Project Acronym:	HosmartAI
Grant Agreement number:	101016834 (H2020-DT-2020-1 – Innovation Action)
Project Full Title:	Hospital Smart development based on AI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101016834

DELIVERABLE

D8.5 – SELP Continuous Monitoring Report 2

Dissemination level:	PU -Public
Type of deliverable:	R -Report
Contractual date of delivery:	31 May 2024
Deliverable leader:	VUB
Status - version, date:	Final – v1.0, 2024-06-03
Keywords:	ELSI, ELSA, AI risk

This document is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under agreement No 101016834. The content of this document reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.
The document is the property of the HosmartAI consortium and shall not be distributed or reproduced without the approval of the HosmartAI Project Coordination Team. Find us at www.hosmartai.eu.

Executive Summary

This Report, entitled D8.5 SELP Continuous Monitoring Report 2 (D8.5), provides the findings and results of the second half of Task 8.4 SELP Continuous Compliance Report (T8.4). Here, "SELP" stands for **S**ocial, **E**thical, and **L**egal **P**erspectives. The primary objective of SELP/WP8 is to assess the impact of HosmartAI technologies through 8 Lighthouse Pilots from social, ethical, and legal perspectives, with the aim of minimizing potential negative impacts or risks by complying with relevant regulations, as well as taking proactive measures in light of cutting-edge discourse regarding AI technology.

To this end, WP8 has formulated a questionnaire to gather relevant information and insights from 8 Lighthouse Pilots. The questionnaire was designed to capture a broad spectrum of issues regarding SELP, taking into account various factors: previous tasks (primarily T8.3 and the first half of T8.4), deliverables (mainly D8.3 and D8.4), subsequent updates, the current stage of the HosmartAI project, and the review report by the European Commission (EC). The questions in the questionnaire were categorized into four groups, each addressing specific issues: (1) Medical and Research Ethics; (2) Data Protection/Privacy and Data Security; (3) Ethical and Societal Impact/Risks of HosmartAI technology; and (4) AI Bias and Explainable AI. Based on the responses from pilot partners, we have conducted an impact assessment in the context of SELP.

We determined that there are no issues requiring further attention or discussion regarding (1) Medical and Research Ethics, and (2) Data Protection/Privacy and Data Security.

In Section (3) Ethical and Societal Impact/Risks of HosmartAI Technology, which addresses potential risks added or heightened due to the use of HosmartAI technology in the research study, we did not identify any issues that require further attention or discussion. The responses from all pilot partners fall into two categories: (1) there are no added or heightened risks due to the HosmartAI project, or (2) there are added or heightened risks, but the pilot partner anticipates these risks and has designed the pilot study with appropriate measures to detect and mitigate them.

In Section (4) AI Bias and Explainable AI, which focuses on measures and initiatives by pilot partners to address AI bias issues and to improve their HosmartAI technology from the perspective of explainability and transparency, we found that pilot partners took sufficient measures and initiatives to address these issues appropriately. These efforts include working towards ensuring diversity and inclusivity in datasets, taking various steps to make their AI systems more explainable, and further preparing to take proactive measures in anticipation of the phase when their HosmartAI technology will be introduced into the European market and utilized in healthcare

At the same time, we recognize that fulfilling the necessary conditions does not guarantee that the sufficient conditions are met. Considering the different levels of progress among pilot partners, it is especially crucial for each partner to learn from the insights and experiences of more advanced partners. This becomes even more critical when HosmartAI technology is launched in the European market and integrated into healthcare practices.

In light of the above, this Report includes chapters devoted to The Artificial Intelligence Act, as well as AI Biases, Explainable AI, and AI Risk Management.

The AI Act was also covered in D8.1, the first deliverable of WP8. At that time, however, the status of the Act was a 'Proposal', and there have been numerous changes and developments since then. In brief, the AI Act categorizes AI systems according to their risk level and either prohibits them or sets out different obligations for each category. AI systems deemed to fall within the 'Unacceptable risk' category are prohibited. A significant portion of the AI Act is devoted to regulating the second category, referred to as 'High-risk AI systems.' If an AI system does not fall within these two categories, it will be deemed either 'Limited risk' or 'Minimal risk.' The AI Act imposes various transparency requirements on Limited risk AI systems to ensure individuals are provided with necessary information. Minimal-risk AI systems are not regulated by the AI Act.

This report also covers issues related to AI biases, explainable AI, and AI risk management. Numerous studies demonstrate that AI bias -- occurrence of biased results due to human biases that skew the original training data or AI algorithm -- is found in AI systems used in healthcare. Examples include: (1) skin-cancer detection algorithms, many of which are trained primarily on light-skinned individuals, perform worse at detecting skin cancer affecting darker skin; (2) a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias; (3) the AI system performs worse for that underrepresented gender when the images for training datasets are insufficient for one gender; and (4) women are more likely to be misdiagnosed with heart disease because many studies focus primarily on male symptoms. The common sources of AI bias include: (1) algorithm bias; (2) cognitive bias; (3) confirmation bias; (4) exclusion bias; (5) measurement bias; (6) out-group homogeneity bias; (7) prejudice bias; (8) recall bias; (9) sample/Selection bias; and (10) stereotyping bias.

The AI Act mandates the implementation of management systems to ensure the safe and ethical deployment of AI technologies. Specifically, Article 9 requires the establishment of a risk management system, while Article 17 mandates that providers of high-risk AI systems have a quality management system in place. The Artificial Intelligence Risk Management Framework (AI RMF) serves as a useful starting point for organizations to build their management systems to address AI risks and comply with the AI Act.

Deliverable leader:	Paul QUINN (VUB)
Contributors:	Hideyuki MATSUMI (VUB)
Reviewers:	Sascha Marschang (VUB), Enrico Dal Pozzo (IRCCS)
Approved by:	Athanasios Poulakidas, Anastasia Panitsa (INTRA)

Document History

Version	Date	Contributor(s)	Description
0.1	03 May 2024	Hideyuki MATSUMI	First draft version
0.2	04 May 2024	Hideyuki MATSUMI	Revised first draft
0.3	18 May 2024	Hideyuki MATSUMI	First draft for internal review within VUB
0.4	19 May 2024	Hideyuki MATSUMI	Revised final draft
0.5	22 May 2024	Hideyuki MATSUMI	First draft for internal review within the consortium
0.6	26 May 2024	Hideyuki MATSUMI	Incorporating the first feedback from the internal review
0.7	28 May 2024	Hideyuki MATSUMI	Incorporating the second feedback from the internal review
0.8	3 June 2024	Hideyuki MATSUMI	Final draft for QA
1.0	3 June 2024	Athanasios Poulakidas, Anastasia Panitsa	Final version for submission after QA

Table of Contents

List of Tables

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	Title
AI	Artificial Intelligence
AI Act	The EU's Artificial Intelligence Act
AI RMF	Artificial Intelligence Risk Management Framework by NIST
CDSS	Clinical Decision Support System
EC	European Commission
XAI	Explainable AI
ML	Machine Learning
NIST	National Institute of Standards and Technology
PD	Personal data
SELP	Social, Ethical, and Legal Perspectives

Term	Definition
AI system	The AI Act defines an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments." While this deliverable follows this definition, we also use "AI system" as a synonym to the technology commonly referred to as "AI."
AI player	In this document, AI player refers to various actors regulated by the AI Act, defined under Article 3. They are, namely: provider, deployer, importer, distributor,[1] and operator.[2] The AI Act defines these actors depending on their roles in relation to AI system. This document refers to these actors collectively as "AI player" unless it refers to a specific "AI player."

[1] Article 3(7), AI Act.

[2] Article 3(8), AI Act.

1 Introduction

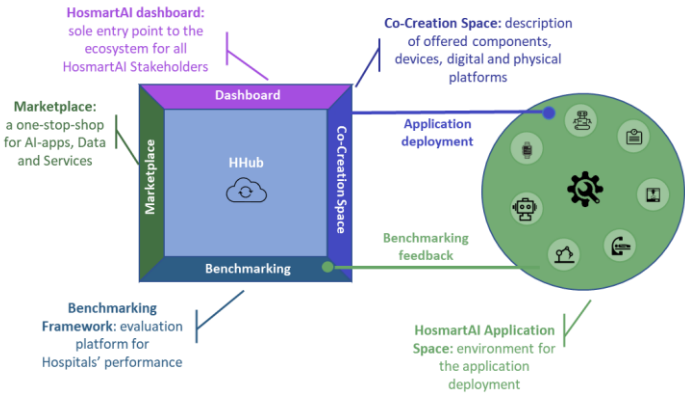
1.1 Project Information

VISION

The HosmartAI vision is a strong, efficient, sustainable and resilient European **Healthcare system** benefiting from the capacities to generate impact of the technology European Stakeholders (SMEs, Research centres, Digital Hubs and Universities).

MISSION

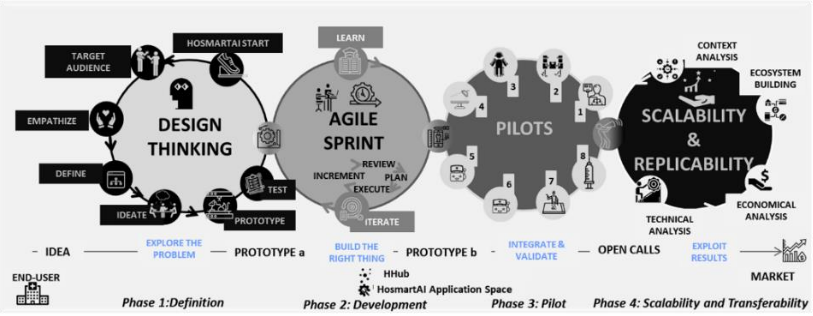
The HosmartAI mission is to guarantee the **integration** of Digital and Robot technologies in new Healthcare environments and the possibility to analyse their benefits by providing an **environment** where digital health care tool providers will be able to design and develop AI solutions as well as a space for the instantiation and deployment of a AI solutions.


HosmartAI dashboard: sole entry point to the ecosystem for all HosmartAI Stakeholders
Co-Creation Space: description of offered components, devices, digital and physical platforms
Marketplace: a one-stop-shop for AI-apps, Data and Services
Dashboard
HHub
Marketplace
Co-Creation Space
Benchmarking
Application deployment
Benchmarking feedback
Benchmarking Framework: evaluation platform for Hospitals' performance
HosmartAI Application Space: environment for the application deployment

HosmartAI will create a common open Integration **Platform** with the necessary tools to facilitate and measure the benefits of integrating digital technologies (robotics and AI) in the healthcare system.

A central **hub** will offer multifaceted lasting functionalities (Marketplace, Co-creation space, Benchmarking) to healthcare stakeholders, combined with a collection of methods, tools and solutions to integrate and deploy AI-enabled solutions. The **Benchmarking** tool will promote the adoption in new settings, while enabling a meeting place for technology providers and end-users.

Eight Large-Scale Pilots will implement and evaluate improvements in medical diagnosis, surgical interventions, prevention and treatment of diseases, and support for rehabilitation and long-term care in several Hospital and care settings. The project will target different **medical** aspects or manifestations such as Cancer (Pilot #1, #2 and #8); Gastrointestinal (GI) disorders (Pilot #1); Cardiovascular diseases (Pilot #1, #4, #5 and #7); Thoracic Disorders (Pilot #5); Neurological diseases (Pilot #3); Elderly Care and Neuropsychological Rehabilitation (Pilot #6); Fetal Growth Restriction (FGR) and Prematurity (Pilot #1).



To ensure a user-centred approach, harmonization in the process (e.g. regarding ethical aspects, standardization, and robustness both from a technical and social and healthcare perspective), the **living lab** methodology will be employed. HosmartAI will identify the appropriate instruments (**KPI**) that measure efficiency without undermining access or quality of care. Liaison and co-operation activities with relevant stakeholders and **open calls** will enable ecosystem building and industrial clustering.

HosmartAI brings together a **consortium** of leading organizations (3 large enterprises, 8 SMEs, 5 hospitals, 4 universities, 2 research centres and 2 associations – see Table 1) along with several more committed organizations (Letters of Support provided).

Table 1: The HosmartAI consortium.

Number[3]	Name	Short name
1 (CO)	INTRASOFT INTERNATIONAL SA	**INTRA**
1.1 (TP)	INTRASOFT INTERNATIONAL SA	**INTRA-LU**
2	PHILIPS MEDICAL SYSTEMS NEDERLAND BV	**PHILIPS**
3	VIMAR SPA	**VIMAR**
4	GREEN COMMUNICATIONS SAS	**GC**
5	TELEMATIC MEDICAL APPLICATIONS EMPORIA KAI ANAPTIXI PROIONTON TILIATRIKIS MONOPROSOPIKI ETAIRIA PERIORISMENIS EYTHINIS	**TMA**
6	ECLEXYS SAGL	**EXYS**
7	F6S NETWORK IRELAND LIMITED	**F6S**
7.1 (TP)	F6S NETWORK LIMITED	**F6S-UK**
8	PHARMECONS EASY ACCESS LTD	**PhE**
9	TERAGLOBUS LATVIA SIA	**TGLV**
10	NINETY ONE GMBH	**91**
11	EIT HEALTH GERMANY GMBH	**EIT**
12	UNIVERZITETNI KLINICNI CENTER MARIBOR	**UKCM**
13	SAN CAMILLO IRCCS SRL	**IRCCS**
14	SERVICIO MADRILENO DE SALUD	**SERMAS**
14.1 (TP)	FUNDACION PARA LA INVESTIGACION BIOMEDICA DEL HOSPITAL UNIVERSITARIO LA PAZ	**FIBHULP**
15	CENTRE HOSPITALIER UNIVERSITAIRE DE LIEGE	**CHUL**
16	PANEPISTIMIAKO GENIKO NOSOKOMEIO THESSALONIKIS AXEPA	**AHEPA**
17	VRIJE UNIVERSITEIT BRUSSEL	**VUB**
18	ARISTOTELIO PANEPISTIMIO THESSALONIKIS	**AUTH**
19	EIDGENOESSISCHE TECHNISCHE HOCHSCHULE ZUERICH	**ETHZ**
20	UNIVERZA V MARIBORU	**UM**

[3] CO: Coordinator. TP: linked third party.

Number[3]	Name	Short name
21	INSTITUTO TECNOLÓGICO DE CASTILLA Y LEON	**ITCL**
22	FUNDACION INTRAS	**INTRAS**
23	ASSOCIATION EUROPEAN FEDERATION FORMEDICAL INFORMATICS	**EFMI**
24	FEDERATION EUROPEENNE DES HOPITAUX ET DES SOINS DE SANTE	**HOPE**

1.2 Document Scope

This Report, entitled D8.5 SELP Continuous Monitoring Report 2 (D8.5), provides the findings and results of Task 8.4 SELP Continuous Compliance Report (T8.4). Building upon D8.4 SELP Continuous Monitoring Report 1 (D8.4), it is the second deliverable in T8.4, and is the fifth and final deliverable of Work Package 8 (WP8).

In this document, "SELP" stands for **S**ocial, **E**thical, and **L**egal **P**erspectives,[4] and it derives from the Ethical, Legal, and Social Implications (ELSI) research,[5] originally conceived in 1988 as part of the Human Genome Project. The primary objective of SELP/WP8 is to assess the impact of HosmartAI technologies by 8 Lighthouse Pilots from the social, ethical, legal perspectives, and to minimize the potential negative impacts, or risks, by complying with relevant regulations, as well as taking proactive measure in light of cutting-edge discourse regarding AI technology. Thus, the overarching question that the WP8 deals with is, what is the impact of HosmartAI technology from social, ethical, and legal perspectives.

To this end, the Task of WP8 involves identifying, analysing, and addressing complex social, ethical, and legal challenges that may arise from the development and deployment of HosmartAI technology. Specifically, in Task 8.3 SELP Impact Assessment (T8.3), WP8 has conducted the first impact assessment of HosmartAI technology of 8 pilot in the context of SELP. WP8 has formulated a questionnaire to collect necessary information regarding all 8 pilot studies, and based on the responses by pilot partners, we have conducted the assessment of impact in the context of SELP. The assessment was done against the first deliverable of WP8, entitled D8.1 SELP Benchmark Report (D8.1). D8.1 provided the regulatory landscape as well as the ethical and social norms and issues that are potentially relevant to HosmartAI technology. It summarized applicable or relevant legal frameworks as well as ethical and social norms, and discussed the potential issues that may be relevant.

T8.4 builds upon T8.4, by providing a continuous monitoring report. The idea behind it is similar to a "fixed point observation," a method in observational research where data is collected from a specific, unchanging location over a period of time.

D8.4, the first deliverable of T8.4, documents the findings and results of the first half of T8.4. By formulating a different questionnaire in response to D8.3, we have provided the first continuous monitoring report, and discussed and focused on topics or issues that have come

[4] It is sometimes also referred to as Social, Ethical, Legal, and Privacy, which basically captures the same meaning.
[5] It is also referred to as Ethical, Legal, and Social Aspects (ELSA) research in, for example, Europe, while it is often referred to as Ethical, Legal, and Social Implications (ELSI) research, for example, in the US.

to light. This report, D8.5, documents the second half of T8.4. By formulating yet another different questionnaire in response to D8.4, we have conducted the second continuous monitoring report, and discussed some topics or issues that are noteworthy.

Furthermore, we decided to include some topics and issues that are not stated in the Grant Agreement, mainly because of important developments that have occurred since delivering D8.4. In short, we expanded D8.5 to cover: (1) The AI Act; and (2) AI Bias, Explainable AI, and AI risk management system.

First, the AI Act was covered in D8.1, which provided the regulatory landscape and summarized applicable or relevant legal frameworks. The European Commission (Commission) issued the Proposal of the AI Act on 21st April 2021, and thus the version covered in D8.1 was the initial proposal.[6] Since then, there have been numerous changes and developments, and most notably, the final text of the AI Act has been approved by all EU legislative institutions: Endorsed by the MEPs on 13th March 2024; CORRIGENDUM issued on 19th April; Final and formal approval by the Council of the EU on 21st May.[7]

It is certainly too early to make any assertions as to how the Act will apply to specific facts/technologies. As the Commission is tasked to develop the first guidelines, we would have to wait for the Commission's guidelines. Nevertheless, this document provides a high-level overview of the AI Act based on the CORRIGENDUM version.[8]

Second, AI related topics and issues, such as AI bias and Explainable AI, are similarly touched in D8.1. This deliverable D8.5, however, covers these again in more detail. One of the reasons for this is to incorporate and respond to the feedback provided by the Commission reviewers (Review Report[9]), which advised the consortium to further address issues "related to transparency and algorithm generalization to promote equitable healthcare for diverse population," "AI explainability for better algorithm transparency," and "potential biases in the algorithms."

While these issues are also addressed in the Continuous Monitoring and Reporting part, D8.5 also offers further information with the aim of providing a helpful resource for HosmartAI partners moving forward.

D8.5 also touches upon standardized and established management systems purported to address AI related risks. This is because, first, the AI Act requires various management systems to be implemented, and we believe providing HosmartAI partners with guidance on risk management systems focusing on AI risks would be helpful. Second, these systematic approaches to risks unique to AI systems, such as AI bias, are helpful for HosmartAI partners to address the issues and further achieve transparent, equitable AI systems.

[6] 2021 Proposal is available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.
[7] Press release available at https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/pdf/.
[8] "CORRIGENDUM to the position of the European Parliament adopted at first reading on 13 March 2024," available at https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.
[9] See the final review report issued on the 26th of September 2023.

1.3 Document Structure

This document comprises the following chapters:

Chapter 1 presents an introduction to the project and the document.

Chapter 2 provides the findings and results of Task 8.4 SELP Continuous Compliance Report, which corresponds to the description in the Grant Agreement. First, it provides the responses by each of the 8 Lighthouse Pilots to the questionnaire. The results are presented systematically, which allows comparing the results of one pilot to another. Second, we offer our analyses on each topic or issue. This chapter also includes issues on AI ethics issues, such as algorithmic transparency, explainable AI, and AI biases.

Chapter 3 offers a high-level summary of the AI Act based on the final adopted text. It covers some essential concepts, definitions, issues, etc. to grasp the overview of the Act. It outlines, *inter alia*, the categorization of AI systems depending on its risk level, various “AI stakeholders” (or “AI players”) and a brief description of obligations pertaining to each AI player.

Chapter 4 discusses AI ethics issues or topics, such as AI biases and Explainable AI. First, it addresses the issue concerning AI bias. It covers, *inter alia*, some examples of AI bias in healthcare, as well as Explainable AI in healthcare, sources of AI bias, and guidance on how to avoid AI bias. Second, it addresses the concept of Explainable AI as a solution to the AI bias problem, and as part of a broader topic on transparency.

Chapter 5 concludes this Report by providing a succinct summary of this document, and also by offering some ideas and thoughts that can be helpful for HosmartAI partners to develop or deploy their AI system that is more advantageous and competitive from the social, ethical, or legal perspectives.

2 Continuous Monitoring Report

This chapter presents the results and analyses derived from Task 8.4 SELP Continuous Compliance Report (T8.4). It provides the responses from each pilot study, along with the questions that we formulated, and our analyses from the Social, Ethical, and Legal Perspectives (SELP).

2.1 Background and Context

In T8.4, we have formulated a questionnaire to gather relevant information and insights from 8 Lighthouse Pilots. The questionnaire was designed to capture a broad spectrum of issues regarding SELP, considering various factors: the issues addressed in past tasks (mainly T8.3 and T8.4) and deliverables (mainly D8.3 and D8.4), updates thereafter, the stage of the HosmartAI project, and the review report by the European Commission (EC).

Structure. The questions in the questionnaire were categorized into four groups: (1) Medical and Research Ethics; (2) Data Protection/Privacy and Data Security; (3) Ethical and Societal Impact/Risks of HosmartAI technology; and (4) AI Bias and Explainable AI. Hence, the following sections are divided accordingly. Each section, which corresponds to each category, has a few subsections focusing on each issue. Each subsection is structured as follows: (1) a brief description of the issue and the objective behind our inquiry; (2) the actual text of the question asked; and (3) the responses provided by each HosmartAI pilot partner. Our analyses are presented either in each subsection, when the findings and analysis are brief and straightforward, or in the final section (Findings and Analyses) when analysing multiple issues together provides better insight.

Respondents. The HosmartAI project includes eight Lighthouse Pilots, but this questionnaire has nine respondents because Pilot #1 consists of two pilot studies, each of which provided a separate response. Below is the list of respondents and a brief description of each.

Table 2: Brief description of pilot studies.

Pilot #	Q. 2. Brief description of pilot studies
Pilot 1 (ECHO)	Clinical decision support for cardiologists
Pilot 1 (VCE)	Clinical decision support for gastroenteroligists
Pilot 2	Algorithm for scheduling management
Pilot 3	Model imprinting for robots in the case of neurological rehabilitation
Pilot 4	AI-based navigation system for catheter tip ablation
Pilot 5	Study 1: **Patient admission supported with assistive humanoid robot and artificial intelligence** Study 2: **A study protocol on the effects of interactive digital assistance on patient engagement and perceived quality of care of surgery patients and self-efficacy and workload of staff** Study 3: **Evaluating the clinical impact of integrating a computerized clinical decision support system and a social robot into grand rounds and pre/post-operative care of patients in abdominal and thoracis surgery department: Study protocol**
Pilot 6	Virtual assistant for cognitive rehabilitation
Pilot 7	Registry for coronary angiograms
Pilot 8	Digital research platform for glioma management combining genetic and image data

2.2 Medical and Research Ethics

This section addresses issues regarding medical and research ethics. Past tasks (T8.3 and the first half of T8.4) and deliverables (D8.3 and D8.4) covered these issues comprehensively in both scope and detail. During this task (i.e., the second half of T8.4), however, the objective was to gather information on any feedback related to SELP from various stakeholders or any deviations from the original study protocol, primarily because the pilot studies are in the final stage of completion.

2.2.1 Human participants

The issue here is whether any of the participants (i.e., individuals who participated or collaborated in the pilot study) expressed comments, concerns, questions, etc. in relation to SELP. The objective of asking this question is twofold: first, to inquire if there was any communication, and second, to learn from the content of those communications.

We did not identify any issues that need further attention or discussion. Pilot 5, for example, noted that "[s]ome patients were not particularly keen on wearing the study monitors (CDSS)." This feedback, however, was properly incorporated, and Pilot 5 adjusted their research accordingly, which demonstrates how the communication with their participants has been effective.

No pilots responded that they became aware of the ethical, legal, or social issues due to human participation to their pilot study.

The actual text of the question and the responses from pilot partners follow (the same for all subsections).

Q. 3. **Human participants**. Please describe if any of the two applies:
(1) Were there any comments, concerns, opinions, questions, thoughts, or anything else that was communicated or expressed to your Pilot by the participants. If there were none, please state so.
(2) Were there any ethical, legal, or social issue that your Pilot noticed or became aware of, due to human participation. If there were none, please state so.

Pilot #	Q. 3. Human participants
Pilot 1 (ECHO)	(1) The only feedback acquired by the participants was via the SUS (System Usability Scale) questionnaire that was answered after the end of the participation. (2) **No** ethical, legal, or social issues were raised.
Pilot 1 (VCE)	(1) The only feedback acquired by the participants was via the SUS (System Usability Scale) questionnaire that was answered after the end of the participation. (2) **No** ethical, legal, or social issues were raised.
Pilot 2	(1) None (2) None
Pilot 3	(1) In the pilot trial, we had No.4 drop-out (5% of enrolled patients), distributed in two organizational models: • No.2 patients dropped out from group with ratio 1:1, due to the following reason: ○ No.1 patient: symptoms of vertigo linked to the use of technological device for balance recovery ○ No.1 patient for hospitalization problem, not linked to the research project. • No.2 patients dropped out from group with ratio 1:2, due to the following reasons: ○ No.1 patient for cardiac complications ○ No.1 patient for hospitalization problem, not linked to the research project. Thus, no reasons were linked to the organizational model of the patients. Only one patient dropped-out for adverse symptoms related to the technological treatment. Other patients' comments, both positive and negative, were used as feedback during the co-design of the solution. (2) No ethical, legal or social issue was raised by participants.
Pilot 4	(1) None (2) None

Pilot 5	(1) Yes, the patient participants were concerned about the AI technology and the loss of human aspect (empathy, closeness, warmth). However, most of the participants were positive about the administrative part that the AI had presented. We had mostly elderly patients (more than 50 % of the participants) which are not as acquainted with the technology as the younger population. Even though the elderly expressed interest in the AI technology, some concerns could be exhibited. The nurse participants also exhibited concerns regarding the human aspect and the quality of the sound (robot's speech). The doctor participants showed positive attitude regarding the CDSS technology and the health data display record. (2) No. The study is non-invasive. Only one issue came along, which was comfort wise. Some patients were not particularly keen on wearing the study monitors (CDSS). Therefore, we had to adjust the research to the participants and measure the vital signs for a shorter time frame. This proved to be a better solution for the participants and they were more likely to agree to wear the study monitor.
Pilot 6	(1) No comments, concerns, or issues expressed by the participants. (2) No ethical, legal, or social issues were identified during the pilot.
Pilot 7	(1) The feedback provided by the participants was about the usability and algorithm performance – not about medical and research ethics. (2) None
Pilot 8	(1) No concerns have been communicated by the participants (2) Access to pseudonymized genetic data is possible only in the hospital environment therefore, a putative computer will used for the deployment of the genetic and image analysis tools developed for Glioma data.

2.2.2 Informed consent

The issue is whether the informed consent procedure, described in their study protocol, has been conducted appropriately. The objective of this question is to identify if there were any deviations from the study protocol, as it serves as the essential foundation of the research study.

We do not find any issues that need further attention or discussion. Although Pilot 5 noted that some patients declined to participate, this exemplifies informed consent, where individuals choose not to participate after considering the information.

Q. 4. **Informed consent**. Please describe if any of these applies:
(1) Were there any deviation from the informed consent procedure your Pilot described in the previous deliverable or questionnaire?
(2) Were there any comments, concerns, opinions, questions, thoughts, or anything else that was communicated or expressed to your Pilot by the participants during the informed consent procedure?
(3) Were there any ethical, legal, or social issue that your Pilot noticed or became aware of, because of the informed consent procedure? For each sub-question, please state so if there was none.

Pilot #	**Q. 4. Informed consent**
Pilot 1 (ECHO)	(1) **No** deviation (2) **No** issues were communicated/expressed (3) **None**
Pilot 1 (VCE)	(1) **No** deviation (2) **No** issues were communicated/expressed (3) **None**
Pilot 2	(1) None (2) None (3) 5 informed consents have to be signed by the project coordinator instead of the Principal Investigator because he was on medical leave.
Pilot 3	(1) No deviations, consent procedure remained the same as expressed in the clinical protocol approved by ethical committee. (2) None (3) None
Pilot 4	(1) No deviations were encountered (2) No comments nor concerns by participants (3) None
Pilot 5	**(**1) No. The informed consent process for patient participants had been carried out and signed letters collected by the medical doctor, nurse or the coordination team. The process of the study had been explained according to the patient's understanding. All of the participants had been aware of what is expected of them, their rights and the aims of the research. All of the nurses and doctors that had been working on the HosmartAI project had agree to be a part of the nurse and doctor participants. (2) Mostly no. If the participant were not willing to participate in the research, they have declined their participation (mostly due to lack of interest or the fear factor before the medical operation). Those who were willing to participate were cooperative and agreed to most of the process. As we have stated in the previous Q.3. question the only alternation had been the duration of the monitor measurements. Some participants declined the measurements after the first day even if the measurements had been abbreviated. Only 1 nurse participant had declined to fill in the questionnaire (reason: she had not seen the immediate purpose of the project that would significantly affect her everyday work). (3) We noticed that the patients were dissatisfied with the monitor measurements. Some were angry and expressed their discomfort. Which also led to the declining of the monitor measurement (we have taken into account their decision and stopped with the measurements). Some patients showed signs of discomfort but had not expressed it.
Pilot 6	(1) No deviation from the informed consent procedures as described in earlier deliverables. (2) Participants did not express any concerns during the informed consent procedure. (3) No ethical, legal, or social issues were detected relating to the informed consent process.

Pilot 7	(1) None (2) None (3) None
Pilot 8	(1) No deviation from informed consent procedure pilot 8 has communicated in the previous questionnaire (2) No comments, concerns, opinions, questions, thoughts or anything else was communicated or expressed to our pilot by the participants. (3) No ethical, legal or social issue noticed by our pilot.

2.2.3 Ethics Committee, legal department, or DPO

The issue is whether the pilot received any comments, concerns, questions, etc. in relation to human participants from external parties, such as the ethics committee approved their research, or the legal department or DPO of their organization. The objective of this question is twofold: to inquire if there was any communication and to learn from the content of those communications.

We do not find any issues that need further attention or discussion. The response by Pilot 5 demonstrates how communication with and feedback from the DPO provides an additional layer of safeguarding patient participants' right to data protection.

Q. 5. Did you receive any comments, opinions, questions, suggestions from your Ethics Committee and/or legal department/DPO of your institution that might be relevant to this questionnaire? If so, please describe below. If there was none, please state so.

Pilot #	**Q. 5. Comments, opinions, questions, suggestions by Ethics Committee/DPO**
Pilot 1 (ECHO)	None
Pilot 1 (VCE)	None
Pilot 2	None
Pilot 3	DPO checked the data flow of cloud dashboard used by clinicians. No issues raised.
Pilot 4	None
Pilot 5	As by practice in our organization the entire documentation was submitted to the DPO, who proposed the form Permission for the use and publication of photographs. The DPO suggested that we should add the segment that all the video and photo material shall be taken in a discrete way. Therefore, the patient participant cannot be recognized. In accordance with our legal legislation the research application had been considered by the internal hospital ethics committee.
Pilot 6	No relevant comments or suggestions were received from the Ethics Committee, legal department, or Data Protection Officer.
Pilot 7	None
Pilot 8	None

2.3 Data Protection/Privacy and Data Security

This section addresses issues regarding data protection and data security. Similar to the previous section, past tasks (T8.3 and the first half of T8.4) and deliverables (D8.3 and D8.4) covered these issues comprehensively in both scope and detail. During this task, however, the objective is to gather information on any deviations from the original study protocol, any feedback related to SELP from various stakeholders, or if there were any data security incidents.

2.3.1 Processing of personal data

The main issue is whether there has been unintended processing of personal data, not described in the original study protocol. The primary objective of this question, coupled with the next question, is to ensure that all personal data have been processed legally.

No issues that need further attention or discussion are identified. Pilot 5 listed various personal data they processed, with appropriate legal basis.

Q. 6. **Personal data you processed** (e.g., **collected, used**, etc). (1) Please tell us about the period -- from when until when -- in which you collected personal data for your pilot study that you used for your HosmartAI technology. (2) Also, was there any **type of personal data you processed** (e.g., **collect, use**, etc), **which was not originally planned to process** (e.g., **collect, use**, etc)? If you processed (e.g., collect, use, etc) only the types of personal data you stated in the past Deliverable, please state so.
Note: Under law (e.g., GDPR), "**process**" means differently from what "data process" means, for example, in computer or data science. It means any operation which is performed on personal data. It can mean, for example, collecting, recording, organizing, structuring, storing, altering, retrieving, using, disclosing, transferring, disseminating, or otherwise making available.

Pilot #	Q. 6. (1) Period of collection of PD; (2) Unintended processing of PD
Pilot 1 (ECHO)	(1) From: 15/04/2022 Until: 30/04/2024 (2) We process **only** the types of data stated in the past deliverable.
Pilot 1 (VCE)	(1) From: 30/05/2022 Until: 30/04/2024 (2) We process **only** the types of data stated in the past deliverable.
Pilot 2	(1) From: 15 March 2024 Until: 24 April 2024 (2) Recording video with CHUL staff. All of them signed an informed consent for image sharing.
Pilot 3	(1) Patients data were collected From: 1 February 2023 Until: 12 April 2024 (2) No deviations, data collection remained the same as expressed in the clinical protocol approved by ethical committee.
Pilot 4	(1) Data from remote and manual navigation was collected From: 1st December 2023 Until: 15th December 2023 (2) No additional personal data was processed
Pilot 5	(1) **CDSS** and **IDA** (recruitment) From: 23. 10. 2023 Until: 31. 5. 2024 **PICOS** (recruitment) From: 25 April 2023 Until: 6 October 2023 (2) **CDSS** and **IDA**: We have processed: pseudonymized medical data of the patient participant, vital body measurements (EKG, SPO2, temperature), date of the patient admission and discharge, demographic data (medical procedure, sex, age, education), types and quantity of patient’s calls, questionnaires: • PHE (personal assessment of one’s disease) • PQMC (personal assessment of medical care) • 5Q-5D-3L (health of the patient today) • VAS (personal assessment of one’s pain) • VZP (personal opinion about AI in the medical care) • AES (personal opinion about the AI technology) • UEQ (personal opinion about the use of AI technology) • Rating of nursing education and physiotherapy for the staff and for the AI technology From the nursing and physiotherapy staff participants we have processed 2 types of questionnaires: • NGSE (assessment of one’s work) • NASA-TLX (assessment of the workload) From the doctor’s participants we have processed 2 types of questionnaires: • SUS (the usefulness of CDSS system) • UTAUT (various aspects of acceptance of the CDSS system) **PICOS**: We have processed: demographic data (sex, age, education), triage questionnaire (admission form, medical conditions, prescriptions), permission for the presence of the students, permission to use the medical data for education and research, questionnaire about the participant’s opinion on AI technology.
Pilot 6	(1) From: October 2023 Until: April 2024 (2) Same types of personal data defined (accepted study protocol).

Pilot 7	(1) The setup of the reader study was such that an evaluation was done with data already acquired by the hospital and not collected specifically for the project. The reader study was executed From: 11 Dec 2023 Until: 12 Dec 2023 (2) None
Pilot 8	(1) From: 26/12/2022 Until: TBD (2) The data originally planned in the previous questionnaire was collected during the pilot.

2.3.2 Legal basis and informed consent

The issue is whether there is an appropriate legal basis for processing of personal data. The objective of this question, coupled with the previous question, is to ensure that all personal data have been processed legally.

No issues that need further attention or discussion are identified.

Q. 7. **Legal basis** and **informed consent for processing of personal data**. Was there any type of personal data that you processed (e.g., collect, use, etc) **WITHOUT informed consent** to individuals participating your pilot research?
If no, please simply state No.
If yes, please provide information regarding: (1) what kind of personal data? (2) what is the legal basis, under the GDPR, that you processed those personal data? (3) Also, would you please describe the background, context, and reason why?

Pilot #	**Q. 7. Legal basis and informed consent**
Pilot 1 (ECHO)	No
Pilot 1 (VCE)	No
Pilot 2	No
Pilot 3	No, we collected only data outlined in the clinical protocol, as approved by ethical committee
Pilot 4	We didn’t collect inform consent from participants because they were the researchers themselves
Pilot 5	No.
Pilot 6	No
Pilot 7	No
Pilot 8	No

2.3.3 Breach of personal data

The issue is whether there was any breach of personal data. The goal is to determine if any personal data breach occurred and, if so, whether appropriate measures have been taken in response to the security incident.

There was no breach of personal data.

Q. 8. **Breach of personal data**. Was there any breach of personal data during the HosmartAI project? If no, please simply state No. If yes, please provide information regarding: (1) Detailed of the incident, including when and where it happened, the types of personal data breached, including the scope; (2) the response measures you took, including notification requirement under Art. 33 of the GDPR; and (3) specifically the communications you took with the affected individuals, the supervisory authority, and any other stakeholders.
Note: A breach of personal data occurs when the data for which your company/organisation is responsible suffers a data security incident resulting in a breach of confidentiality, availability, or integrity.

Pilot #	Q. 8. Breach of personal data
Pilot 1 (ECHO)	No
Pilot 1 (VCE)	No
Pilot 2	No
Pilot 3	No
Pilot 4	No
Pilot 5	No.
Pilot 6	No breaches of personal data during the project.
Pilot 7	No
Pilot 8	No

2.3.4 Comments, opinions, questions, suggestions from DPO

The issue is whether the pilot received any comments, concerns, questions, etc. in relation to the processing of personal data from external parties, such as the legal department or DPO of their organization. The objective of this question is twofold: to inquire if there was any communication and to learn from the content of those communications. The response by pilots slightly overlaps with Q.5, *supra*.

No issues that need further attention or discussion are identified.

Q. 9. Did you receive any comments, opinions, questions, suggestions from your DPO, legal department, or any other department/team of your institution regarding data protection or processing of personal data that might be relevant to this questionnaire? If so, please describe below. If there was none, please state so.

Pilot #	Q. 9. Comments, opinions, questions, suggestions from DPO on processing of PD
Pilot 1 (ECHO)	None
Pilot 1 (VCE)	None
Pilot 2	No
Pilot 3	None
Pilot 4	None
Pilot 5	No.
Pilot 6	No relevant comments or suggestions were received regarding data protection or processing of personal data. Discussions with DPO took place during the initial consulting phase for the study protocol definition.
Pilot 7	No
Pilot 8	None

2.4 Ethical and Societal Impact/Risks of HosmartAI technology

This section addresses the potential risks that are added or heightened by the use of HosmartAI technology in the research study. As emphasized in the questions, the focus is on added or heightened risks, rather than any risks in general. For example, the use of X-rays in healthcare entails certain risks. If the risk level deriving from the use of X-rays is the same with or without the HosmartAI project, then we consider there to be no added or heightened risk for the purpose of this analysis. These issues were addressed in past tasks (T8.3 and the first half of T8.4) and deliverables (D8.3 and D8.4). Nevertheless, we continue to address them as part of the continuous monitoring report in the second part of T8.4.

Sections 2.4.1, 2.4.2, and 2.4.3 are discussed together in Section 2.6, *infra*, because they are interconnected and integral to each other.

2.4.1 Added or heightened risks

The issue here is whether there are added or heightened risks in relation to SELP, and if any, what are the risks. The objective of asking this question is twofold: first, to clarify and identify if there are any added or heightened risks due to the use of HosmartAI technology in the research study, and second, to analyse how each pilot partner have conducted their risk assessment. The responses to this question serve as the basis for the following questions and discussions.

Q. 10. **Potential Risks**. Throughout the HosmartAI pilot study, did you become aware of any **added or heightened risks** in the AI technology of your Pilot Study? What is the worst-case scenario, if any, that can happen from the AI technology of your pilot study? If there are any, please explain the foreseeable risks and their scenario(s), presuming that AI technology makes mistakes.
Note: This question emphasizes on added or heightened risks, which means it aims to compares between the standard clinical/medical practice (without HosmartAI) and the pilot study research (with HosmartAI). For example, using electricity can be a risk in various ways, but generally, the risk level is the same whether it is in a regular medical practice

context or in HosmartAI research context. If the risk is largely the same regardless of inside/outside HosmartAI, please indicate so.

Pilot #	Q. 10. Added or heightened risks
Pilot 1 (ECHO)	All patients involved in the pilot study received the standard care regardless the AI outcome. Thus, **no** added or heightened risk due to AI existed. In future, if the developed AI technology is incorporated in the clinical practice and if it makes a mistake, then the worst-case scenario is to have an inconclusive diagnosis which will lead to manual measurements by a senior physician.
Pilot 1 (VCE)	All patients involved in the pilot study received the standard care regardless the AI outcome. Thus, **no** added or heightened risk due to AI existed. In future, if the developed AI technology is incorporated in the clinical practice and if it makes a mistake, then the worst-case scenario is to have an inconclusive diagnosis which will lead to manual further unnecessary follow-up examinations.
Pilot 2	Any risk. Pilot 2 is the creation of a software coupled to a chatbot
Pilot 3	Our AI-based monitoring technology is designed to observe the environment, such as a room or bathroom, and provide feedback on anomalous situations. For instance, it can detect if lights are on when they should be off or if shutters are closed during the day when they should be open. Possible errors in our AI solution would still result in incorrect anomaly detection or missed alerts. Hence, we do not foresee any potential risks to humans arising from this AI-based monitoring solution.
Pilot 4	No added or heightened risks were noted during our Pilot development

Pilot 5	The AI used in the pilots was mostly related to speech synthesis, speech recognition, emotion recognition and motion tracking. **(1) Added Risks of Using Speech Synthesis and Speech Recognition** AI-powered speech recognition can misinterpret patient statements, especially in noisy environments or with patients having accents or speech impairments. Similarly, speech synthesis might produce misleading or incorrect information if not accurately programmed or if it misinterprets input data. **Worst-Case Scenario** Misdiagnosis or miscommunication could occur, leading to inappropriate medical advice or delays in critical care. In extreme cases, this could result in adverse health outcomes. The triage part and the support for the CDSS were carried out using traditional (rule-based) ML approach. **(2) Added Risks of Using Emotion Recognition** Emotion recognition technologies might not accurately interpret the emotional state of patients due to cultural, individual, or contextual factors, leading to inappropriate responses or care decisions. **Worst-Case Scenario:** Inaccurate assessment of a patient's emotional state could lead to a lack of appropriate care in situations where understanding emotional cues is crucial, such as in mental health assessments. **(3) Added Risks of Using Chatbots** Chatbots may struggle with understanding and responding accurately to complex health queries or when patients use colloquial language, jargon, or have speech impairments. Unlike human doctors, chatbots might miss contextual cues that are crucial for accurate diagnosis or health advice. Chatbots may fail to properly escalate urgent health issues to human professionals, possibly leading to neglect of severe conditions. **Worst-Case Scenario** If a chatbot misunderstands a patient's symptoms or provides incorrect medical information, it could lead to the patient taking harmful actions or neglecting necessary medical care. Reliance on chatbots for initial patient interaction might delay the diagnosis of serious conditions that require immediate human medical attention. **(4) Added Risks of Using Motion Tracking** Motion tracking technologies might inaccurately record movements, particularly in complex scenarios involving multiple individuals or subtle movements, leading to incorrect assessments or treatments. **Worst-Case Scenario** Inaccurate motion tracking could lead to mismanagement of physical therapies or incorrect assessments of patient mobility, potentially exacerbating injuries or leading to incorrect treatments. **(5) Comparison to Standard Practice**

	Traditional clinical settings involve direct human interactions, which naturally allow for a higher degree of empathy, contextual understanding, and immediate response to emergencies. Moreover, in standard clinical settings without AI, the risks primarily focus on human error, confidentiality breaches through non-digital means, and less systematic but still possible misinterpretations of speech, emotion, and motion. The introduction of AI technologies heightens the risk by introducing systematic errors that could occur at scale and by adding layers of complexity regarding data security and privacy. Moreover, the introduction of chatbots introduces automation that can scale patient interactions but also brings the risk of systematized errors, reduced personal connection, and potential gaps in handling complex health situations.
Pilot 6	No added or heightened risks were detected in the AI technology used during the pilot compared to standard clinical or medical practices. Minor technical issues were noted with some devices, such as limited battery life of the smartwatch and operational limitations, but these did not elevate the overall risk profile.
Pilot 7	The pilot study did not introduce any increased risk, the AI-based tool was used in an offline evaluation. When the tool will once be introduced in clinical practice, it will serve as a support system for the physician, but final decision making for the treatment will be done by the physician.
Pilot 8	The worst case scenario is same as previous questionnaire that the AI will not be able to pinpoint anything useful in the patient data that might help the clinician's decision. The final responsibility for clinical decisions is always with the clinician, and the AI only serves to highlight possible useful connections in the patient data.

2.4.2 Detection of risks

The main issue here is what measures were (or should be, beyond HosmartAI) taken to detect the added or heightened risks, and what the limitations of these measures.

The objective of this question is to clarify and identify what measures are effective to detect added or heightened risks due to HosmartAI, including their limitations.

Q. 11. **Detection**. If there are any **added or heightened risk(s)**, (1) what are the measures you took or safeguards you implemented to **detect** the risk? (2) Or are these risks difficult to detect? (3) When your AI technology is used outside HosmartAI research setting (e.g., actual healthcare setting), what measures or safeguards should or can be implemented to detect the added or heightened risks associated with the use of AI technology of your pilot study? Your response can be technical, operational, organizational, etc. Please share your thoughts and explanations.

Pilot #	**Q. 11. Detection of risks**
Pilot 1 (ECHO)	(1) N/A (no added or heightened risks(s)) (2) N/A (no added or heightened risks(s)) (3) It needs a testing period during which the AI will be tested against known conditions in order to detect conditions for which the AI is not reliable.
Pilot 1 (VCE)	(1) N/A (no added or heightened risks(s)) (2) N/A (no added or heightened risks(s)) (3) It needs a testing period during which the AI will be tested against known conditions in order to detect conditions for which the AI is not reliable.
Pilot 2	(1) N/A (2) N/A (3) N/A
Pilot 3	(1) Not applicable, we do not foresee any additional risks (2) Not applicable, we do not foresee any additional risks (3) Not applicable, we do not foresee any additional risks
Pilot 4	(1) None (2) N/A (3) N/A

Pilot 5	(1) **Error Logging and Analysis: We implemented comprehensive logging of all interactions and decisions made by the AI systems. Using Block Chain**. **Continuous Performance Monitoring**: A trained operator was present during all the sessions to carry out real-time monitoring and to assess the AI's performance continuously, ensuring that outputs remain within expected parameters. **Validation and Testing**: Extensive testing against control groups and varied scenarios at pilot site was carried out over period of 2 years to validate the AI's decision-making and interaction capabilities under diverse conditions. (2) AI systems, especially those involving deep learning, can be "black boxes" making it difficult to understand why certain decisions are made. This opacity can make error detection challenging. (3) The performance of AI systems can vary significantly based on the data they were trained on, the environment in which they are deployed, and the specificity of the tasks they perform. Detecting errors due to these variances requires sophisticated analytical tools and expertise. To be used outside HosmartAI and research the following measures should be considered. **Technical Measures**: Tools that can diagnose issues in AI behavior by comparing it against expected outcomes and historical performance metrics. The enhanced logging process can significantly contribute. AI systems need to be integrated with existing healthcare IT systems to leverage holistic monitoring and management tools. **Operational Measures**: Standard Operating Procedures (SOPs) for using AI technologies need to be developed, including clear guidelines on when to trust AI decisions and when to seek human intervention. A comprehensive risk management framework that include risk assessment, mitigation, and continuous monitoring specific to AI technologies needs to be implemented. **Organizational Measures**: Create clear channels for feedback on AI performance from healthcare professionals and patients, using this feedback to fine-tune AI behaviors and responses need to be established. **Educational and Training Initiatives**: Ongoing education and training for all healthcare staff involved with AI tools, focusing on understanding AI capabilities, limitations, and the importance of reasoning over AI recommendations and maintaining a supervisory role.

Pilot 6	(1) Incidences detection and management file. Regular team meetings. There are no safeguards in our AI models because they only make recommendations that are always within the possibilities proposed by a qualified physician. These same healthcare professionals periodically review the system to ensure that they have a personalized plan for each patient. (2) Issues regarding the migrations of the data repositories. There was data lost and the process for recovering took time for re-establish the dataset (which included reestablishing profiles necessary for continuing pilot activities). (3) **Technical**: automating further the systems for data collection and quality verification **Operational**: pilot team and professionals trained to use the system and identify faults in the implementation in the different care settings that can be considered (e.g. home; clinical individual use; group sessions). **Organizational**: traceability plan tailored to each service entity's characteristics, including how to handle tool failures and compliance with ethical and legal requirements
Pilot 7	(1) N/A (2) N/A (3) N/A
Pilot 8	(1) The AI based models provided confidence levels based on the training data. (2) No (3) The AI based models generate an indication or prediction and it is crucial to verify and validate the results by the clinicians. Clinicians apply their expertise to evaluate the AI models' prediction in the context of the patients' unique circumstances.

2.4.3 Mitigation of risks

The main issue here is what measures were (or should be, for beyond HosmartAI) taken to mitigate the added or heightened risks, and what the limitations of these measures.

The objective of this question is to clarify and identify what measures are effective to mitigate added or heightened risks due to HosmartAI, including their limitations.

Q. 12. **Mitigation**. If there are any **added or heightened risk(s)**, (1) what are the measures you took or safeguards you implemented to **mitigate** the risks? (2) Or are these risks difficult to [mitigate]? (3) When your AI technology is used outside HosmartAI research setting (e.g., actual healthcare setting), what measures or safeguards should or can be implemented to detect the added or heightened risks associated with the use of AI technology of your pilot study? Your response can be technical, operational, organizational, etc. Please share your thoughts and explanations.

Pilot #	Q. 12. Mitigation of risks
Pilot 1 (ECHO)	(1) N/A (no added or heightened risks(s)) (2) N/A (no added or heightened risks(s)) (3) In a new setting, the AI should be fine-tunned and tested in data collected from the new setting. Since different ultrasound devices and data formats may be used in the new setting, it is important to ensure that the developed AI maintains its performance. Moreover, always a senior physician should make the final verdict. In other words, the AI does not decide but makes recommendations to the expert. If there is not a senior echocardiographer in the new setting, then some kind of training should be provided to the users of the AI in order to ensure that they will not fully rely on the AI outcomes.
Pilot 1 (VCE)	(1) N/A (no added or heightened risks(s)) (2) N/A (no added or heightened risks(s)) (3) Always a senior physician should make the final verdict. The AI does not decide but makes recommendations to the expert. If there is not a senior echocardiographer in the new setting, then some kind of training should be provided to the users of the AI in order to ensure that they will not fully rely on the AI outcomes.
Pilot 2	(1) N/A (2) N/A (3) N/A
Pilot 3	(1) Not applicable, we do not foresee any additional risks (2) Not applicable, we do not foresee any additional risks (3) Not applicable, we do not foresee any additional risks
Pilot 4	(1) None (2) N/A (3) N/A

Pilot 5	(1) We implemented stringent data validation, cleaning, and augmentation practices to ensure the AI is trained on high-quality, diverse datasets. This reduced the risk of biased or inaccurate outputs. We used traditional ML systems to verify and cross-check some of the AI-classifications related to Triage and DSS. We delivered a Human-in-the-loop system to ensure that critical decisions are reviewed or made with human oversight, especially in cases where the AI's recommendations could have significant consequences. We establish a multidisciplinary team combining tech. experts and clinicians, to oversee AI implementations and ensure they meet clinical standards and ethical considerations. (2) N/A (3) To use Pilot 5 AI technology outside the pilot study, the following measures should be considered **Technical Measures**: • **Advanced Monitoring Tools: Advanced monitoring and visualization tools to track AI performance in real-time, providing alerts for any deviations from expected behavior.** • Systematic Updates and Patching: Regular update of AI systems to address newly discovered vulnerabilities or errors, much like software patches in IT security. **Operational Measures**: • **Clear Escalation Pathways: Clear protocols for escalating issues from AI decision making to human medical professionals should be established.** • **Continuous Learning and Adaptation: Continuous learning systems where the AI can adapt and improve over time based on new data and feedback without compromising initial training stability should be implemented.** **Organizational Measures**: • **Regular Training and Simulations: Regular training sessions and simulations for medical staff to stay familiar with AI tools and their integration into clinical workflows.** • Compliance and Auditing: Regular audits to ensure AI applications comply with medical regulations and standards; involve external auditors for unbiased assessments.
Pilot 6	(1) We adhered to the same mitigation strategies as stated in D8.3. Periodic verification of data collected by the management system and of data stored in FHIR. Usability surveys were also used, although answering was optional. (2) Data visualization challenges should be addressed to improve the operational efficiency of data collection. (3) The mitigation strategies used in the pilot can be applicable outside of the research setting, focusing on technical and operational measures to ensure system reliability and adherence to ethical standards.
Pilot 7	(1) N/A (2) N/A (3) N/A

Pilot 8	(1) The clinicians will consider the results from the AI in relation to their normal diagnosis and feedback their interpretation on the usefulness of the AI results to the researchers. (2) We have tried to address all the risks in our DPIA report which is submitted to the ethical review committee. (3) The measures that we mentioned in our DPIA reports related to ICT, data collection, Data anonymization would be applicable in that case too.

2.4.4 Comments, opinions, questions, suggestions on AI ethics

The issue is whether the pilot received any comments, concerns, or questions regarding AI ethics issues from external parties. The objective of this question is twofold: to inquire if there was any communication and to learn from the content of those communications.

No issues that need further attention or discussion are identified. Pilot 5 noted that they received concerns by healthcare providers about the reliability of AI systems and accountability for decisions made by CDSS. Nevertheless, the way in which Pilot 5 responded and handled these concerns demonstrate how their regular engagement sessions with stakeholders as well as the co-designing process were effective in addressing those concerns.

Q. 13. Did you receive any comments, opinions, questions, suggestions, or anything similar from anybody (for example, pilot study participants, your DPO, legal department, or any other department/team of your institution) regarding **AI ethics issues** that may be relevant to this questionnaire? If yes, please describe and share your thoughts.

Pilot #	Q. 13. Comments, opinions, questions, suggestions on AI ethics
Pilot 1 (ECHO)	No
Pilot 1 (VCE)	No
Pilot 2	N/A
Pilot 3	None
Pilot 4	None
Pilot 5	We carried out regular engagement sessions with stakeholders, including pilot study participants, healthcare professionals, and ethicists, to gather and address feedback systematically. The focus of these exercises was to design and deliver a trustworthy and ethically acceptable set of services and functionalities offered by the robotic nurse. Co-designing the robotic nurse with clinicians and nurses ensured that the technology augments rather than disrupt clinical workflows. Continuous feedback loop was established with between UM (system developers) and UKCM (healthcare providers) to refine the tools. Concerns about the reliability of AI systems and accountability for decisions made by AI were frequently raised by healthcare providers. Thus**,** implemented robust testing protocols, a robust logging process and human-in-the-loop oversight process, as the baselines to establishing clear accountability when the system is used in patient care.
Pilot 6	There were no specific comments or feedback received related to AI ethics issues during the pilot study.
Pilot 7	No
Pilot 8	None

2.4.5 Ethical, Legal, or Social issues to be shared with WP8

The issue is whether the pilot had issues in relation to SELP that they wish to share with WP8, and the objective is to identify those issues, if any. No issues that need further attention or discussion are identified.

Q. 14. If you have any ethical, legal, or social issues that you were concerned about while conducting your pilot study, please share with us.

Pilot #	Q. 14. Ethical, Legal, or Social issues to be shared with WP8
Pilot 1 (ECHO)	No
Pilot 1 (VCE)	No
Pilot 2	N/A
Pilot 3	None
Pilot 4	None
Pilot 5	N/A
Pilot 6	No new ethical, legal, or social issues were identified during the conduct of the pilot study.
Pilot 7	None
Pilot 8	N/A

2.5 AI Bias and Explainable AI

This section addresses issues related to AI bias and Explainable AI. Specifically, it covers three topics: datasets quality, potential AI biases, and transparency and explainable AI. These issues are discussed in the context of both during and beyond the HosmartAI project. These topics were added in response to the feedback provided by the Commission reviewers (Review Report[10]).

The analyses of all subsections are discussed together in Section 2.6, *infra*, because they are interconnected and integral to each other.

2.5.1 Datasets quality (during HosmartAI)

There are two main issues here: (1) Whether pilot partners were able to collect datasets that sufficiently represent various groups to avoid AI bias; and (2) If not, what measures have they taken to mitigate AI bias.

There are two main objectives: (1) to analyse whether the pilot partners appropriately addressed and paid attention to the AI bias issue at the dataset collection level, and (2) to learn from the insights each pilot gained in addressing the dataset quality issue and share them within the HosmartAI project (this objective is common to all in this section).

Q. 15. **Datasets quality (during HosmartAI)**. (1) Were you able to collect and use datasets (i.e., training set, validation set, and test set) that are sufficiently representative of various groups -- e.g., gender/sex, race or ethnic origin, age, etc -- during HosmartAI pilot study? Did you observe any imbalance in your datasets, or datasets being skewed towards some subset of the group? If so, what group is underrepresented?
(2) If the datasets were insufficiently representative of various groups, what measures have you taken to overcome some of the potential negative consequences (e.g., decrease in accuracy, insufficient generalization, potential biases, etc)?

[10] See the final review report issued on the 26th of September 2023.

Pilot #	Q. 15. Datasets quality (during HosmartAI)
Pilot 1 (ECHO)	(1) The AI developed using a big publicly available dataset that contains 10,030 echocardiography videos. The dataset is balanced regarding the demographics, health conditions, and devices. Thus, no imbalance or any underrepresented group have been observed. A small new dataset collected during the pilot was used only for testing. Although small, it is balanced regarding the health conditions. (2) Since the datasets are considered balanced, no measures for bias mitigation took place.
Pilot 1 (VCE)	(1) The AI developed using a big publicly available dataset that consists of 117 videos which can be used to extract a total of 4,741,504 image frames. The main bias issue regards the pathologies exist in the dataset, i.e., some rare conditions are underrepresented in the dataset. (2) To mitigate bias imbalance, we apply data augmentation to balance the dataset.
Pilot 2	(1) We were supposed to collect data from 20 lung cancer patients and we collected data from only 3 of them (2) The only way to compensate for this low recruitment rate would have been to start the clinical study at least 6 to 8 months before.
Pilot 3	(1) Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it. (2) Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it.
Pilot 4	(1) We were able to collect and use datasets (always anonymized) that were sufficiently representative of various groups, no gender preferences nor any other distinction was made. However, enlarging dataset would be convenient to test better the HosmartAI technology. (2) N/A
Pilot 5	(1) yes, we have collected and used multiple datasets openly and publicly available for research. Overall, jointly they were sufficiently representative of the targeted group. Please consider that no AI was used for clinical decision support. (2) N/A

Pilot 6	(1) Pilot 6 included three implementation scenarios with a number of participants per scenario: **Scenario A**: older adults using the system at home independently (interface: tablet // n=60) **Scenario B**: older adults with mild cognitive impairment using the system in clinical sessions with therapists (interface tablet and robot// n=25) **Scenario C**: older adults in AHA group program (interface robot// n=50 planned, but just 22 ensured for the moment) **Control Group** (n=25) Limitation of number of social robots available. (2) Pilot 6 has limited use of AI. The sample size was statistically determined as sufficient for the pilot study hypothesis and study conditions. The study is unbalanced in terms of age but because it is aimed at older people who require care. The data will only be used to determine if any feature specifically affects the usability responses of the application, so it is considered appropriate.
Pilot 7	(1) Sufficient data has been collected, of good quality and with sufficient variation. (2) N/A
Pilot 8	(1) The data is limited in our pilot as Glioma is a rare disease. This project serves as a pilot study to examine the feasibility and usefulness of the AI algorithm. (2) N/A

2.5.2 Datasets quality (beyond HosmartAI)

The same question is asked in anticipation of a situation where HosmartAI technology is introduced to the European market and used in healthcare practice, following the conclusion of the HosmartAI project. Thus, the question is hypothetical in nature, and the objective is to learn from the insights of each pilot partner and to share them within the HosmartAI project as organisational knowledge.

Q. 16. **Datasets quality (beyond HosmartAI)**. (1) In your view, what measures can, and should, be taken in terms of collecting and using datasets (i.e., training set, validation set, and test set), ensuring that they are sufficiently representative of various groups -- e.g., gender/sex, race or ethnic origin, age, etc -- when the HosmartAI technology of your pilot study is actually used in healthcare and placed in the European market? Please share your expertise/explanation in terms of various layers (e.g., technical, organizational, etc) or otherwise you see fit.

(2) In your view, what other quality criteria should datasets meet, so the HosmartAI technology of your pilot study would be more responsible, accountable, transparent, and trustworthy, when the HosmartAI technology is actually used in healthcare and placed in the European market? Please share your expertise/explanation in terms of various layers, or otherwise you see fit.

Pilot #	Q. 16. Datasets quality (beyond HosmartAI)
Pilot 1 (ECHO)	(1) Regardless the size of the dataset, a full documentation of the dataset's characteristics, i.e. demographics, health conditions, device types and data acquisition-related conditions, should be provided in order the AI developer/engineer to assess the value of the dataset and decide the proper AI methodology to develop a respective solution. (2) Our AI should be tested in a much larger multi-center randomised clinical study in order to have more robust and persuasive results.
Pilot 1 (VCE)	(1) Regardless the size of the dataset, a full documentation of the dataset's characteristics, i.e. demographics, health conditions, device types and data acquisition-related conditions, should be provided in order the AI developer/engineer to assess the value of the dataset and decide the proper AI methodology to develop a respective solution. (2) Our AI should be tested in a much larger multi-center randomised clinical study in order to have more robust and persuasive results.
Pilot 2	(1) Does not apply to appointment scheduling software in radiotherapy units where all patients, regardless of gender, race, cancer or ethnicity, must be treated according to law. (2) The dataset can be enhanced by increasing more variables regarding the patient personal/medical situation and hospital logistics as human and machine resources.
Pilot 3	(1) Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it. (2) Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it.
Pilot 4	(1) Should be no issues after HosmartAI considering that in our European Healthcare scenario that takes care of the entire population, no distinction among genders nor any groups is expected (2) Specially, it should comply with European Health Institutions and regulations as usual for any new medical device or pharmacotherapy

Pilot 5	(1) • Diverse Data Collection: Collect data from a wide variety of sources, including different geographic locations, healthcare settings, and demographics. This helps ensure that the data encompasses a broad spectrum of patient characteristics such as gender, age, race, and ethnicity. • Stratified Sampling: Use stratified sampling techniques to ensure that all relevant subgroups are adequately represented in the training, validation, and test sets. This approach helps in maintaining the proportionality of each subgroup within the dataset. • **Stakeholder Collaboration: Collaborate with a wide range of stakeholders, including hospitals, clinics, and patient advocacy groups, to gather comprehensive data and insights. This collaboration can help identify and address gaps in data collection.** • **Continuous Monitoring and Auditing: Regularly review and audit datasets for representativeness and bias. This should be an ongoing process as the model may drift over time due to changes in population demographics and disease patterns.** • Documentation and Transparency: Maintain thorough documentation about data sourcing, criteria for inclusion, and methods used for data processing. This documentation should be accessible to regulators and stakeholders to ensure transparency. (2) **The researchers should verify that the data is complete, consistent, and accurate. techniques to validate the accuracy of the data, including cross-validation and external validation using independent datasets should be carried out before the final release of the models. If a bias is detected, techniques such as re-sampling, re-weighting, and algorithmic fairness interventions can be used to mitigate bias.** Beyond tech. considerations, impact assessments to understand how AI decisions affect different groups should be carried out to understand unintended consequences of AI and addressing them proactively.
Pilot 6	(1) For future implementations beyond the pilot, enhancing dataset quality can involve extending the application period of core tools and modules, and potentially conducting sub-studies focused on specific modules. (2) Improvements could include enhanced data visualization tools on the dashboard to better monitor and analyse data, ensuring data quality and representativeness in ongoing and future research settings. (3) Provide accessibility to the datasets by making them public as they are anonymized.

Pilot 7	(1) Data collection should be facilitated via a scalable solution that can be easily deployed at clinical sites, such that sufficient variability in data points can be achieved. (2) No immediate suggestion for additional quality criteria, but in the process of training and evaluating AI-based applications, a thorough evaluation process should be in place, ensuring good quality annotations and a proper reader study to check the AI-application's performance against human experts.
Pilot 8	(1) Our pilot study faces limitation due to the rarity of Glioma, which impacts the availability of data. (2) We need to test and validate this technology further with more patients in coming years to answer this question. We need to ensure data relevance to meet the requirements of the intended application. Make sure that the data collection methods follow the ethical guidelines.

2.5.3 Potential AI biases and algorithm generalization (during HosmartAI)

The overarching issues are: what measures did pilot partners take (1) to detect and correct AI bias problem; and (2) to improve generalization and prevent overfitting to training data.

One of the two objectives is to analyse whether pilot partners paid appropriate attention to and addressed the AI bias problem and overfitting to training data, which both can result in inaccuracy of output generated by the AI system.

Q. 17. **Potential AI biases and algorithm generalization (during HosmartAI)**.
(1) What kind of measures have you taken (e.g., test, mechanism, etc) to **detect** if the HosmartAI technology of your pilot study has or manifests any "AI bias" -- based on gender/sex, race or ethnic origin, age, etc -- during the pilot study of HosmartAI? Also, please describe the AI bias(es) if you observed any during HosmartAI pilot study, whether potential or apparent?
(2) What kind of measures have you taken -- e.g., technical (including computational or statistical), organizational (including human or systematic), or otherwise -- to correct the AI bias you **detected** during HosmartAI pilot study? Please share your insight and story and as to how you corrected the AI bias you detected.
(3) What measures -- technical, organizational, or otherwise -- have you taken to improve **generalization** and prevent **overfitting** to training data? Is the **generalization performance** during HosmartAI pilot study satisfactory in your view that it provides equitable healthcare for diverse (e.g., in terms of gender/sex, race or ethnic origin, age, etc) population?
Note: A few examples of AI bias in healthcare:
(1) algorithms trained with gender imbalanced data do worse at reading chest x-rays for an underrepresented gender (https://doi.org/10.1073/pnas.1919012117); or
(2) skin-cancer detection algorithms, many of which are trained primarily on light-skinned individuals, do worse at detecting skin cancer affecting darker skin (https://doi.org/10.1001/jamadermatol.2018.2348).

Pilot #	Q. 17. Potential AI biases and algorithm generalization (during HosmartAI)
Pilot 1 (ECHO)	(1) **Detecting AI bias**. The distributions of the different characteristics, e.g., age, sex, etc., of the dataset were plotted to visualise if there is some evident bias. Given that the datasets used were created to be balanced no significant imbalanced was anticipated, nevertheless this visual analysis took place. One bias that was considered possible was the one created by the device type. Different devices create ultrasound images of different image characteristic, e.g., resolution, contrast, etc. Thus, a device type is not very common, e.g. a portable ultrasound device, it might not be included in the datasets used. To detect this type of bias, we collected data from portable ultrasound devices, which were not included in the training set, and checked the distribution of the AI outcomes. (2) **Correcting AI bias**. To correct the possible distribution shift that can be observed in the AI outcomes when the input comes from new device types, the mechanisms used were, on the one hand, to equalise the histogram of every new image to be same as the images used during training and, on the other hand, to use a domain adaptation technique to validate how unbiased the model is. Specifically, the domain adaptation technique keeps all layers of the deep neural network except for the last one are frozen and the last one is replaced by a new one with the number of output neurons to be same as the number of different types of devices exist in the dataset, including the uncommon new ones. Then, we train the network to classify the device type, if it fails then the frozen layers do not comprise any device-specific information, i.e., no device type bias. (3) **Improving generalization and preventing overfitting**. Techniques to ensure generalization and preventing overfitting are: -data augmentation before training -early stopping regularization -use of dropout layers - carefully split the dataset into training/validation/testing to avoid data leakage
Pilot 1 (VCE)	(1) **Detecting AI bias**. From the documentation of the datasets (public available and collected videos), it is evident that the pathology categories are imbalanced. Thus, if no measure is taken, possible AI bias will be created. Given the nature of the data, endoscopic videos of the small bowel, we do not expect significant demographic bias. (2) **Correcting AI bias**. We train the AI using weighted loss functions. (3) **Improving generalization and preventing overfitting**. Techniques to ensure generalization and preventing overfitting are: -data augmentation before training -k-fold cross-validation -use of dropout layers -carefully split the dataset into training/validation/testing to avoid data leakage

Pilot 2	(1) **Detecting AI bias**. None (2) **Correcting AI bias**. N/A (3) **Improving generalization and preventing overfitting**. N/A
Pilot 3	(1) **Detecting AI bias**. Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it. (2) **Correcting AI bias**. Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it. (3) **Improving generalization and preventing overfitting**. Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it.
Pilot 4	(1) **Detecting AI bias**. The entire dataset that was used during HosmartAI was representative of a regular European Health System. No bias was detected regarding groups. (2) **Correcting AI bias**. There was no need for a correction (3) **Improving generalization and preventing overfitting**. We tried to keep as general as possible the utilization of our dataset in order to avoid overfitting. In this regard, no special data was collected, all the collected data were considered regular in terms of subjects pathology

Pilot 5	(1) **Detecting AI bias**. **We Conducted performance evaluations of speech recognition and chatbot systems across different demographic groups.** We measured and compared error rates in speech recognition for in the wild for different, dialects, and accents. Similar analysis was applied to chatbots in terms of their ability to understand and correctly respond to diverse linguistic and cultural expressions and to factor in the word recognition errors and misspellings performed by speech recognition. **Cross-validation during model training for emotion/distress recognition using multiple models was carried out to ensure the AI system generalizes well to unseen data.** We deployed the system beyond HosmartAI and implemented feedback mechanisms where users can report misunderstandings or dissatisfaction, providing direct insights into potential areas of bias. (2) **Correcting AI bias**. **We augmented the training datasets with a range of accents, dialects, and speech patterns. For chatbots, include diverse linguistic and cultural scenarios in the training data.** **We adjusted the AI models to penalize bias.** The speech recognition system is implemented as a continually learning system also exploiting the [r]eal-world interactions. The newly generate datasets are manually reviewed and errors identified through user feedback corrected. (3) **Improving generalization and preventing overfitting**. **We implemented cross-validation techniques to ensure the models generalize well beyond the training data, particularly for handling a wide variety of inputs in the wild. The models are regularly evaluated against the system's performance in real-world settings, especially focusing on how well it handles inputs from underrepresented groups.** We continuously update and retrain the models using newly collected data that reflect ongoing changes in language and communication styles.
Pilot 6	(1) **Detecting AI bias**. The pilot study did not specifically detect AI biases due to the limited use of AI technologies. However, standard testing mechanisms were in place to monitor for any potential biases that could arise. (2) Correcting AI bias (2) **Correcting AI bias**. Given the limited detection of biases, no specific measures were taken to correct AI biases during the pilot study. (3) **Improving generalization and preventing overfitting**. The generalization performance was considered satisfactory for the scope of this pilot. Overfitting does not apply in this case.

Pilot 7	(1) **Detecting AI bias**. During the last period in the HosmartAI project, the focus in pilot 7 has been on creating a thorough Quality System for Data Annotation at scale, because it is essential for the development of AI-based applications to have good quality annotations. Three main phases have been identified: Preparation, Quality Assurance, Quality Control. During the Quality Control phase, drift in annotation quality is checked. (2) **Correcting AI bias**. This step is part of the overall data annotation quality system (3) **Improving generalization and preventing overfitting**. This step is part of the overall data annotation quality system
Pilot 8	(1) **Detecting AI bias**. One of the known biases on genomic driver mutation prediction is the fact some cancer genes are very frequently studied and much more prevalent in the public databases. This can create bias in the training data: predictors often do not interpret the mutation itself, but rather decide on whether the mutation occurs in a well-known cancer related gene. (2) **Correcting AI bias**. The genomic D2Deep model is trained on a balanced training set comprised the same amount of classes for all genes studied and so effectively mitigates biases related to hotspot mutations compared to state-of-the-art techniques. (3) **Improving generalization and preventing overfitting**. For the genomic D2Deep model we mitigated overfitting during training by implementing various generalization techniques such as dropout regularisation, Lasso Regression (L2 Regularization) and early stopping. Cross-validation was used to assess the generalization performance of the model.

2.5.4 Potential AI biases and algorithm generalization (beyond HosmartAI)

The same question is asked in anticipation of a situation where HosmartAI technology is introduced to the European market and used in healthcare practice, following the conclusion of the HosmartAI project. Thus, the question is hypothetical in nature, and the objective is to learn from the insights of each pilot partner and to share them within the HosmartAI project.

Q. 18. **Potential AI biases and algorithm generalization (beyond HosmartAI)**. (1) In your view, what kind of measures can, and should, be taken to **detect** AI biases when the HosmartAI technology of your pilot study is actually used in healthcare and placed in the European market? Please share your expertise/explanation in terms of: various layers (e.g., technical, organizational, etc), AI life cycle (e.g., design, development, deployment, use/operation, monitor, train/validate/test, etc), or otherwise you see fit.

(2) In your view, what kind of measures can and should be taken to **correct** the AI biases you **detected** when the HosmartAI technology of your pilot study is actually used in healthcare and placed in the European market? Please share your expertise/explanation in terms of various context, layers, AI life cycle, as you see fit.

(3) In your view, what measures can and should be taken to improve **generalization** and prevent **overfitting** to training data, so it provides equitable healthcare for diverse (e.g., gender/sex, race or ethnic origin, age, etc) population when the HosmartAI technology of

your pilot study is actually used in healthcare and placed in the European market? Please share your expertise/explanation in terms of various context, layers, AI life cycle, as you see fit.

Pilot #	Q. 18. Potential AI biases and algorithm generalization (beyond HosmartAI)
Pilot 1 (ECHO)	The same approaches used during HosmartAI can be used beyond HosmartAI too.
Pilot 1 (VCE)	The same approaches used during HosmartAI can be used beyond HosmartAI too.
Pilot 2	(1) **Detecting AI bias**. N/A (2) **Correcting AI bias**. N/A (3) **Improving generalization and preventing overfitting**. N/A
Pilot 3	1) **Detecting AI bias**. Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it. (2) **Correcting AI bias**. Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it. (3) **Improving generalization and preventing overfitting**. Aspects related to diversity, non-discrimination and fairness do not apply to Pilot 3. The AI-based solution monitors the environment and does not collect and/or use any information about the people who use it.
Pilot 4	(1) **Detecting AI bias**. Probably, regular validation should be planned to detect AI bias (2) **Correcting AI bias**. We strongly believe that, as stated before, regular validation and regular review by clinical staff should probably be planned if an AI technology goes on to further development (3) **Improving generalization and preventing overfitting**. By keeping always in mind that further improvement of the technology should ideally always be supported by data from regular clinical scenarios

Pilot 5	(1) **Detecting AI bias**. • **Performance evaluations across different demographics to identify any discrepancies in AI behavior or outcomes. Both the validation phase and continuous monitoring after deployment.** • **Explore tools that can automatically flag potential biases by analysing the AI's decisions across various segments of data.** • **Bias detection as a core part of the development phase, using tools and methodologies that can identify bias in training data and model output.** • Post-Deployment: Continuous Human monitoring to detect biases as they emerge in real-world settings, and regular review of AI performance. (2) **Correcting AI bias**. If biases are detected, retrain the model using more representative data or modify the model architecture to mitigate biases. Regular updates and patches should be carried out to address any emerging biases or disparities in AI performance. (3) **Improving generalization and preventing overfitting**. • Cross-validation during model training to ensure the AI system generalizes well to unseen data. • Use of dropout, L1 or L2 regularization to prevent the model from learning noise in the training data. • Systems to continuously monitor the model's performance in real-world applications to quickly identify and address overfitting. • Adaptive learning where the model can update itself from new data under strict privacy and ethical guidelines, ensuring it remains relevant and effective across diverse patient groups.
Pilot 6	(1) **Detecting AI bias**. To detect AI biases when the technology is deployed in healthcare settings, continuous monitoring and regular audits of AI systems are generally recommended, specifically because pilot 6 solution is developed for ensuring incremental integration of new modules that can include further AI aspects. Technical safeguards should include routine checks for data integrity and bias detection. (2) **Correcting AI bias**. Although there is limited use of AI technologies in the current prototype, future version for the market can be upgraded and, in such case, for each module in which this applies, corrective measures should involve recalibrating the AI models periodically with updated, diversified datasets to mitigate any detected biases. (3) **Improving generalization and preventing overfitting**. Although there is limited use of AI technologies in the current prototype, future versions for the market can be upgraded and, in such case, it would be good if AI systems can be tested across diverse demographic settings.
Pilot 7	The same methods applied during HosmartAI will be applied beyond the project.

Pilot 8	(1) **Detecting AI bias**. Detecting AI bias: Analysis of possible bias, use of balanced datasets or gender-specific algorithms can be considered. (2) **Correcting AI bias**. N/A (3) **Improving generalization and preventing overfitting**. Improving generalization and preventing overfitting. Here, initiatives for collecting and sharing large, representative datasets are of importance. Federated learning could aid in overcoming data privacy issues.

2.5.5 Transparency and Explainable AI (during HosmartAI)

The overarching issue here is explainable AI. One objective is to analyse how the pilot partners aimed to implement their HosmartAI technology to provide meaningful information, particularly in situations where the accuracy of the output is questionable.

Q. 19. **Transparency and Explainable AI (during HosmartAI)**.
Explainable AI (XAI) -- the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes -- is a key to ensuring or striving for responsible, accountable, transparent, or trustworthy AI, as well as to avoiding or mitigating AI biases.
(1) Is the HosmartAI technology of your pilot study capable of providing or generating evidence, support, or reasoning (“explanations” or “information”) related to an outcome from or a process of the AI system?
(2) Who is the intended recipient(s) of these explanations/information? Would you please describe/explain how these explanations or information help the recipient(s) understand why and/or how the AI system generated the output? Does the explanation/information help recipient(s) to contest, challenge, or falsify the output of the AI system?
(3) Does the HosmartAI technology of your pilot study identify cases in which they were not designed or approved to operate, or in cases for which their outputs are unreliable? In such cases, does the AI system provide or generates explanations/information regarding technical limitations and potential risks, such as its level of accuracy and/or error rates?

Pilot #	Q. 19. Transparency and Explainable AI (during HosmartAI)
Pilot 1 (ECHO)	(1) Yes. Part of the AI algorithm is the segmentation of specific cardiac chamber. When the algorithm fails in segmentation then the final AI outcomes are wrong. This is why the segmentation outcome is displayed in order the physician to self-assess if he/she can trust the outcome. (2) The physician who is in charge of conducting the respective measurements. The visualization of the segmentation explains what cardiac area was used for producing the final measurements. Thus, it helps them to self-assess if the AI makes some mistake. (3) During the pilot study, there was not any cases other than what was designed.
Pilot 1 (VCE)	(1) The AI model used is one that applies explainability by design. Specifically, the developed AI is based on RetinaNet neural network which both classifies an image in one category and returns a bounding box which indicates the region that activates the respective classification outcome. Thus, the gastroenterologist is capable of self-assessing if the outcome is valid or the AI makes a mistake. (2) The gastroenterologist who is in charge of reading the capsule endoscopy video. In some gastroenterology departments, the nurses make a first reading of the video, which is called pre-reading, in order to specify the areas of interest for the gastroenterology in charge. They can also benefit by the developed AI. (3) During the pilot study, there was not any cases other than what was designed.
Pilot 2	(1) The software created in the Pilot 2 is not AI-based but an optimization software. Users (doctors and appointment coordinators) know that the appointments generated by this optimization software are the result of transferring patient data from different hospital sources to the software. (2) The users are doctors and appointment coordinators. Without knowing what other hospital sources the appointments are based on, users will not be able to react in the event of an inconsistency in the system. (3) No

Pilot 3	(1) Our virtual sensors (VS) detect anomalies, which can be explained by analysing the logs of our notification service. For instance, if there is an anomaly related to lights being on, we can verify whether the lights were actually on and cross-reference this with historical usage patterns. We have also developed tools and services to explore historical data, providing valuable insights. Currently, our VSs are in the experimentation phase. Once the product is released to the market, it will be possible to include additional explanations for users, helping them understand what is happening and what caused the reported anomalies. (2) Our AI solution is designed for physiotherapists, caregivers, and medical personnel. The users have received training on the purpose and functionality of the sensors. Additionally, we have developed tools that allow the users to explore data easily. Thanks to these visualization tools, users can verify the accuracy of anomalies reported by VSs by checking patterns in historical data. (3) Vimar internal tests have revealed that infrequently used environments are not suitable for monitoring with VSs because extracting usage patterns is challenging. During the training phase, we communicate this information to users, explaining which environments are conducive to VS installation. When the application is released to the market it will be possible to include in the app the explanation about possible limitations, so the users can be informed before activating the sensors. Additionally, during operators training, we emphasize that the AI solution may produce incorrect outputs, however data can be verified using the visualization tools we have developed, so that the users can identify potential mistakes.
Pilot 4	(1) It is not (2) NA (3) NA
Pilot 5	(1) AI models for DSS in Pilot 5 are based on decision trees and RF, which inherently offer more transparency about how decisions are made. For more complex models of distress classification that use AI, LIME and SHAP were employed to provide insights into the decision-making process. Explanations were aimed to make the AI's decisions transparent, helping clinicians, researchers and developers to understand why certain recommendations or decisions were made. The main aim was to build trust and allow healthcare professionals to better integrate AI assistance into clinical decision-making. (2) • Doctors, nurses, and other clinical staff who participated in the use of the AI system to make or support clinical decisions. Explanations targeted to help validating the AI-generated insights and integrating them with their clinical judgment. They also enable to visualize the rationale behind AI-supported decisions to patients. • Patients, since patient interaction with the AI was direct. The explanations target to increase transparency and trust, making AI tools more acceptable to patients and explaining how their data is used and why certain recommendations are made. (3) No

Pilot 6	(1) The pilot focuses on patient care and evaluations, so any conclusions drawn must be the result of a professional analysis of the data, not an automatic one. (2) For the reason explained in the previous answer, there is no automatically generated output, so there cannot be a falsehood derived from an automatic response. (3) No
Pilot 7	(1) XAI was not the primary objective of the pilot 7 activities. The focus was on setting up a proper data annotation environment and performing reader studies to evaluate the performance of a QCA algorithm by means of collecting feedback from expert physicians. Transparency is embedded by having the clinical experts evaluate the segmentations performed by the QCA algorithm and gather their feedback on the algorithm results. (2) Interventional Cardiologists are invited to provide their feedback on the algorithm’s performance. (3) Likely yes, but need to check.
Pilot 8	(1) One algorithm, that provided automatic segmentation based on MRI, provided probabilities maps for regions belonging to a certain tissue type, allowing for a more interpretable output compared to deterministic segmentations. During the implementation of cancer driver predictor, we paid significant attention to the explanation of predictions. The selected features have been demonstrated to elucidate the rationale behind each prediction and link novel mutations with previously validated ones. (2) Clinicians: The recipient of this information is the clinician. It allows a better understanding of the certainty of the automatically predicted segmentation and where mistakes might have been made by the network. Moreover, it allows an interpretation of the mixed nature of these tissues and gradual transition of one tissue into another (e.g. tumorous tissue into healthy tissue). The confidence score provided by the genomic mutation predictor can help clinicians and scientists to challenge the prediction output. Moreover, the features offer insights into which specific parts of the protein influenced the prediction. (3) Same as above.

2.5.6 Transparency and Explainable AI (beyond HosmartAI)

The same question is asked in anticipation of a situation where HosmartAI technology is introduced to the European market and used in healthcare practice, following the conclusion of the HosmartAI project. Thus, the question is hypothetical in nature, and the objective is to learn from the insights of each pilot partner and to share them within the HosmartAI project.

Q. 20. **Transparency and Explainable AI (beyond HosmartAI)**.

(1) In your view, can or should the HosmartAI technology of your pilot be capable of providing or generating any other explanations/information (e.g., evidence, support, or reasoning) related to an outcome from or a process of the AI system, when it is actually used in healthcare and placed in the European market, so it can further achieve trustworthy AI?

(2) In your view, can or should there be any other intended recipient(s) of these explanations/information, when the HosmartAI technology of your pilot study is actually

used in healthcare and placed in the European market, so it can further achieve trustworthy AI?

(3) In your view, can or should there be any other cases in which the HosmartAI technology of your pilot study provides or generates explanations/information of its technical limitations and potential risks (e.g., level of accuracy, error rates, etc), when it is actually used in healthcare and placed in the European market, so it can further achieve trustworthy AI?

Pilot #	Q. 20. Transparency and Explainable AI (beyond HosmartAI)
Pilot 1 (ECHO)	(1) If the objective is to automatically measure the metrics targeted in the HosmartAI project, i.e. ejection fraction and global longitudinal strain of left ventricle, then no change in the explanation should take place. However, if other objectives are targeted, e.g., to detect specific pathologies, then other explainability approaches should be used. If the detection of a pathology is in focus, then class activation mapping techniques should be used. (2) Apart from the cardiologists/echocardiographers, nurses of cardiology and/or emergency departments are possible recipients of the developed AI and its explanations. (3) The developed AI and its explanation provide a subset of metrics measured in a typical echocardiography examination. There are other metrics which can also be automatically measured if transfer learning takes place. Nevertheless, beyond echocardiography (the ultrasound of heart), we cannot foresee other opportunities for the developed AI.
Pilot 1 (VCE)	The same approaches used during HosmartAI can be used beyond HosmartAI too.
Pilot 2	(1) We can always explain more if necessary. It depends on the public (2) During the project we tried to join industries specializing in radiotherapy, SMEs and hospitals. However, efforts still need to be made. (3) This technology can be developed in all other services that require serial appointments such as oncology.
Pilot 3	(1) Currently, our AI system is constrained to a testing environment with fixed parameters. Specifically, the interaction with the AI occurs via a Telegram Bot. However, for a European market release, a more customized solution will be created, such as a mobile app. This app could inform users about any limitations, associated risks, and provide additional context related to the system's outcomes. For example, if a user receives a notification about a potential anomaly with lights being on, the app could offer insights into the usual behavior of lights during that specific hour. (2) Explanations and information should be available to all the following subjects involved in the care process: clinicians, patients, caregivers. (3) We know that our AI system have limitations, for example, it does not perform well in rarely used environments. The mobile app allowing users to interact with VSs should guide them during their activation to limit usage to environments where the sensors reach their maximum potential. It is also a good idea to guide users during service activation to enable only sensors truly useful for that particular context: this way, our AI system will function better and result in fewer errors. However, it is essential to inform users that all AI solutions can make mistakes, so it is crucial to verify the feedback provided.
Pilot 4	(1) Not at this time (2) NA (3) NA

Pilot 5	(1) Additional possible Explanations beyond XAI: • **Causal Reasoning: optimality based on causal relationships. This requires advancements and modifications in AI architectures to include causal inference models.** • Contextual Information: contextualized explanations based on the clinical scenarios, including how similar cases have been handled and the outcomes of those cases. (2) • **Clinical Decision Support Teams: Teams that focus on integrating AI insights into broader clinical decision-making processes could benefit from detailed explanations to coordinate care more effectively.** • Healthcare Administrators and Policy Makers: Providing them with explanations can help in resource allocation, policy formulation, and strategic planning. • Ethics Bodies and DPOs: They need to understand AI processes to ensure ethical guidelines and compliance standards. (3) • **Adverse Event Prediction: When predicting adverse events, the system should explain the factors leading to such predictions and the associated uncertainty to allow pre-emptive actions to be taken.** • Longitudinal Patient Care: As AI systems increasingly play a role in managing chronic conditions, providing ongoing explanations about how patient data trends over time influence the AI's recommendations could be crucial. **Limitations**: • **Information on how the AI's performance varies across different clinical environments and patient conditions.** • **As AI models evolve, explanations regarding how updates affect model performance and decision-making criteria should be transparently communicated to all users.**
Pilot 6	(1) The pilot focuses on patient care and evaluations, so any conclusions drawn must be the result of a professional analysis of the data, not an automatic one. (2) The pilot focuses on patient care and evaluations, so any conclusions drawn must be the result of a professional analysis of the data, not an automatic one. (3) The pilot focuses on patient care and evaluations, so any conclusions drawn must be the result of a professional analysis of the data, not an automatic one.
Pilot 7	(1) To achieve a trustworthy solution, transparency aspects should be embedded in the product development. Setting up a co-creation environment that allows customers to evaluate algorithms at an early stage and interact with the AI-model using own data sets will increase the trustworthiness of the solution. (2) Main recipient should be the interventional cardiologist. (3) The trained algorithm is very specific for interventional cardiology and cannot be applied in any other clinical domain. However, the general approach to co-create with customers and follow the thorough data annotation process and algorithm evaluation can be applied in other clinical domains that involve image based diagnosis and treatment.

Pilot 8	(1) The cancer driver mutation predictor uses features that capture the effect of the mutation throughout the protein and can be used for the interpretation of results in the clinical setting, complemented by the provided confidence score. For image segmentation tool we calculate DICE score. Both the metrics refers to deployed AI model's confidence in its prediction or decision. (2) No, the main recipient should be the user, thus the clinician. However, the data produced in this analysis could be useful for further analysis and improvement of new technologies (e.g. focus on improving where the algorithm is not sure about its prediction), so sharing it with researchers could be of benefit. (3) We provide Dice scores for image segmentation model and confidence levels for cancer driver mutation predictions (genetic analysis). The uncertainty scores are important for use in clinics, allowing insight on when manual intervention is necessary.

2.6 Findings and Analyses

This section presents our analyses on (1) Section 2.4 (Ethical and Societal Impact/Risks of HosmartAI technology) and (2) Section 2.5 (AI Bias and Explainable AI), respectively, in the following Subsections: (1) 2.6.1 Added or heightened risks; and (2) 2.6.2 AI Bias and Explainable AI.

2.6.1 Added or heightened risks

Section 2.4 focused on the added or heightened risks associated with HosmartAI technology when used as part of pilot studies. It also covered measures for detecting and mitigating identified risks, if any.

In summary, responses from all pilot partners fall into two categories: (1) there are no added or heightened risks due to the HosmartAI project; (2) there are added or heightened risks, but the pilot partner anticipates these risks and has designed the pilot study with appropriate measures to detect and mitigate them.

All pilots except for Pilot 5 fall into the first category. In Pilot 1, for example, "[a]ll patients involved in the pilot study received the standard care regardless [of] the AI outcome." According to their study protocol, the pilot study is designed so that diagnoses by the physicians are treated as the ground truth, and the output by their HosmartAI technology is evaluated against this ground truth. In Pilot 7, for example, their "AI-based tool was used in an offline evaluation," and even if it is introduced in clinical practice in the future, "it will serve as a support system for the physician," and the "final decision-making for the treatment will be done by the physician."

Pilot 5, which conducted an in-depth and well-developed examination of additional or heightened risks, provided comprehensive and sophisticated measures to detect and mitigate the risks they identified.

The measures to detect included initiatives at various levels and from different aspects: technical, operational, organisational, as well as educational and training. Among the many remarkable examples, three of them are commented here are: (1) Error Logging and Analysis;

(2) Continuous Performance Monitoring; and (3) Validation and Testing. First, by using Block Chain technology, they "implemented comprehensive logging of all interactions and decisions made by the AI systems." Second, they implemented the so-called human-in-the-loop approach, by placing a trained operator during "during all the sessions to carry out real-time monitoring and to assess the AI's performance continuously, ensuring that outputs remain within expected parameters." Third, they have conducted "[e]xtensive testing against control groups and varied scenarios at pilot site was carried out over period of 2 years to validate the AI's decision-making and interaction capabilities under diverse conditions."

Furthermore, Pilot 5 proposed specific measures to address potential risks when their HosmartAI technology is placed on the market and used in actual healthcare settings. Among the many remarkable examples, four of them are: (1) implementing "[a]dvanced monitoring and visualization tools to track AI performance in real-time, providing alerts for any deviations from expected behavior" (technical measure); (2) Establishing "clear protocols for escalating issues from AI decision-making to human medical professionals" (operational measure); (3) implementing [c]ontinuous learning systems where the AI can adapt and improve over time based on new data and feedback without compromising initial training stability" (operational measure); and (4) providing "[r]egular Training and Simulations: Regular training sessions and simulations for medical staff to stay familiar with AI tools and their integration into clinical workflows" (organizational measures).

Pilot 6, although indicated that there are not additional or heightened risks, also provided effective measures including at various levels: (1) technical: automating "the systems for data collection and quality verification"); (2) operational: personnel involved were "trained to use the system and identify faults in the implementation in the different care settings that can be considered"); and (3) organisational: implemented "traceability plan tailored to each service entity's characteristics, including how to handle tool failures and compliance with ethical and legal requirements".

2.6.2 AI Bias and Explainable AI

Section 2.5 focused on AI bias and Explainable AI. AI bias is referred to as "biased results due to human biases that skew the original training data or AI algorithm,"[11] and Explainable AI (XAI) is referred to as "a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms."[12] For further details, please see Chapter 4, as it devoted to cover these issues and explains the two concepts. In this subsection, we present our analysis in two parts: (1) for the duration of the HosmartAI project (during HosmartAI); and (2) for the period following the conclusion of the HosmartAI project (beyond HosmartAI).

[11] IBM, *What Is AI Bias?*, https://www.ibm.com/topics/ai-bias.

[12] IBM, *What is Explainable AI (XAI)?*, https://www.ibm.com/topics/explainable-ai.

2.6.2.1 During HosmartAI

We asked three sets of questions to examine whether and how the AI bias issue was appropriately addressed by pilot partners. In summary, we found that pilot partners paid sufficient attention to AI bias and appropriately addressed the issue. This includes working towards ensuring diversity and inclusivity in datasets, as well as taking various steps to make their AI systems explainable. Our reasoning is outlined below.

The first set of questions focused on the issue of **dataset quality**. As discussed in subsection 4.1.2, AI bias can infiltrate the system during data collection phase. Therefore, it is crucial to ensure that datasets are sufficiently representative of various groups. Additionally, appropriate measures must be taken to mitigate potential negative consequences, such as decreased accuracy, insufficient generalization, and potential biases, which are discussed next.

As a result of the analysis, we determined the following:

- Most partners responded that they were able to collect datasets that are sufficiently representative of various groups (e.g., gender/sex, race or ethnic origin, age, etc) during HosmartAI project (Pilot 1 ECHO & VCE, Pilot 4, Pilot 5, Pilot 7, and Pilot 8) or datasets are "unbalanced in terms of age" but it is justifiable because their HosmartAI technology is "aimed at older people who require care" (Pilot 6).
- One pilot partner affirmed that the issue regarding "diversity, non-discrimination and fairness" does not apply to their HosmartAI technology because "AI-based solution monitors the environment and does not collect and/or use any information about the people who use it" (Pilot 3).
- Additionally, one pilot partner took extra measure to address bias imbalance: Data augmentation to balance the dataset (Pilot 1 VCE)

The second set of questions focused on issues with regard to **detecting and correcting AI bias** as well as **improving algorithm generalization and preventing overfitting**. Based on the analysis, we arrived at the following two findings.

(1) Anticipating potential AI biases, many pilot partners took appropriate measures to detect and correct them.

- The distributions of the different characteristics, e.g., age, sex, etc., of the dataset were plotted to visualise if there is some evident bias (Pilot 1 ECHO).
- Equalised the histogram of every new image to be same as the images used during training and, and used a domain adaptation technique to validate how unbiased the model is (Pilot 1 ECHO).
- Trained the AI using weighted loss functions (Pilot 1 VCE)
- Conducted performance evaluations of speech recognition and chatbot systems across different demographic groups. Specifically, measured and compared error rates in speech recognition for in the wild for different, dialects, and accents (Pilot 5).

- Cross-validation during model training for emotion/distress recognition using multiple models was carried out to ensure the AI system generalizes well to unseen data (Pilot 5).
- Augmented the training datasets with a range of accents, dialects, and speech patterns. For chatbots, include diverse linguistic and cultural scenarios in the training data (Pilot 5).
- Adjusted the AI models to penalize bias (Pilot 5).
- Created and used Quality System for Data Annotation (Pilot 7).
- Trained the genomic D2Deep model on a balanced training set comprised the same amount of classes for all genes studied and so effectively mitigates biases related to hotspot mutations compared to state-of-the-art techniques (Pilot 8).

(2) Furthermore, the following measures were implemented or introduced to **improve algorithm generalization and prevent overfitting to training datasets**.

- Techniques, such as data augmentation before training, early stopping regularization, use of dropout layers, carefully split the dataset into training/validation/testing to avoid data leakage (Pilot 1).
- Cross-validation techniques to ensure the models generalize well beyond the training data, particularly for handling a wide variety of inputs in the wild (Pilot 5, Pilot 8).
- Regularly evaluated the models against the system's performance in real-world settings, especially focusing on how well it handles inputs from underrepresented groups (Pilot 5).
- Use of data annotation quality system (Pilot 7)
- For the genomic D2Deep model, dropout regularisation, Lasso Regression (L2 Regularization) and early stopping (Pilot 8).

The third set of questions focused on transparency and **Explainable AI (XAI).** As discussed in Section 4.2, XAI is not a complete solution to the problem of AI bias. However, developers can incorporate XAI to effectively address and mitigate the problem.

In summary, the responses from pilot partners fall into two categories: (1) their HosmartAI technology demonstrates explainability; or (2) their HosmartAI technology does not exhibit explainability, or explainable AI was deemed less relevant due to factors such as the design of their pilot studies.

Pilot 1 ECHO, Pilot 1 VCE, Pilot 3, Pilot 5, and Pilot 8 fall into the first category:

- Segmentation outcome is displayed in order the physician to self-assess if he/she can trust the outcome (Pilot 1 ECHO).
- The developed AI is based on RetinaNet neural network which both classifies an image in one category and returns a bounding box which indicates the region that activates the respective classification outcome. Thus, the gastroenterologist is capable of self-assessing if the outcome is valid or the AI makes a mistake (Pilot 1 VCE).
- Virtual sensors (VS) detect anomalies, which can be explained by analysing the logs of our notification service. If there is an anomaly related to lights being on, pilot partner

can verify whether the lights were actually on and cross-reference this with historical usage patterns (Pilot 3).

- AI models for DSS are based on decision trees and RF, which inherently offer more transparency about how decisions are made. For more complex models of distress classification that use AI, LIME and SHAP were employed to provide insights into the decision-making process (Pilot 5).
- One algorithm that provides automatic segmentation based on MRI is capable of providing probabilities maps for regions belonging to a certain tissue type, allowing for a more interpretable output compared to deterministic segmentations (Pilot 8).

Pilot 2, Pilot 4, Pilot 6, and Pilot 7 fall into the second category. For example, XAI was not the primary objective of Pilot 7 because the "focus was on setting up a proper data annotation environment and performing reader studies to evaluate the performance of a QCA algorithm by means of collecting feedback from expert physicians." For Pilot 2, the software created is "not AI-based but an optimization software."

Upon comprehensive review and holistic analysis of all the above elements, we found that pilot partners paid sufficient attention to AI bias and appropriately addressed the issue, including working to ensure diversity and inclusivity in datasets and taking various steps to make their AI systems explainable where necessary. However, we also note that it does not imply that it is already perfect and that there is no room for further improvement. Even if the *necessary conditions* are met, it does not mean that the *sufficient conditions* are fulfilled. Thus, we also asked the same three sets of questions in anticipation of a situation where HosmartAI technology is introduced to the European market and used in healthcare practice, following the conclusion of the HosmartAI project. This is discussed further below.

2.6.2.2 Beyond HosmartAI

Anticipating the scenario where HosmartAI technology is placed in the EU market and actively utilized in healthcare, multiple pilot partners suggested significant measures and raised critical viewpoints.

Specifically, pilot partners suggested the following measures to further ensure the quality, diversity, and inclusivity of datasets:

- Regardless the size of the dataset, a full documentation of the dataset's characteristics, i.e. demographics, health conditions, device types and data acquisition-related conditions, should be provided in order the AI developer/engineer to assess the value of the dataset and decide the proper AI methodology to develop a respective solution (Pilot 1 ECHO & VCE).
- Stakeholder Collaboration: Collaborate with a wide range of stakeholders, including hospitals, clinics, and patient advocacy groups, to gather comprehensive data and insights. This collaboration can help identify and address gaps in data collection (Pilot 5).
- Continuous Monitoring and Auditing: Regularly review and audit datasets for representativeness and bias. This should be an ongoing process as the model may drift over time due to changes in population demographics and disease patterns (Pilot 5).

- The researchers should verify that the data is complete, consistent, and accurate. techniques to validate the accuracy of the data, including cross-validation and external validation using independent datasets should be carried out before the final release of the models. If a bias is detected, techniques such as re-sampling, re-weighting, and algorithmic fairness interventions can be used to mitigate bias (Pilot 5).
- Data collection should be facilitated via a scalable solution that can be easily deployed at clinical sites, such that sufficient variability in data points can be achieved (Pilot 7).
- Our pilot study faces limitation due to the rarity of Glioma, which impacts the availability of data (Pilot 8).

While all of the above are insightful suggestions, we also note that the limitation indispensable to their objective pointed out by Pilot 8 is very thoughtful. Recognizing such unavoidable limitations and challenges is an essential step towards solving them.

Pilot partners also suggested the following measures to further detect and correct AI bias or improve generalization and prevent overfitting:

- Performance evaluations across different demographics to identify any discrepancies in AI behavio[u]r or outcomes. Both the validation phase and continuous monitoring after deployment (Pilot 5).
- Explore tools that can automatically flag potential biases by analy[s]ing the AI's decisions across various segments of data (Pilot 5).
- Bias detection as a core part of the development phase, using tools and methodologies that can identify bias in training data and model output (Pilot 5).
- Continuous monitoring and regular audits of AI systems are generally recommended, specifically because pilot 6 solution is developed for ensuring incremental integration of new modules that can include further AI aspects (Pilot 6).
- Initiatives for collecting and sharing large, representative datasets are of importance. Federated learning could aid in overcoming data privacy issues (Pilot 8)

Finally, pilot partners also suggested the following measures aimed at improving transparency and explainable AI:

- Causal Reasoning: optimality based on causal relationships. This requires advancements and modifications in AI architectures to include causal inference models (Pilot 5).
- Clinical Decision Support Teams: Teams that focus on integrating AI insights into broader clinical decision-making processes could benefit from detailed explanations to coordinate care more effectively (Pilot 5).
- Adverse Event Prediction: When predicting adverse events, the system should explain the factors leading to such predictions and the associated uncertainty to allow pre-emptive actions to be taken (Pilot 5).
- Setting up a co-creation environment that allows customers to evaluate algorithms at an early stage and interact with the AI-model using own data sets will increase the trustworthiness of the solution (Pilot 7).

The examples above demonstrate how the pilot partners are preparing to take proactive measures to address various AI ethics issues, including AI bias and Explainable AI, in anticipation of the phase when their HosmartAI technology will be introduced into the European market and utilized in healthcare.

At the same time, we humbly reiterate that meeting the necessary conditions does not imply that the sufficient conditions are fulfilled. Given the varying levels of progress among pilot partners, it is particularly important for each partner to learn from the insights and experiences of the more advanced partners. The importance of this becomes even greater when HosmartAI technology is introduced to the European market and integrated into healthcare practices.

In light of the above, we further created chapters devoted for The Artificial Intelligence Act and AI Biases, Explainable AI, and AI Risk Management. These chapters were prepared with the intention of providing a helpful resource for the pilot partners as they address AI bias and explainable AI.

3 The Artificial Intelligence Act

This chapter covers the EU's Artificial Intelligence Act (AI Act). The AI Act was also covered in D8.1, the first deliverable of WP8.[13] Back then, however, the status of the Act was a "Proposal." It was on the 21st of April 2021 when the Commission issued the initial Proposal of the AI Act.[14]

There have been numerous changes and developments since then. Members of the European Parliament (EP) approved the text of the Act on the 13th of March 2024. Shortly thereafter, on the 19th of April, the EP issued the CORRIGENDUM version, following final edits by lawyers and linguists.[15] On the 21st of May, the Council of the EU gave its final and formal approval to the AI Act.[16] The AI Act will enter into force 20 days after its publication in the Official Journal. The Commission is charged with the mission to issue guidelines on how the AI Act applies in practice.

We reiterate that it is too soon to make any assertions as to how the Act will apply to specific facts/technologies and work in practice. Nevertheless, this document provides a high-level overview of the AI Act based on the CORRIGENDUM version.

There are many reasons why we included a chapter on the AI Act, despite it not being mentioned in the pertinent part of D8.5 in the Grant Agreement. The primary reason is that the topics and issues discussed in this document as an addition -- i.e., the AI Act; AI bias and Explainable AI; and AI Risk Management System -- are all intertwined and relevant to each other. The ways in which they are intertwined and related to each other is discussed at the end of this chapter.

3.1 Categories of AI System

The AI Act takes a so-called "risk-based approach." Instead of regulating all technologies that might fall within the definition of "AI system," the AI Act categorizes AI systems according to its risk level. There are 4 levels of risks, and the AI Act prohibits or lays down different obligations for each.

3.1.1 Unacceptable Risk and Prohibited AI Practices

The first level, which is considered to pose the highest risk, is referred to as "Unacceptable risk." Article 5 of the AI Act enumerates "Prohibited AI Practices," and AI systems that are deemed to fall within these practices are prohibited.[17] Some prohibited AI practices include:

[13] See 5.4.3 of the D8.1 SELP Benchmark Report.

[14] 2021 Proposal is available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[15] "CORRIGENDUM to the position of the European Parliament adopted at first reading on 13 March 2024," available at https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138-FNL-COR01_EN.pdf.

[16] Press release available at https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/pdf/.

[17] Art. 5(1), Chapter II, AI Act.

1. **AI systems that use subliminal, manipulative, or deceptive techniques** to distort behavior and impair informed decision-making, causing significant harm;[18]
2. **AI systems that exploit vulnerabilities** related to age, disability, or socio-economic circumstances to distort behavior, causing significant harm.[19]
3. **Social scoring**. I.e., evaluating or classifying individuals or groups based on social behavior or personal traits, causing detrimental or unfavourable treatment of those people.[20]
4. **Predicting criminality**. I.e., assessing the risk of an individual committing criminal offenses **solely based on profiling or personality traits**, except when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity.[21]
5. Compiling facial recognition databases by untargeted scraping of facial images from the internet or CCTV footage.[22]
6. **Emotion recognition system**. **Inferring emotions in workplaces or educational institutions**, except for medical or safety reasons.[23]
7. **Biometric categorization systems** capable of inferring sensitive attributes (e.g., race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation), except labelling or filtering of lawfully acquired biometric datasets or when law enforcement categorizes biometric data.[24]

3.1.2 High-risk AI systems

The second level is referred to as "High-risk AI systems." A significant portion of the AI Act is devoted to high-risk AI systems.[25] For example, Articles 8 to 25 of the AI Act impose various obligations on providers of such AI systems,[26] *infra*.

Thus, whether or not a particular AI system falls within the category of "high-risk AI system" is a critical question. An AI system is considered to be a high-risk AI system if:

1. Both of the following conditions are met:[27]
    a. It is intended to be used as a safety component of a product, or the AI system is itself a product, covered by legislation listed in Annex I;[28] and

[18] Art. 5(1)(a), AI Act.
[19] Art. 5(1)(b), AI Act.
[20] Art. 5(1)(c), AI Act.
[21] Art. 5(1)(d), AI Act.
[22] Art. 5(1)(e), AI Act.
[23] Art 5(1)(f), AI Act.
[24] Art. 5(1)(g), AI Act.
[25] Entire Chapter III of the AI Act is one high-risk AI systems.
[26] Future of Life Institute, *High-level summary of the AI Act*, https://artificialintelligenceact.eu/high-level-summary/.
[27] Article 6(1), AI Act.
[28] Annex I: List of Union Harmonisation Legislation.

   b. the product whose safety component pursuant to point (a) is the AI system, or the AI system itself as a product, is required to undergo a third-party conformity assessment.[29]

2. The AI system is referred to in Annex III.[30]
   a. Unless it falls within the exceptions.

Article 6(3) provides two types of exceptions: (1) exception in general; or (2) specifically enumerated exceptions. Even if the AI system is referred to in Annex III, it is not a high-risk AI system, if:

1. It "does not pose a significant risk of harm to the health, safety or fundamental rights of natural persons, including by not materially influencing the outcome of decision making";[31] or
2. It meets any of the following conditions below:[32]
   a. the AI system is intended to perform a narrow procedural task;
   b. the AI system is intended to improve the result of a previously completed human activity;
   c. the AI system is intended to detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review; or
   d. the AI system is intended to perform a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III.

Note that if an AI system referred to in Annex III performs any **profiling of individuals**, these exceptions do NOT apply, and it is always considered to be a high-risk AI system.[33]

3.1.3 Limited-risk and Minimal-risk

If an AI system does not fall within the previous two categories, they will be deemed either in the category of Limited risk or Minimal risk. "Limited risk refers to the risks associated with lack of transparency in AI usage,"[34] and thus the AI Act imposes various transparency requirements[35] to ensure individuals are provided with necessary information.

Examples include,[36]

[29] With an intent. . . pursuant to the legislation listed in Annex I.

[30] ANNEX III High-risk AI systems referred to in Article 6(2).

[31] The first subparagraph of Article 6(3), AI Act.

[32] The second subparagraph of Article 6(3), AI Act.

[33] The third subparagraph of Article 6(3), AI Act.

[34] The European Commission, *AI Act | Shaping Europe's digital future*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[35] Article 50, AI Act. Please note that requirements under Article 50 are nevertheless applicable to high-risk AI systems regulated under Chapter III.

[36] The European Commission, *AI Act | Shaping Europe's digital future*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

- When using AI systems such as chatbots, humans should be made aware that they are interacting with a machine, so they can make an informed decision about whether or not to continue their interaction.
- Providers are required to ensure that humans are able to identify if the content is AI-generated.
- AI-generated text published with the purpose of informing the public on matters of public interest must be labelled as artificially generated. The same requirement applies to audio and video content constituting so-called "deepfakes."

The AI Act does not regulate minimal-risk AI systems. AI-enabled video games or spam filters are examples of such minimal-risk AI systems. "The vast majority of AI systems currently used in the EU fall into this category."[37]

3.2 Definitions

3.2.1 AI system

While definitions are important in any law, this is especially true in the AI Act. Definitions, especially what constitutes "AI systems," are critical because whether a particular technology amounts to an 'AI system' or whether it only amounts to a traditional software system,[38] for example, determines whether the AI Act applies in the first place.

Article 3(1) defines the 'AI system' as:

> a **machine-based system** that is **designed to operate with varying levels of autonomy** and that **may exhibit adaptiveness after deployment**, and that, for explicit or implicit objectives, **infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions** that **can influence physical or virtual environments**.

While it is certainly too early to conclude what this actually means, and how the definition applies to specific facts, it is noteworthy to mention that this definition comprises multiple elements.

1. **Autonomy**. The wording of "is designed to operate with varying levels of autonomy" suggests there should be some level of autonomy. I.e., a machine capable of completing tasks autonomously, at least to some extent.
2. **Adaptability**. The wording "may exhibit adaptiveness after deployment" indicates that, although not necessarily, it is capable of adapting itself to its context, environment, or situation in which it is deployed. This means that the AI system can perform differently depending on where it is deployed; systems deployed in setting A can perform differently from systems deployed in setting B.
3. **Inference**. While the objectives can be explicit or implicit, "infers how to generate outputs from the input it receives" signals that AI systems should be capable of making

[37] The European Commission, *AI Act | Shaping Europe's digital future*, https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai.

[38] See e.g., Recital 12.

inferences and generating various outputs. The definition enumerates non-exhaustive examples, such as predictions, content, recommendations, or decisions. This element is particularly important because it also triggers another important piece of digital legislation of the EU: the General Data Protection Regulation (GDPR).

4. **Influence**. The wording of "can influence physical or virtual environments" suggests that, although not necessarily, it is capable of exerting influence on its physical surroundings (e.g., in the case of robots) or on the virtual environment.

As the second and the fourth elements are not mandatory, it is reasonable to assume that the core of the definition is, "AI system that is capable of **making inferences** or **generating output**, with some degree of **autonomy**." This will be further discussed in the section "Relevance to HosmartAI", *infra*.

3.2.2 AI players

The AI Act also defines various stakeholders, or "players," of AI systems. Namely, 'provider,'[39] 'deployer,'[40] 'importer,'[41] 'distributor,'[42] 'operator,'[43] etc. Defining and distinguishing each "AI player" is critical for various reasons. Notably, one of the primary reasons is that the AI Act lays down different obligations for different "AI players."

This is important in the HosmartAI context because certain obligations imposed on providers (developers) and deployers are related to Explainable AI, *infra*. For example, the text of Article 4 (AI literacy) reads:

> **Providers and deployers** of AI systems **shall take measures to ensure**, to their best extent, **a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems** on their behalf, taking into account their technical knowledge, experience, education and training and the context the AI systems are to be used in, and considering the persons or groups of persons on whom the AI systems are to be used.

3.3 Obligations

Many, if not most, of the obligations under the AI Act are imposed on the providers (i.e., developers) of high-risk AI systems.[44] **Providers** of high-risk AI systems must, *inter alia*:[45]

1. Establish a **risk management system** throughout the high-risk AI system's lifecycle.[46]

[39] Article 3(3) AI Act.
[40] Article 3(4) AI Act.
[41] Article 3(6) AI Act.
[42] Article 3(7) AI Act.
[43] Article 3(8) AI Act.
[44] Future of Life Institute, *High-level summary of the AI Act*, https://artificialintelligenceact.eu/high-level-summary/.
[45] Id.
[46] Article 9 AI Act.

2. Conduct **data governance**, ensuring that training, validation and testing datasets are relevant, sufficiently representative and, to the best extent possible, free of errors and complete according to the intended purpose.[47]
3. Draw up and keep up-to-date **technical documentation** to demonstrate compliance, and to provide authorities with the necessary information to assess that compliance.[48]
4. Design their high-risk AI system for **record-keeping** to enable it to automatically record events relevant for identifying national level risks and substantial modifications throughout the system's lifecycle.[49]
5. Provide **instructions for use** to downstream **deployers** to enable the latter's compliance.[50]
6. Design their high-risk AI system to allow **deployers** to implement **human oversight**.[51]
7. Design their high-risk AI system to achieve appropriate levels of **accuracy, robustness, and cybersecurity**.[52]
8. Establish a **quality management system** to ensure compliance.[53]

3.4 Relevance to HosmartAI project

- It is critical for HosmartAI partners to avoid their AI systems fall within the category of "unacceptable risk" when they are developing or deploying an AI system different from HosmartAI after the project. Just to clarify, none of the HosmartAI technologies of pilot studies are even close to this category.
- Whether or not a particular AI system falls within the category of high-risk AI system depends on specific facts, including whether it poses a significant risk of harm to the health, safety, or fundamental rights of individuals, the intent of the developer/deployer, and the like. However, it is a good rule of thumb to start from the assumption that it is a high-risk AI system[54] because AI technology in healthcare is likely to trigger conditions under Article 6(1)(2), or fall under the exceptions under Article 6(3), and then carefully assess if indeed any of the exceptions do apply. In the event that a provider thinks their AI system is not high-risk AI, they need to "document its assessment before" placing it on the market or putting it into service.[55]
- In essence, one of the subject matters (i.e., technologies) that the AI Act aims to regulate is an "AI system that is capable of **making inferences and generating output**, with some degree of autonomy," as seen in the section discussing the definition, *supra*. This very likely implicates the profiling regulations under the GDPR[56] because the GDPR defines **profiling** as ". . . processing of personal data. . . to **evaluate** certain

[47] Article 10, AI Act.
[48] Article 11, AI Act.
[49] Article 12, AI Act.
[50] Article 13, AI Act.
[51] Article 14, AI Act.
[52] Article 15, AI Act.
[53] Article 17 AI Act.
[54] Or even to start from assuming it is an unacceptable risk AI system.
[55] Article 6(4), AI Act.
[56] Needless to say, there are cases where one is sufficed, while the other is not; thus, they are not identical.

personal aspects relating to a natural person, in particular to **analyse or predict** aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."[57] The two concepts overlap in terms of wording because the concept of **profiling** under the GDPR is equated or deemed to be very similar with the concept of **inference** which is found under, for example, privacy law in the United States.[58] HosmartAI project has submitted deliverables D10.1,[59] D10.2,[60] and D10.3.[61] Of these, D10.2 provided explanations regarding profiling, which are required by the GDPR. Similarly, when HosmartAI partners are developing or deploying an AI system different from HosmartAI after the project, they must consider doing the same as part of compliance with the profiling regulation under the GDPR.

- Finally, Article 9[62] and Article 17[63] of the AI Act require **risk management systems** and **quality management systems**, respectively, to be implemented. Management systems purported to address AI related or specific risks are not only necessary to comply with the obligations imposed by the AI Act, but also helpful to address issues concerning AI bias and to enhance AI explainability. As such, this report will cover AI Risk Management System in the following chapter, *infra*.

[57] Article 4(4), GDPR.

[58] "Inferences drawn" under the California Consumer Privacy Act.

[59] D10.1: H - Requirement No. 1.

[60] D10.2: POPD - Requirement No. 5.

[61] D10.3: GEN - Requirement No. 6.

[62] Article 9(1) of the AI Act reads, "A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems."

[63] Article 17(1) of the AI Act reads, "Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation…"

4 AI Biases, Explainable AI, and AI Risk Management

This chapter covers topics or issues related to AI biases, Explainable AI, and AI risk management system.

4.1 AI Biases

AI bias[64] refers to the "occurrence of biased results due to human biases that skew the original training data or AI algorithm—leading to distorted outputs and potentially harmful outcomes."[65] If AI bias is insufficiently or inadequately addressed, it can have a negative impact at various levels on various stakeholders. Not only does it negatively affect at a technical level because bias reduces the accuracy of an AI system, but it can also inflict negative impacts on the organizations involved, society at large or on individuals, especially those belonging to particular minority or marginalized groups (e.g., gender wise, racially, ethnically, disabilities, sexual orientation, etc.).

4.1.1 AI Biases in Healthcare

4.1.1.1 Introduction: Bias and AI Systems in Healthcare

Numerous studies, discussed below, demonstrate that AI systems used in healthcare can exhibit biases, and that it is critical to have diversity in datasets to avoid perpetuating biases and inequalities, according to, e.g., an article entitled Health Care AI Systems Are Biased[66] published in Scientific American.

It is no doubt that state-of-the-art AI systems are increasingly used in healthcare, and they are revolutionizing healthcare practices in many ways. For example, those AI systems are capable of conducting complex tasks like diagnosing skin cancer and detecting strokes with accuracy comparable to, or in some instances surpassing, specialists.

At the same time, however, there are significant concerns related to biases. AI systems can perpetuate or reinforce existing biases unless developers build them with inclusivity or diversity of data in mind.[67] For example, as discussed below, skin-cancer detection algorithms often perform worse on darker skin due to their training on datasets of predominantly light-skinned individuals, as discussed in the following sections.

Ensuring diversity and inclusivity in datasets can be challenging in healthcare because medical data are often siloed and difficult to share due to various reasons: e.g., data protection law, economic reasons, technical barriers, etc.

Furthermore, historically, it has been pointed out that underrepresentation (i.e., lack of diversity and inclusiveness) in clinical trials has long been an issue, which results in disparities in treatment effectiveness across different demographic groups. Additionally, similar

[64] Also referred to as machine learning bias or algorithm bias.
[65] IBM, *What Is AI Bias?*, https://www.ibm.com/topics/ai-bias.
[66] Amit Kaushal, Russ Altman, & Curt Langlotz, *Health Care AI Systems Are Biased*, Scientific American (2021), https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/.
[67] This may sound quite similar to the importance of diversity and inclusiveness for a society to be democratic, healthy, safe, secure, sustainable, etc.

concerns are raised regarding the use of AI systems in other domains, such as criminal sentencing and loan approvals.[68]

The following sections provide some actual examples of how AI biases can manifest in healthcare.

4.1.1.2 Machine Learning and Health Care Disparities in Dermatology

In an article entitled *Machine Learning and Health Care Disparities in Dermatology*,[69] the authors presented their concern that skin-cancer detection algorithms, many of which are trained primarily on light-skinned individuals, perform worse at detecting skin cancer affecting darker skin.[70]

Specifically, a major concern is that most AI systems they studied are trained predominantly on images of fair-skinned individuals, leading to potential biases against people with darker skin. This bias has the potential to lead to less precise diagnoses and less favourable health outcomes for these populations. The lack of representation of diverse skin types in ML training datasets is a critical issue that must be addressed to ensure that ML technology benefits all patients regardless of skin colour.

4.1.1.3 Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations

In fact, another study demonstrates that this is not just a "concern" and that AI systems can be racially biased. In an article entitled *Dissecting racial bias in an algorithm used to manage the health of populations*, published in Science, Obermeyer et al. presented their study confirming that "a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias."[71]

According to this study, the examined AI system erroneously provided the same level or risk score due to racial bias. In this context, the risk scores relevant to patients to indicate how much care they might need. Their study demonstrates that even if patients were assigned the same risk score, Black patients had more severe health problems compared to White patients. In other words, even if the AI system evaluated Black and White patients with the same risk score, the Black patients actually had more serious health issues that require attention. If this bias was corrected, the AI system would be able to identify more Black patients who need additional help. The increase in the percentage of Black patients receiving additional help would be from 17.7 to 46.5%.

According to the authors, this bias arises because the AI system predicts health care costs rather than illness. Due to societal problems, such as unequal access to healthcare, less

[68] Julia Angwin Mattu Jeff Larson,Lauren Kirchner,Surya, *Machine Bias*, ProPublica (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[69] Adewole S. Adamson & Avery Smith, *Machine Learning and Health Care Disparities in Dermatology*, 154 JAMA Dermatology 1247 (2018), https://doi.org/10.1001/jamadermatol.2018.2348.

[70] See also Amit Kaushal, Russ Altman, & Curt Langlotz, *Health Care AI Systems Are Biased*, Scientific American (2021), https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/.

[71] Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 Science 447 (2019), https://www.science.org/doi/10.1126/science.aax2342.

money is spent on caring for Black patients compared to White patients, even if their health status is the same or worse.

This is also a result of the proxy issue. Humans who developed the AI system considered healthcare costs as a proxy (or stand-in) for measuring health conditions. Although healthcare costs may have appeared to be an effective proxy for health by some measures of predictive accuracy, it can introduce large scale racial biases, as spending does not equally reflect health needs across different racial groups.

This research suggests that using an easy and seemingly effective proxy (i.e., healthcare costs in this instance) for the actual health condition can cause significant biases in many situations.

4.1.1.4 Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

AI systems can also exhibit gender bias. In an article entitled *Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis*, the authors of the study concluded that when the images for training datasets are insufficient for one gender, the AI system performs worse for that underrepresented gender. The authors provided evidence from a large study using three different deep learning models (types of AI) and two well-known X-ray image datasets, which these datasets are used to diagnose various lung and chest diseases.

The study focuses on the importance of having a balanced number of images from both males and females in the datasets used to train AI systems. This finding serves as a warning to organizations responsible for regulating and approving such AI systems that they should ensure that the dataset includes a good balance to avoid gender bias. The study also points out a challenge for researchers. Improved algorithms that can handle gender imbalances better and still perform well are needed.

4.1.1.5 Neglecting sex and gender in research is a public-health risk

An article recently published in Nature on 15th May 2024, entitled *Neglecting sex and gender in research is a public-health risk*,[72] argues that overlooking sex and gender differences in scientific research also poses significant public-health risks. Specifically, it contends that ignoring sex and gender in research can lead to misdiagnoses, inappropriate treatments, and poor health outcomes. The article introduces several examples or cases supporting the argument that (1) clinical trials often underrepresent women, resulting in less effective treatments for them; and (2) women are more likely to be misdiagnosed with heart disease because many studies focus primarily on male symptoms.

The article calls for inclusive research. Specifically, it calls on researchers and funding agencies to consider sex and gender as critical variables in studies, and it contends that incorporating these factors can improve the accuracy and applicability of research findings.

[72] Sue Haupt, Cheryl Carcel & Robyn Norton, *Neglecting Sex and Gender in Research Is a Public-Health Risk*, 629 Nature 527 (2024), https://www.nature.com/articles/d41586-024-01372-2.

Furthermore, the article advocates for policies that require inclusion of sex and gender analysis in research and also suggests that training programs for researchers should emphasize the importance of these variables.

4.1.2 Sources or Types of Bias

Biases can seep into AI systems in different ways at different stages. This section provides common sources of AI bias.[73]

1. **Algorithm bias**: Bias can creep into an AI system when the problem or question posed is not sufficiently accurate or specific, or if the feedback to the machine learning algorithm is ineffective in guiding the search for a solution.
2. **Cognitive bias**: Because AI technologies always involve "humans in the loop,"[74] and because humans are fallible, cognitive bias can sneak in without practitioners even knowing it. This can affect how the data or model works.
3. **Confirmation bias**: An output of an AI system can be biased when there is overreliance on pre-existing beliefs or trends in the data. As a result, it would worsen existing biases and would render identifying new patterns or trends more difficult.
4. **Exclusion bias**: Exclusion bias occurs when important data is excluded from the dataset being used. This can happen, for example, when the developer has overlooked new and important factors.
5. **Measurement bias**: Measurement bias is caused by incomplete or insufficient data. This typically stems from an oversight or insufficient preparation, resulting in the dataset not representing the entire population that should be included. For example, an AI system purported to predict which students are likely to successfully graduate would render an inaccurate prediction when a dataset of students graduated successfully is included and a dataset of other students is excluded.
6. **Out-group homogeneity bias**: This bias may be characterized as "the wisdom of knowing that one does not know." Humans tend to have a better understanding of their ingroup members (the group one belongs to), and to think they are more diverse than outgroup members. As a result, there would be a risk of building AI systems that are less capable of distinguishing between individuals who are not part of the majority group in the training data. This can eventually cause racial bias, misclassification and incorrect output.
7. **Prejudice bias**: Prejudice bias occurs when stereotypes and faulty societal assumptions infiltrate the dataset of the AI system. Suppose an AI system provides an output showing doctors as males, and all nurses as females. This can happen if the developer does not question the misconception that doctors are predominantly males and nurses are predominantly female, even if the datasets reflected an actual situation where all doctors are males, and all nurses are females.

[73] This section provides common sources of AI bias, primarily, based on expertise by IBM. IBM, *What Is AI Bias?*, https://www.ibm.com/topics/ai-bias.

[74] E.g., Crootof et al., Humans in the Loop. 76 Vanderbilt Law Review 429 (2023), U of Colorado Law Legal Studies Research Paper No. 22-10, U of Michigan Public Law Research Paper No. 22-011.

8. **Recall bias**: This occurs during the process of data labelling, wherein labels are inconsistently applied by subjective observations. In other words, recall bias occurs when participants in a study fail to recall previous events or experiences accurately or omit details. This type of bias can often occur in retrospective studies, in which individuals are asked to recall past behaviours, exposures, or experiences. Recall bias can lead to systematic differences among the groups being compared, thus distorting the results and conclusions of the research. For instance, in a study examining the correlation between diet and disease, individuals with a disease may recall their dietary habits differently from those who do not have the disease, resulting in inaccurate or biased reporting.
9. **Sample/Selection bias**: This type of bias occurs when the data used to train the model is not sufficiently large, not sufficiently representative, or is too incomplete to adequately train the system. Suppose a researcher wants to study the average level of physical fitness among adults in a large city and chooses to collect data by surveying people at a local gym. However, the people at the gym could be more physically active and health-conscious compared to the general population, and thus there is a sample bias issue because the sample is not representative of the entire city's adult population, many of whom may not exercise regularly or at all.
10. **Stereotyping bias**: This bias occurs when an AI system reinforces harmful stereotypes, usually accidentally. Consider an AI language translation system that consistently translates phrases involving certain professions with gender biases. For example, if the AI system translates the English phrase “The doctor said. . .” to French, for example, it might use the male form “Le docteur a dit. . .” by default, regardless of the actual gender of the doctor. Likewise, translating “The nurse said. . .” might default to the female form “L'infirmière a dit. . .” This reflects and reinforces the stereotype that doctors are typically male, and nurses are typically female, even though all genders practice both professions. See also McKinsey’s article on this.[75]

Please note that each kind of bias is not mutually exclusive. One bias can also be another bias as well (e.g., cognitive bias also being prejudice bias, etc.). Although this is not an exhaustive list, this list can serve as a watchlist to prevent AI bias from infiltrating AI systems.

4.1.3 How to avoid AI bias: a checklist

While there is no one-size-fit-all solution or “silver bullet” to address the AI bias problem, this section introduces a checklist consisting of six process steps developed by IBM[76] that can help reduce AI bias.

1. Is the machine learning model correct for the intended purpose?

✓ Selecting the correct machine learning model is the first step to reduce AI bias.

[75] Jake Silberg & James Manyika, *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*, https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.

[76] IBM, *How to avoid bias - What Is AI Bias?*, https://www.ibm.com/topics/ai-bias.

- ✓ For supervised models, stakeholders select the training data, and therefore it is also important that the stakeholder team be diverse, and that they have had training to help prevent unconscious bias.
- ✓ For unsupervised models, bias prevention tools need to be built into the neural network so that it learns to recognize what is biased.

2. Is the training dataset correct and adequate for the intended purpose?

- ✓ Machine learning trained on the wrong data will produce wrong results. Whatever data is fed into the AI should be complete and balanced to replicate the actual demographics of the group being considered.

3. Is the team developing AI system balanced and diverse?

- ✓ The more diverse the individuals developing the AI system are -- i.e., racially, economically, by educational level, by gender, and by job description -- the more likely it is that AI bias would be recognized.
- ✓ Similarly, the talents and viewpoints of the team should include various individuals with different roles: e.g., AI business innovators, AI creators, AI implementers, and a representation of the beneficiaries of this particular AI effort.

4. Perform data processing mindfully

- ✓ AI bias can seep into AI systems not only at the data selection phase, but also at the data processing phase.
- ✓ Thus, being mindful and careful at each step of the data processing phase is also essential: during pre-processing, in-processing, or post-processing.

5. Continuous monitoring

- ✓ Ongoing monitoring and testing with real-world data can help detect and correct bias being baked into the AI system.
- ✓ AI developers should consider assessments either by an independent team within the organization or a trusted third-party.

4.2 Explainable AI

Explainable AI (XAI)[77] is referred to as “a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms.”[78] For more details about the concept and its history, please refer to a concise article entitled *Explainable AI: A Brief History of the Concept* published in ERCIM News.[79]

[77] Interpretable AI or Explainable Machine Learning (XML) are also similar concepts.

[78] IBM, *What is Explainable AI (XAI)?*, https://www.ibm.com/topics/explainable-ai.

[79] Mihály Héder (SZTAKI), *Explainable AI: A Brief History of the Concept*, ERCIM News, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.

XAI is a concept and an approach with the primary objective of making AI systems more transparent, interpretable, and understandable to humans by providing information or insights into how AI systems generate output. While XAI is not a complete solution[80] to the problem of AI bias, developers of AI systems can incorporate XAI to address the problem.[81]

XAI can be helpful to address issues related to AI bias in several ways. First, XAI helps developers and users examine the factors influencing an AI system's output. By analysing the importance of different features and the relationships between inputs and outputs, developers would be able to detect biases that may have infiltrated the AI system due to biased training data or algorithmic design.[82]

Second, XAI can help ensure fairness. XAI can help organizations to assess whether their AI systems are generating fair and unbiased outputs. By understanding how the model arrives at its conclusions, it becomes easier to identify instances where the model may be perpetuating or amplifying societal biases. This understanding can help guide efforts to mitigate bias and ensure more equitable outcomes.[83]

Third, XAI contributes to greater accountability in output generated by AI systems. When the reasoning behind AI predictions is transparent and explainable, it would help developers/deployers/users of AI systems to trace the source of any biases and hold the relevant parties accountable. This accountability can incentivize more responsible AI development practices.[84]

Fourth and finally, XAI can facilitate human oversight. XAI allows human experts to review and validate AI systems' output generation processes. By providing a clear understanding of how the model works, XAI enables domain experts to spot potential biases and intervene when necessary. This human oversight is crucial for ensuring that AI systems operate in an unbiased and trustworthy manner.[85]

While XAI is not a complete solution to the problem of AI bias, it is an important tool in the ongoing effort to develop more transparent, accountable, and fair AI systems. By promoting understanding and enabling human oversight, XAI can help mitigate the risks of biased AI and foster greater trust in these technologies.

[80] Not a sufficient condition

[81] But a necessary condition

[82] IBM, *What is Explainable AI (XAI)?*, https://www.ibm.com/topics/explainable-ai.

[83] Simon Chandler, *How Explainable AI Is Helping Algorithms Avoid Bias*, Forbes, https://www.forbes.com/sites/simonchandler/2020/02/18/how-explainable-ai-is-helping-algorithms-avoid-bias/.

[84] See e.g., Alaa Marshan, *Artificial Intelligence: Explainability, Ethical Issues and Bias*, Annals of Robotics and Automation 034 (2021), https://www.researchgate.net/publication/353952148_Artificial_intelligence_Explainability_ethical_issues_and_bias.

[85] *An Introduction to the Four Principles of Explainable AI*, https://www.linkedin.com/pulse/iintroduction-four-principles-explainable-ai-algolia.

4.3 Risk Management for AI Systems

As discussed in the previous chapter, the AI Act requires implementation of management systems to ensure the safe and ethical deployment of AI technologies. Specifically:

Article 9 requires a **risk management system** to be established, implemented, documented, and maintained.[86] The management system is expected to run through the entire lifecycle of a high-risk AI system. The Act requires systematic review and update on a regular basis.

Article 17 requires providers of high-risk AI systems to have a **quality management system** in place.[87] The Act requires, *inter alia*, the following aspects included in the quality management system:

(a) a strategy for regulatory compliance (including compliance with conformity assessment procedures and procedures for the management of modifications to the high-risk AI system);
(b) techniques, procedures and systematic actions to be used for the design, design control and design verification of the high-risk AI system;
(c) techniques, procedures and systematic actions to be used for the development, quality control and quality assurance of the high-risk AI system;
(d) examination, test and validation procedures to be carried out before, during and after the development of the high-risk AI system, and the frequency with which they have to be carried out.

The Commission has issued the Ethics Guidelines for Trustworthy AI, which includes risk management principles to ensure ethical AI development and deployment. As the Ethics Guidelines were published on April 8, 2019, and are discussed already in the previous deliverable D8.1 - SELP Benchmark Report,[88] the following sections provide other similar AI risk management frameworks that can be helpful for HosmartAI partners to comply with the obligations under the AI Act.

4.3.1 Artificial Intelligence Risk Management Framework (AI RMF)

Artificial Intelligence Risk Management Framework (AI Risk Management Framework or AI RMF) is a risk management framework to better manage risks to individuals, organizations, and society associated with AI systems.[89]

Issued on January 26, 2023, the AI Risk Management Framework was developed in the United States by NIST, the National Institute of Standards and Technology, in collaboration with the private and public sectors, and it is intended for “voluntary use and to improve the ability to

[86] Article 9(1) of the AI Act reads, “A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems.”
[87] Article 17(1) of the AI Act reads, “Providers of high-risk AI systems shall put a quality management system in place that ensures compliance with this Regulation. . . .”
[88] See 5.4.1 of the deliverable entitled D8.1 - SELP Benchmark Report, delivered in May 2021.
[89] NIST, *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework.

incorporate trustworthiness considerations into the design, development, use, and evaluation of AI products, services, and systems."[90]

The AI Risk Management Framework is composed of two parts.[91] In Part 1, it discusses how organizations can frame the risks related to AI systems and describes the intended audience. Following, it analyses AI risks and trustworthiness, outlining the characteristics of trustworthy AI systems. These include being valid and reliable, safe, secure and resilient, accountable and transparent, explainable and interpretable, privacy-enhanced, and fair with managing harmful biases.

Part 2 forms the "Core" of the Framework, detailing four functions to help organizations manage AI system risks. It further breaks these functions (i.e., GOVERN, MAP, MEASURE, and MANAGE) into categories and subcategories. "GOVERN" applies to all stages of AI risk management processes and procedures. In contrast, "MAP," "MEASURE," and "MANAGE" are applied in specific AI system contexts and stages of the AI lifecycle.

While there are many standards and best practices helpful for organizations to mitigate the risks of traditional software, Appendix B articulates the unique risks raised by AI systems that can be helpful for organizations to identify their risks in relation to their AI system.

The AI Risk Management Framework is accompanied by other various useful resources further helpful for organizations to manage risks unique to AI systems: e.g., NIST AI RMF Playbook; AI RMF Roadmap; AI RMF Crosswalk; Perspectives; and video explainer. For these documents, please refer to the webpage.[92]

4.3.2 ISO/IEC 23894:2023 - AI - Guidance on risk management

Another standardized framework to manage risks associated with AI systems is ISO/IEC 23894.[93] ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management (ISO/IEC 23894),[94] issued by the ISO[95] and the IEC,[96] is an international standard that provides guidelines for AI system risk management. It offers guidance for organizations that develop, produce, deploy, or use products, systems, and services utilizing AI systems to manage AI-specific risks. It aims to help organizations integrate risk management into their AI-related activities and functions, and outlines processes for effectively implementing and integrating AI risk management.

ISO/IEC 23894 offers a systematic approach to AI system risk management, covering various aspects such as: (1) Principles and framework for AI risk management; (2) Risk assessment

[90] Id.

[91] NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, https://doi.org/10.6028/NIST.AI.100-1.

[92] NIST, *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework.

[93] Additionally, there is ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system, https://www.iso.org/standard/81230.html.

[94] ISO/IEC 23894:2023 - Information technology - Artificial intelligence - Guidance on risk management, https://www.iso.org/standard/77304.html.

[95] International Organization for Standardization.

[96] International Electrotechnical Commission.

methodology; (3) Risk treatment and mitigation strategies; (4) Monitoring and review processes; (5) Communication and consultation with stakeholders.

Because ISO/IEC 23894 is designed to be applicable to a wide range of organizations, regardless of their size, sector, or the nature of their AI systems, it can be a great starting point for HosmartAI partners. By following the guidelines in ISO/IEC 23894, HosmartAI partners can show their commitment to responsible AI development and deployment, build trust with stakeholders, and ensure compliance with relevant regulations and ethical principles.

ISO/IEC 23894 refers to ISO 31000:2018[97] in terms of principles, frameworks, and processes, and relies on these principles to provide an international perspective on how to manage risks and on associated best practices in the context of AI.

Additionally, ISO/IEC 23894 references ISO/IEC 22989[98] for AI related concepts and terminology. While the AI Act requires relevant organizations to draw up technical documentation or provide instructions for use to downstream deployers, this can also be a helpful resource in providing concise documentation with consistent terminologies.

[97] ISO 31000:2018 - Risk management - Guidelines, https://www.iso.org/standard/65694.html.

[98] ISO/IEC 22989:2022 - Information technology - Artificial intelligence - Artificial intelligence concepts and terminology, https://www.iso.org/standard/74296.html.

5 Closing Remarks: Towards Ethical AI

This Report documented the findings and results of the second half of Task 8.4 SELP Continuous Compliance Report (T8.4). By formulating questions to gather relevant information and insights from 8 Lighthouse Pilots, and through comprehensive review and holistic analysis, we concluded that there were no issues requiring further attention or discussion regarding medical and research ethics, as well as data protection/privacy and data security. Moreover, as we articulated in Section 2.6 Findings and Analyses, we found that pilot partners paid sufficient attention to AI bias and appropriately addressed the issue. This includes working towards ensuring diversity and inclusivity in datasets, as well as taking various steps to make their AI systems explainable. We noted, however, that fulfilling the necessary conditions does not guarantee that the sufficient conditions are met.

In light of the above, we humbly wish to underscore one key concept that is fundamentally common to all of the above chapters as we close this report: ***Continuous iterative process***

In fact, the AI Act puts a lot of emphasis on processes to be *continuous and iterative*. For example, Article 9(2)[99] and Recital 65[100] state that risk management systems should consist of '*continuous iterative process* planned and run throughout the entire lifecycle of a high-risk AI system. . .' This is in part because the unique benefits, as well as the risks, of AI systems stem from their nature of continuous learning,[101] even after being placed on the market or put into service, and their likely different behaviour as time progresses. Various articles[102] and recitals[103] aptly observe this characteristic of AI systems. For example, Article 15(4) -- an article on accuracy, robustness, and cybersecurity -- states: 'High-risk AI systems that *continue to learn after being placed on the market or put into service* shall be developed in such a way

[99] Article 9(2), AI Act. The pertinent part reads, "The risk management system shall be understood as a *continuous iterative process* planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. . ." (emphasis added).

[100] Recital 65, AI Act. The pertinent part reads, "The risk-management system should consist of a *continuous, iterative process* that is planned and run throughout the entire lifecycle of a high-risk AI system. That process should be aimed at identifying and mitigating the relevant risks of AI systems on health, safety and fundamental rights. The risk-management system should be regularly reviewed and updated to ensure its *continuing effectiveness*, as well as justification and documentation of any significant decisions and actions taken subject to this Regulation. . ." (emphasis added).

[101] In terms of HosmartAI, Pilot 5 was also explicit as to this point, and explained that their speech recognition system is "implemented as a *continually learning system* also exploiting the [r]eal-world interactions" (emphasis added).

[102] E.g., Article 15(4), AI Act. The pertinent part reads, "Art 15(4) *High-risk AI systems that continue to learn after being placed on the market or put into service* shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures" (emphasis added).

[103] See e.g., Recital 128 (". . . changes occurring to the algorithm and the performance of AI systems which *continue to 'learn' after being placed on the market or put into service*, namely automatically adapting how functions are carried out. . .") and Recital 155 (". . . This system is also key to ensure that the possible risks emerging from AI systems which *continue to 'learn' after being placed on the market or put into service* can be more efficiently and timely addressed. . .") (emphases added).

as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations. . .'

The measures and initiatives by pilot partners also stress this key aspect. For example, the measures taken by Pilot 5[104] during the HosmartAI project emphasises that they were *continuous processes*: (1) They established a *continuous feedback loop* between system developers (UM) and healthcare providers (UKCM) to refine the tools with the aim of co-designing the robotic nurse with clinicians and nurses, ensuring that the technology augments rather than disrupts clinical workflows. (2) To detect potentially increased risk due to the use of HosmartAI technology in the research study, they implemented *continuous performance monitoring*, with a trained operator present during all sessions to carry out real-time monitoring and continuously assess the AI's performance, ensuring that outputs remained within expected parameters. (3) To tackle the issues concerning AI bias and overfitting to training datasets/algorithm generalization, they, *inter alia*, continuously update and retrain the models using newly collected data that reflect ongoing changes in language and communication styles.

Moreover, Pilot 5 raised insightful measures to be taken when HosmartAI technology is placed in the EU market and actively utilized in healthcare, following the conclusion of the HosmartAI project. To detect and mitigate potentially increased risk, they suggested: (1) A comprehensive risk management frameworks that include risk assessment, mitigation, and *continuous monitoring* specific to AI technologies needs to be implemented; (2) *Continuous Learning and Adaptation*, i.e., *continuous learning systems* where the AI can adapt and improve over time based on new data and feedback without compromising initial training stability should be implemented. To ensure quality datasets, they also suggested that *continuous monitoring and auditing* is a key, emphasizing to "[r]egularly review and audit datasets for representativeness and bias," and that "[t]his should be an ongoing process as the model may drift over time due to changes in population demographics and disease patterns." To address potential AI bias issues as well as to improve algorithm generalization and prevent overfitting to training datasets, they reaffirm the importance of continuous monitoring at different phases (e.g., validation phase, post-deployment).[105]

Therefore, we invite and recommend all pilot partners to *continue* their dedicated efforts in addressing AI bias issues, improving AI performance, and enhancing AI transparency. By maintaining rigorous standards of continuous and proactive initiatives throughout the lifecycle of their AI systems, we believe that the benefits of HosmartAI technology will be maximized and appreciated by the market in the future, while minimizing the associated risks. We also believe that partners will greatly benefit from actively learning from each other's

[104] Other pilots also mentioned continuous measures, but we highlighted Pilot 5 here because they *continuously* emphasized the importance of continuous measures and activities.

[105] E.g., "Performance evaluations across different demographics to identify any discrepancies in AI behavior or outcomes. Both the validation phase and *continuous monitoring after deployment*"; "Post-Deployment: *Continuous Human monitoring* to detect biases as they emerge in real-world settings, and regular review of AI performance"; and "Systems to *continuously monitor* the model's performance in real-world applications to quickly identify and address overfitting" (emphases added).

experiences and expertise, fostering a synergy effect within the HosmartAI project. Finally, we sincerely hope that we were also an integral part of this collaboration and that this report contributes to these efforts by providing valuable insights and resources.

6 References

[REF-01]	Future of Life Institute, *High-level summary of the AI Act*, https://artificialintelligenceact.eu/high-level-summary/.
[REF-02]	World Economic Forum, *The EU's Artificial Intelligence Act, explained*, https://www.weforum.org/agenda/2023/06/european-union-ai-act-explained/.
[REF-03]	World Economic Forum, *Research shows AI is often biased. Here's how to make algorithms work for all of us*, https://www.weforum.org/agenda/2021/07/ai-machine-learning-bias-discrimination/.
[REF-04]	Amit Kaushal, Russ Altman, & Curt Langlotz, *Health Care AI Systems Are Biased*, Scientific American (2021), https://www.scientificamerican.com/article/health-care-ai-systems-are-biased/.
[REF-05]	Julia Angwin Mattu Jeff Larson,Lauren Kirchner,Surya, *Machine Bias*, ProPublica (2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[REF-06]	Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 Science 447 (2019), https://www.science.org/doi/10.1126/science.aax2342.
[REF-07]	Agostina Larrazabal et al., *Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis*, https://www.pnas.org/doi/10.1073/pnas.1919012117.
[REF-08]	Adewole S. Adamson & Avery Smith, *Machine Learning and Health Care Disparities in Dermatology*, 154 JAMA Dermatology 1247 (2018), https://doi.org/10.1001/jamadermatol.2018.2348.
[REF-09]	Sue Haupt, Cheryl Carcel & Robyn Norton, *Neglecting Sex and Gender in Research Is a Public-Health Risk*, 629 Nature 527 (2024), https://www.nature.com/articles/d41586-024-01372-2.
[REF-10]	Jake Silberg & James Manyika, *Notes from the AI Frontier: Tackling Bias in AI (and in Humans)*, https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans.
[REF-11]	Mihály Héder (SZTAKI), *Explainable AI: A Brief History of the Concept*, ERCIM NEWS 134 July 2023, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.
[REF-12]	Antonio Bruno, Giacomo Ignesti and Massimo Martinelli, *Explaining Ensemble Models for Lung ultrasound Classification*, ERCIM NEWS 134 July 2023, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.
[REF-13]	Luigi Briguglio, Francesca Morpurgo, & Carmela Occhipinti, *A Governance and Assessment Model for Ethical Artificial Intelligence in Healthcare*, ERCIM NEWS 134 July 2023, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.
[REF-14]	Nikolaos Rodis, Christos Sardianos, & Georgios Th. Papadopoulos, *Current Challenges and Future Research Directions in Multimodal Explainable Artificial*

	Intelligence, ERCIM NEWS 134 July 2023, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.
[REF-15]	Alexandre Lädermann, Philippe Collin, & Patrick J. Denard, *Predictive Model for Functional Outcome after Orthopaedic Surgery using Machine Learning Methods*, ERCIM NEWS 134 July 2023, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.
[REF-16]	Michaela Areti Zervou, Effrosyni Doutsi, & Panagiotis Tsakalides, *Unleashing the Power of Artificial Intelligence for Personalised Drug Design*, ERCIM NEWS 134 July 2023, https://ercim-news.ercim.eu/images/stories/EN134/EN134-web.pdf.
[REF-17]	P Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence*, NIST IR 8312 (2021), https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf.
[REF-18]	NIST, *AI Risk Management Framework*, https://www.nist.gov/itl/ai-risk-management-framework.
[REF-19]	NIST, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, https://doi.org/10.6028/NIST.AI.100-1.