the context of XAI explanations, these heuristics may cause confirmation biases, where the user only searches for explanations that are consistent with his existing beliefs [51].

Control: The control that people have on decision making in an application domain often has a role on how they might trust the AI application. The higher the autonomy of AI decision making is in an application, the user aspects play a significant role. For example, in [49], the authors found out that the users did not trust the system, even with the interpretable models since they did not want the autonomy to be completely given at he hands of an AI algorithm. On the other hand, in [52], for a recommendation system for healthy diet options, even placebic explanations improved the trust in the users. Thus, the users trusted the recommendation systems more easily, than the decision making autonomous systems, where the users had to give the control to the AI systems.

IV. Conceptualizing Accountability for Healthcare IoT Systems

A. Conceptualizing Accountability

Accountability can be interpreted as 'the obligation to explain and justify conduct' [53], [54]. It is often necessary when an entity in power does not behave as expected, causing a need to understand the reason behind the actions and identify the responsible person or organisation. Thus, ensuring accountability also inherently motivates actors to behave in a better way [55]. Boven [54] defines accountability as a relationship between an actor and a forum, where he is obligated to explain and justify his actions and also faces consequences of the judgement from the forum. Thus Boven also relates 'accountability' as 'answerability'. All these definitions emphasise accountability as a socio-technical concern where the actor may also face legal consequences. Since AI also has a high societal impact, the accountability of AI should also be dealt from a more socio-technical approach. The EU ethics guidelines [22], emphasises that the AI should be accountable for its decisions both before and after their development, deployment and use. Additionally it also mentions the requirements of accountability to be the various performance parameters (like robustness and non-discrimination), and provide auditability, minimize negative impacts, address trade-offs, and provide means to redress. From a system's safety perspective this could mean to provide prevention for hazards, and could be achieved by process models [56]. In this paper, we have a lens from the AI perspective. Thus, we focus on the features of AI that could help us achieve the accountability and go a step further by finding the design elements needed in AI to achieve this.

From a regulatory perspective, accountability may further relate to 'auditability' [54] or 'reviewability' [42]. Figure 4 depicts various features associated with accountability. Further, it is also observed that causality, a study of causes and effects, may be used for understanding the causes behind the events detected by AI algorithms, and thus, may contribute to review its decision making process. Since accountability in current AI systems is hindered by their opaque nature, transparency generated by interpretable methods is also an important factor. Responsibility, the notion of commitment to a task and the ability to change the outcome of an event helps define a responsible person, or organisation for the occurrence of an event [53]. We discuss these aspects in detail in the next section.

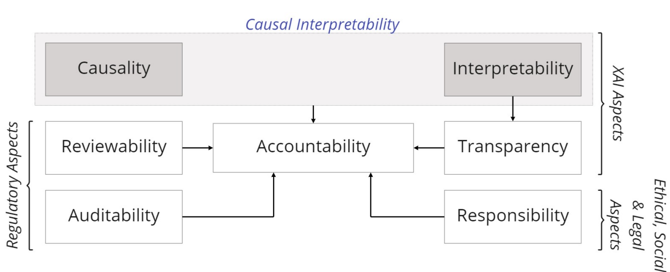

Fig. 4. Conceptualizing accountability factors for healthcare IoT systems

B. Accountability Features

Reviewability and Auditability: From the definition of Boven [54], we understand accountability as a obligation to justify conduct to a forum. Thus, in the context of algorithmic decision making, the algorithms would be the actors making the decision. Reference [55], mentions the three phases of algorithmic accountability, the information phase where the information pertaining to the action (or event) is provided, the discussion phase, where the forum discusses the answers provided by the actor, and the concerns and consequences phase, where the actor faces the consequences. Thus, the whole process of accountability is a socio-technical process [42], [55]. In fact the process of understanding the information from the XAI methods, in case of algorithmic accountability is also a socio-technical process. Cobbe et al. [42] emphasise reviewability as an approach of improving accountability. They define reviewability as a process of contextual record keeping to define the purposes and role of the whole decision making process. They also define accountability as process of commissioning, designing, decision making and auditing. Reviewability in this whole process may help in building accountability. Specifically, it could help in providing answers in the auditing phase.

Interpretability and Transparency: In the context of algorithmic accountability for opaque machine learning algorithms, Wieringa [55] defines accountability as the obligation to justify the use, the design, and the decision made by the algorithm. Thus, the XAI methods may justify this conduct in various ways. Global explanations may help justify for the use and design of the model, whereas local explanations may provide justifications for particular decisions. This kind of transparency in the system may be helpful, but in itself does not provide accountability. Moreover, Wieringa contrasts transparency with accountability, where the former is a passive concept and the latter is active.

Responsibility: Responsibility is referred to as the willingness of an actor to act in a transparent, fair and equitable way. This is a non-analytical factor, but can be used by the authorities for drawing conclusions while performing accountability [54]. As in human decision making systems, AI decision making systems can also face the problem of many hands [57]. In this context, multiple inputs, data points, architectures, and parties providing the data could be held accountable for any given event. Thus, defining clear responsibility can help resolve such issues.

Causality: Causality is referred to as study of causes and effects to understand the different causes behind particular events [44]. Kacianka et al. [53] reflects on causality as a retrospective approach to find out the causes for an given event under scrutiny. Since, in accountability, we often use retrospective answers, causality could be very much suitable for such methods. Moreover, since causality are ontological structures, meaning that they imbibe the environmental constraints of the world, these might produce stable means of explanations for achieving accountability. The author also proposes structural causal models as suitable for this purpose.

C. Towards Causal Interpretability for Accountability

Properties of Explanations Generated by Causal Models: Since accountability is a socio-technical concept, interpretability of the explanations plays an important role for making it accountable. We focus on interpretability by the users having the most interest and power in decision making, i.e. the group of people falling in quadrant II of the interest-power matrix in Figure 2. According to the figure, we have two groups of people for whom the explanations should be tailored, the data scientists and the healthcare professionals. Since, the theories from psychology and cognitive science provide evidences that humans are able to make causal inferences [58], generating explanations from such models can help us enhance the interpretability of the explanations. Moreover, Miller [59] suggests that, to have interpretability, explanations should also explain in terms of contrasting events (why event X happened instead of Y). 'What if' statements and explanations based on such contrasts are known as counterfactuals. Judea Pearl [60] has also emphasised on causal models to generate explanations, where associations being the first in the ladder, intervention at the second stage, and counterfactual being at the third stage. Investigating on how and when to use these three kinds of explanations in the accountability process would be a part of our future study.

Generation of causal models may be as a result of medical study such as the WHO-UMC [61], or may even be developed by using machine learning approach [62], [63]. In our research, we refer to the causal graphs generated by experts representing their expertise and domain knowledge in a cause-effect manner [64]. Thus, explanations generated by such knowledge may be more consistent with the domain knowledge, thereby making it more interpretable for the healthcare professionals. For enhancing the accountability, such explanations may also be used to verify the performance of the given model. The explanations contrasting with the expert models may be used to scrutinize the models. These are the events where the model is either expected to be erroneous, or is able to generate new knowledge. Thus this method may be helpful for understanding and predicting the behaviour of the model as well as to understand under which circumstances the model might need expert intervention. Since such causal models imbibe the domain knowledge, the explanations created by using them are also expected to be consistent and stable. Thus, three properties of causal explanations that make it suitable for accountability are: interpretability by the high stake users, stability in explanation generation, and consistency with the domain knowledge.

Limitations and Challenges:

Causal models may induce selection bias, i.e. the bias created while choosing the sample for the study. This bias is induced due to the biases present in the population of the study and in the process of choosing this group [64].

One of the reasons causal reasoning has gained so much attention is due to its power to eliminate confounding bias. Confounders are factors that confound the causality between an intervention and its outcome. Since, in healthcare, there are a number of identified variables, creating information on confounders through real time experiments might be a tedious task [64]. Thus the task of generating meaningful data is high.

Even with such limitations, the medical domain has been studying causal reasoning for ontological and epistemological studies. The causal reasoning is expected to answer the ”what”, ”how”, and ”why” questions. However, the current causal models are created by real time experimental data either by randomised or non-randomised control trials. Thus, they are based mostly on ”what” is observed by the patients and ”what” the experts think. Thus to the best of our knowledge, we can answers these ”what” questions and try to map them to the ”how” and ”why” questions. In this process, the individualistic biases such as confirmation biases, observational biases, and publication biases might also play a role. Therefore, realistically framing the healthcare causal models is one of the major challenges faced by the experts [65].

Moreover, the explanations generated by the causal models may face challenges to create real world examples. Specifically, for counterfactuals, where the explanations may help to provide interventions or recourse, creating a real-world scenario is necessary for taking actions. For example in a loan application scenario, a recourse action item for a person cannot be to lower their age, or to generate an unfeasible amount of salary in a short time. Thus, generating meaningful explanations is still a challenging task even with causal explanations [66]–[68]. In the same vein, explaining for accountability should also consider all plausible scenarios based on such real-world constraints.

Our Vision: It is important to investigate how explanations generated by causal models enhance the interpretability for high stake users, and how these explanations can answer the accountability questions by performing quantitative as well as qualitative studies. For healthcare AI systems we justify that causal interpretability has a high potential to achieve this. However, it is limited by the existence of expert models. Thus, it is not a sufficient condition for achieving accountability, and we plan to address this issue in our future studies.

Since we plan to use expert generated causal models, we do not plan to induce them directly in the AI models, but rather use them for the accountability studies with the experts. Our vision is to enhance accountability of the model by having interpretability, to understand the different scenarios where the model complies and diverges from the expert understanding and to account model for such scenarios. For healthcare IoT systems, this should also be extended to understand the architecture of the model and its influence on the decision making. Doing so might help the understand when to trust the AI model. Thus, for accountability, we envision to make the model answerable to such expert models. Moreover, accountability of the AI decisions that are non-compliant with the expert models would be an open research area for future work.

V. Trade-offs

While using explanations for achieving accountability, there are other aspects of the system that can be affected and need careful considerations. We illustrate in this section various trade-offs that are important for designing accountable AI.

A. Interpretability vs Accuracy

Post-hoc explanations may cause a drop in the complexity of the models to make them interpretable and in the process also negatively affect the accuracy of the models. Thus, it is also observed that for complex systems, interpretability could come at a cost of the performance of the AI models [69]. For healthcare systems, both interpretability and accuracy plays a very significant role as they work together to provide the needed performance and avoid any mishaps. Thus, a well scrutinized model during the development process and an ante-hoc explainability [21], [70] may help in such applications. This tradeoff further extends to other values for which interpretability is key. For example, insights from system safety show that too complex models may inject various safety hazards in situations where operators should act based on a model or are ultimately responsible, due to the limits of human cognition. This "curse of flexibility" should be actively prevented by taking a broader systems lens and asking what is ultimately important and needed to assure safety in the context of use, and designing the model and use of the model and necessary failsafe mechanisms integrally [56], [71].

B. Soundness and Completeness vs User Comprehension

Generating explanations from complex machine learning algorithms involves providing detailed information to users. However, generating too much information can have the drawback of reducing a user's comprehension of the explanation. For example, for some quantitative models, such as Partially Observable Markov Decision Processes, the sheer size and complexity of the model can become informationally overwhelming [72]. This also aligns with the three principles put forth by Kulesza et al. [73] - (1) be sound, (2) be complete and (3) don't overwhelm. There is a natural tension between (3) and the other two principles. Soundness means that each component of an explanation must be truthful to the underlying system, and thus should not be oversimplified or made out to be less complex than it actually is. Completeness means that an explanation cannot omit important information about the model. But the more focus is put on the principles of soundness and completeness, the more strain is put on user comprehension and attention. In other words, in addition to the model, the curse of flexibility (the well-known challenge from system safety mentioned for the previous tradeoff) also holds for the construction of explanations [71].

C. Expert Bias vs Autonomy

Machine learning models autonomously model the real time data. Comparing their decisions with the experts for accountability may result in introduction of expert biases like the confirmation biases. In cases of non-compliance of the explanations with the experts, a careful examination can help determine the action plan for the usage and modifications of the AI model. However, careful scrutiny is time consuming, and the time needed from the experts is also quite expensive. Thus, the design for accountability should consider these constraints to determine an optimum solution for the trade-off.

D. Accountability vs Resource Consumption

Answering the accountability questions after the deployment, especially on IoT devices might introduce extra overheads. Since in healthcare applications, there is a high demand of deployment of the AI algorithms in a distributed manner by edge or fog computing, making such algorithms may introduce even larger overheads. Thus, dealing with the processing constraints such as processing power and bandwidth requirements for communications should also be considered in IoT based systems [11]–[13].

VI. Concluding Remarks and Outlook

In this paper we discussed how opaque models pose accountability issues, and how critical decision making and trust in healthcare are affected by those. We advocate accountability as a key element contributing to the trustworthiness of an AI system. Additionally, we put emphasis on accountability through the socio-technical lens, where the explanations for AI must be interpretable to the people developing it and the domain experts capable of scrutinizing it. Therefore, accountable AI can lead to a verifiable system for experts with the domain knowledge. We envision to facilitate this by using expert generated causal models as the knowledge representation against which the AI model should justify its decision making. We envision such explanations to be interpretable for the experts. Moreover, such an approach could also be used to detect if the model is compliant or non-compliant with the expert knowledge. This will lead to either improving the model in case of erroneous explanations, or generating new knowledge for AI systems in the healthcare IoT domain.

Acknowledgment

This research is supported by SPATIAL project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.101021808.

References

[1] Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436 – 444, 2015. [Online]. Available: https://doi.org/10.1038/nature14539

[2] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of Field Robotics*, 2020.

[3] G. Sukanya and J.Priyadarshini, "A meta analysis of attention models on legal judgment prediction system," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 2, 2021. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2021.0120266

[4] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, no. 3, 2021.

[5] R. C. Cavalcante, R. C. Brasileiro, V. L. Souza, J. P. Nobrega, and A. L. Oliveira, "Computational intelligence and financial markets: A survey and future directions," *Expert Systems with Applications*, vol. 55, pp. 194–211, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095741741630029X

[6] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, “Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis,” *IEEE J Biomed Health Inform*, vol. 22, no. 5, pp. 1589–1604, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29989977

[7] S. Tian, W. Yang, J. M. L. Grange, P. Wang, W. Huang, and Z. Ye, “Smart healthcare: making medical care more intelligent,” *Global Health Journal*, vol. 3, no. 3, pp. 62–65, 2019.

[8] A. Monteiro, H. Dubey, L. Mahler, Q. Yang, and K. Mankodiya, “Fit: A fog computing device for speech tele-treatments,” in *2016 IEEE International Conference on Smart Computing (SMARTCOMP)*, 2016, pp. 1–3.

[9] M. Abdel-Basset, G. Manogaran, A. Gamal, and V. Chang, “A novel intelligent medical decision support model based on soft computing and iot,” *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4160–4170, 2020.

[10] S. U. A. 1 and I. M. SHAMIM HOSSAIN 2, (Senior Member, “Edge intelligence and internet of things in healthcare: A survey,” *IEEE : SPECIAL SECTION ON EDGE INTELLIGENCE FOR INTERNET OF THINGS*, 2020.

[11] A. Y. Ding, E. Peltonen, T. Meuser, A. Aral, C. Becker, S. Dustdar, T. Hiessl, D. Kranzlmuller, M. Liyanage, S. Magshudi, N. Mohan, J. Ott, J. S. Rellermeyer, S. Schulte, H. Schulzrinne, G. Solmaz, S. Tarkoma, B. Varghese, and L. Wolf, “Roadmap for Edge AI: A Dagstuhl Perspective,” *ACM SIGCOMM Computer Communication Review*, vol. 52, no. 1, pp. 28 – 33, 2022. [Online]. Available: https://doi.org/10.1145/3523230.3523235

[12] B. Varghese, E. De Lara, A. Y. Ding, C. H. Hong, F. Bonomi, S. Dustdar, P. Harvey, P. Hewkin, W. Shi, M. Thiele, and P. Willis, “Revisiting the Arguments for Edge Computing Research,” *IEEE Internet Computing*, vol. 25, no. 5, pp. 36–42, 2021.

[13] A. Y. Ding, M. Janssen, and J. Crowcroft, “Trustworthy and Sustainable Edge AI: A Research Agenda,” in *Third IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 2021, pp. 164–172.

[14] S. M. Preum, S. Munir, M. Ma, M. S. Yasar, D. J. Stone, R. Williams, H. Alemzadeh, and J. A. Stankovic, “A review of cognitive assistants for healthcare: Trends, prospects, and future directions,” *ACM Comput. Surv.*, vol. 53, no. 6, feb 2021. [Online]. Available: https://doi.org/10.1145/3419368

[15] C. Y. Hui, B. McKinstry, O. Fulton, M. Buchner, and H. Pinnock, “Patients’ and clinicians’ perceived trust in internet-of-things systems to support asthma self-management: Qualitative interview study,” *Jmir Mhealth and Uhealth*, vol. 9, no. 7, p. 12, 2021. [Online]. Available: ¡Go to ISI¿://WOS:000692249000026

[16] Y. P. Ongena, M. Haan, D. Yakar, and T. C. Kwee, “Patients’ views on the implementation of artificial intelligence in radiology: development and validation of a standardized questionnaire,” *Eur Radiol*, vol. 30, no. 2, pp. 1033–1040, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31705254

[17] D. Yakar, Y. P. Ongena, T. C. Kwee, and M. Haan, “Do people favor artificial intelligence over physicians? a survey among the general population and their view on artificial intelligence in medicine,” *Value Health*, vol. 25, no. 3, pp. 374–381, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/35227448

[18] W. T. Hutiri and A. Y. Ding, “Bias in automated speaker recognition,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’22. ACM, 2022, pp. 230–247.

[19] I. C. Office, “Big data, artificial intelligence, machine learning and data protection,” 2017. [Online]. Available: https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

[20] A. Hanif, X. Zhang, and S. Wood, “A survey on explainable artificial intelligence techniques and challenges,” pp. 81–89, 2021.

[21] M. F. H. Nadia Burkart, “A survey on the explainability of supervised machine learning,” *Journal of Artificial Intelligence Research 70*, 2020.

[22] European-Commision, “Ethics guidelines for trustworthy ai,” 2019.

[23] t. C. (2019-2020), “S.1108 - algorithmic accountability act of 2019.” [Online]. Available: https://www.congress.gov/bill/116th-congress/senate-bill/1108

[24] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, “Formalizing trust in artificial intelligence,” pp. 624–635, 2021.

[25] O. Vereschak, G. Bailly, and B. Caramiaux, “How to evaluate trust in ai-assisted decision making? a survey of empirical methodologies,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–39, 2021.

[26] E. Toreini, M. Aitken, K. Coopamootoo, K. Elliott, C. G. Zelaya, and A. van Moorsel, “The relationship between trust in ai and trustworthy machine learning technologies,” pp. 272–283, 2020.

[27] L. Thornton, B. Knowles, and G. Blair, “Fifty shades of grey,” pp. 64–76, 2021.

[28] P. Flach, “Performance evaluation in machine learning: The good, the bad, the ugly and the way forward,” *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[29] H. Qiu, M. Qiu, M. Liu, and G. Memmi, “Secure health data sharing for medical cyber-physical systems for the healthcare 4.0,” *IEEE J Biomed Health Inform*, vol. 24, no. 9, pp. 2499–2505, 2020, qiu, Han Qiu, Meikang Liu, Meiqin Memmi, Gerard eng Research Support, Non-U.S. Gov’t 2020/02/20 IEEE J Biomed Health Inform. 2020 Sep;24(9):2499-2505. doi: 10.1109/JBHI.2020.2973467. Epub 2020 Feb 12. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32071015

[30] R. M. Seepers, W. Wang, G. de Haan, I. Sourdis, and C. Strydis, “Attacks on heartbeat-based security using remote photoplethysmography,” *IEEE J Biomed Health Inform*, vol. 22, no. 3, pp. 714–721, 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28391214

[31] Y. Qian, J. Shen, P. Vijayakumar, and P. K. Sharma, “Profile matching for iomt: A verifiable private set intersection scheme,” *IEEE J Biomed Health Inform*, vol. 25, no. 10, pp. 3794–3803, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/34111016

[32] R. Ghasemi, M. M. Al Aziz, N. Mohammed, M. H. Dehkordi, and X. Jiang, “Private and efficient query processing on outsourced genomic databases,” *IEEE J Biomed Health Inform*, vol. 21, no. 5, pp. 1466–1472, 2017, ghasemi, Reza Al Aziz, Md Momin Mohammed, Noman Dehkordi, Massoud Hadian Jiang, Xiaoqian eng R21 LM012060/LM/NLM NIH HHS/ U01 EB023685/EB/NIBIB NIH HHS/ R00 LM011392/LM/NLM NIH HHS/ R01 HG007078/HG/NHGRI NIH HHS/ R13 HG009072/HG/NHGRI NIH HHS/ R01 GM118609/GM/NIGMS NIH HHS/ Research Support, Non-U.S. Gov’t Research Support, N.I.H., Extramural 2016/11/12 IEEE J Biomed Health Inform. 2017 Sep;21(5):1466-1472. doi: 10.1109/JBHI.2016.2625299. Epub 2016 Nov 4. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27834660

[33] H. D. Hesar and M. Mohebbi, “An adaptive particle weighting strategy for ecg denoising using marginalized particle extended kalman filter: An evaluation in arrhythmia contexts,” *IEEE J Biomed Health Inform*, vol. 21, no. 6, pp. 1581–1592, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28541230

[34] O. M. Staal, S. Salid, A. Fougner, and O. Stavdahl, “Kalman smoothing for objective and automatic preprocessing of glucose data,” *IEEE J Biomed Health Inform*, vol. 23, no. 1, pp. 218–226, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/29994742

[35] K. Stacke, G. Eilertsen, J. Unger, and C. Lundstrom, “Measuring domain shift for deep learning in histopathology,” *IEEE J Biomed Health Inform*, vol. 25, no. 2, pp. 325–336, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33085623

[36] K. Wickstrom, K. O. Mikalsen, M. Kampffmeyer, A. Revhaug, and R. Jenssen, “Uncertainty-aware deep ensembles for reliable and explainable predictions of clinical time series,” *IEEE J Biomed Health Inform*, vol. 25, no. 7, pp. 2435–2444, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33284756

[37] G. Athanasiou, G. C. Anastassopoulos, E. Tiritidou, D. Lymberopoulos, G. Athanasiou, G. C. Anastassopoulos, E. Tiritidou, and D. Lymberopoulos, “A trust model for ubiquitous healthcare environment on the basis of adaptable fuzzy-probabilistic inference system,” *IEEE J Biomed Health Inform*, vol. 22, no. 4, pp. 1288–1298, 2018, athanasiou, Georgia Anastassopoulos, George C Tiritidou, Eleni Lymberopoulos, Dimitrios eng 2017/08/03 IEEE J Biomed Health Inform. 2018 Jul;22(4):1288-1298. doi: 10.1109/JBHI.2017.2733038. Epub 2017 Jul 28. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/28767375

[38] T. Hashem, R. Hasan, F. Salim, and M. T. Mahin, “Crowd-enabled processing of trustworthy, privacy-enhanced and personalised location based services with quality guarantee,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 4, p. Article 167, 2018. [Online]. Available: https://doi.org/10.1145/3287045 https://dl.acm.org/doi/pdf/10.1145/3287045

[39] X. Liu, P. Zhou, T. Qiu, and D. O. Wu, “Blockchain-enabled contextual online learning under local differential privacy for coronary

heart disease diagnosis in mobile edge computing," *IEEE J Biomed Health Inform*, vol. PP, 2020, liu, Xin Zhou, Pan Qiu, Tie Wu, Dapeng Oliver eng 2020/08/06 IEEE J Biomed Health Inform. 2020 Jun 2;PP. doi: 10.1109/JBHI.2020.2999497. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32750921

[40] C. Esposito, A. De Santis, G. Tortora, H. Chang, and K.-K. R. Choo, "Blockchain: A panacea for healthcare cloud-based data security and privacy?" *IEEE Cloud Computing*, vol. 5, no. 1, p. 31 – 37, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8327543

[41] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[42] J. Cobbe, M. S. A. Lee, and J. Singh, "Reviewable automated decision-making," pp. 598–609, 2021.

[43] R. Nassih and A. Berrado, "State of the art of fairness, interpretability and explainability in machine learning," pp. 1–5, 2020.

[44] L. Cheng, "Socially responsible ai algorithms: Issues, purposes, and challenges," *Journal of Artificial Intelligence Research*, 2021.

[45] R. Calegari, G. Ciatto, A. Omicini, M. Baldoni, F. Bergenti, S. Monica, and G. Vizzari, "On the integration of symbolic and sub-symbolic techniques for xai: A survey," *Intelligenza Artificiale*, vol. 14, no. 1, pp. 7–32, 2020.

[46] M. Clinciu and H. Hastie, "A survey of explainable ai terminology," *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*, 2019.

[47] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: An initial review of publicly available ai ethics tools, methods and research to translate principles into practices," *Sci Eng Ethics*, vol. 26, no. 4, pp. 2141–2168, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31828533

[48] E. Puiutta and E. M. S. P. Veith, "Explainable reinforcement learning: A survey," in *Machine Learning and Knowledge Extraction*, A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, Eds. Cham: Springer International Publishing, 2020, pp. 77–95.

[49] H.-F. Cheng, R. Wang, Z. Zhang, F. O'Connell, T. Gray, F. M. Harper, and H. Zhu, "Explaining decision-making algorithms through ui," pp. 1–12, 2019.

[50] S. S. Sundar and J. Kim, "Machine heuristic," pp. 1–9, 2019.

[51] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry, "Human-centered tools for coping with imperfect algorithms during medical decision-making," pp. 1–14, 2019.

[52] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, "The impact of placebic explanations on trust in intelligent systems," pp. 1–6, 2019.

[53] S. Kacianka and A. Pretschner, "Designing accountable systems," pp. 424–437, 2021.

[54] M. Bovens, "Analysing and assessing accountability: A conceptual framework," *European Law Journal*, 2007.

[55] M. Wieringa, "What to account for when accounting for algorithms," pp. 1–18, 2020.

[56] R. I. J. Dobbe, "System Safety and Artificial Intelligence," in *The Oxford Handbook of AI Governance*. Oxford University Press, 2022. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780197579329.013.67

[57] I. Poel, van de, L. Royakkers, and S. Zwart, Eds., *Moral responsibility and the problem of many hands*. United Kingdom: Routledge Taylor & Francis Group, 2015.

[58] M. Chen, "A tale of two deficits: Causality and care in medical ai," *Philosophy & Technology*, vol. 33, no. 2, pp. 245–267, 2019.

[59] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *arXiv:1706.07269v3*, 2018. [Online]. Available: https://arxiv.org/pdf/1706.07269.pdf

[60] J. P. Mackenzie and Dana, *The book of why: the new science of cause and effect*. Basic Books, Hachette Book Group, 1290 Avenue of the Americas, New York, NY 10104, 2018.

[61] T. U. M. Centre, "The use of the who-umc system for standardised case causality assessment."

[62] Y. Zhu, Y. Sha, H. Wu, M. Li, R. A. Hoffman, and M. D. Wang, "Proposing causal sequence of death by neural machine translation in public health informatics," *IEEE J Biomed Health Inform*, vol. 26, no. 4, pp. 1422–1431, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/35349461

[63] H. Q. Yu, "Extracting and representing causal knowledge of health conditions," 2020. [Online]. Available: http://hdl.handle.net/10547/624766

[64] X. Wu, J. Li, Q. Qian, Y. Liu, and Y. Guo, "Methods and applications of causal reasoning in medical field," pp. 79–86, 2021.

[65] S. Fox and O. Aranko, "Healthcare framing: Critical realist framing for causal interdependencies and uncertainties within healthcare," *Technology in Society*, vol. 50, pp. 66–72, 2017.

[66] S. Barocas, A. D. Selbst, and M. Raghavan, "The hidden assumptions behind counterfactual explanations and principal reasons," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 80–89. [Online]. Available: https://doi.org/10.1145/3351095.3372830

[67] A.-H. Karimi, B. Schölkopf, and I. Valera, "Algorithmic recourse: From counterfactual explanations to interventions," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 353–362. [Online]. Available: https://doi.org/10.1145/3442188.3445899

[68] A. Kasirzadeh and A. Smart, "The use and misuse of counterfactuals in ethical machine learning," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 228–236. [Online]. Available: https://doi.org/10.1145/3442188.3445886

[69] A. Rawal, J. McCoy, D. B. Rawat, B. Sadler, and R. Amant, "Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives," *IEEE Transactions on Artificial Intelligence*, pp. 1–1, 2021.

[70] Y. Zhang, P. Tino, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.

[71] N. G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*. Cambridge, MA, USA: MIT Press, 2012. [Online]. Available: http://ebookcentral.proquest.com/lib/delft/detail.action?docID=3339365

[72] D. V. Pynadath, N. Wang, and M. J. Barnes, "Transparency communication for reinforcement learning in human-robot interaction," *XAI 2018*, p. 123, 2018.

[73] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 126–137.