Formalizing Ethical Design in Prostate Cancer Image Analysis: A Preliminary Case Study

Sadie Lee
Cognitive Systems and Data Science
University of British Columbia
Vancouver, Canada
slee103@student.ubc.ca

Adam Resnick
College of Medicine and Science
Mayo Clinic
Rochester, United States
Resnick.Adam@mayo.edu

Nasibeh Zanjirani Farahani
College of Medicine and Science
Mayo Clinic
Rochester, United States
ZanjiraniFarahani.Nasibeh@mayo.edu

***Abstract*—Although artificial intelligence has shown potential to transform healthcare, adoption in clinical practice has been limited due to concerns regarding systems' trustworthiness. Several ethical guidelines have been issued to address these concerns, however, implementation is often challenging. We describe a preliminary case study to implement the Coalition for Health AI ethical guidelines in the design of a prostate abnormality detection system, using formal specifications to address challenges of conflicting principles and translation into design requirements. Initial evaluation indicates that formal specifications are effective in addressing these challenges.**

***Index Terms*—Formal specification, ethical design, medical image analysis, artificial intelligence, translation to practice**

I. Introduction

Artificial intelligence (AI) has shown immense potential in the transformation of healthcare by enabling the analysis of large volumes of medical data. However, adoption in clinical practice has been limited due to concerns regarding the trustworthiness (i.e. ethical design and behavior) of AI systems. Issues with data privacy, bias, lack of transparency, safety, and reliability have been of concern [1]. Given the high-risk nature of a healthcare setting, there is a prominent ethical imperative for AI systems [2] and thus the implementation of ethical guidelines is necessary.

Several guidelines for the ethical development of AI have recently been published in various domains [3], although few have been implemented in practice due to challenges at every stage of the AI lifecycle [4]. Different stakeholders may also apply different ethical frameworks, which often lack formal specifications and verification, leading to vaguely defined responsibilities and limited implementation. Specifically within healthcare, the Coalition for Health AI (CHAI) has collaboratively developed practical quality standards for the ethical development and deployment of AI solutions with the aim of improving trustworthiness and increasing adoption of these solutions in clinical practice [5].

As such, this work addresses challenges of implementing ethical guidelines within a healthcare setting by applying the CHAI guidelines in the design and preparation stages for a prostate abnormality detection system, outlining formal specifications for the implementation of ethical design, and considering scenarios in the context of the case study. The scope of this paper focuses only on the design and preparation stages of the detection system, specifically regarding solution planning, cohort identification, and annotation, given its preliminary nature.

II. Related Work

A. Coalition for Health AI Guidelines

The aim of the CHAI quality standards is to increase the reliability, safety, and trustworthiness of AI systems in healthcare throughout the end-to-end lifecycle [5]. The six-stage lifecycle for AI systems in healthcare encompasses processes from initial problem identification and solution planning to evaluation and deployment, and how it is integrated in the clinical workflow. This lifecycle is built on a set of core principles integrated at each stage and are 1) usefulness, usability, and efficacy; 2) fairness and equity; 3) safety and reliability; 4) transparency, accountability, intelligibility; and 5) security and privacy [5].

B. Challenges to Implementation

Despite the many sets of ethical AI guidelines, challenges to their implementation in practice have been categorized by [4] into five levels: ethical principles, design, technology, organizational, and regulatory. We focus on the first two levels, ethical principles and design, given the scope of the paper.

The challenge of ethical principles is the relationship between them: conflicts and contradictions may occur. For example, if an AI system is trained on retrospective data, historical biases are embedded such that certain groups are naturally favored. Other groups then need to be protected to ensure fairness in the dataset and thus in the decision-making of the system. However, fairness is often achieved by an optimal balance of impact, performance, and resources which inherently requires trade-offs. Optimal performance, for example, may lead to unequal outcomes for a group with certain protected characteristics due to a lack of resources for testing in the population [4].

At the design level, the translation of abstract ethical principles into tangible design requirements for features and functionalities of an AI system can be challenging to implement [4]. Abstract ethical principles, such as fairness or transparency, are open to interpretation and formal definitions are

978-1-6654-7345-3/22/$31.00

uncommon. Conflicts may also occur, whereby ensuring full transparency for example could compromise patient privacy.

III. Case Study

A. Project Background

While designing and implementing processes for AI development, we were tasked with exploring ethical practices and how these could be addressed specifically for medical image analysis at a large healthcare organization in the Midwest. We viewed this as an opportunity to apply the CHAI ethical guidelines and develop formal specifications to assist in translating these guidelines into design requirements. The use case focuses on an abnormality detection system for prostate magnetic resonance (MR) images. The processes in the preparation and design stages are outlined in Figure 1.

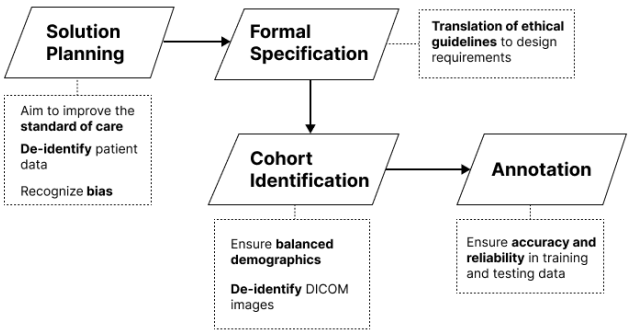

Fig. 1. Stages of development integrating the CHAI ethical guidelines in the context of the prostate abnormality detection use case.

B. Prostate Image Analysis

Prostate cancer (PCa) is one of the most commonly diagnosed malignant carcinomas, and second leading cause of mortality from cancer in men worldwide [6]. Diagnostic accuracy is thus significant to the prognosis of patients with PCa. While systematic tissue biopsies have previously remained the standard of care for diagnosis, advances in MRI have enabled non-invasive, lower-risk procedures to be performed in addition or replacement to systematic biopsies [6].

Algorithms have been developed to assist diagnostic radiologists who interpret MR images looking for prostate abnormalities. These algorithms can perform image analysis at various stages and have the potential to improve the PCa screening workflow where the AI system can indicate abnormalities that may not have been identified by a radiologist or augment clinician decision-making [7].

C. Solution Planning

The solution planning stage intended to address the CHAI principles of usefulness and efficacy by considering the necessity of the AI system in relation to current standards of care and potential integration into the clinical workflow. It was assessed that the prostate abnormality detection system aimed to improve the standard of care by enabling radiologists as the intended end users to augment their decision-making and verify identification of abnormalities. Considering potential sources of bias, as per CHAI guidelines in this stage, we identified that our dataset primarily contained patients from one region (Midwest) and the majority identified as white, likely due to the demographics of patients who have received care from the healthcare organization. This has the potential to reduce overall generalizability, and thus ensuring fairness was necessary in the cohort identification stage.

Regarding data requirements, the detection system was assessed to require clinical unstructured data (MR images and radiology reports), as well as structured data (e.g. patient demographics and diagnoses) to identify relevant images. Given the requirement of structured patient data from the electronic health record, albeit retrospectively, the CHAI core principle of privacy was identified as necessary to ensure that patient data was protected. This was done through a comprehensive data de-identification process to ensure personal health information was not revealed. Personal health information includes but is not limited to name, medical record number, date of birth, age, ethnicity, date of visits, and locations where services were received according to the Health Insurance Portability and Accountability Act [5]. Patients from the healthcare organization gave consent to de-identified data collection.

As the use case aimed to detect prostate abnormalities, risks to patients were recognized particularly in forms of cognitive bias such as automation bias [8], which could occur due to over-reliance on indications made by the AI system and less attention to images not flagged. Assessing integration into the clinical workflow, we determined that the detection system should assist radiologists in interpreting images from multiparametric MRI procedures alongside AI output in the form of image annotation where the AI system annotates images from the procedure to indicate potential abnormalities.

D. Formal Specification

With the challenge of translating abstract ethical principles into tangible design requirements leading to limited implementation in practice, we define base case formal specifications for the prostate abnormality detection system. These can be expanded based on different scenarios as examined in the following section. Formal specifications use mathematical notation to unambiguously define the characteristics an information system must have [9]. Z notation was chosen for its schema calculus, which allows for greater modularity [9].

1) Usefulness, usability, and efficacy: Usefulness, usability, and efficacy are specified where the detection system must improve the standard of care (i.e. *ImprovedCare* must be true), the set of annotated abnormalities cannot be empty, and the set of users who are assisted (e.g. radiologists) must be the same set for end users, meaning that all users receiving assistance are the end users.

UsefulnessUsabilityEfficacy

$[ABNORMALITY, USER]$
$ImprovedCare : BOOL$
$AnnotatedAbnormalities : \mathbb{P}\, ABNORMALITY$
$AssistedUsers : \mathbb{P}\, USER$
$EndUsers : \mathbb{P}\, USER$

$ImprovedCare = TRUE$
$AnnotatedAbnormalities \neq \emptyset$
$AssistedUsers = EndUsers$

2) Fairness and equity: Fairness and equity are specified as treating all users without bias based on protected attributes such as race, gender, or their equivalents in the identification of a cohort; i.e. for every user x and y, and for every pair of protected attributes a and b associated with x and y, if x and y have different protected attributes a and b, then x and y must receive the same treatment.

FairnessEquity

$[USER, ATTRIBUTE]$
$ProtectedAttribute : USER \nrightarrow ATTRIBUTE$
$Treatment : USER \rightarrow TREATMENT$

$\forall x, y : USER;\ a, b : ATTRIBUTE \bullet$
$(x \in \mathrm{dom}\, ProtectedAttribute \land y \in \mathrm{dom}$
$ProtectedAttribute \land$
$ProtectedAttribute(x) = a \land ProtectedAttribute(y)$
$= b \land a \neq b) \Rightarrow (Treatment(x) = Treatment(y))$

3) Safety and reliability: Safety and reliability are specified such that for every dataset d, if d is part of the data used for training and validation, then d must be diverse and representative.

SafetyReliability

$[DATASET]$
$TrainedOn : \mathbb{P}\, DATASET$
$ValidatedOn : \mathbb{P}\, DATASET$
$Diverse : DATASET \rightarrow BOOL$
$Representative : DATASET \rightarrow BOOL$
$AccuracyEnsured : BOOL$
$BiasMinimized : BOOL$

$\forall d : DATASET \bullet (d \in TrainedOn \lor$
$d \in ValidatedOn) \Rightarrow$
$(Diverse(d) \land Representative(d))$
$AccuracyEnsured = TRUE$
$BiasMinimized = TRUE$

4) Transparency, accountability, and intelligibility: Transparency, accountability, and intelligibility are specified where for every user u who is an intended user of the system, user oversight must be allowed (i.e. *Oversight* must be true), and the system must implement mechanisms to mitigate risks to patients (i.e. *RiskMitigation* must be true).

TransparencyAccountabilityIntelligibility

$[USER]$
$IntendedUsers : \mathbb{P}\, USER$
$Oversight : BOOL$
$RiskMitigated : BOOL$

$\forall u : USER \bullet (u \in IntendedUsers) \Rightarrow$
$(Oversight = TRUE)$
$RiskMitigated = TRUE$

5) Privacy and security: Privacy and security are specified such that patient data must be kept confidential, cannot be shared without consent, and must be de-identified in both images and structured data; i.e. for every patient p, and for every piece of structured data d or image i associated with p, if d or i is to be shared (i.e. included in $Consent(p)$), then d and i must be de-identified.

PrivacySecurity

$[PATIENT, DATA, IMAGE]$
$PatientData : PATIENT \leftrightarrow DATA$
$PatientImages : PATIENT \leftrightarrow IMAGE$
$Consent : PATIENT \rightarrow \mathbb{P}(DATA \cup IMAGE)$
$DeIdentified : (DATA \cup IMAGE) \rightarrow BOOL$

$\forall p : PATIENT;\ d : DATA;\ i : IMAGE \mid$
$(d \in PatientData[p] \lor i \in PatientImages[p]) \bullet$
$(d \in Consent(p) \lor i \in Consent(p)) \Rightarrow$
$(DeIdentified(d) = TRUE \land$
$DeIdentified(i) = TRUE)$

E. Cohort Identification

As identified in the solution planning stage, mitigation of bias was necessary to ensure fairness based on the CHAI core principles. Data bias was minimized when identifying the patient cohort to be used as the dataset by ensuring balanced demographics with the available structured patient data. Patients were identified with Structured Query Language from a harmonized, de-identified database containing retrospective structured and unstructured clinical data. With the onset of prostate cancer typically occurring in men ages forty and older [10], criteria for inclusion were patients who identified as male, were forty and older, and had a prostate MRI examination completed at the healthcare organization. Potential outliers were identified who had prostate MRI exams but did not meet demographic criteria, although given the preliminary nature of the case study, only a small subset of patients could be used as part of the dataset.

The final dataset for the preliminary case study in Table 1 included 5 patients, 7 studies where studies are all the images acquired in a given imaging protocol, 47 T2-weighted imaging series where imaging series are the specific type of data captured by an imaging modality (in this case, MR) given a predetermined set of acquisition parameters, and 1181 images where an image is defined as one slice in an imaging series.

TABLE I
COHORT IDENTIFICATION

Patients	Studies	Series	Images
5	7	47	1181

In addition to metadata in Digital Imaging and Communications in Medicine images, personal health information may also be embedded in images themselves and should be removed to ensure patient privacy [8]. Once images were identified as relevant based on the inclusion criteria, images were de-identified.

F. Annotation

With the de-identified prostate MR images, abnormalities were annotated by radiologists to ensure the CHAI principles of reliability and accuracy. Abnormalities were defined as markedly hypointense or a potential area of extraprostatic extension. For imaging data, the region of interest is typically labeled either by 1) marking the approximate centroid of the target; 2) drawing a bounding box around the target; or 3) drawing the contour of the target (pixel-based annotation) depending on the modeling task. Annotation with the prostate MR images consisted of manually drawing binary mask contours (Method 3) to overlay the specific location of the abnormality.

IV. Scenarios

In order to demonstrate the method in action for the detection system, we examined three scenarios that may occur in which formal specifications can assist in addressing the challenges to implementation outlined by [4]: managing conflicts between the CHAI ethical principles and translating the ethical principles into design requirements.

1) Guiding Trade-off Decision-Making: This scenario is based on situations where trade-offs for ethical principles in the detection system's design must be evaluated and managed. If a trade-off between fairness and efficacy of system performance occurs (e.g. improving fairness reduces a metric such as accuracy), the formal specification for fairness can be expanded to define a directive where fairness is given priority unless accuracy falls below a defined acceptable threshold. Making decisions for trade-offs is then transparent and not left to subjective judgment.

$$
\begin{array}{|l}
\textit{FairnessEquity}_1 \\
[\textit{USER}, \textit{ATTRIBUTE}] \\
\textit{ProtectedAttribute} : \textit{USER} \pfun \textit{ATTRIBUTE} \\
\textit{Treatment} : \textit{USER} \fun \textit{TREATMENT} \\
\textit{Accuracy} : \mathbb{R} \\
\textit{AccuracyThreshold} : \mathbb{R} \\
\hline
\forall x, y : \textit{USER};\ a, b : \textit{ATTRIBUTE} \bullet \\
(x \in \operatorname{dom} \textit{ProtectedAttribute} \land y \in \operatorname{dom} \\
\textit{ProtectedAttribute} \land \\
\textit{ProtectedAttribute}(x) = a \land \textit{ProtectedAttribute}(y) \\
= b \land a \neq b) \Rightarrow \\
(\textit{Accuracy} \geq \textit{AccuracyThreshold} \Rightarrow \\
\textit{Treatment}(x) = \textit{Treatment}(y)) \\
\hline
\end{array}
$$

2) Constraints for Ethical Principles: In the scenario that patient privacy conflicts with usability, constraints within the formal specification for privacy could be made where no personal health information can be exposed, even if it reduces system efficiency. This constraint ensures that usability can be optimized within the limits of protecting patient privacy. Additional constraints can be defined such as for any abnormality detected, the system must generate an explainable output that highlights the specific regions of the image that led to its conclusion.

$$
\begin{array}{|l}
\textit{PrivacySecurity}_1 \\
[\textit{PATIENT}, \textit{DATA}, \textit{IMAGE}] \\
\textit{PatientData} : \textit{PATIENT} \rightarrow \textit{DATA} \\
\textit{PatientImages} : \textit{PATIENT} \rightarrow \textit{IMAGE} \\
\textit{Consent} : \textit{PATIENT} \rightarrow \mathbb{P}(\textit{DATA} \cup \textit{IMAGE}) \\
\textit{DeIdentified} : (\textit{DATA} \cup \textit{IMAGE}) \rightarrow \textit{BOOL} \\
\textit{Efficiency} : \textit{BOOL} \\
\hline
\forall p : \textit{PATIENT};\ d : \textit{DATA};\ i : \textit{IMAGE} \mid \\
(d \in \textit{PatientData}[p] \lor i \in \textit{PatientImages}[p]) \bullet \\
(d \in \textit{Consent}(p) \lor i \in \textit{Consent}(p)) \Rightarrow \\
(\textit{DeIdentified}(d) = \textit{TRUE} \land \textit{DeIdentified}(i) = \textit{TRUE}) \\
\land (\textit{Efficiency} = \textit{TRUE} \Rightarrow \\
\textit{DeIdentified}(d) = \textit{TRUE} \\
\land \textit{DeIdentified}(i) = \textit{TRUE}) \\
\hline
\end{array}
$$

3) Conditions for Prioritization: Our final example looks at the scenario where a priority structure can be embedded in a formal specification to mitigate conflicts between ethical principles and other considerations to provide criteria to make decisions. For example, in the typically non-critical setting of a prostate MRI exam, we can define patient safety as the top priority, followed by efficacy, resource efficiency, and timeliness. Since the scenario is routine, the detection system can take longer to process images, ensuring accuracy while conserving resources. If the setting was critical, however, timeliness could take greater priority.

$$
\begin{array}{|l}
\textit{SafetyReliability}_1 \\
[\textit{DATASET}, \textit{PRIORITY}] \\
\textit{Priority} : \operatorname{seq} \textit{PRIORITY} \\
\textit{Safety}, \textit{Efficacy}, \textit{Efficiency}, \textit{Timeliness} : \textit{PRIORITY} \\
\textit{TrainedOn} : \mathbb{P}\, \textit{DATASET} \\
\textit{ValidatedOn} : \mathbb{P}\, \textit{DATASET} \\
\textit{Diverse} : \textit{DATASET} \rightarrow \textit{BOOL} \\
\textit{Representative} : \textit{DATASET} \rightarrow \textit{BOOL} \\
\textit{AccuracyEnsured} : \textit{BOOL} \\
\textit{BiasMinimized} : \textit{BOOL} \\
\hline
\forall d : \textit{DATASET} \bullet (d \in \textit{TrainedOn} \lor d \in \textit{ValidatedOn}) \\
\Rightarrow (\textit{Diverse}(d) \land \textit{Representative}(d)) \\
\textit{AccuracyEnsured} = \textit{TRUE} \\
\textit{BiasMinimized} = \textit{TRUE} \\
\textit{Priority} = \langle \textit{Safety}, \textit{Efficacy}, \textit{Efficiency}, \textit{Timeliness} \rangle \\
\hline
\end{array}
$$

4) Remark on the Scenarios: Our example scenarios are of course only fragments of a complete integration into the clinical workflow and in each case we have chosen only a single property to demonstrate different ways the formal specifications can be adapted. A full formal verification of these specifications would want to examine the entire system, different scenarios, and edge cases, however, our aim has been only to demonstrate how formal specifications for reasoning about ethical concerns can be utilized to address challenges to the implementation of ethical guidelines, specifically in a healthcare context.

V. Discussion

The formal specifications assisted in understanding and implementing the CHAI ethical guidelines in the design and preparation of the prostate abnormality detection system. With this scope, the ethical principles primarily focused on were usefulness, usability, and efficacy in solution planning, fairness and equity in cohort identification, privacy and security in

de-identification of patient data, and safety and reliability in annotation.

Furthermore, the study highlighted the need to address potential sources of bias throughout the AI lifecycle, including data bias and automation bias. Particularly in the high-risk setting of diagnostic radiology, it is crucial to mitigate these risks. Identifying bias in the solution planning stage and the utilization of structured patient data to identify MR images relevant to the detection task enabled understanding of performance in varying populations.

Given the preliminary nature of the case study, the primary limitation was scoping–the CHAI ethical guidelines were unable to be implemented in the modeling and deployment stages of the AI lifecycle. Another major challenge in evaluating the implementation of the ethical guidelines in the case study was the absence of clearly defined measures of success, as also noted by [4]. Moreover, there are obviously limitations to what formal specifications can tell us, particularly since many simplifications are involved. Further work is needed to establish clear measures of success for implementation of ethical guidelines, and formally verify the specifications.

VI. Conclusion

The CHAI ethical guidelines were implemented in a preliminary case study for a prostate abnormality detection system. We addressed the challenges to implementation of ethical guidelines by defining base case formal specifications for the CHAI principles, and examined scenarios in which the specifications could be further extended to guide trade-off decision making and establish constraints for the principles. Using these specifications may thus assist in ensuring the ethical use of AI systems in medical imaging workflows, although further work is needed to evaluate its impact in clinical practice.

References

[1] 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.

[2] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI—Two Frameworks for Ethical Design Practice," IEEE Transactions on Technology and Society, vol. 1, no. 1, pp. 34–47, Mar. 2020, doi: 10.1109/TTS.2020.2974991.

[3] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nat Mach Intell, vol. 1, no. 9, pp. 389–399, Sep. 2019, doi: 10.1038/s42256-019-0088-2.

[4] M. Goirand, E. Austin, and R. Clay-Williams, "Implementing Ethics in Healthcare AI-Based Applications: A Scoping Review," Sci Eng Ethics, vol. 27, no. 5, p. 61, Sep. 2021, doi: 10.1007/s11948-021-00336-3.

[5] N. Economou, M. Elmore, A. Callahan, J. McCall, and R. Baig, "Assurance Standards Guide Coalition for Health AI (CHAI)," CHAI, https://chai.org/assurance-standards-guide/ (accessed Aug. 2, 2024).

[6] F.-J. H. Drost et al., "Prostate MRI, with or without MRI-targeted biopsy, and systematic biopsy for detecting prostate cancer," Cochrane Database of Systematic Reviews, vol. 2019, no. 4, Apr. 2019, doi: 10.1002/14651858.CD012663.pub2.

[7] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, "An overview of clinical decision support systems: benefits, risks, and strategies for success," npj Digit. Med., vol. 3, no. 1, pp. 1–10, Feb. 2020, doi: 10.1038/s41746-020-0221-y.

[8] B. Kocak et al., "Bias in artificial intelligence for medical imaging: fundamentals, detection, avoidance, mitigation, challenges, ethics, and prospects," Diagnostic and interventional radiology (Ankara, Turkey), Jul. 2024, doi: 10.4274/dir.2024.242854.

[9] J. M. Spivey, "An introduction to Z and formal specifications," Software Engineering Journal, vol. 4, no. 1, p. 40, 1989. doi:10.1049/sej.1989.0006

[10] P. H. Gann, "Risk Factors for Prostate Cancer," Rev Urol, vol. 4, no. Suppl 5, pp. S3–S10, 2002.