Call: HORIZON-HLTH-2021-DISEASE-04
Topic: HORIZON-HLTH-2021-DISEASE-04-04 - Clinical validation of artificial intelligence (AI) solutions for treatment and care
Funding Scheme: HORIZON-RIA

MELISSA

Deliverable No. D3.3

Ethical framework for a trustworthy implementation of ABBA

Grant Agreement no.: 101057730

Project Title: Mobile Artificial Intelligence Solution for Diabetes Adaptive Care

Contractual Submission Date: 28.02.2023

Actual Submission Date: 29.03.2023

Resubmission following IEA recommendations: 02.07.2024

Responsible partner: P9-TUM

Funded by the European Union

Grant agreement no.	101057730
Project full title	MELISSA- Mobile Artificial Intelligence Solution for Diabetes Adaptive Care

Deliverable number	**D3.3**
Deliverable title	**Ethical framework for a trustworthy implementation of ABBA**
Type[1]	R
Dissemination level[2]	PU
Work package number	WP3
Author(s)	Alexander Kriebitz, Lameck Amugongo, Auxane Boch and Raphael Max (P9-TUM)
MELISSA reviewers	Bastiaan de Galan (P1-UM), Stavroula Mougiakakou (P12-UBERN), Stephan Proennecke (P11-DEBIOTECH)
Keywords	Ethics; Trustworthy AI, Ethical Implementation of AI

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA).

Neither the European Union nor the granting authority can be held responsible for them.

[1] **Type**: Use one of the following codes (in consistence with the Description of the Action):
R: Document, report (excluding the periodic and final reports)
DEM: Demonstrator, pilot, prototype, plan designs
DEC: Websites, patents filing, press & media actions, videos, etc.

[2] **Dissemination level**: Use one of the following codes (in consistence with the Description of the Action)
PU: Public, fully open, e.g., web
SEN: Sensitive, limited under conditions of the Grant Agreement

Table of Contents

Summary

Deliverable D3.3 refers to the design of an "ethical framework for a trustworthy implementation of ABBA". Over the course of the project, we decided to widen the focus of the ethical framework developed for D3.3 including all Artificial Intelligence (AI) solutions developed over the course of the MELISSA project to prevent violations of ethical principles in aspects relevant to the project. The key task has been to develop recommendations for the trustworthy implementation of AI-driven diabetes management apps, while addressing ethical obstacles and proposing insights to mitigate them. The approach underlying this ethical framework is based on existing discourses on bio-ethics and AI ethics as well as corresponding legislation.

The deliverable has been resubmitted following the observations and recommendations made by the independent ethics advisor on February 27, 2024, and in response to the changes in the legal landscape following the promulgation of the EU AI Act in May 2024[3].

1 Introduction

A substantial number of individuals worldwide are currently affected by diabetes. In 2021, the International Diabetes Federation estimated that 537 million adults, aged 20 to 79 years old, suffered from diabetes, which represent about 1 in 10 persons.[4] Furthermore, data suggest that diabetes is on the rise, primarily due to demographic change and aging societies.[5,6] Consequently, enhancing technologies that improve the situation of people with diabetes (PwD), but also that relieve stress on the medical system, would be a timely contribution to healthcare as outlined by a joint OECD/EU study[7]. Furthermore, addressing diabetes matters not only from the perspective of common international objectives such as the United Nations Sustainable Development Goals (UN SDGs), but also for improving the quality of life of PwD, their families and communities.

The MELISSA project aims to meet this challenge by providing PwD with AI-based solutions, which are meant to "become a game changer in the self-management of diabetes"[8]. While the MELISSA project has significant potential to improve the quality of life of PwD, the use of AI in health and its specific use for diabetes self-management sparks ethical questions. The relevance of ethical considerations is reinforced by the aim articulated in the MELISSA mission statement to provide insulin-treated PwD with a trustworthy AI solution. Trustworthiness implies that AI solutions developed by MELISSA satisfy societal expectations related to AI solutions in the health sector such as depicted in the EU High Level Experts' Group on trustworthy AI[9], or the AI4People's work on good AI society[10]. Relevant ethical questions that traditionally relate to trustworthy AI solutions are as follows:

[3] European Parliament. (2024). Artificial Intelligence Act. In TEXTS ADOPTED. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf

[4] Factsheets | IDF Diabetes Atlas. (n.d.). https://diabetesatlas.org/regional-factsheets/?dlmodal=active

[5] *Facts & figures*. (n.d.). https://idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html

[6] Choudhury, H., Pandey, M., Hua, C. K., Mun, C. S., Jing, J. K., Kong, L., ... & Kesharwani, P. (2018). An update on natural compounds in the remedy of diabetes mellitus: A systematic review. *Journal of traditional and complementary medicine*, *8*(3), 361-376.

[7] OCDE/Union européenne (2020), *Health at a Glance: Europe 2020 : State of Health in the EU Cycle*, Éditions OCDE, Paris, https://doi.org/10.1787/82129230-en.

[8] *Better Quality of Life for People Living with Diabetes Through Innovative Artificial Intelligence Applications: Launch of EU Research Project MELISSA | EURICE GmbH.* (n.d.). https://eurice.eu/news/better-quality-of-life-for-people-living-with-diabetes-through-innovative-artificial-intelligence-applications-launch-of-eu-research-project-melissa

[9] AIHLEG (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, 6.

[10] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707.

- What is the ethical benefit of using AI for the self-management of PwD?
- What negative side effects could emerge from the over- or misuse of AI in diabetes management?
- How does the use of AI solutions ramify decisions made by PwD?
- Are there imbalances in the performance of the AI solutions, especially when it comes to age, gender, biological sex, or ethnicity?
- Are PwD and medical personnel able to understand AI solutions decision making process / predictions / recommendations?

While these questions are representative of a larger set of underlying legal and ethical implications, particularly reinforced by most recent legislation on the level of the European Union, they highlight the complexity of the normative challenge when developing, deploying and finally using AI solutions in the health context[11]. Moreover, previous cases of over- and misuse of AI have shown violations of the rights of patients or situations aggravating patterns of existing discrimination[12]. Ethics seeks to address such challenges by establishing action-guiding principles for the entire AI lifecycle. As mentioned in the Ethics Guidelines for Trustworthy AI, "trustworthiness is a prerequisite for people and societies to develop, deploy, and use AI systems"[13]. In this sense, ethics constitutes a set of principles that serve as an enabler of trust in AI solutions but also as a means to mitigate both technical error and human misconduct. On this basis, the normative framework formulated in the following seeks to address both purposes of AI ethics and to support the development and implementation of trustworthy AI solutions in the context of diabetes management.

[11] EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA). (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts. In *https://www.europarl.europa.eu/* (PE 729.512). Retrieved January 23, 2023, from https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf

[12] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

[13] AI, HLEG. (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI*, p.4.

2 Description of Activities

The deliverable is concerned with the ethical problem statement of AI use in the context of diabetes management and the formulation to address these ethical challenges. Owing to the absence of a particular framework on AI ethics in health, the methodology behind the ethical recommendations is based on the interpretation of existing normative principles derived from bioethics as well as international conventions on human rights in the health context and AI ethics and corresponding regulation relating to AI (e.g., High-level Expert Group on Artificial Intelligence - AI HLEG, AI4People, EU AI Act). When choosing between different principles of similar hierarchy, bioethical principles were given precedence over general AI ethics implications (*lex specialis derogat legi generali* – special law repeals general laws). Moreover, stricter principles were given precedence over less strict principles. The recommendations were developed in a discourse among the authors of the paper and based on the majority decision of the researchers involved in the ethics team of MELISSA at TUM. In case of dissent, the minority view is stated explicitly in the document, if requested by the researcher.

2.1 Analysis of ethical issues pertaining to AI

The ethical evaluation of AI hinges on the properties traditionally associated with this family of technologies. While the academic discourse on AI provides us with different overlapping and competing definitions of AI, we decided to orient ourselves to the definition provided by the most recent OECD Definition of AI[14], which has been also harmonized with the EU AI Act. The definition reads as follows:

> *"*An AI system is a machine-based system that for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*" (OECD Definition of AI).*

Moreover, the EU AI Act specifies in ANNEX I of the Act the following approaches and techniques as AI:

- Machine learning approaches, including supervised, unsupervised, and reinforcement learning, using a wide variety of methods including deep learning;
- Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning, and expert systems;
- Statistical approaches, Bayesian estimation, search, and optimization methods.

A wider discourse on the legal and ethical application of AI reveals certain characteristics that render the technology distinct from conventional solutions[15]. This difference matters from a normative

[14] Regulation 2021/0106. *Regulation of the European Parliament and of the council laying down harmonized rules on Artificial Intelligence*. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206

[15] EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA). (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts. In *https://www.europarl.europa.eu/* (PE 729.512). Retrieved January 23, 2023, from https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf

perspective as some of the inherent properties of AI challenge traditional legal and moral concepts such as responsibility, liability, autonomy, or accountability[16].

The specificities that characterize AI solutions are identified as follows:

1. AI substitutes (partially or entirely) activities that have been associated as human such as the task of the recruiter in the pre-selection of candidates for a given job or a physician that analyses the data of a patient. This does not mean, however, that the entire process is replaced or that the final decision is made by an AI solution but that certain actions are performed by an AI solution[17]. That being said, AI reduces the number of human decisions made in specific processes (what we call automation). The extent to which this happens is determined by the design and human machine interaction.
2. When performing a task, AI relies on a given data input. While this data input can be dynamic and based on real-time data, the decision or conclusion that AI arrives at is determined by the received input. This leads to different epistemology between human and AI, as human decision-making can be based on the transfer of knowledge from one area to another area but also be based on empathy, emotions, and instinct, also called gut feelings[18].
3. Consequently, the quality of AI depends on the input data used, but also on the statistical model underlying the AI solution. This owes to the fact that AI processes data with the means identified in ANNEX I to the EU AI Act, which are mostly statistical in nature. Biases refer here to the structural deviations of the factual from the intended result[19]. However, in the discourse on AI ethics, algorithmic bias is framed as a conflict between the actual output produced by an AI solution and the principle of fairness. In the following, we use therefore the normative definition of bias used by Friedman and Nissenbaum (1996): "*we use the term bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others.*" (p. 332). Biases that overlap with certain characteristics of human beings such as gender, sex, or nationality result in discrimination are the main result of unfairly discriminated individuals or groups of individuals[20] and fall under the scope of AI regulation, for instance Art. 10 of the EU AI Act
4. Depending on the design of the system, decisions made by AI might not be completely understood in each single case. The "black box character" of AI, particularly self-learning AI, renders it difficult to understand what has caused a specific result[21]. The key reason is that outcomes generated by an AI solution depend on the combination between "dynamic data input" and machine learning which is constantly updating the decision-making process that underlies an AI solution. While "input" and "output" can be specified, the workings between are often obfuscated. The problem of algorithmic opacity applies in specific to artificial neural networks and/or deep learning approaches[22].

[16] Smith, H. (2021). Clinical AI: opacity, accountability, responsibility and liability. *AI & SOCIETY*, *36*(2), 535-545.
[17] Kriebitz, A., & Lütge, C. (2020). Artificial intelligence and human rights: a business ethical assessment. *Business and Human Rights Journal*, *5*(1), 84-104.
[18] Dzobo, K., Adotey, S., Thomford, N. E., & Dzobo, W. (2020). Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. Omics: a journal of integrative biology, 24(5), 247-263.
[19] Piedmont, R. L. (2014). Bias, Statistical. *Encyclopedia of Quality of Life and Well-Being Research*, 382-383.
[20] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.
[21] Zednik, C. (2021). Solving the black box problem: A normative framework for explainable artificial intelligence. Philosophy & technology, 34(2), 265-288.
[22] Rudin, C., & Radin, J. (2019). Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review, 1(2), 10-1162.

Furthermore, the design of AI solutions matters for the occurrence of the following problems:

- "Personalization" refers to the use of AI to tailor operations to specific individuals by leveraging personal data and understanding their unique characteristics. This process involves data, which falls under the scope of the GDPR and presents certain risks for personality rights. Additionally, personalized AI solutions tend to be affected by the cold start phenomenon.[23] This phenomenon arises when there is insufficient initial data for the AI to function effectively.
- "Amplification" refers to situations where the standardized application of an AI solution at scale aggravates pre-existing biases[24]. In contrast to biases of an individual recruiter or physician, the same bias that characterises an AI solution is rolled out over many use cases. The phenomenon occurs in specific cases when technical biases create more inequality than cognitive biases of human beings[25,26]. However, biases can enter data sets and thus algorithms trained with them through the use of historical data sets based on earlier medical decisions made by human actors. Human actors often tend to be biased in their judgments and guided by stereotypes. Such biases have been reported widely in AI ethics. Finally, these biases can be unintentional and impede the robustness of an AI solution[27].
- "Interconnectivity": An AI system can receive data from other devices with which it is connected or share information with other devices in a network[28]. This network character is not an inherent feature of AI solutions, as designers and developers of an AI solution as well as devices connected to the network can decide how data is stored and with which entities it is shared. Higher levels of interconnectivity go along with increased risks from the cyber security perspective and enhance the dependency on specific hardware solutions.

The AI ethics discourse has established normative frameworks to address challenges posed by these features. A key concept here is "explainability" or alternatively "explicability", which requires AI systems to be transparent and understandable for users[29,30]. The normative implications of the different approaches in AI ethics are covered in subchapter 3.2.

2.2 Analysis of ethical challenges pertaining to "AI Ethics in Health"

When compared to other AI use cases, we can encounter specificities that render AI in health distinct from other discussed use cases. This owes to normative arguments, for example, inequities and

[23] Ishakian, V., Muthusamy, V., & Slominski, A. (2018, April). Serving deep learning models in a serverless platform. In 2018 IEEE International conference on cloud engineering (IC2E) (pp. 257-262). IEEE. Feng, J., Xia, Z., Feng, X., & Peng, J. (2021). RBPR: A hybrid model for the new user cold start problem in recommender systems. Knowledge-Based Systems, 214, 106732

[24] Holstein, K., & Doroudi, S. (2021). Equity and Artificial Intelligence in Education: Will" AIEd" Amplify or Alleviate Inequities in Education?. arXiv preprint arXiv:2104.12920.

[25] Samorani, M., Harris, S. L., Blount, L. G., Lu, H., & Santoro, M. A. (2022). Overbooked and overlooked: machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management*, *24*(6), 2825-2842.

[26] Samorani, M., & Blount, L. G. (2020). Machine learning and medical appointment scheduling: creating and perpetuating inequalities in access to health care. *American journal of public health*, *110*(4), 440-441.

[27] Compare: AI, HLEG. (2019). High-level expert group on artificial intelligence. *Ethics guidelines for trustworthy AI,* p.7.

[28] German Data Ethics Commission. (2019). Opinion of the Data Ethics Commission, p.14. Retrieved from https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2

[29] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and machines*, *28*(4), 689-707.

[30] Jongepier, F., & Keymolen, E. (2022). Explanation and Agency: exploring the normative-epistemic landscape of the "Right to Explanation". Ethics and Information Technology, 24(4), 49.

inequalities in access to healthcare could be amplified by algorithmic bias and thus violate the right to health[31]. Furthermore, the constitution of the WHO refers to the right to health as a fundamental right of "*every human being without distinction of race, religion, political belief, economic or social condition non-discrimination*"[32]. From a wider ethical perspective, the specificities of AI solutions have been generally discussed in terms of criticality and under the prism of "high-risk" as understood by the EU AI Act. The EU AI Act refers here to high risk as situations that "*pose significant risks to the health and safety or fundamental rights of persons*"[33]. Moreover, criticality is closely tied to human rights and the fundamental rights enumerated in the EU Charter[34].

The development and deployment of AI in health constitute a "high-potential" but also "high-risk" case for different normative reasons.

- The use of AI in health constitutes one of the most promising use cases of AI in general[35]. The potential of AI in these areas is closely related to human rights - the right to health - but also to other normative concepts including the UN SDGs (in particular UN SDG 3 and 5). The potential reduction of physical pain by AI solution is particularly relevant for utilitarian approaches or ethical views influenced by effective altruism.
- The AI solutions in the context of health do not constitute a monolithic block[36]. In fact, there are fundamentally different types of AI applications in healthcare (patient-, physician-, research-, - drugs-, administration-centric). These AI solutions involve different types of data and generate different types of consequences (relating to the physical or mental well-being of patients, the distribution of risks among individuals, the allocation of time and resources, financial consequences, etc.).
- In application cases of AI in health, there is a higher likelihood of life and death-related decisions, especially when it comes to decisions regarding diagnosis and course of treatments. Incorrect output of AI decisions can here lead to irreversible consequences impeding the rights of patients, but also leading to physical impairments of patients or to significant harm or mental stress for patients[37]. Moreover, as AI is limited to its statistical reasoning, some variables such as subjective experience of the patient and its possible impact on their current state, and other contexts related variations might be overlooked or misunderstood by the system[38,].

[31] German Data Ethics Commission. (2019). Opinion of the Data Ethics Commission, p.19. Retrieved from https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2

[32] *Constitution*. (n.d.). https://www.who.int/about/governance/constitution

[33] Regulation 2021/0106. *Regulation of the European Parliament and of the council laying down harmonized rules on Artificial Intelligence*. European Parliament, Council of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (see section 1.1.)

[34] EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA). (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts, p.32. In *https://www.europarl.europa.eu/* (PE 729.512). Retrieved January 23, 2023, from https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf

[35] Kriebitz, A., & Lütge, C. (2020). Artificial intelligence and human rights: a business ethical assessment. *Business and Human Rights Journal*, *5*(1), 84-104.

[36] Ramesh, A. N., Kambhampati, C., Monson, J. R., & Drew, P. J. (2004). Artificial intelligence in medicine. Annals of the Royal College of Surgeons of England, 86(5), 334.

[37] See: German Data Ethics Commission. (2019). Opinion of the Data Ethics Commission, p.19. Retrieved from https://www.bmj.de/SharedDocs/Downloads/DE/Themen/Fokusthemen/Gutachten_DEK_EN.pdf?__blob=publicationFile&v=2

[36] The words "their" or "they" are here used for gender-neutral purposes and not as plurals.

- The use of AI in health involves different stakeholder groups including patients, families of patients, medical teams, but also the entire ecosystem of health including commercial actors such as insurance companies and privately owned hospitals. Stakeholder groups have not always compatible interests, for example when it comes to the use or evaluation of data[39].
- The application of AI in health might unfold different implications based on the time horizon. Short-term gains such as reduction in pain can be followed by long-term issues such as addiction to specific medications. Moreover, data is dynamic in the sense of constantly changing. AI solutions themselves can also have an impact on changes of the input data.
- In contrast to other areas, designers of AI solutions are confronted with a multitude of variables. Moreover, there are many concepts which are based on subjective perceptions, such as "pain", the mental state of the patient, etc.

2.3 Ethical challenges pertaining to diabetes management

The use of AI in diabetes management faces specific conditions that transcend the general challenges of applying AI to the medical field. The management of diabetes takes place in a setting where technical error or human miscalculation might cause significant physical pain or discomfort to the PwD. Consequently, AI solutions developed in the diabetes context is likely to fall under the high-risk definition of the EU AI Act as defined in Art. 6 of the EU AI Act. This implies that the project management needs to consider the requirements laid down the in the high risk-requirements of the EU AI Act, particularly in terms of risk management as well as mitigation measures to prevent bias. Further still, there are several conditions that underpin the societal and ethical relevance of the development and deployment of AI solutions in the specific diabetes context.

1. One key factor is that the target population is people with a chronic health condition, which implies that the benefits of diabetes-related AI solutions are particularly high. The longitudinal aspect (problem of many implementations in healthcare) constitutes a major challenge of the project[40].
2. Insulin injection and food consumption happen on a daily basis. Consequently, everyday use entails high risks of 'misuse'. This applies in specific to the potential "worst case" namely hypo- or hyperglycaemia caused by a wrong recommendation from the system. The symptoms coming along with such conditions entail high risks for the patient ranging from physical harm and/or mental discomfort to possible coma[41]. Furthermore, there might be general limitations on the use of specific hardware solutions employed such as limitations posed by quality of a camera or a sensor's ability to cope with temperature. The project is therefore situated not only relevant to health and safety considerations, but also might produce irreversible consequences, which feature prominently in Art. 6 of the EU AI Act.
3. The data used is personal, particularly when it comes to food or beverage consumption, physical activity as well as sleeps patterns and parameters related to glucose control[42]. These concerns apply to other variables such as hormones that correlate with the biological sex of the person. The curation of these data is a big concern from the perspective of cybersecurity

[39] Ballantyne, A. (2019). Adjusting the focus: a public health ethics approach to data research. Bioethics, 33(3), 357-366.
[40] Silva, J. A. D., Souza, E. C. F. D., Echazú Böschemeier, A. G., Costa, C. C. M. D., Bezerra, H. S., & Feitosa, E. E. L. C. (2018). Diagnosis of diabetes mellitus and living with a chronic condition: participatory study. BMC Public Health, 18, 1-8.
[41] Harding, J. L., Pavkov, M. E., Magliano, D. J., Shaw, J. E., & Gregg, E. W. (2019). Global trends in diabetes complications: a review of current evidence. Diabetologia, 62, 3-16.
[42] Kovatchev, B. (2019). A century of diabetes technology: signals, models, and artificial pancreas control. Trends in Endocrinology & Metabolism, 30(7), 432-444.

and data privacy. Moreover, it falls under Art. 4(5) of the GDPR since involves health data and not just merely personal data.

4. The design of the project including the different AI solutions developed suggest specific risks along the lines of algorithmic bias, accuracy as well as accessibility considerations. Certain variables are not measured directly, but rather indirectly, for example, when analysing the amount of carbohydrates in a meal through AI-based dietary assessment ("carb counting"). The same holds true for glucose measurement, which might be based on one signal only. From an epistemological perspective, the margin of error depends on the methodology of calorific estimation[43]. In addition to the risks of algorithmic bias and accuracy concerns, it is crucial to address issues at the intersection of accessibility and digital literacy. Specifically, hardware design choices—such as access to smartphones with cameras, connectivity, digital skills, and supplementary devices like smartwatches—must account for the unique impacts on specific groups, such as the elderly or individuals with lower socio-economic status. This consideration is essential in the design and deployment of the AI solutions proposed in the project.
5. The quality of the AI solutions developed in the project partly depends on the design and implementation of human-machine interaction (HMI) considerations. This applies specifically to the self-management aspect of the tools proposed. It is therefore important to find measures to enhance user-friendliness of the app and explainability, so that users make decisions based on informed consent. Ethical frameworks relating to AI use in health are therefore increasingly considering the importance of strong user engagement.[44] Moreover, the relevance of HMI reinforces the issue of accessibility and privacy concerns mentioned in section 2.1.

[43] Amugongo, L. M., Kriebitz, A., Boch, A., & Lütge, C. (2023, January). Mobile Computer Vision-Based Applications for Food Recognition and Volume and Calorific Estimation: A Systematic Review. In Healthcare (Vol. 11, No. 1, p. 59). Multidisciplinary Digital Publishing Institute.

[44] Xu, W. 'Toward human-centered AI: a perspective from human-computer Interactions', 26(4), pp.42-46, 2019; Barda, A.J., Horvat, C.M. and Hochheiser, H., 'A qualitative research framework for the design of usercentered displays of explanations for machine learning model predictions in healthcare.', BMC medical informatics and decision making, 20(1), pp.1-16, 2020.

3 Results and Discussion

In the following, we derive implications for the MELISSA project that are informed by two discourses on bioethics and on ethics in AI. Not all recommendations can be turned into "metrics", however developers and AI experts should be aware of different epistemological and normative challenges as well as diverse stakeholder demands when developing and deploying AI solutions in the context of MELISSA. Here, we will present our recommendations with a gradient of strength: from must (highest priority) to strongly recommend (important), to recommend (to consider). Prioritization is necessary in this context, due to technical, financial, and time-related constraints, but also suggested in literature on AI ethics as well as AI and human rights.[45] Our aim is therefore to get as close as possible to the recommendations identified here with the available resources and capacities.

Furthermore, the ethical framework is subdivided in different sections, which refer to the different areas of concern at the intersection of bioethics and AI ethics. Existing ethical frameworks on AI and bioethics have operated with terms such as beneficence, non-maleficence, autonomy, justice and explainability, for instance European frameworks such as AI4People or HLEG AI.[46] However, the specificities of the project, but also the most recent legal innovations of the EU AI Act also warrant the consideration of human rights in health, requiring a contextualization of said principles along the lines of established legal and ethical frameworks such as the Oviedo Convention.[47] The principles of patient centricity present a contextualization of beneficence, linking to overarching dignity considerations as expressed in the Oviedo Convention, particularly Article 1, which reaffirms the focus on the human being in the context of health and medicine.[48] The principle of prevention and minimization of harm contextualizes non-maleficence, as found in Article 2, which aims to protect the welfare and interests of individuals in this context. Given the relevance of data in the context of MELISSA, this section addresses violations of the principle of data privacy. The principle of autonomy and self-determination draws from the principle of autonomy, especially reinforced in Article 5, which emphasizes the principle of consent in the medical context. The principle of equality, non-discrimination, and accessibility is a contextualization of justice considerations addressed in Article 3, also referring to different sources of unequal treatment that could occur in the context of the project. Finally, the principle of moral integrity and transparency concretizes the principle of explainability. Since the ethics guideline does not solely address the development of AI solutions in isolation, it is warranted to establish overarching principles of transparency and communication based on common ethical principles and stakeholder expectations, particularly addressing the issue of algorithmic opacity.[49]

[45] Kieslich, K., Keller, B., & Starke, C. (2021). AI-ethics by design. Evaluating Public Perception on the Importance of Ethical Design Principles of AI. arXiv preprint arXiv:2106.00326. Götzmann, N. (2017). Human rights impact assessment of business activities: Key criteria for establishing a meaningful practice. Business and Human Rights Journal, 2(1), 87-108.

[46] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Minds and machines, 28, 689-707.

[47] den Exter, A. (2022). Artificial Intelligence in Health Care and the Oviedo Convention. *Medicne pravo*, (2 (30)), 9-23.

[48] Council of Europe. (1997). Convention on Human Rights and Biomedicine. (Ovideo Convention). https://www.coe.int/de/web/bioethics/oviedo-convention.

[49] Amugongo, L. M., Kriebitz, A., Boch, A., & Lütge, C. (2023). Operationalising AI ethics through the agile software development lifecycle: a case study of AI-enabled mobile health applications. *AI and Ethics*, 1-18.

"Patient Centricity"

1. Patient centricity means that patients' interest takes precedence over other utilitarian, ecological or financial considerations[50]. PwD are defined as the users of the MELISSA app. The principle of patient centricity is closely related to the principle of human dignity, as outlined in the recent "Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law" by the Council of Europe. This principle is further substantiated in frameworks such as the Convention on Human Rights and Biomedicine of the Council of Europe, which specifies in the implications of human dignity and the primacy of the human being in the context of health.[51]
    a. We *strongly recommend* a focus on the physical health of PwD. In accordance with the mission statement, this implies the optimisation and personalisation of insulin treatment and achievement of optimal glucose control.
    b. As a second priority, we *recommend* enhancing the overall life quality of people with diabetes. This includes enhancing positive mental and cognitive effects of the use of MELISSA when possible, and controlling the impact of using the technology through surveys, comparative studies, interviews, and the integration of strong feedback loops from users to developers. This monitoring is to be ongoing as long as the tool is being used.
    c. As a further priority, we *recommend* focusing on benefits for PwD related to life other than the original purpose of the app (such as more time for other activities, less dependency, etc.)
    d. We *recommend* tackling anxiety-related side effects (e.g., panic attack, enhancement of pre-existing conditions, etc.) induced by the use of the app and considering the potential loss of self-treatment abilities.[52]

2. We *recommend* considering the interests of other parties, in particular the patient ecosystem in the design, deployment, and use of the app. This includes:
    a. The interests of doctors, carers, and other medical personnel to be freed from certain actions that consume a lot of time.
    b. Improving access to the service in the meaning of financial aspects (it is cheaper than conventional services).
    c. Considering the community of people with diabetes and the impact the app might have on their current balance, life quality, etc.
3. In case of conflict and allocation of resources between 1 and 2, we *strongly recommend* prioritizing on the needs of PwD and most importantly the original purpose of the app as defined in the mission statement.

[50] We derive this principle from the MELISSA mission statement, the principle of beneficence in bioethics and robustness of the Trustworthy AI Framework (HLEG). Further literature is: Health Care Ethics | Internet Encyclopedia of Philosophy. (n.d.). https://iep.utm.edu/h-c-ethi/.

[51] Council of Europe. (1997). Convention on Human Rights and Biomedicine. (Ovideo Convention). https://www.coe.int/de/web/bioethics/oviedo-convention.

[52] Westermann, T., Möller, S., & Wechsung, I. (2015, August). Assessing the relationship between technical affinity, stress and notifications on smartphones. In proceedings of the 17th international conference on human-computer interaction with mobile devices and services adjunct (pp. 652-659).

"Prevention and Minimization of Harm"[53]

1. Prevention of harm means that the developers or other crucial decision makers are considering potential harm to patients, their ecosystem, and the community as whole, and thus not being influenced by other factors such as financial considerations or reputational gains when making decisions in respect to the project[54].
2. The design, deployment, and use of the app *must not* conflict with legal principles including standards emerging on data management from the General Data Privacy Regulation (GDPR).[55] Developers and everyone involved in the AI lifecycle must consider all legal restrictions when developing the application. Moreover, we recommend alignment with recommendations and guidance on the treatment of health data made by governmental organizations.[56]
    a. The app *must* be protected from misuse of external parties according to best practice in cybersecurity.
    b. Particular emphasis lies on the prevention of data breaches, especially when it comes to data that is revealing very personal information of the PwD or data that is indicative of the biological sex of the user or religious habits such as specific diets.
3. Furthermore, we recommend strict adherence to the principle of data minimization as outlined in the GDPR, Art. 5(c).[57] Every use of personal data *must* be justified in relevant documents, such as the data management plan. Apart from obedience to law, MELISSA *must* prevent harm done to the patient in every way.
    a. This means that MELISSA *must not* worsen the mental, physical, or cognitive condition of PwD.
    b. Irreversible harm *must* be prioritized over other considerations.
    c. We *strongly recommend* that the app is developed and updated based on feedback by PwD or based on externally or internally generated data[58].
    d. We *recommend* that feedback (emails, but also any performance-related data from the feedback loop) is taken into due account during the entire lifecycle and life time of the application.
    e. We would further recommend that instances of harm or concerns and complaints (e.g. via user reports, internal feedback) be kept in a separate log, where said log is fed back into the product improvement process and lifecycle.
4. When taking unavoidable harms and benefits into account, the team *must* find an appropriate balance between avoidance of "risks" for PwD and "data privacy". The assessment needs to be lawful, proportionate, and explainable. Therefore, a broader understanding of risk,

[53] The framework addresses harm originating in technical error and human misconduct. The Panel for the Future of Science and Technology has treated this in the "misuse of biomedical tools" sections. However, it is also relevant from the perspective of loopholes in terms of accountability identified in the paper. See: EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA) [EPRS STOA]. (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf, p. 36; but also p. 15. (Keyword: "Patient harm due to AI errors").

[54] Acting at the expense of stakeholders due to financial considerations or negligence and acts of omission fall under that principle.

[55] Regulation 2016/679. General Data Protection Regulation. European Parliament, Council of the European Union. https://eur- lex.europa.eu/legalcontent/EN/TXT/PDF/?uri =CELEX:32016R0679.

[56] For example: Bundesministerium für Wirtschaft und Energie: „Orientierungshilfe zum Gesundheitsdatenschutz." Berlin, November 2018, p. 61 – 67.

[57] Regulation (EU) 2016/679. General Data Protection Regulation. (2016) https://eur-lex.europa.eu/eli/reg/2016/679/oj .

[58] Here, external data refers to data generated outside of the use case for example statistical data on the use of the app and surveys, whereas internal data describes data used in the model.

including ethical considerations, needs to be integrated into the project management level. This is relevant particularly from the perspective of Art. 9 of the EU AI Act.

"Autonomy and Self-Determination"[59]

1. Self-determination means that all steps of a patient's course of disease management are voluntary and based on enlightened consent[60]. The principle has been reinforced in most recent approaches on AI regulation, particularly, the Council of Europe's Convention of Artificial Intelligence, Human Rights, Democracy and Rule of Law, for instance in Art. 7 referring directly to human dignity and individual autonomy.[61] This implies that patients always have the choice to follow or not a course of treatment, opt out of it, and follow or not the decisions of the AI. Autonomy concerns the informed and uninfluenced choice when, how, and whether to use AI. Autonomy is maximized by giving users choices that are more individual in or using an AI solution concerning consent, confidentiality, and privacy[62]. The more the decision affects the life of a patient, the more important is autonomy.[63]
2. Technologies aiming at nudging, priming, influencing PwD, or inducing behavioural changes are facing ethical limitations. These limitations have been defined in the EU AI Act proposal, and thus we derive *strong recommendations* from this.
    a. The data and notifications/alerts shown to the patient or medical person *must* be as accurate and precise as technically possible. Data output delivered to the user of the app *must not* deceive the user, even if deception aims at improving the condition of the patient.[64]
    b. Exceptions to this rule *must* be well reasoned and based on existing practices by human actors in the specific use case. In other words, exceptions have to be clearly explained, with reasonable arguments based on the case-by-case scenarios.
    c. We *recommend* that MELISSA points out behaviour of the patient considered as positive without penalising the individual for behaviour regarded as unhealthy or unethical (such as consuming food considered to be unhealthy, etc.).
3. We *recommend* considering potential overuse by the patient such as too much time spent on the app, and the possible reduction of the initial goal – to have the patient manage safely - when it comes to well-being by the patient.[65]
4. We *recommend* taking account for specific eating times, conditions, or habits that individual patients are facing, including cultural aspects that are relevant to the specific deployment context.

[59] The words self-determination and autonomy carry different meanings. Consequently, we provide definitions of what we mean by the concepts mentioned.
[60] Cambridge Dictionary. (2023). self-determination definition: 1. the ability or power to make decisions for yourself, especially the power of a nation to decide. Learn more. https://dictionary.cambridge.org/dictionary/english/self-determination.
[61] Council of Europe. (2024). Convention on Artificial Intelligence, Human Rights, Democracy and Rule of Law. https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence
[62] Ewuoso, C. (2021). An African Relational Approach to Healthcare and Big Data Challenges. *Science and Engineering Ethics*, *27*(3), 1-18.
[63] The implications of autonomy are manyfold and treated not exclusively in the "autonomy and self-determination" chapter.
[64] EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA) [EPRS STOA]. (2022). Table 2, p. 38.
[65] See: ibid. p.39.

5. There *must be* a possibility to rectify input of the app, which is perceived as false- by the patients themselves or by a health care professional, depending on the setting.
    a. We *strongly recommend* the existence of a contact hotline or other contact mechanisms preferably reaching a human helper to rectify issues that are unclear or even problematic once the solutions enter the market.
    b. We *strongly recommend* engaging users in the design and development of the AI tool.
    c. Furthermore, we *strongly recommend* precising the feedback loop also in individual cases, so that patients can forward observations that are relevant for the constant improvement of the app e.g., certain types of food are not recognized by the app, etc.
    d. There *must* be due consideration in the development of how much can the patient go back to non-digital support when necessary, such as a fall-back option.

"Equity, Non-discrimination and Accessibility"

1. The principle of equity and the right to health require non-discriminatory treatment of - and within the population of PwD. Thus, equity must be realized in the entire AI lifecycle consisting of design, deployment and use[66] A major aim of the project is to prevent structural differences in the treatment of patients that could be reinforced by algorithmic bias or other design choices made in the project. Algorithmic bias may arise during aggregation, learning, and deployment processes, along with other related concerns.[67] This also involves configuring accessibility considerations to address hardware shortcomings or potential misuse in such situations:
    a. The development of the app *must* aim at preventing biases in the treatment of PwD of the European Union based on traditional discrimination criteria such as phenotypic traits and ethnicity, race[68], gender, nationality, socio-economic status, and sex. The MINMAR (MINimum Information for Medical AI Reporting) principles can here be used for pre-testing procedures; however, the diabetes use case might require additional demographic characteristics.
    b. Apart from societal biases, the app *must* address specific medical conditions faced by PwD such as individuals with disabilities (e.g., blindness, deafness, general learning intellectual disabilities, etc.). Moreover, we *strongly recommend* checking what kind of medically relevant criteria that may overlap with discriminatory factors (hormones, age, body mass index, number of pregnancies, body temperature, genetic aspects, pre-existing conditions and comorbidities, and interference with other treatments). Recognizing these underlying criteria in a broader context goes beyond protected criteria outlined in anti-discrimination legislation.
2. The consideration of the mentioned sources of bias should include potential sources such as aggregation, learning, and deployment biases. We strongly recommend aligning the project's

[66] This point has been addressed in earlier literature such as: High Level Experts' Group on Trustworthy AI; p. 11. The EPRS report has dealt wit.h the implication of fairness in health: EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA) [EPRS STOA]. (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729512/EPRS_STU(2022)729512_EN.pdf page 34.

[67] Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388-409.

[68] The use of the term "race" is controversial, particular when translated to German and French. In these cases, we recommend the use of alternative concepts.

existing practices with Article 10 of the EU AI Act, which pertains to data and data management. The principle of equity entails consequences for the design and modalities of the use case. While it is difficult to take all individual diets and use environments into account, we *strongly recommend* designing the app as inclusive as possible to cover vulnerable groups. This includes individuals that have a special diet for medical reasons or migrant groups residing in Europe.

a. The development of the app *must* consider limitations of the data set and check for so called "historical biases" that entered data which are relevant to the performance of the app[69].
b. In settings for which the app does not include certain use cases, limitations in the use of the app *must* be stated[70].
c. The development of the app *must* consider the accessibility for different types of patients' population within the European Union, which includes economic and social but also potential geographical conditions that affect the performance of the app such as temperature. A specific emphasis is here to reduce disparities between different European countries[71].

3. Moreover, we *recommend* that publicly available reports and/or explanations handed out to patients/health practitioners include information on whether and to which extent the performance of the app varies based on demographic factors (age, biological sex, ethnicity, etc.). In the development and deployment of MELISSA, developers might encounter conflicts of allocating resources such as time, money, or effort to:
   - either representing as many groups as possible in the target population and reducing potential biases between different groups (sex and gender, ethnicity, age, etc.)
   - or to widen the use case (e.g., including different types of national cuisines/ beverages).

   In such settings, we *recommend* prioritizing covering a target population within the European Union as inclusive as possible. We *strongly recommend* tracking decisions that have been made in such situations, and making sure to have strong arguments behind each decision that can be explained to anyone requesting clarification.
4. When having to choose between different types of inequities or biases, we *strongly recommend* focusing on fixing the issue with the highest relevance for PwD in the European Union[72]. The patients that receive the gravest consequences arising from the disparity or the ones that would rely most on an AI solution should be prioritized[73]. Moreover, the judgment should be guided by factors such as the number/proportion of individuals affected and the discomfort created by the bias, inequity in access, or the reduction of quality for the affected group[74].

[69] Example: The data used in the model comes mainly from countries situated in a specific climate zone or with a specific population that is not representative for the target population.
[70] Example: The food recognition app fails to detect the amount of glucose in beverages.
[71] This view is also espoused in EPRS, European Parliamentary Research Service Scientific Foresight Unit (STOA). (2022). Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts, p.52. In https://www.europarl.europa.eu/ (PE 729.512).
[72] Here, developers need to consider which individuals are the most in need for the specific AI solution.
[73] Example: Individuals with impaired sight profit mostly from the recognition of nutrients by an AI solution. However, the design of the app bears major implications in terms of accessibility for such individuals.
[74] Here, we orient ourselves to moral traditions that point out to prioritize the "weakest" in a given situation. Refer to: Desai, P., Shook, J. R., & Giordano, J. (2022). Addressing and Managing Systemic Benefit, Burden and Risk of Emerging

"Trust and Moral Integrity"

1. The solutions and processes developed in the context of the MELISSA project should aim at enhancing trust of PwD in AI, but also show the example of the implementation of trustworthy AI in this specific context. Trust is to be understood as the precondition for the adoption of AI solutions in healthcare by patients and practitioners.
2. To enhance this individual trust in the system, communication policies and explainable AI *must* meet ethical standards of trust and moral integrity. Furthermore, we *strongly recommend* considering trust building and adoption rates as important metrics in the development of MELISSA.
3. Integrity implies that all communications regarding the project are transparent and that promises made to stakeholders and to the general public are realized in practice. If keeping promises is not possible, moral integrity requires accountable parties to inform individuals affected as soon as possible and, if necessary, the public. We thus set as a *must* the need for transparency in this case.
4. When developing the app, a focus *must* lie on providing explainability for all relevant stakeholder groups (customized for patients, caregivers, providers). This includes the following points:
    a. We *strongly recommend* explaining how the app works in understandable language for each population targeted[75]. This is particularly relevant from the perspective of existing digital divides.
    b. We *strongly recommend* explaining how, why and which data has been used as training data.
    c. We *strongly recommend* considering vulnerable populations such individuals with disabilities or the elderly when communicating relevant information to stakeholders, for instance in respect to the workings of the algorithms.
    d. We recommend considering relevant industry ethical design standards such as ISO/IEC 42001 or IEEE P7002-2022.
5. We *strongly recommend* a manual accessible by all patients - maybe consider having it in multimodal ways.
6. We *strongly recommend* that statistical data be published with occasional audits of the system. Furthermore, we *recommend* pre- and post-implementation risk assessment with obligatory transparent publication of how the model works but also how the decisions are made.[76]
    a. We *recommend* that the publication includes statistical data on the patient demographic characteristics including age, sex and gender, and race/ethnicity.
    b. We *recommend* that the publication includes statistical data on the socio-economic status of individuals.
    c. We *recommend* that publications refer to methods that have been employed in order to prevent, mitigate, or resolve technical bias.

Neurotechnology. *AJOB neuroscience*, *13*(1), 68-70.; Ewuoso, C. (2021). An African Relational Approach to Healthcare and Big Data Challenges. *Science and Engineering Ethics*, *27*(3), 1-18.

[75] The passage "how the model works" implies method-related information as understood in common practice.

[76] The 2021 EC proposal for the EU AI Act lists several requirements for AI solutions. One obligation is to conduct post-marketing monitoring. Likewise, one requirement is to ensure "appropriate degree of transparency". This implies providing users with information on capabilities and limitations of the system. This links up to performance variations of a given tool across the entire spectrum of demographic features (age , sex etc.). (ref.: EPRS Report (2022). p. 48. "5.3. Create an AI passport and traceability mechanisms for enhanced transparency and trust in medical AI")

7. We *strongly recommend* establishing education programs to disseminate information concerning the app to health care professionals and to enhance their skills. This links up to most recent pushes towards enhancing digital literacy as a joint mission for all entities in the AI lifecycle.
8. We *strongly recommend* creating a final ethics report relating to the project covering the points discussed here, particular in terms of how ethical risks were resolved in the project We further recommend alignment of the ethics report with fundamental rights, given the focus of the final version of the EU AI Act on fundamental rights and the need of performing a fundamental rights impact assessment along the lines of Art. 29 a of the EU AI Act.

4 Conclusion

The ethical framework developed serves the purpose of detecting potential ethical risks appearing throughout the project's lifecycle and of deriving conclusions on how to mitigate these. The ethical risks stem from the general features of AI, the limitations of using AI in health, and, finally, the specificities of diabetes management. Consequently, the framework expressed here identifies bio-ethics and AI ethics as two distinct ethical frameworks that both entail partly conflicting implications for AI solutions developed in the context of MELISSA.

One major contribution of the framework here is to present a way to navigate such situations, especially when it comes to conflicting principles or trade-off. This explains the strong focus on prioritization, which is a major component of the ethical framework.

Further still, the framework articulates based on existent literature five overarching ethical principles for the conduct of the project. These are "patient centricity", "prevention and minimization of harm", "autonomy and self-determination", "equity and non-discrimination" as well as "trust and moral integrity". The stated ethical principles have been substantiated by clear ethical recommendations.

The framework distinguishes here between different levels of criticality, in order to facilitate the prioritization of different ethical topics. Strong recommendations are for example to concentrate on realizing statements made in public concerning MELISSA, to ensure that the use of MELISSA does not violate the principle of non-discrimination, and to disclose relevant information concerning the solutions to the public.

The recommendations presented here are subject to constant improvement. They serve as an instrument to raise awareness to critical issues in the development of MELISSA, but also function as a voluntary code of conduct for the parties involved in the project. In cases of conflict between the guidelines and legal principles, the application of legal principles takes precedence.