**Paper presented at the conference "Artificial Intelligence and Social Equity"
at the Université Bordeaux Montaigne, May 14-16, 2023, Doi:10.24406/publica-3667**

Trustworthy AI in the health sector: Challenges and solutions illustrated by three real-world examples

Bernd Beckert[1], Julia Hoxha[2], Philipp Kellmeyer[3], Patrick Philipp[4]

ABSTRACT

AI promises efficiency gains and innovations in the health sector. In the future, AI-based apps will be able to recognise symptoms of illness and warn patients before an emergency occurs, doctors will be able to have medical data, such as EEG data, interpreted in an automated way, or they will be able to make therapy recommendations based on a large number of studies that are automatically evaluated by the AI.

But what about the reliability of such applications? How robust is the data being used for training? What about transparency and explainability? Do the data reinforce established patterns of discrimination? What happens when the system "learns" new patterns based on new data and changes its output? Questions like these are being discussed under the heading of "Trustworthy AI ." The concept, which been developed by a high-level expert group in the EU in 2019 comprises of seven dimensions ranging form "human oversight" to "accountability". Since the concept was formulated, there have been many attempts to translate it into practical guidelines to make it applicable for the development and implementation of AI.

However, practical examples and best practices are still rare. This contribution presents three exemplary implementations of AI in the health sector: A chatbot app used as a digital patient companion (1), a diagnostic pattern recognition system (2), and a clinical decision support system (3).

[1] Fraunhofer-Institute for System and Innovation Research ISI, Karlsruhe, email: bernd.beckert@isi.fraunhofer.de; [2]Zana Technologies, Karlsruhe/ Berlin; [3]University Freiburg, Medical Center; [4]Fraunhofer IOSB, Karlsruhe.

It turns out that in the projects analyzed, certain dimensions of trustworthiness were in the foreground, while others turned out to be not relevant. Against this background, it does not seem reasonable to expect AI projects to fulfil all seven dimensions of the concept equally. Furthermore, the search for successful implementation strategies has recently given way to a discussion related to the practical requirements of the planned AI Act of the EU.

Keywords: Trustworthy AI, implementation, AI-based health applications, AI chatbot, diagnostic pattern recognition, clinical decision support, empirical analysis of implementation, best practices, MDR, AI Act

IA de confiance dans le secteur de santé : Défis et solutions illustrés par trois exemples concrets

ABSTRACT IN FRENCH

L'IA promet des gains d'efficacité et des innovations dans le secteur de la santé. À l'avenir, les applications basées sur l'IA seront capables de reconnaître les symptômes d'une maladie et d'avertir les patients avant qu'une urgence ne survienne, les médecins pourront faire interpréter les données médicales, telles que les données EEG, de manière automatisée, ou ils pourront faire des recommandations thérapeutiques sur la base d'un grand nombre d'études évaluées automatiquement par l'IA.

Mais qu'en est-il de la fiabilité de ces applications ? Quelle est la solidité des données utilisées pour la formation ? Qu'en est-il de la transparence et de l'explicabilité ? Les données renforcent-elles les schémas de discrimination établis ? Que se passe-t-il lorsque le système "apprend" de nouveaux modèles sur la base de nouvelles données et modifie ses résultats ? Les questions de ce type sont débattues sous le titre "Trustworthy AI" (IA digne de confiance). Ce concept, élaboré par un groupe d'experts de haut niveau au sein de l'UE en 2019, comprend sept dimensions allant de la "surveillance humaine" à la "responsabilité". Depuis que le concept a été formulé, de nombreuses tentatives ont été faites pour le traduire en lignes directrices pratiques afin de le rendre applicable au développement et à la mise en œuvre de l'IA.

Toutefois, les exemples concrets et les meilleures pratiques sont encore rares. Cette contribution présente trois exemples de mise en œuvre de l'IA dans le secteur de la santé : Une application de chatbot utilisée comme compagnon numérique du patient (1), un système de reconnaissance des formes de diagnostic (2) et un système d'aide à la décision clinique (3).

Il s'avère que dans les projets analysés, certaines dimensions de la fiabilité étaient au premier plan, tandis que d'autres se sont révélées non pertinentes. Dans ce contexte, il ne semble pas raisonnable de s'attendre à ce que les projets d'IA remplissent les sept dimensions du concept de la même manière. En outre, la recherche de stratégies de mise en œuvre réussies a récemment cédé la place à une discussion relative aux exigences pratiques de la loi sur l'IA prévue par l'UE.

Keywords in French: IA de confiance, mise en œuvre, applications de santé basées sur l'IA, chatbot d'IA, reconnaissance des schémas de diagnostic, aide à la décision clinique, analyse empirique de la mise en œuvre, MDR, législation sur IA.

Vertrauenswürdige KI im Gesundheitssektor: Herausforderungen und Lösungen anhand von drei Beispielen aus der Praxis

ABSTRACT IN GERMAN

KI verspricht Effizienzgewinne und Innovationen im Gesundheitssektor. Künftig werden KI-basierte Apps in der Lage sein, Krankheitssymptome zu erkennen und Patienten zu warnen, bevor ein Notfall eintritt, Ärzte können medizinische Daten, wie z. B. EEG-Daten, automatisiert interpretieren lassen oder Therapieempfehlungen auf der Grundlage einer Vielzahl von Studien geben, die von der KI automatisch ausgewertet werden.

Doch wie steht es um die Zuverlässigkeit solcher Anwendungen? Wie robust sind die Daten, die für das Training verwendet werden? Wie steht es um Transparenz und Erklärbarkeit? Verstärken die Daten etablierte Muster der Diskriminierung? Was passiert, wenn

das System auf der Grundlage neuer Daten neue Muster "lernt" und seine Ergebnisse ändert? Fragen wie diese werden unter der Überschrift "Vertrauenswürdige KI" diskutiert. Das Konzept, das 2019 von einer hochrangigen Expertengruppe in der EU entwickelt wurde, umfasst sieben Dimensionen, die von "menschlicher Aufsicht" bis zu "Rechenschaftspflicht" reichen. Seit der Formulierung des Konzepts gab es viele Versuche, die Anforderungen in praktische Richtlinien zu übersetzen, um es für die Entwicklung und Implementierung von KI anwendbar zu machen.

Praktische Beispiele und Best Practices sind jedoch immer noch rar. In diesem Beitrag werden drei beispielhafte Implementierungen von KI im Gesundheitssektor vorgestellt: Eine Chatbot-App als digitaler Patientenbegleiter (1), ein diagnostisches Mustererkennungssystem (2) und ein klinisches Entscheidungsunterstützungssystem (3).

Es zeigt sich, dass bei den analysierten Projekten bestimmte Dimensionen der Vertrauenswürdigkeit im Vordergrund standen, während andere sich als nicht relevant herausstellten. Vor diesem Hintergrund erscheint es nicht sinnvoll, von KI-Projekten zu erwarten, dass sie alle sieben Dimensionen des Konzepts gleichermaßen erfüllen. Darüber hinaus ist die Suche nach erfolgreichen Implementierungsstrategien in letzter Zeit einer Diskussion gewichen, die sich auf die praktischen Anforderungen des AI Acts der EU bezieht.

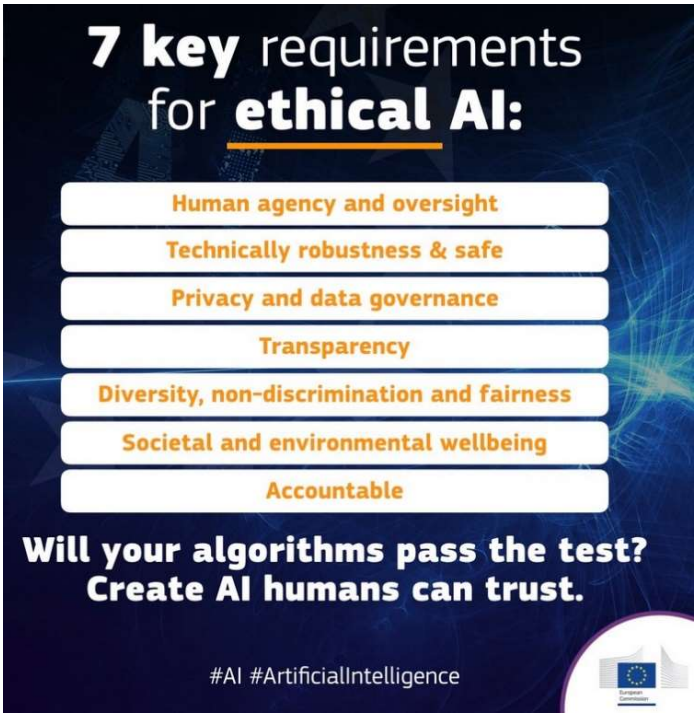
Keywords in German: Vertrauenswürdige KI, Implementierung, KI-basierte Gesundheitsanwendungen, KI-Chatbot, diagnostische Mustererkennung, klinische Entscheidungshilfe, empirische Analyse der Implementierung, MDR, KI-Gesetz

INTRODUCTION

Artificial Intelligence in the health sector promises to make healthcare more efficient, to increase the quality of health services and to make results of current research and therapies available to physicians via telemedicine in remote areas (see for example Aung; Wong; Ting 2021 or Suen; Scheinker; Enns 2022).

Because AI-based technologies could in the future replace the work of doctors and other medical personnel and make decisions with far-reaching consequences, the question of trustworthiness is of particular importance. For AI systems in general, the European Commission has recognized this early on and has formulated requirements for "AI made in Europe" with the framework of "Trustworthy AI". The framework was developed in 2019 by the EU's High-Level Expert Group on Artificial Intelligence and has dominated the discourse in Europe since then. The framework has seven dimensions (see figure 1).

Figure 1: The seven dimensions of the EU concept of Trustworthy AI



Source: High-Level Expert Group on Artificial Intelligence 2019

The first of the seven dimensions in the framework is the primacy of human action and human supervision. Here, it is demanded that the AI's decision should not be adopted uncritically by humans, but that humans should regularly check that the AI is still proposing what was initially intended by its inventors. The second dimension is the technical robustness and safety, resp. security. Robustness, on a technical level, refers to whether the AI does what it is supposed to do. And safety and security refer to the request that the system does not cause damage to patients or the users in general and that the system is secure against manipulation from the outside.

The dimension of privacy and data governance addresses the questions how personal data is handled, who has access to it, and what it is used for. These are central issues in healthcare. When it comes to transparency, the most relevant categories are comprehensibility and explainability of AI. In addition, there is also the question of whether users of AI systems are informed that they are dealing with an AI system or not. The next dimension comprises diversity, non-discrimination and fairness. The aim here is to prevent prejudices and discrimination that are in today's data from being reproduced and perpetuated in AI.

The sixth dimension comprises of societal and environmental wellbeing. This dimension refers to sustainability and the resource consumption of AI. It is well known that AI consumes a lot of electricity. But in a broader sense, it is also about the question of what kind of society we want to live in, and the question of what degree of automation or control should be given to an AI. The seventh dimension is accountability of the AI system. Here, the question is dealt with, who is liable if something goes wrong. This is a central question, especially with self-learning AI, i.e. when AI-systems come to conclusions that were not programmed by humans.

The High-Level Expert Group has conceded that the framework is a rather abstract one summarizing very different requirements. Nevertheless, it became the central point of reference in the discussion about ethical AI in the EU.

To put the concept into practice, a series of concretizations has been proposed in recent years. Guidelines and checklists have been worked out (for an overview see Hagendorff 2020 or OECD 2021), traffic light systems have been developed (AI Ethics Impact Group 2020), and consideration has been given as to what degree of trustworthiness an AI has achieved if certain criteria are met (see for example Poretschkin et al. 2021).

However, the attempts at concretization did not lead to the expected implementation successes and to date there is a lack of examples and best practices that could illustrate what good implementation of trustworthy AI could look like. This has led to a number of speculations about the reasons for the implementation deficit (Shneiderman 2020, Kusner et al. 2020, IEEE; ANE, University Copenhagen 2021, Schiff et al. 2021, Beckert 2021). Above all, the lack of depth of concretization of the concept against the background of the heterogeneous application fields has been criticized (see for example Stahl et al. 2021 or Hagendorff 2022).

This paper follows on from these analyses but takes the opposite approach by presenting three implementation examples in the healthcare sector to show concrete implementations of trustworthy AI. It will be shown that not all dimensions of the concept are relevant in the examples and it that some of the formulated requirements are not free from contradictions and thus make a complete implementation of the concept difficult.

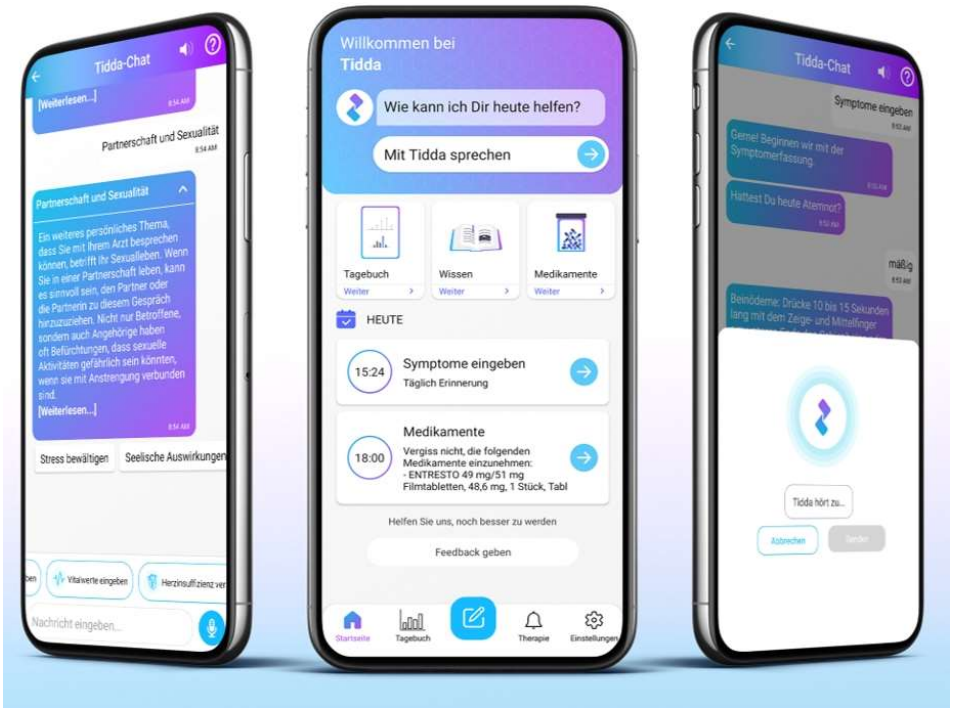
The three examples presented are a chatbot app used as a digital patient companion, a diagnostic pattern recognition system, and an AI-based clinical decision support system. The paper is a result of the symposium "Trustworthy AI in healthcare: Challenges and solutions" held in Karlsruhe, Germany on October 5, 2022.

2 CHATBOT APP AS A DIGITAL PATIENT COMPANION

The first example of an AI project in which the implementation of trustworthiness played a role is the chatbot system developed by German start-up Zana (www.zana.com). The chatbot app can be used either for digitally collecting patient entries in clinical trials or for actively managing their chronic diseases in patients´ everyday life. Here, we focus on a digital instrument for self-management ofchronic disease.

The core of the app is a conversational AI that enables speech and chatbot assistance to leverage communication and interaction between patients and physicians. The app is called Tidda Herz (https://tidda.care/tidda-herz) (see figure 2). Heart failure patients have a reduced life expectancy and depend on the continuous supervision of their condition and on regular examinations. Heart failure disease is currently number one in terms of the costs and medical specialization in the German health system.

Figure 2: The Tidda Herz App as a digital companion for Heart Failure patients



Source: Own screenshot

Since heart failure patients have to monitor their health conditions every day, they usually have a paper book which they fill out at home to track their symptoms. After the disease is diagnosed, patients receive instructions from their physician on what symptoms and conditions to look out for. They are also given information material on the background of the disease and what to do if they have certain symptoms.

Although permanent monitoring of symptoms and a regular review of data is required, usually heart disease patients see their physician only 2-3 times a year. What the Tidda Herz app offers here is to give heart disease patients a companion for their everyday life with which they can track their conditions digitally by speaking to the app and get spoken feed-back based on the data they have provided. Users of the app can track vitals as well as document symptoms as they occur. The app also reminds the users to take their medication on time. And when visiting their physician, data from the Tilda Herz app can be used to monitor the course of the disease over time.

To give the correct answers to questions of patients and to recommend appropriate steps in case of incidents, ZANA has fed the system system with information from several sources. The back-end system uses NLP technologies and methods of Entity Recognition, for which the correctness and reliability of the answers and recommendations had to be ensured. Robustness and repeatability is another focus of the system engineering, given that the end product has to pass an official validation according to Medical Device Regulation (MDR) procedure before being qualified as a reimbursable medical device. Of particular importance during the development were also questions of data protection and data security: To comply with the relevant regulations but also to build trust with patients, personal data and health data are stored in separate databases. Furthermore, the company attaches importance to the statement that personal data will not be passed on to third parties without consent. All data transferred to the platform is encrypted, making it impossible for the communication data to be exploited by external security

threats. According to ZANA, the anonymised data is stored on a server in Germany that meets the highest security standards (https://tidda.care).

For the developers of the Tidda app it was important to find out what information patients were willing to share with the chatbot in the first place. To find out about this, test users were involved in the trial phase and explorative interviews with test users were carried out.

Before being able to offer products like Tidda Herz to patients there is the need that these products comply with the European Medical Device Regulation (MDR). The MDR imposes a long list of requirements, including requirements for clinical investigations on medical devices. Also, it requires surveillance and management of the medical device during the entire life cycle. Most of the dimensions of the concept of trustworthy AI seem to be already covered by these sector-specific regulations.

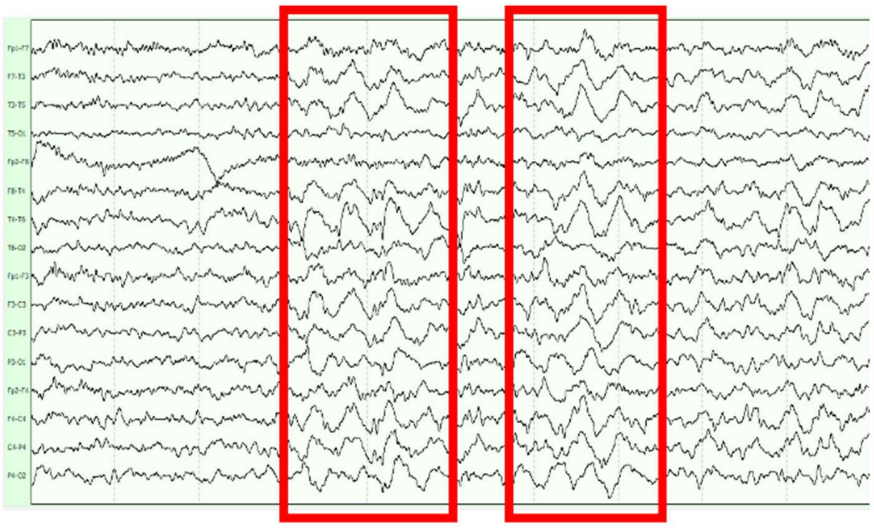
For example the validation process according to the MDR requires external experts and an audit at a notified body. The clinical validation encompasses audits by an ethical committee, audits concerning data protection, data privacy and patient safety. The positive effects of the app for patients have to be proven according to concrete measurables and endpoints.

3 DIAGNOSTIC PATTERN RECOGNITION

The second example is a research project at the University of Freiburg that deals with the automated, AI-based evaluation of EEG data for clinical practice. Electroencephalogram (EEG) as a diagnostic tool is used for various clinical pictures, especially epilepsy sleep medicine and for assessing brain function in comatose patients. Classical surface EEG measured from the scalp represents the bioelectric sum potential of millions of neurons on the order of milliseconds (see fig. 3). Even in such a comparatively short period of time, the bioelectrical activity in the brain can change considerably. Interpretating the resulting

patterns, the “EEG curves”, takes expert training and clinicians, especially neurologistis, often spend a considerable amount of time evaluating the EEG diagrams.

Figure 3: EEG diagrams



Source: Own diagram

Automating this work with the help of AI methods was not possible for a long time because the signals often contain artefacts and no intuitive understanding is possible that would facilitate the interpretation of the deviations. In the meantime, however, possibilities have been found to automate the evaluation to such an extent that an algorithm can distinguish normal courses from deviations (Gemein; Schirrmeister, Chrabąszcz et al. 2020). For this purpose, Convolutional Neural Networks (CNNs) can analyse the EEG data and recognise whether the patterns are inconspicuous or have a relevant anomaly. Accordingly, such an automated EEG analysis tool can issue the labels "healthy" or "pathological". The AI would thus save the doctor the time-consuming work of differentiating between healthy and pathological EEGs.

However, a downside of approaches based on artificial neural networks is of course that the internal learning dynamics are opaque: the well-known “black box problem” of deep neural networks.

Thus, many physicians as intended users for such systems may remain sceptical whether they can trust the results of such a diagnostic tool.

This is where the project "Interpretable Artificial Intelligence Systems for Trustworthy Applications in Medicine" (AI Trust) at the University of Freiburg comes in (https://responsible-ai.org/ai-trust). To understand what exactly the Convolutional Neural Networks do, on which features of the EEG data the system actually "learns", and which weightings it carries out, Invertible Neural Networks are used to identify the relevant features in the training phase of the system. Because even poorly programmed algorithms or incomplete data usually produce outputs (the problem of “garbage in, garbage out”), it is not sufficient to only analyse the input-output relationship; invertible neural networks use bijective functions that can evaluate the learning process in both directions to better understand the internal dynamics of the system. The aim of the procedure was to obtain interpretable results and to understand and ultimately control the training process of the Neural Network. At the output level, this could give the users, i.e. neurologists and other physicians with EEG training, understandable representations about which part of the patterns were recognized by the system.

Heat maps and other forms of visualisation contribute to the understanding of the learning process. In fact, the „AI Trust“-project goes one step further and tries to create an ethics-by-design system using a co-creation method integrating different stakeholders. The aim is to develop a reliable, trustworthy diagnostic system. Findings from human-technology interaction research and expertise from computer science, neurology, psychology and law guides the project in developing a successful prototype that takes the needs of doctors and patients into account.

Although the interpretability and explainability of AI decisions are initially established by technical means, non-technical factors are also important for the practical use of a clinical decision system. These include the question of regulatory and legal requirements, practical applicability and the question

of how the system changes the doctor-patient relationship. Ultimately, it is a “question of trust”. However, the concept of „trust“ being used here is not the same as when humans trust each other. In interactions with technology, and especially with AI-based systems, it is rather „reliability“ that developers should strive for and ensure.

4 CLINICAL DECISION SUPPORT SYSTEM

The third example is a clinical decision support system for use in consultations. It was developed by seven Fraunhofer institutes as a prototype in the MED2ICIN project (https://websites.fraunhofer.de/med2icin). The basis for the active support of the doctor is a digital patient model in which various data from different sources are linked. The sources are genome or RNA data or biomarkers of the patient, MRI or CT slice images, and clinical data such as blood values or lifestyle data.

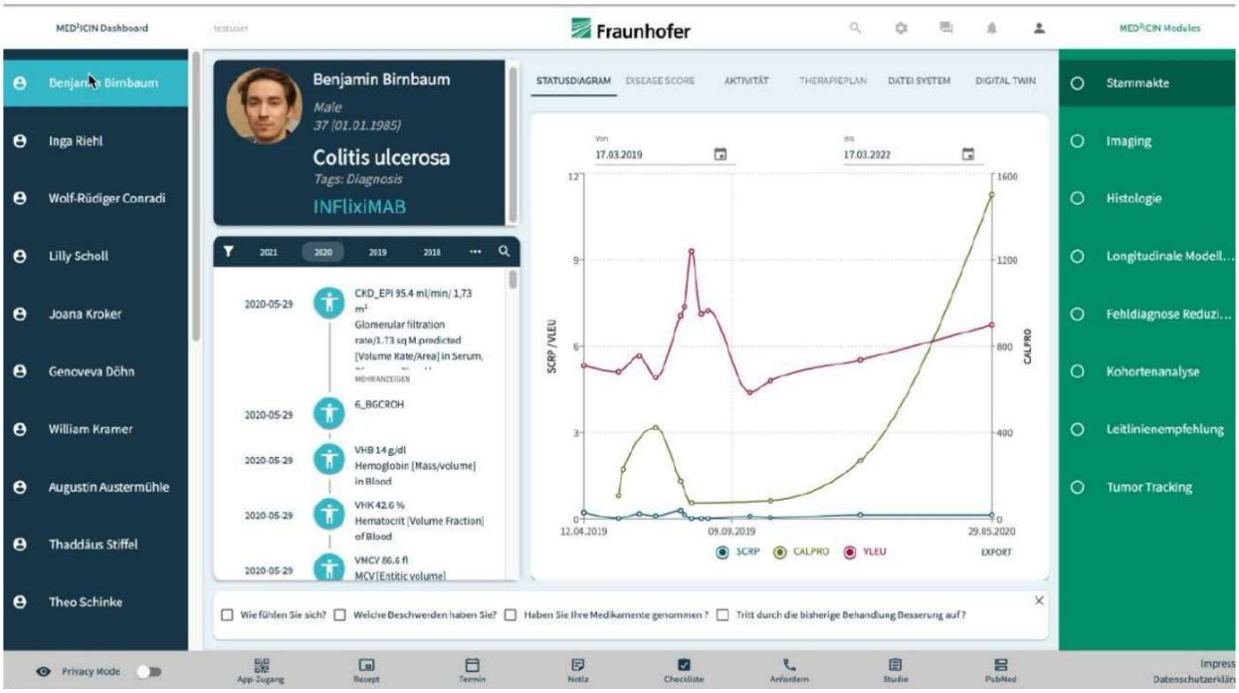
In addition to patients´ data, the system also contains research information on disease progression, cohort behaviour and medication administration as well as information from medical guidelines, how to treat the respective diseases. The AI-based system compares the patient's individual data with the stored expert data and suggests a specific therapy or course of action to the doctor.

An important aspect in the development of the system was to maintain the decision-making autonomy of the doctor. The system should not be perceived as an automatic decision-making system whose decisions are adopted 1:1. On the contrary, the system is programmed in such a way that there is an interplay between the AI and the human expert. The medical professional retains the final decision on which therapy to prescribe.

To establish this interactivity between the AI and humans, an interactive user interface was developed (figure 4). The Med2icin dashboard provides the doctor with all relevant data processed for the

individual treatment of a patient. From the basic menu, the doctor can successively access further information levels that are of interest for the corresponding question and then call up the corresponding analyses.

Figure 4: Interactive User Interface for the clinical decision support system in the projekt „Med2icin“



Source: Own screenshot

The AI can show whether there are indications that the patient could be treated within his or her cohort. To do this, the algorithm compares the phenotypic characteristics of this patient with those of other patients in the same cohort and suggests the medication that has proven to be the most successful treatment option in this group.

The medical information stored in the system can also support the doctor during the initial diagnosis. If a certain diagnosis is entered into the interface, the system points out aspects that may be important in the context of this diagnosis, e.g. that the diagnosed disease often occurs together with another disease, which must then be checked accordingly. In this case, the AI assists with the diagnosis and points

out aspects that are known from the literature but that the attending physician may not be aware of because he or she cannot have read all the specialist articles on the subject.

For the developers of the "Med2icin"-system, data protection and the adaptability of the system to future needs were important in addition to preserving human decision-making autonomy. Therefore, a modular software architecture was used in which the specific modules only have access to the data they really need. This means that the individual function modules and the data are stored separately from each other. The function modules can also be developed separately, i.e. if new modules are created, they can easily be fitted into the system.

The "Med2icin" system, and in particular the interactive dashboard, was developed in close consultation with potential end users and were iteratively adapted to their needs. Special human-centred design methods were used for this purpose.

According to Fraunhofer, the system can be used especially in large consultations or where telemedicine is relied on, i.e. in regions where there is no specialised care but where specialised therapeutic decisions nevertheless have to be made (Fraunhofer 2022).

5 SUMMARY AND OUTLOOK

The analysis of the three sample implementations has shown that different dimensions of the concept of trustworthy AI are of relevance in each case. The technical robustness of the additional benefit of the AI-supported system compared to other solutions were important dimensions in all three cases. Apart from these similarities, different dimensions came to the fore in the three cases: Privacy and data protection in the chatbot system, transparency and explainability in the diagnostic system, and human supervision in the recommendation system for therapies.

These examples show, that many implementations can deliver on several of the dimensions highlighted in the EU's trustworthy AI framework. On the other hand, the search for examples in which all seven dimensions of the framework are equally tested and implemented in an exemplary manner will often be disappointing. Model implementations that fully satisfy all dimensions may be possible in research projects for demonstration purposes. In practice and under market conditions, however, this is often not possible, and maybe it is also not necessary. This is because, as shown, the different applications each require their own focus.

The desire for AI to be perfect and trustworthy and to fulfil all criteria reflects a normative top-down impetus and shows that society as a whole has become more sensitive and that better systems are required in the future. However, whether this is feasible to implement systems satisfying all trustworthy dimensions in a scalable, yet context-sensitive manner, remains an important question. Perhaps, against this background, one should speak of "acceptable" rather than „trustworthy" AI. „Acceptable AI" also includes testing, validation, certification, etc. These are complicated and difficult processes to implement, but as users we assume that AI systems have been tested, just as we assume that our drinking water is clean and not contaminated.

Analysing implementations in health care, the concept of trustworthy AI also reveals inconsistencies. For example concerning the dimension of human oversight: The concept of trustworthy AI foresees that humans make the final decision, not the AI. But not in all cases, this is a viable option. There are also AI based medical devices where reducing human variability and fallibility is important. There, humans are intentionally taken out of the loop to minimize errors. The other example is fairness: The definition of fairness is tricky because the definition of what is fair depends on subjective factors, on time, on region, etc. The pandemic has shown that what is considered as fair during an acute emergency may not be considered fair during other times.

Against this background, the original concept of trustworthy AI changes its character from a quasi-binding list of dimensions to be worked through in a consistent manner to a reference model that can help to point to the relevant dimensions for the respective applications. For developers, companies and users, the concept of trustworthy AI remains important because it points to the critical aspects of an AI app.

Following the current AI debate, it can be observed that the exploratory phase of dealing with the concept of trustworthy AI is replaced by a discussion of the specifications of the planned AI Act of the EU (see Burri 2022). The draft AI Act provides for different AI risk classes for which different regulatory requirements are to apply. Therefore, researchers and companies are now asking themselves which concrete specifications their AI systems must fulfil, how they could carry out the self-certifications, and what consequences the AI Act will have for the further development of their AI activities.

The draft AI Act of the EU has also caused some uncertainty among manufacturers of AI-based medical devices. As the example of the development of a heart disease monitoring app in our analysis has shown, there is already a dense regulatory web in the health sector that governs the approval of such new technologies. It seems that the sector-specific regulation for health care already covers many aspects of the proposed AI Act (Benjamens; Dhunnoo; Mesko 2020). We suggest that in the future, aspects such as acceptability, appropriate use, and added value of AI systems will become the dominant aspects in healthcare.

6 REFERENCES

AI Ethics Impact Group (2020). From Principle to Practice. An interdisciplinary framework to operationalise AI ethics. VDE & Bertelsmann Stiftung. Online: www.ai-ethics-impact.org/en

Aung, Y.M., Wong, D. , Ting, D. (2021). The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. British Medical Bulletin, 2021, 139: 4–15. Online: https://doi.org/10.1093/bmb/ldab016.

Beckert, B. (2021). The European way of doing Artificial Intelligence: The state of play implementing Trustworthy AI. Paper presented at FITCE/IEEE, September 29-30, Vienna. Online: https://ieeexplore.ieee.org/document/9588560.

Benjamens, S., Dhunnoo, P., Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. npj Digital Medicine (2020) 3:118. Online: https://doi.org/10.1038/s41746-020-00324-0.

Burri, T. (2022). The New Regulation of the European Union on Artificial Intelligence. Fuzzy Ethics Diffuse into Domestic Lay and Sideline International Law. In: Vöneky, S, Kellmeyer, P., Müller, O. (eds.): Cambridge Handbook of Responsible Artificial Intelligence. Cambridge University Press. p. 104-122.

Fraunhofer (2022): Sprechstunde mit KI-Assistenz. 16 May 2022, Fraunhofer IGD. Online: https://websites.fraunhofer.de/med2icin/sprechstunde-mit-ki-assistenz/

Gemein, L., Schirrmeister, R., Chrabąszcz, P. et al. (2020). Machine-learning-based diagnostics of EEG pathology. NeuroImage 220 (2020) 117021.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. Minds & Machines 30, p. 99–120.

Hagendorff, T. (2022). A Virtue-Based Framework to support Putting AI Ethics into Practice. Philosophy & Technology 25:55, June 21

High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines for Trustworthy AI. Brussels: European Commission. Online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Institute of Electrical and Electronics Engineers (IEEE), Association of Nordic Engineers (ANE), Department of Computer Science at the University of Copenhagen, and Data Ethics ThinkDoTank (2021). Addressing Ethical Dilemmas in AI: Listening to Engineers. January. Online: https://dataethics.eu/listen-to-the-engineers-the-algorithm-does/

Kusner, M. J. and Loftus, J. R. (2020). The long road to fairer algorithms. Nature, Vol. 578, 6 February, p. 34ff. Online: https://media.nature.com/original/magazine-assets/d41586-020-00 274-3/d41586-020-00274-3.pdf.

OECD (2021). Tools for trustworthy AI. A framework to compare implementation tools for trustworthy AI systems. OECD Digital Economy Papers, No. 312, OECD Publishing: Paris. Online: https://doi.org/10.1787/008232ec-en.

Poretschkin, A., Schmitz, A., Akila, M. et al. (2021). KI-Prüfkatalog. Leitfaden zur Gestaltung vertrauenswürdiger Künstlicher Intelligenz. Sankt Augustin: Fraunhofer IAIS. Online: www.iais.fraunhofer.de/de/forschung/kuenstliche-intelligenz/ki-pruefkatalog.html#KI-Pruefkatalog-kostenfrei-erhalten

Schiff, D., Rakova, B., Ayesh, A. et al. (2021). Explaining the Principles to Practice Gap in AI. IEEE Technology & Society Magazine, vol 40, Issue 2, June.

Shneiderman, B. (2020). Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. ACM Transactions on Interactive Intelligent Systems, Vol. 10, No. 4, October, Article 26.

Stahl, B. C., Antoniou, J. Ryan, M. et al. (2021). Organisational responses to ethical issues of artificial intelligence. AI & Society (2022) 37:23-37.

Suen, S., Scheinker, D., Enns, E. (eds.) (2022). Artificial Intelligence for Healthcare: Interdisciplinary Partnerships for Analytics-driven Improvements in a Post-COVID World. Cambridge: Cambridge University Press.

VDE; Confiance.ai (2022): Franco-German alliance develops label for trustworthy artificial intelligence (AI). VDI press release Oct 10, 2022, Frankfurt/ M.: VDE. Online: www.vde.com/en/press/press-releases/deutsch-franzoesisches-ki-label.