Received 13 December 2022
Revised 24 April 2023
Accepted 17 May 2023

# Application of ethical AI requirements to an AI solution use-case in healthcare domain

Zohreh Pourzolfaghar
*School of Business, National University of Ireland Maynooth, Maynooth, Ireland, and*
Marco Alfano and Markus Helfert
*Innovation Value Institute, National University of Ireland Maynooth, Maynooth, Ireland*

## Abstract

**Purpose** – This paper aims to describe the results of applying ethical AI requirements to a healthcare use case. The purpose of this study is to investigate the effectiveness of using open educational resources for Trustworthy AI to provide recommendations to an AI solution within the healthcare domain.

**Design/methodology/approach** – This study utilizes the Hackathon method as its research methodology. Hackathons are short events where participants share a common goal. The purpose of this to determine the efficacy of the educational resources provided to the students. To achieve this objective, eight teams of students and faculty members participated in the Hackathon. The teams made suggestions for healthcare use case based on the knowledge acquired from educational resources. A research team based at the university hosting the Hackathon devised the use case. The healthcare research team participated in the Hackathon by presenting the use case and subsequently analysing and evaluating the utility of the outcomes.

**Findings** – The Hackathon produced a framework of proposed recommendations for the introduced healthcare use case, in accordance with the EU's requirements for Trustworthy AI.

**Research limitations/implications** – The educational resources have been applied to one use-case.

**Originality/value** – This is the first time that open educational resources for Trustworthy AI have been utilized in higher education, making this a novel study. The university hosting the Hackathon has been the coordinator for the Trustworthy AI Hackathon (as partner to Trustworthy AI project).

**Keywords** Artificial intelligence (AI), Healthcare, Ethical requirements, AI solutions

**Paper type** Case study

## Introduction

Artificial intelligence (AI) as a science which studies and develops the theory, methodology, technology and application systems for simulating, extending and expanding human intelligence had revolutionary effect on contemporary human society (Liu *et al.*, 2021). In recent years, AI systems have been developed in ways inconsistent with the proclaimed values of their developers. This has increased concern, research and activism regarding the effects of AI systems (Whittaker *et al.*, 2018; Crawford *et al.*, 2019). As stated by Brundage *et al.* (2020), there is a growing concern about how to ensure that development and deployment of AI is beneficial to humanity, in light of AI's rapid technical advancement and proliferation of AI-based applications in recent years.

In this manner, prominent institutions across the political, commercial and academic strata of society have created ethics guidelines for trustworthy AI (Floridi and Cowls, 2021).

American Journal of Business
Vol. 38 No. 3, 2023
pp. 112-128
Emerald Publishing Limited
1935-5181
DOI 10.1108/AJB-12-2022-0201

In April 2018, in response to a request from the European Council to present a European approach to AI, the European Commission presented its AI strategy in the "Artificial Intelligence for Europe" [1]. Then High-Level Expert Group on AI (AI HLEG) presented Ethics Guidelines for Trustworthy Artificial Intelligence (Smuha, 2019). This guideline introduced a set of seven essential requirements that AI systems must satisfy. However, the industry lacks the tools and incentives to translate high-level ethics principles to verifiable and actionable criteria for designing and deploying AI (Raji *et al.*, 2020).

In response to this challenge, the "Trustworthy AI" Erasmus + project aimed to facilitate the introduction of the High-Level Expert Group's Guidelines on Trustworthy AI [2] into Higher Education across disciplines. This project's objective was to utilize EU Ethical requirements for the introduction of ethical and socio-legal competencies in AI-related Higher Education topics. In order to achieve this objective, the Trustworthy AI project developed a framework that describe the principles and learning strategies that must be used to develop students' competencies. The project's first intellectual output consists of recommendations for educators, educational materials requirements and policy incentives. The second intellectual output of this project was creation of open educational resources to enhance the AI-related knowledge and skills of Higher Education Institutes (HEIs) students. As a partner in this project, our university was responsible for coordinating the Hackathon in the three universities partners to the project.

This paper seeks to present the outcomes of the Hackathon for the healthcare use case developed by a university research project. As the result of this study a prototype for the healthcare domain has been developed to empower the patients.

The remainder of this paper is organized as follows. In the subsequent section (Section 2), the selected research methodology for this study is described. In section 3, there will be a concise summary of the "*Ethical AI Hackathon*" plan, including the recruitment of team members/mentors and the materials. Section 4 will then introduce the healthcare use case. As a consequence of the outcomes of the first and second days of the Hackathon, a summary of the AI-related challenges identified by the winning team will be provided and analysed (Section 5). In section 6, a framework of recommendations to the CHAPE AI solution will be presented. In section 7 the recommendations will be evaluated based on predetermined criteria of the Hackathon.

## Section 2: Research method

Hackathon methodology has been used as the research technique for this study. According to Gama *et al.* (2018), Hackathons are short events in which participants have a common goal. As Maaravi (2020) stated Hackathons is one methodology of experiential learning that is becoming more and more common to enhance student and employee learning and motivation. Experiential learning is widely accepted as an effective approach for quality learning and education (Kolb and Kolb, 2005). Similarly, Towhidi and Pridmore (2022) argued that Hackathons are an effective experiential learning tool for higher education to help develop soft skills and to prepare students for the job market. They stressed that Hackathon events help students apply their knowledge and skills in real-world settings and develop the required technical and soft skills for industry needs.

In this study, the authors intended to investigate efficacy of students' educational resource utilization. As Nandi and Mandernach (2016) emphasized, collegiate Hackathons feature peer learning. In addition, Warner and Guo (2017), who focused their research on collegiate Hackathons, validated the importance of learning from peers. Also, Hackathons are an excellent illustration of an educational technique that combines the practical, contextual and social parts of this modern pedagogical paradigm into a compelling learning experience (Rys, 2021).

The second objective for this research was to assess the efficacy of applying the educations to actual use cases. As Mtsweni and Abdullah (2015) highlighted Hackathons are used to bring together students and experts to develop software focused on socially relevant challenges. To attain the second objective, each Hackathon team comprised of an academic mentor and four to five students. In this investigation, the application domain was healthcare. During the first and second days of the Hackathon, students learned about ethical requirements for trustworthy AI by utilizing the provided educations resources. Then, they made recommendations for the healthcare use case based on the acquired knowledge. This use-case was chosen since the research team responsible for developing this prototype was headquartered at our university. The research team for the healthcare use case participated in the Hackathon by introducing the use case and analysing and evaluation of Hackathon teams recommendations.

## Section 3: Hackathon plan

The EU Ethics Guidelines for Trustworthy AI outline four ethical principles f trustworthy AI, including: (1) Respect for Human Autonomy, (2) Prevention of Harm, (3) Fairness and Explicability. From these principles, derives the seven key requirements have been derived that AI systems should consider. The requirements are: (1) Human agency and oversight: Including fundamental rights, human agency, and human oversight; (2) Technical robustness and safety: Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility; (3) Privacy and data governance: Including respect for privacy, quality and integrity of data, and access to data; (4) Transparency: Including traceability, explainability and communication; (5) Diversity, non-discrimination, and fairness Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation; (6) Societal and environmental wellbeing: Including sustainability and environmental friendliness, social impact, society, and democracy; and (7) Accountability: Including auditability, minimization and reporting of negative impact, trade-offs, and redress. The plan for the user test included three main phases including:

(1) Familiarization Phase: The teams become acquainted with the available resources, such as AI card deck and knowledge clips,

(2) Introduce the healthcare use case and have mentors facilitate brainstorming sessions to identify potential challenges and

(3) Propose a framework of recommendations for the identified challenges and submission.

### *Recruiting teams/mentors*

Participants in the Hackathon were recruited from academic lectures and students. These two categories were selected in accordance with the recommendations. In this regard Mtsweni and Abdullah (2015) proposed bringing together students and experts to develop software focused on socially pertinent issues.

To satisfy the preceding structure, eight faculty members from the School of Business and School of Computing have joined the Hackathon. In addition, 24 students from our university and other universities in Europe, United States and were participating to the event. The lecturers attended a pre-event meeting in order to learn about the event. Additionally, there have been further informal conversations with individual lecturers. As a result of these dialogues the event's guideline were enhanced. This Hackathon was open to all students around the world. This is due to the fact that the open educational resources have been made openly available to everyone, and the ethical issues might be of interest to people all over the world.

*Resources*

To assist students and mentors in learning about the Hackathon, the following open educational resources have been developed. The teams had access to the resources over the course of Hackathon's three days. The resources consist of (1) Knowledge clips introducing Trustworthy AI and the seven Requirements for Trustworthy AI; (2) Trustworthy AI Card Deck; and (3) an Exercise on the seven steps towards Trustworthy AI Exercise. The teams have been using Padlet tool for their brainstorm sessions.

*Analysis of the results and selection of the winning team*

The results have been reviewed by a panel of three experts. Dr. Marco Alfano, the Research Fellow with expertise in the healthcare domain and the other two authors of this paper have served as a panellist. The panel have been reviewing the results of brainstorming sessions over the Padlet platform for each team. In addition, they have been analysing the reports containing teams' recommendations.

## Section 4: Healthcare use case

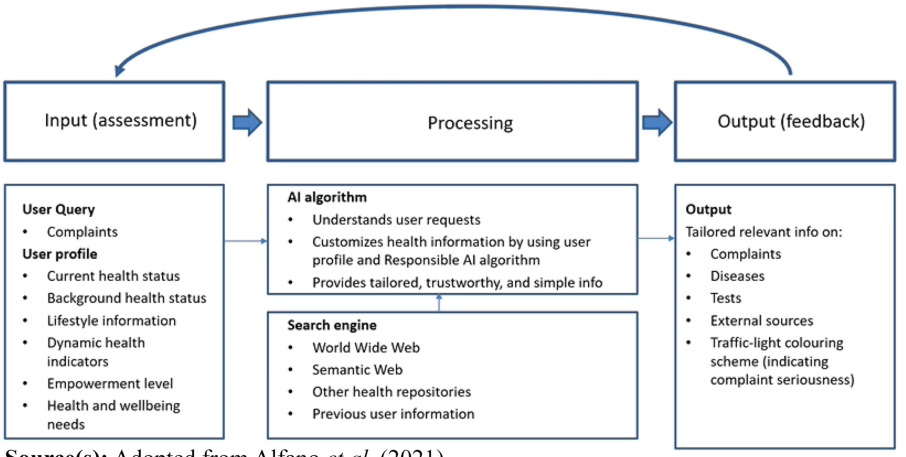
Empowerment is a process by which people acquire the knowledge and self-awareness necessary to comprehend their health conditions and treatment options in order to self-manage them and make informed decisions with healthcare professionals. Empowerment enables people/patients to communicate with medical professionals more effectively and ensures that care is provided in accordance with their needs, values and best interests. Responsible Intelligent Empowering Agents (RIEAs) employ Trustworthy AI to assist people/patients in the comprehending health information regarding specific complaints or health in general.

CHAPE is a Conversational Agent for Health Empowerment on Health and Wellbeing (Alfano *et al.*, 2021). CHAPE engages the user in natural language, collects health data from heterogeneous sources, and provides tailored, easily understood, and reliable information, thereby empowers users to create their own comprehensive and objective opinion on health matters of concern (see Figure 1). This use case was selected because the research team responsible for developing this prototype was headquartered at our university.

CHAPE is an application that raises knowledge and awareness on disorders and diseases and enables improved health management. CHAPE [3] has been developed in collaboration with a medical practitioner. However, it is intended solely for informational purposes only. It is not a substitute for professional medical advice, diagnosis or treatment provided by trained professional. In the form of an ethical framework, the participants to ethical AI Hackathon were required to propose how CHAPE can be modified and adapted to address each of the seven ethical requirements of Trustworthy AI in the form of an ethical framework.

## Section 5 – Recognized challenges

This section is a summary of the results for the brainstorming sessions. For the healthcare use case, participants had an hour-long session. The research fellow involved in the healthcare application has been providing detailed information on the problem that prompted the application's development and initial solution concept for the application ability to empower patients. Additionally, the healthcare expert introduced some challenges for patients and some challenges for AI in healthcare. The expert then described the entire process for CHAPE application, including inputs (from patients), input processing using AI algorithms and search engines, and outputs (feedback, recommendations and suggestions to patients). Participants were given the link to access the application (https://cohealth.ivi.ie/chape/). After this session the participants met their team members and mentors to identify




**Source(s):** Adopted from Alfano *et al*. (2021)

**Figure 1.**
Conversational agent for health empowerment on health and wellbeing

the potential challenges associated with the use of AI based on seven key requirements. The following table is a summary of the discussions in the form of queries pertaining to each requirement.

## Section 6 – Brainstorming results and recommendations

This section will introduce the outcomes of discussions and provide recommendations in response to the identified challenges. At all points throughout the decision-making process, the quote "AI systems should empower human beings allowing them to make informed decisions and fostering their fundamental rights," from the WHO guidelines "Ethics and Governance of Artificial Intelligence for Health", has been considered as a central tenet to how this type of framework should be developed. The following are decisions regarding the seven key requirements.

### *Discussions on requirements #1 – human agency and oversight*

According to the introduced Ethics guidelines by EU Commission [2] for trustworthy AI, this requirement expresses that: "*AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.*" The following are the outcomes of the winning group's brainstorming sessions.

The healthcare market for AI expanded by 167.1% between 2019 and 2021 (Grand Review Research, 2019). While 40 million new health-sector jobs are expected to be created by 2030, 9.9 million physicians, nurses and midwives will be needed (EIT Health, 2021). For a medical system such as this, the impact of an erroneous diagnosis or incorrect treatment information on an end-user's decision making must be carefully considered. Even if the system makes it clear that it does not recommend seeking medical attention, it could potentially help them to determine if they need to seek medical attention.

It is necessary to have specialists involved in the process to prevent end-users from developing an unsafe level of reliance on the system. There must be opportunities for human intervention and verification throughout the decision-making process of the system. This can

take many forms, but it must always be simple to use, efficient and human decisions must always take precedence over those of AI. Human agency must always come first; otherwise, the AI system is not trustworthy.

*Discussions on requirements #2 – technical robustness and safety*

As it has been stated in the EU Ethics guidelines for trustworthy AI, "*AI systems need to be resilient and secure. They need to be safe, ensuring a fall-back plan in case something goes wrong, as well as being accurate, reliable, and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.*"

For this requirement the team determined that AI systems can be technically trusted if they attain a high and demonstratable level of technical robustness and safety (Chatila *et al.*, 2021). Additionally, they discussed potential attacks, such as adversarial attacks. Some methods, such as adversarial training have been created to defend against adversarial attacks. These methods could enhance the adversarial robustness of a model by incorporating adversarial sample into the training set (Li *et al.*, 2021). In addition, they discussed a particular application of an AI tool. It has been reported that stress testing against extreme cases or unusual environments monitoring increased prescriptions and other atypical decisions (IBM, 2021).

*Discussions on requirements #3 – privacy and data governance*

According to the introduced EU Ethics guidelines Privacy and data governance express that: "*besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.*"

In relation to this requirement, the team believed that any sensitive or identifying information about a user that is not required for the system's functionality should never be used. The data should be encrypted and anonymized to prevent any potential data breaches or leaks from causing damage to individuals whose data has been collected. There are already standards for how information should be processed, shared and stored in the medical field. These requirements should serve as the bare minimum for an AI system in the field. Anything less would be unacceptable, but more should be sought in order to increase confidence in a developing technology. Internal and external audits should be able to access detailed records of data sourcing, who has access to it and who has used it. Any recommendations made by the AI model, as well as the data used to generate those recommendations, should be saved. These documents are required to guarantee that all data used is stored correctly, securely and privately. Not only are these practices necessary, but they are also of little use if all parties involved in handling information management do not comprehend them and know how to follow them correctly. All developers, data scientists, medical personnel and other relevant service providers and evaluators must therefore receive training in data quality, governance and cybersecurity. Gallagher (2022) reports that the attackers sent a malicious email to a workstation. This granted access to HSE systems to the intruders. A few days later, the HSE antivirus software detected activity but was unable to suppress it. This example illustrates the significance of data privacy by demonstrating how it could be compromised.

*Discussions on requirement #4 – transparency*

As it has been explained in EU Ethics guideline for Trustworthy AI, Transparency is: "*the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner*

*adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system and must be informed of the system's capabilities and limitations.*"

Related to this requirement, the team discussed the significance of clearly communicating any variable used to influence the system's output. Whether a simple regression model or a deep learning model was employed, this information should be accessible to all users. The purpose of the system, i.e. empowerment individuals to improve the medical skills of the general population, should be stated explicitly. In addition, they believed it necessary to provide explicit visualizations of how the weight of each user input (personal info, medical history and current symptoms) contributes to the decision.

Juravle *et al.* (2020) investigated the possibility of increasing confidence in AI diagnoses by informing the participants that AI outperforms the human physicians and nudging them to favour AI diagnoses when choosing between AI and human doctors. The results of these experiments indicated a general decrease in trust in AI and in its ability to diagnose diseases with high risk. Participants were less inclined to trust a second opinion from an AI doctor regarding high-risk diseases. The results of this experiment demonstrated that individuals are informed that AI outperforms the human doctors (Juravle *et al.*, 2020).

*Discussions on requirements #5 – diversity, non-discrimination and fairness*

As stated in EU Ethics guideline for Trustworthy AI Diversity, non-discrimination and fairness says that: "*Unfair bias must be avoided, as it could have multiple negative implications, from the marginalisation of vulnerable groups to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle*".

For this requirement, the following has been discussed during the brainstorming session. The team determined that any dataset used to train a machine learning algorithm must endure extensive analysis to ensure its objectivity. The team stated that there are limitations to the data currently collected in the Ireland healthcare system (e.g. patients aged 18 and older, historical data), and these data are only accessible for Irish healthcare services. They believed that during the design phase, consideration should be made for people with disabilities, special needs or who are at the risk of exclusion, such as colour palettes that are accessible to colourblind users, variable text size and font, text to speech services. According to a report by The National Institute for Health Care Management Foundation (NIHCM, 2021), it has been suggested to develop a mechanism to include the participation of end users, physicians, AI experts/development.

*Discussions on requirements #6 – societal and environmental well-being*

As stated in EU Ethics guideline Societal and environmental well-being means: "*AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered*".

In relation to this requirement, the team discussed a variety of topics concerning the propriety and confidence of the patients' responses. The team argued that probabilities and level must be communicated in application output. For instance, if the app is uncertain about a diagnosis, it must request additional information from the user. In addition, they recommended utilizing A/B testing on critical aspects of the user interface to provide detailed feedback on the clarity of system's communication of recommendations. The team believed that provisions should be made to procure the materials and energy required to store the data in a sustainable manner. In conclusion, they emphasized that the deployment of an AI

solution is anticipated to have a smaller environmental impact than the equivalent amount of human labour.

*Discussions on requirements #7 – accountability*

According to the EU Ethics guideline for trustworthy AI, Accountability explains that "*Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured*".

The following is the discussion regarding this requirement. The team believed that a system requiring sensitive information and dealing with sensitive topics such as medical information must be held to a high accountability standard. A representative sample of stakeholders should be involved in an iterative process that provides feedback on model recommendations, according to their statement. Priority should be given to the individual and their rights when weighting compromises. In a decision between model efficacy and data privacy, this should favour data privacy. In addition, they emphasized that routine risk assessments for misdiagnosis, overreliance on the system and insufficient system robustness must be conducted. The argument was that if there is a lack of confidence in any aspect of the system's functionality or if a risk assessment identifies a significant risk, the system cannot continue to be accessible until these issues are resolved.

*Summary of recommendations*

Table 1 summarizes the recommendations for the CHAPE application regarding the seven requirements:

## Section 7 – Analysis of the recommendations

The participants in the ethical AI Hackathon have been given access to a variety of resources and guidance. In one of the resources, there have been number of guiding questions. During brainstorming sessions, these questions have been utilized by the team members. The outcomes of these sessions were identified obstacles for the CHAPE application for each question (see Table 2). Subsequently, the team members held additional sessions to develop recommendations. In this section, we will introduce the guiding questions and then discuss the recommendations for addressing CHAPE's identified challenges.

*Recommendations for requirements #1 – human agency and oversight*

The guiding questions [4] for this requirement have been: "Does the AI system enhance or augment? Is this AI system human-centric? Does it leave meaningful opportunity for human choice? Does it enable individuals to have more control over their lives or does it limit their freedom and autonomy?" Considering these guiding questions, the team members have recognized three specific challenges facing CHAPE application.

In response to the first question, the winning group provided evidence from healthcare-related national reports. Given that statistics indicate a shortage of professionals. The response explains that a shortage of healthcare providers anticipated. The team discussed the possibility that this deficiency may result in additional misdiagnoses or misleading patient information. For this area they have suggested that CHAPE (healthcare application), "*consider how a false diagnosis could affect the end user's decision.*" To answer the second question the team determined that the involvement of professionals is necessary so that patients do not rely solely on their own judgement. Regarding the second question, their recommendation was as follows: "*Keep humans in the loop to prevent overconfidence in or*

| # | Requirements | Recommendations |
|---|---|---|
| 1 | Human agency and oversight | • Consider how a false diagnosis could affect the end user's decision making<br>• Keep humans in the loop to prevent overconfidence in or overreliance on the AI system<br>• Ensure there are opportunities for human intervention and verification in the system's decision process |
| 2 | Technical robustness and safety | • To consider extreme cases or unusual environments<br>• Monitoring increased prescriptions and other anomalous decisions<br>• Log of procedures and decisions made<br>• Try to approach accuracy achieved by junior medical professionals<br>• Consider the (human) cost of an incorrect medical recommendation |
| 3 | Privacy and data governance | • If including sensitive/proprietary patient data does not significantly improve the effectiveness of the system, it should not be used<br>• Any of this confidential data that is used should be encrypted and anonymized<br>• Keep detailed logs of data sourcing and access/use, as well as citing info in the recommendations given by the model<br>• Provide the relevant data quality, governance and cybersecurity training for developers, data scientists, medical personnel and other service evaluators |
| 4 | Transparency | • The importance of any variables used to inform the system's output should be communicated clearly<br>• The purpose of the system, people empowerment to enhance the medical skills of the general populace, should be clearly stated<br>• Provide clear visualizations of how the strength of each user input (personal info, medical history, current symptoms) contribute to the decision |
| 5 | Diversity, non-discrimination, and fairness | • Any dataset used to train a machine learning involved must undergo thorough study to ensure it is unbiased<br>• Consider some limitations to the datasets, (e.g. patients 18+, historical data), or make them available them only for specific organizations<br>• Considerations should be made at the design phase for people with disabilities, special needs or who are at the risk of exclusion, e.g. colourblind friendly palettes, adjustable text size and font, text to speech service<br>• Develop a mechanism to incorporate the involvement end users, doctors, AI experts/development |
| 6 | Societal and environmental wellbeing | • Ensure probabilities/confidence in model output is communicated. If it is unsure of any diagnosis, request more info from the user<br>• A/B testing on critical aspects of the UI to give rich feedback on the clarity of communicating suggestions made by the system<br>• Provision should be made to sustainably source the materials and energy required to store the data and host the system on a server<br>• The deployment of an AI solution is envisaged have a lower environmental impact than the human labour equivalent |

**Table 1.**
Summary of recommendations to CHAPE

(*continued*)

| # | Requirements | Recommendations |
|---|---|---|
| 7 | Accountability | • Design an iterative process whereby trained evaluators (representative sample of stakeholders) provide feedback on model recommendations<br>• Consider the trade-offs to be made:<br>○ Model performance vs data privacy<br>○ Disease severity and available medical resources<br>• Routine risk assessments for o end user's misdiagnosis<br>○ Medical professionals-confirmation bias (over reliance on the system)<br>○ Developers of the AI system lack of robustness/stress testing of model |

**Source(s):** Table created by author

**Table 1.**

*overreliance on the AI system*". In order to provide a response to the question, the team stated that professional intervention opportunities should be considered. With this question in mind, their recommendation was: "Ensure there are opportunities for human intervention and verification in the system's decision process".

Based on the responses from the winning team to the guiding questions, it appears that they had a clear understanding of certain aspects of this requirement from the national reports (in Ireland), such as the necessity of AI systems in healthcare domain, and the significance of professional interventions. However, additional research is required to address the first challenge identified in relation to the effects of AI systems on human autonomy. Regarding the professional interventions, additional research is suggested to determine if the current version of CHAPE includes the recommended capabilities.

*Recommendations for requirements #2 – technical robustness and safety*

The guiding questions for this requirement have been including: "*Can you identify any potential forms of attacks which the AI-system could be vulnerable to? Is there a probable chance that the AI-system may cause damage or harm to users or third parties?*"

In order to answer the guiding questions for this requirement, the team members have been discussing potential attacks and countermeasures by referencing to research works. In addition, based on discussion on a particular stress test case, they provided CHAPE with the following recommendation: "Consider extreme cases or unusual environments." The second question has been about "Fallback plan and general safety" asking: "Is there a probable chance that the AI-system may cause damage or harm to users or third parties?" To answer this question, the team has utilized the same case. Referring to this instance, they suggested "*to monitor increased prescriptions and other anomalous decisions*". Further recommendations in this regard have been: "Keep log of procedures and decisions made (e.g. if a patient has had multiple X rays this year)"; "Try to approach accuracy achieved by junior medical professionals"; and "Consider the (human) cost of an incorrect medical recommendation".

In conclusion, the recommendations for the first guiding question have received stronger scientific backing. The majority of suggestions for the second set of guiding questions have been mostly based on the individual perceptions of the team member. For this area, the CHAPE developers were advised to conduct additional research.

*Recommendations for requirements #3 – privacy and data governance*

The guiding questions for this requirement have been: Are there ways to develop the AI-system or train the model without or with minimal use of potentially sensitive or personal

| # | Requirements | Challenges |
|---|---|---|
| 1 | Human agency and oversight | Could the AI system affect human autonomy by interfering with the (end) user's decision-making process in an unintended way?<br>Did you take safeguards to prevent overconfidence in or overreliance on the AI system for work processes?<br>Who is the "human in control" and what are the opportunities for human intervention? |
| 2 | Technical robustness and safety | Did you verify how your system would react in unexpected situations or environments?<br>Did you ensure that your system has a sufficient fallback plan in the case of adversarial attacks or other unexpected situations?<br>Did you assess whether there is a probable chance that the AI system may cause damage, or harm to users or third parties? Did you assess the likelihood, potential damage, impacted audience and severity?<br>Did you assess what level and definition of accuracy would be required in the context of the AI-system and use case?<br>Did you verify what harm would be caused if the AI-system makes inaccurate predictions? |
| 3 | Privacy and data governance | Did you consider the ways to develop the AI system or train the model without or with minimal use of potentially sensitive or personal data?<br>Did you take measures to enhance privacy, such as encryption, anonymization and aggregation?<br>Did you establish oversight mechanisms for data collection, storage, processing and use?<br>Did you ensure that people working with data are qualified and required to access the data, and that they have the necessary competencies to understand the details of data protection policy?<br>Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed? |
| 4 | Transparency | Did you ensure an oversight mechanism to log when, where, how, by whom and for what purpose data was accessed?<br>Why was this particular system deployed in this specific area?<br>Did you establish mechanisms to inform (end-)users on the reasons and criteria behind the AI system's outcomes? |
| 5 | Diversity, non-discrimination, and fairness | Did you assess and acknowledge the possible limitations stemming from the composition of the used data sets?<br>Did you assess whether the AI system usable by those with special needs or disabilities or those at risk of exclusion? How was this designed into the system and how is it verified?<br>Did you assess whether there could be persons or groups who might be disproportionately affected by negative implications?<br>Did you consider a mechanism to include the participation of different stakeholders in the AI system's development and use? |
| 6 | Societal and environmental wellbeing | Did you assess whether the AI-system encourages humans to develop attachment and empathy or vice versa?<br>Did you assess whether the logic of the AI-system might simplify and polarize public discourse?<br>Did you assess whether the AI-system could be used to manipulate or confuse people?<br>Did you ensure measures to reduce the environmental impact of your AI-system's life cycle?<br>Did you establish mechanisms to measure the environmental impact of the AI-system's development, deployment, and use (for example the type of energy used by the data centres)? |

**Table 2.**
Summary of the recognized challenges for the use of AI in CHAPE application

(*continued*)

| # | Requirements | Challenges |
|---|---|---|
| 7 | Accountability | Did you establish mechanisms that facilitate the system's auditability, such as ensuring traceability and logging of the AI-systems processes and outcomes?<br>Did you carry out a risk or impact assessment of the AI-system, which considers different stakeholders that are (in)directly affected?<br>How do you decide on trade-offs between ethical principles? Did you ensure that the trade-off decision was documented?<br>Did you establish an adequate set of mechanisms that allows for redress in the case of the occurrence of any harm or adverse impact? |

**Source(s):** Table created by author

**Table 2.**

data? Can you think of oversight mechanisms for data collection, storage, processing and use? What protocols, processes and procedures can you think of to manage and ensure proper data governance? Who should be allowed to access users' data and under what circumstances?

To provide appropriate answers to the guiding questions the team argued that sensitive data should be stored in the system. The study also suggested that anonymizing the patients' information is advantageous when this information must be stored in the system. In this regard, the recommendation is as follows: "*If including sensitive/proprietary patient data does not significantly improve the effectiveness of the system, it should not be used*", and "*any of this confidential data that is used should be encrypted and anonymised"*. Regarding the questions about processes and protocols for data governance they recommended: "To keep detailed logs of data sourcing and access/use". Further recommendation to this area has been: "*To provide the relevant data quality, governance and cybersecurity training for developers, data scientists, medical personnel, and other service evaluators"*.

In brief, the recommendations in this area have been based on team members' knowledge about other systems, such as the privacy settings in various organizations, their shared sentiments and more likely their own personal experience. They were attempting to recommend to CHAPE the common parameters that should be incorporated into any system that handles personal information. The suggestion for CHAPE is to ensure that this application incorporates all of these features.

*Recommendations for requirement #4 – transparency*

The guiding questions for this requirement have been: What mechanisms could you establish that facilitate the system's auditability, such as ensuring traceability and logging of its processes and outcomes? Did you review the outcomes or decisions taken by the system, as well as potential other decisions that would result from different cases? Can you explain why the system will make a certain choice in a way that is understandable for all users? What mechanisms can you put in place to inform (end-)users on the reasons and criteria behind the AI system's outcomes? What is the exact purpose of your AI-system and who or what may benefit from it? Can you specify usage scenarios for the system and clearly communicate them to ensure that the system is understandable and appropriate for the intended audience?

Taking into account the guiding questions, the team argued that appropriate variables must be incorporated into the AI diagnosis process in CHAPE in order to effectively communicate the output to the end-users. In this regard, the team provides the following recommendation: "Communicate system's output to the end-users clearly." To respond the following questions, they provided research evidence about the decision making by patients.

In addition, they emphasized that the main aim of CHAPE – which is to empower patients – must be communicated to the users. In this regard, the recommendation has been as follows: "*The purpose of the system, people empowerment to enhance the medical skills of the general populace, should be clearly stated.*" Thus, users would have a better understanding of why they are required to provide certain inputs to the system (such as confidential information or medical history). "*Provide clear visualisations of how the strength of each user input (personal info, medical history, current symptoms) contribute to the decision,"* was another recommendation in this regard.

In summary, the recommendations for this requirement have provided CHAPE with suggestions for enhancing the application credibility. However, there are some unexplored mechanisms for informing end-users of the rationale and criteria underlying AI systems. The CHAPE application's developers need to conduct additional research into this subject.

*Recommendations for requirements #5 – diversity, non-discrimination and fairness*

The guiding questions for this requirement have been: "*Whether there could be persons or groups who might be disproportionately affected by negative implications? Whether the AI system is useable by those with special needs or disabilities or those at risk of exclusion? How can this be designed into the system and how can it be verified? Can you think of ideas to include the participation of different stakeholders in the AI system's development and use?*"

In response to the guiding questions, the team identifies data-sharing restrictions within the Irish healthcare system. In this regard, the recommendation is as follows: "*Any dataset used to train a machine learning involved must undergo thorough study to ensure it is unbiased.*" and "Consider some limitations to the datasets, (e.g. patients aged 18 and older, historical data), or make them available them only for specific organisations". In addition, they suggested conducting research on the methods to reduce bias in the processes used to analyse datasets. There have been some suggestions to reduce the risk of excluding individuals with special needs. The recommendation for this area has been: "*Considerations should be made at the design phase for people with disabilities, special needs or who are at the risk of exclusion.*" The other suggestion was to involve multiple stakeholders. The recommendation in this area has been: "*To develop a mechanism to incorporate the involvement end users, doctors, AI experts/developers*.

In a nutshell, the majority of the guiding questions have been addressed by the recommendations for this requirement. However, these recommendations lacked scientific evidence. Regarding this requirement, CHAPE is advised to investigate the application processes and procedures in greater depth, per the provided recommendations.

*Recommendations for requirements #6 – societal and environmental well-being*

The guiding questions for this requirement have been: "*Is there a risk of job loss or deskilling of the workforce? What steps to been taken to counteract such risks? Whether the logic of AI might simply and polarise public discourse? Whether the AI-system could be used to manipulate or confuse people? What mechanisms could you establish to measure the environmental impact of the AI-system's development, deployment, and use? What measures can you think of that can reduce the environmental impact of your AI-system's life cycle?*"

As a response to the guiding questions, the team provided some recommendations regarding the end-users' confidence level and other requirements. In this regard, they suggested CHAPE to make probabilities/confidence of the feedback obvious to users. The recommendation in this area has been as follows: "*Ensure probabilities/confidence in model output is communicated. If it is unsure of any diagnosis, request more info from the user.*" They also recommended deploying A/B testing methods which are useful for evaluating numerous aspects of evaluation, such as the promotion of positive outcomes, minimizing unintended

consequences on safety and poor user experience (Austrian *et al.*, 2021). The recommendation in this regard was "A/B testing on critical aspects of the UI to give rich feedback on the clarity of communicating suggestions made by the system". Additional recommendations have been made to make the CHAPE application more environmentally friendly and sustainable. These recommendations have been stated in general manner.

In brief, the recommendations for this requirement included a number of suggestions for the guiding questions. However, additional concerns raised regarding the negative effects of AI solutions in healthcare domain (e.g. reduction in employment opportunities and possible confusion caused by AI solutions). The recommendation for CHAPE developers is to conduct a thorough investigation of the application to ensure that these issues will not negatively affect end-users.

*Recommendations for requirements #7 – accountability*

The guiding questions for this requirement have been: "*What mechanisms could you establish that facilitate the system's auditability, such as ensuring traceability and logging of its processes and outcomes? What are the relevant interests and values impacted by the AI-system? What are the potential trade-offs between them? How do you decide on such trade-offs? What mechanisms can you establish to allow for redress in case of the occurrence of any harm or adverse impact?*"

In order to provide appropriate responses to the guiding questions, the team discussed a variety of topics such as the sensitivity of information in healthcare domain and significance of receiving feedback on the provided recommendations. In this area the team provided some recommendations, such as "design an iterative process whereby trained evaluators provide feedback on model recommendations" and "consider the trade-offs to be made, e.g. for model performance vs data privacy". The last recommendation as the final remedy for potential damages or adverse effects was: "development of a routine risk assessments for end user's misdiagnosis, medical professionals-confirmation bias, and for developers of the AI system lack of robustness/stress testing of model".

To summarize the recommendations for this requirement, the team proposed processes for the continuous evaluation of user feedback and the mitigation of potential damages. Nonetheless the CHAPE developers' research group must conduct research on other essential aspects of this requirement, including ensuring auditability and traceability.

## Conclusion

In recent years, AI systems have been developed in ways inconsistent with the values of their creators. This prompted some concerns regarding the effects of AI systems. Among these concerns is how to assure that the development and deployment of AI will be beneficial. In response to the urgency of these concerns, the world's foremost institutions have developed ethical guidelines for trustworthy AI. In 2019, the European Commission released Ethics Guidelines for Trustworthy Artificial Intelligence. However, the industry lacks necessary instruments and incentives to translate high-level ethical principles into verifiable and actionable criteria for designing and deploying AI systems. To address this challenge, "Erasmus + Trustworthy AI project" aimed to use of EU Ethical requirements for the introduction of ethical and socio-legal competences in Higher Education. Several educational resources have been developed for this purpose. As a partner to Erasmus + Trustworthy AI project, our university was responsible for coordinating a Hackathon to investigate efficacy of applying educational resources for trustworthy AI to a real-world healthcare use case in healthcare domain. The outcomes of the Ethical AI Hackathon were recommendations for an AI solution in healthcare domain. With this research effectiveness and applicability of the open educational resources were demonstrated.

## Notes

1. COM (2018) 237 final, Brussels, 25.4.2018
2. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
3. http://cohealth.ivi.ie/chape/
4. All the guiding questions have been developed by Erasmus + Trustworthy AI project partners. All the outputs of this project are openly accessible (on 31 July 2022 onward), on the project website available at: https://www.trustworthyaiproject.eu/

## References

Alfano, M., Kellett, J., Lenzitti, B. and Helfert, M. (2021), "Proposed use of a conversational agent for patient empowerment", *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021)*, Vol. 5, pp. 817-824, HEALTHINF.

Austrian, J., Mendoza, F., Szerencsy, A., Fenelon, L., Horwitz, L.I., Jones, S., Kuznetsova, M. and Mann, D.M. (2021), "Applying a/B testing to clinical decision support: rapid randomized controlled trials", *Journal of Medical Internet Research*, Vol. 23 No. 4, e16651.

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R. and Maharaj, T. (2020), "Toward trustworthy AI development: mechanisms for supporting verifiable claims", arXiv preprint arXiv:2004.07213.

Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S. and Yeung, K. (2021), "Trustworthy AI", *Reflections on Artificial Intelligence for Humanity*, Springer, Cham, pp. 13-39.

Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A.N. and Raji, D. (2019), *AI Now 2019 Report*, AI Now Institute, New York, NY.

EIT Health (2021), "EIT Health launch new report on AI in health", available at: https://eithealth.eu/news-article/eit-health-launch-new-report-on-ai-in-health/#:~:text=Despite%20global%20economic%20growth%20predicted,globally%20over%20the%20same%20period

Floridi, L. and Cowls, J. (2021), "A unified framework of five principles for AI in society", *Ethics, Governance, and Policies in Artificial Intelligence*, pp. 5-17, Springer, Cham.

Gallagher, H. (2022), "Recognising a right to hack back-Tom and Jerry in cyberspace?", *Trinity CL Rev*, Vol. 25, p. 56.

Gama, K., Alencar, B., Calegario, F., Neves, A. and Alessio, P. (2018), "A hackathon methodology for undergraduate course projects", *In 2018 IEEE Frontiers in Education Conference (FIE)*, IEEE, pp. 1-9.

Grand Review Research (2019), "Artificial intelligence in healthcare market size report, 2019-2025", available at: https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-healthcare-market

IBM (2021), "Stress testing, robustness, adversarial testing", available at: https://research.ibm.com/blog/securing-ai-workflows-with-adversarial-robustness

Juravle, G., Boudouraki, A., Terziyska, M. and Rezlescu, C. (2020), "Trust in artificial intelligence for medical diagnoses", *Progress in Brain Research*, Vol. 253, pp. 263-282.

Kolb, A.Y. and Kolb, D.A. (2005), "Learning styles and learning spaces: enhancing experiential learning in higher education", *Academy of Management Learning and Education*, Vol. 4 No. 2, pp. 193-212.

Li, X., Pan, D. and Zhu, D. (2021), "Defending against adversarial attacks on medical imaging AI system, classification or detection?", *In 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1677-1681, IEEE.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A.K. and Tang, J. (2021), "Trustworthy ai: a computational perspective", arXiv preprint arXiv:2107.06641.

Maaravi, Y. (2020), "Using hackathons to teach management consulting", *Innovations in Education and Teaching International*, Vol. 57 No. 2, pp. 220-230.

Mtsweni, J. and Abdullah, H. (2015), "Stimulating and maintaining students interest in Computer Science using the Hackathon model", *The Independent Journal of Teaching and Learning*, Vol. 10 No. 1, p. 8597.

Nandi, A. and Mandernach, M. (2016), "Hackathons as an informal learning platform", *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 346-351.

NIHCM - National Institute for Health Care Management Foundation (2021), "Racial bias in health care artificial intelligence", available at: https://nihcm.org/publications/artificial-intelligences-racial-bias-in-health-care

Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020), "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing", *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33-44.

Rys, M. (2021), "Invention development. The hackathon method", *Knowledge Management Research and Practice*, pp.1-13.

Smuha, N.A. (2019), "The EU approach to ethics guidelines for trustworthy artificial intelligence", *Computer Law Review International*, Vol. 20 No. 4, pp. 97-106.

Towhidi, G. and Pridmore, J. (2022), "Hackathons for experiential learning in IS higher education", *Issues in Information Systems*, Vol. 23 No. 1, pp. 293-305.

Warner, J. and Guo, P.J. (2017), "Hack. edu: examining how college hackathons are perceived by student attendees and non-attendees", *Proceedings of the 2017 ACM Conference on International Computing Education Research*, pp. 254-262.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J. and Schwartz, O. (2018), *AI Now Report 2018*, AI Now Institute at New York University, New York, pp. 1-62.

## About the authors

Dr Zohreh Pourzolfaghar is an Assistant Professor in Management Information System, at School of Business, National University of Ireland (NUIM). She is member of Lero, the Science Foundation Ireland Research Centre for Software. Zohreh formerly has been Research Fellow at Innovation Value Institute (IVI) at NUIM. She is a senior research member of the Business Informatics Group and the theme lead in a project investigating the application of Enterprise Architecture in Smart Cities, funded by Science Foundation Ireland. She has been Principal Investigator for an Industry Fellowship Programme, funded by Science Foundation Ireland (SFI), and PI for two projects in collaboration with industry, funded by Enterprise Ireland. Zohreh presented her work at international conferences and has published research in Journals and Conference Proceedings. She is also has been leading research within the Cost Action CA16222 (WISE-ACT), a European-wide network that explores the wider impacts of Autonomous and Connected Transport. She has been thematic lead in a successful European Marie Skłodowska Curie (MSCA) proposal, i.e. PERFORM. She is a member of the supervisory board of the EU project PERFORM, that focuses on Digital Retail. She has been awarded the SFI industry fellowship (2018) working with industry on topics related to Smart Cities and Building Information Management. Zohreh is the Co-PI for two European MSCA ENTRUST Doctoral Network (funded on Aug. 2022), and MCSA Co Fund DIGI+ (funded on Jun 2022). Zohreh Pourzolfaghar is the corresponding author and can be contacted at: zohreh.pourzolfaghar@mu.ie

Dr Marco Alfano is a Senior Researcher at the Innovation Value Institute (IVI), Maynooth University and leader of the IVI Digital Health research cluster. He is also affiliated with Lero, the SFI Research Centre for Software and receives SFI funding for his research. He is currently working on responsible use of AI in health and well-being by facilitating person/patient empowerment and seamless communication within the healthcare system (http://cohealth.ivi.ie/). His research interests include Responsible AI, Digital Health Transformation, Patient Empowerment, Human-machine communication, Data analytics, Semantic Web, Smart cities, Cybersecurity and Open Data/Big Data. He has authored more than 50 peer-reviewed articles for journals, books and conferences. He has participated in several European projects

and has received grants from international bodies, such as the European Union (under the FP7 and H2020 framework programs), and national bodies, such as Science Foundation Ireland, Enterprise Ireland and the National Research Council of Italy.

Professor Markus Helfert is the Director of IVI and Professor of Digital Service Innovation at Maynooth University. He is also Director of the Business Informatics Group at Maynooth University. He is a Principal Investigator at Lero – The Irish Software Research Centre and at the Adapt Research Centre. His research is centred on Digital Service Innovation, Smart Cities and IoT-based Smart Environments and includes research areas such as Service Innovation, Intelligent Transportation Systems, Smart Services, Building Information Management, FinTech, Data Value, Enterprise Architecture, Technology Adoption, Analytics, Business Process Management. Prof. Helfert is an expert in Data Governance Standards and is involved in European Standardization initiatives. Markus Helfert has authored more than 200+ academic articles, journal and book contributions and has presented his work at international conferences. Helfert has received national and international grants from agencies such as European Union (FP7; H2020), Science Foundation Ireland and Enterprise Ireland, was project coordinator on EU projects, and is the Project coordinator of the H2020 Projects: PERFORM on Digital Retail and MSCA Doctoral Network EnTrust and Co-fund Digi+.