Sub-project description

dfki Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

Mission AI

Innovation and Quality Center (IQZ) Kaiserslautern

TrustifAI - Health and well-being

Version:	12.04.2024
Term:	01 05.2024 - 31.12.2025
Execution:	German Research Center for Artificial Intelligence GmbH Trippstadter Str. 122 67663 Kaiserslautern

The proliferation of AI has created a growing awareness of the associated social, ethical and technical challenges and has coined the term trustworthy AI, i.e. AI that meets and responds to these challenges. The challenges increase when the applications are in high-risk areas, such as healthcare. One of the responses to this is the movement towards trustworthy AI, which is being driven by academic research institutions as well as industry players, governments and international organizations.

The **Innovation and Quality Center (IQZ) Kaiserslautern** wants to help improve the trustworthiness of AI solutions in the context of AI applications in the areas of health and well-being at different stages of the development cycle with a series of concrete solutions. The proposal focuses on two complementary platforms and **five representative use cases in medicine**. The **quality platform** for the development of trustworthy AI applications will enable users to create efficient and effective data science analysis pipelines through a human-in-the-loop approach, in an interplay of artificial and human intelligence, with the aim of increasing trustworthiness. Based on the analysis of regulations, testing frameworks and testing tools specifically relevant for medical applications, we want to develop a prototype for a **testing platform** for AI applications in healthcare. Both the quality platform and the testing platform contribute to the overall objectives of Mission AI and are thus integral components of the testing approaches envisaged in this context and of an AI quality label based on them. The two platforms are to be understood as a contribution to the overarching Mission AI platform.

The IQZ also aims to set up and operate the **Mission KI contact point** in Kaiserslautern and provide corresponding community services.

Table of contents

1 Goals

1.1 Overall objective of the project

Trustworthy AI systems

The spread of AI has created a growing awareness of the associated social, ethical and technical challenges and has coined the term trustworthy AI, i.e. AI that meets these challenges. The movement towards trustworthy AI has been driven by academic research institutions as well as industrial companies, governments and international organizations. In recent years, numerous initiatives have been launched to structure and develop the requirements for trustworthy AI systems. For example, the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) has identified seven key requirement dimensions for trustworthy AI:

1. **Human action and supervision:** AI systems should complement human action without undermining human autonomy.
2. **Technical robustness and security:** AI systems should be secure, reliable and robust against manipulation.
3. **Data protection and data governance:** protecting privacy and ensuring appropriate data management.
4. **Transparency:** Promoting the traceability of AI systems through clear communication about their capabilities and purpose.
5. **Diversity, non-discrimination and fairness:** avoiding injustice and discrimination through AI systems.
6. **Social and environmental wellbeing:** Consideration of the social and environmental impact of AI systems.
7. **Accountability:** Establish mechanisms for responsibility and accountability in the development and implementation of AI systems.

The requirements for trustworthy AI described above are deliberately designed to be cross-industry and not limited to specific sectors. This universal approach is intended to ensure that the principles for trustworthy AI can be used in all application areas and in all industries. While this is not the only definition of trustworthy AI and other aspects of Mission AI are attempting to reach a common understanding of what trustworthy AI means for AI "Made in Germany", it is a good working definition that encompasses important socio-technical goals.

Advantages of trustworthy AI systems

Working on the trustworthiness of AI systems, especially in sensitive, high-risk applications such as healthcare, is extremely important and beneficial for several reasons:

- **Trustworthy AI is more ethical.** The aspects of trustworthiness concern the key ethical principles that ensure that AI is used for the benefit of all people and the environment. Diversity, fairness and non-discrimination ensure that the systems are protected or

not harm underrepresented groups. Social and environmental wellbeing should ensure that the systems are used and developed for the benefit and not for harm.

- **Trustworthy AI is technically superior.** Ensuring technical robustness and security leads to better systems that are designed to perform better in the long term and minimize the risk of accidental or intentional breaches. New mechanisms and paradigms for data protection and data management are being deployed to harness the potential of big data while ensuring the protection of data and privacy of individuals and organizations.
- **Trusted AI improves compliance with the legal framework.** The concept of trustworthy AI is included in all major current and future legislation (e.g. GDPR, DSA, AI Act). Especially for high-risk or sensitive applications, aspects of trustworthy AI are strict requirements - human oversight, transparency, accountability.
- **Trustworthy AI favors adaptation in practice.** Many organizations are reluctant to implement systems that may have unintended consequences - which trustworthy AI is designed to avoid as much as possible. Trustworthy systems enjoy a higher level of trust among decision-makers and are more likely to be used in practice. A survey conducted by the market research platform "Appinio" in 2023, in which 1,000 Germans were asked about "Health and AI", revealed that only between 27% and 34% of respondents fully trust the results of current AI systems for various application areas of AI in healthcare (Fig. 1). Low trust inhibits the willingness to adopt new and better technologies and benefit from their advantages.

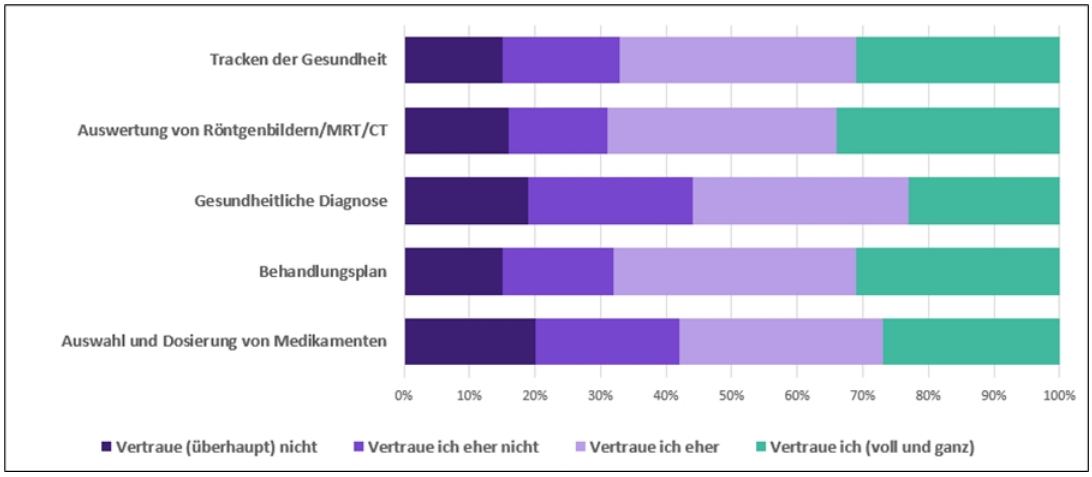


Figure 1: Germans' trust in health-related AI systems (source: https://www.appinio.com

Example of trustworthy AI in the healthcare sector

To illustrate how an AI system in healthcare can meet the exemplary requirements and principles of trustworthy AI, we present a socio-technical measure that considers each principle for an AI-driven diagnostic system:

1. **Human intervention and monitoring**: It must be ensured that the diagnostic system serves as a support for healthcare professionals without replacing or impairing their decision-making ability.

2. **Technical robustness and safety: It** must be ensured that the diagnostic system makes precise and reliable diagnoses and is protected against manipulation and malfunctions.
3. **Data protection and data governance**: Patient data used to train AI procedures must be protected from unauthorized access, and the secure handling of patient data must be ensured throughout the entire diagnostic process.
4. **Transparency**: The diagnostic system must provide clear information on how it arrives at its results, including the data and algorithms used.
5. **Diversity, non-discrimination and fairness**: It must be ensured that the diagnostic system functions without prejudice and is fair to all patient groups, regardless of gender, age, ethnic origin or other demographic factors.
6. **Social and environmental wellbeing**: The diagnostic system must have a positive impact on healthcare.
   Accountability: It must be clear who is liable for the diagnoses made with the help of the AI system.

The establishment of an Innovation and Quality **Center (IQZ) in Kaiserslautern** is intended on the one hand to serve local companies as a community space and contact point for questions relating to the topic of trustworthy AI, and on the other hand, as a competence center, the IQZ is intended to make a real contribution to improving trustworthy AI solutions in various phases of the development cycle in connection with applications in the areas of health and well-being. The sub-project within the Mission AI Innovation and Quality Center (IQZ) is managed internally at DFKI under the project name **TrustifAI - Health and Wellbeing.**

Aim of the TrustifAI project

The goal of TrustifAI is to contribute to trustworthiness for AI applications, especially in the healthcare sector. Our approach aims to provide viable solutions for the development of trustworthy AI systems in healthcare, which can also be transferred to other industries and application areas. To ensure that the solutions are actually applicable and useful, they are motivated and executed using five healthcare use cases that cover a wide range of data modalities, applications and challenges. An important aspect of developing such solutions is engaging end users and ensuring that the AI system fits into the regulatory and organizational framework while meeting the specific needs of healthcare professionals, educators and patients.

Embedded in the overall context of Mission AI, TrustifAI's approach is to develop two platforms that specifically address the principles of trustworthiness throughout the development cycle and testing of AI systems for use in healthcare. The two platforms are complementary to each other and to the testing approaches and minimum standards developed by the other Mission AI project leaders and are intended for integration into the overall Mission AI platform.

The **quality platform** we propose can be used in the development of trustworthy AI applications. It will enable users to create efficient and effective AI and data science products through a human-in-the-loop approach. The primary goal is to increase the trustworthiness of the system to be developed.

The **testing platform** we propose can be used to test the trustworthiness of AI systems before or after their deployment. Based on the analysis of regulations, testing frameworks and testing tools specifically relevant to medical applications, we aim to develop a prototype testing platform for healthcare AI applications that addresses the problem of measuring key quantifiable indicators of trustworthiness.

Use cases of the TrustifAI project

To ensure that the practical, organizational, regulatory and human challenges are taken into account in the development of the overall platform being created as part of Mission AI, we are developing all platform components based on **five use cases from the healthcare sector**. These use cases have been selected to best reflect the variety of data modalities and problems that can occur in different aspects of the healthcare sector. While three use cases involve the development of prototypes in conjunction with the quality platform, the other two use cases are based on existing AI systems that are merely to be tested and improved. These three use cases, designed from the ground up, also involve partners from the healthcare sector - either primary care providers, emergency care providers or educators. Involving real-world partners will improve AI development through real-world insights and help to effectively incorporate the resulting complex healthcare challenges into the development of the overall Mission AI platform. Each of the five use cases will be finally tested against the overall Mission AI platform and will have one or more specific focus areas that reflect the principles of trustworthy AI and are based on specific needs and requirements.

A brief description of the use cases follows, along with specific focus areas of trustworthiness and partners (if applicable):

1. **Skin cancer detection with dermatoscopic images** - As part of various projects, DFKI has developed systems that can solve various problems in **medical image analysis.** This includes the detection of malignant skin lesions. This use case focuses on testing and improving the trustworthiness of image-based AI systems (i.e. "computer vision") and takes particular account of the principles of "transparency" and "diversity, non-discrimination and fairness".
2. **Decision support for tumor treatments** - The system is being developed by DFKI as part of the KITTU project (https://kittu.org/) together with Unimedizin Mainz and Innoplexus AG. This use case concerns **tabular data** (with additional natural language processing) and focuses on the principles of "human agency and oversight" and "accountability".
3. **Demand forecasting and planning for emergency services** - Following a successful pilot project on a related topic, we will work with the German Red Cross Rhineland-Palatinate to develop a system that uses multimodal data, but especially **spatio-temporal data,** to forecast the demand for emergency services and provide resource planning recommendations to the dispatcher in an emergency call center, with the ultimate goal of reducing waiting times for emergency services in emergencies. The project will address a variety of trustworthy AI principles, with a focus on "human action and control", "transparency" and "societal and environmental wellbeing".
4. **Combining text and survey data to improve psychiatric treatment and education**. To develop a system with complex **text and language modeling** challenges, we will work with the Psychiatric

outpatient clinic at the University of Trier to develop systems that help trainee psychiatrists to develop better treatment plans for their psychiatric patients. The principle of "data protection and data governance" is at the heart of this use case.

5. **Using endoscopy video data for safer intubation**. We will work with secondary and tertiary care providers from Saarland University Hospital to develop a system that uses medical **video data to** identify potential airway problems during preoperative intubation. The focus here is on the principles of "social and ecological well-being" and "diversity, non-discrimination and fairness".

1.2 Reference to funding policy objectives

"Mission AI - National Initiative for Artificial Intelligence and Data Economy" (https://mission-ki.de/) is a joint project of acatech - National Academy of Science and Engineering and the Federal Ministry for Digital and Economic Affairs (BMDV). It aims to strengthen the digital competitiveness of the German economy by promoting AI innovations and developing trustworthy, marketable AI applications. The project focuses on the three pillars "Networking data spaces across sectors", "Creating transparent AI quality and testing standards" and "Supporting the growth of AI innovations".

The second pillar of Mission AI, "Creating transparent AI quality and testing standards", focuses on the development and establishment of transparent AI quality and testing standards. The aim is to strengthen trust in AI technologies and thus promote their acceptance. By defining clear criteria and methods for evaluating AI systems, the aim is to create a reliable framework for the development of trustworthy AI solutions.

In this context, the "TrustifAI - Health" project is a critical research contribution that focuses on the healthcare sector. Through the investigation of representative use cases, specific challenges in the area of testing and quality improvement of AI systems are addressed in order to develop innovative solutions. This approach will not only be used in the healthcare sector, but will also serve as a guideline for the development of trustworthy AI applications in other areas with high risk potential.

1.3 Scientific and technical work objectives

The EU-AI Act promotes the adoption of trustworthy AI and provides high-level guidance on the key requirements that AI applications must meet in order to be considered trustworthy. The implementation and systematic application of these guidelines and the creation of corresponding standards and certification programs undoubtedly represent a complex and multi-layered challenge. The conceptual scientific challenge is to close the gap between high-level guidelines and concrete, implementable standards that go beyond raising awareness or working through a checklist. A practical implementation of these standards is particularly relevant in the area of healthcare and well-being, where it is of utmost importance to ensure the trustworthiness of AI systems that often have a direct impact on the lives of individuals or entire societies. Therefore, our goal is to support the process of achieving and verifying trustworthiness in the context of medical AI applications. The development of an AI quality platform is intended to support the development process of medical AI and facilitate the achievement of properties relevant to trustworthiness. In parallel, a testing platform will be developed to verify these properties in the context of the intended medical use case.

Both platforms are being developed with a focus on medical applications and therefore act as complementary components within the overall Mission AI platform to the general testing approaches and minimum standards developed by the other Mission AI project managers.

The AI quality platform focuses on two conceptual challenges: How to define the requirements for trustworthy AI in a way that is both intuitive for humans and readable for machines, and how to ensure that the development process of AI applications fulfills the identified requirements with minimal effort for practitioners and stakeholders. To overcome the first challenge propose we a centralized knowledge base/database/store/ontology to provide (a) known dimensions of trustworthy AI (e.g., fairness, robustness, explainability), (b) possible definitions (e.g., fairness as equality of opportunity, equity, or demographic parity), (c) associated quantifiable metrics and procedures for evaluating the performance of a given AI application in accordance with the requirements and measures (e.g., statistical parity difference, average parity difference, statistical parity difference, average parity difference, etc.), and (d) a centralized knowledge base/database/storage/ontology to provide (e.g., a) known dimensions of trustworthy AI (e.g., fairness, robustness, explainability).(e.g., statistical parity difference, average odds difference), (d) identified weaknesses that existing data pre-processing and modeling techniques might have (e.g., random sampling might bias the training data), and (e) known mitigation strategies (e.g., using stratified sampling instead of random sampling). To address the second challenge, we build on the knowledge curated in step (1) and propose a prototype debugging utility that enables practitioners to either rely on best practices to identify domain-specific (i.e., niche) requirements for trustworthy AI or automatically identify vulnerabilities in the model and, if possible, suggest strategies to address them. In contrast to the AI testing platform, which verifies an already created AI model, the debugging utility implements a human-in-the-loop approach to assisted AI development.

The testing platform focuses on the evaluation of a variety of medical AI systems in terms of their trustworthiness. Therefore, the focus is primarily on the conceptual challenge of translating abstract definitions of trustworthiness into actionable measures that allow to quantify or assess the degree of trustworthiness of a system. Due to the strong heterogeneity of data modalities, AI methods and target variables in healthcare, the definition of trustworthiness criteria requires a high degree of abstraction. This and the high complexity of these criteria make it impossible to translate them directly into applicable testing approaches or procedures. One of the objectives is therefore to develop clearly defined yet universally applicable test conditions for each dimension of trustworthiness. These test conditions should serve as a blueprint for the application of specific test tools in each use case. As part of a gap analysis, all existing testing platforms and testing tools from the literature should be identified, recorded and systematically categorized according to a structured approach developed in this thesis. Missing testing tools in the landscape of medical AI systems are to be identified and their requirements detailed in a requirements analysis. The compatibility and integrability of this expandable and operationalizable medical testing platform with the overall Mission AI platform is the aim of the project. Finally, medical use cases will be investigated on the basis of this overall platform. In order to demonstrate the broad applicability of the platform, different modalities (image, natural language, tabular data), AI methods (convolutional neural networks, transformers, language models, etc.) and target variables (diagnosis, therapy decision) are considered in the context of two specific use cases (skin cancer detection and urological tumor board decisions). Concrete, feasible measures of the trustworthiness requirements and evaluation procedures are to be implemented within the use cases in prototypical test procedures. The experiences and findings from the prototype tests are to be fed back in order to improve the trustworthiness of the AI systems, the

developed platforms themselves and to make recommendations for the development of a final overall platform.

1.4 Indication of relevant indicators for the achievement of objectives in the network

Under the responsibility of the coordinator (see overall project description).

2 State of the art in science and technology

2.1 Other R&D approaches and delimitations

2.1.1 Test platform

The prototype testing platform for trustworthy medical AI developed as part of the sub-project will, on the one hand, build on the minimum standards and testing approaches developed by the others responsible for the Mission AI project and, on the other hand, also examine their compatibility with current regulatory and normative requirements and existing work in this area. The complete recording of the state of the art in science and technology is part of the work packages. The definition and basic structure of the various dimensions of trustworthiness should be compatible with the current status of general, existing work relating to trustworthy AI, such as the *Ethics Guidelines for Trustworthy AI* of the EU High Level Expert Group, the EU AI Act, the GDPR, or the DIN *AI Standardization Roadmap* and similar efforts. The current status and analysis of existing public and commercial test platforms (e.g. https://www.holisticai.com/) and similar test catalogs (e.g. https://www.iais.fraunhofer.de/content/dam/iais/fb/Kuenstliche_intelligenz/ki-pruefkatalog/202107_KI-Pruefkatalog.pdf) should also be taken into account.

In contrast to the work of the others responsible for the Mission KI project, the focus should be on efforts relevant to medical applications (e.g. medical device, CE, EMA, etc.) when considering previous efforts, but in particular regulations and standards. The prototypical testing platform designed on this basis should differ from the previous, theoretical and broadly formulated work in that there is a strong focus on the operationalizability of trustworthiness using the testing tools. On the other hand, the strong reference to the challenges of AI in medicine is also a core aspect of the project.

The detailed development of the state of the art in science and technology with regard to the testing tools that can be used is also an essential part of the project. Currently, there is neither a uniform definition of testing tools suitable for testing, nor are there complete, expandable catalogs that can record and categorize possible tools. For this reason, testing tools appear in the current research literature in various forms, for example as documentation tools, synthetic test data sets, model attacks, tests for the quantification of trustworthiness criteria or as plausibility checks. Examples of documentation tools are among others *explainability fact Sheets* (https://doi.org/10.1145/3351095.3372870) or *datasheets for datasets* (https://doi.org/10.1145/3458723), which facilitate the structured recording of explanatory methods used and datasets within of an AI system enable. *SCDB* (https://github.com/adriano-lucieri/SCDB) and *Causal Conversations* (https://doi.org/10.1109/TBIOM.2021.3132237) are examples of synthetic data sets that can be used for targeted testing of the trustworthiness criteria of an AI system. The use of model attacks serves to quantify vulnerability and is made possible, for example, by *Membership Inference Attacks* (https://doi.org/10.1109/SP.2017.41) or the ProPILE framework (https://proceedings.neurips.cc/paper_files/paper/2023/file/420678bb4c8251ab30e765bc27c3b047-Paper-Conference.pdf). The *System Causability Scale* (https://doi.org/10.1007/s13218- 020-00636-z) offers an initial approach to quantifying the explainability of an AI system, whereas the *Differential Privacy* (https://doi.org/10.1145/2976749.2978318)

Framework aspects of data protection. Current methods for the plausibility check of AI decision statements are for example the *ROAR* (https://proceedings.neurips.cc/paper_files/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf) framework and the *Model & Data Randomization* (https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf) technique. Through the structured collection and cataloging of testing tools using the testing platform to be designed and the new testing tools developed on the basis of the use cases, the project described here contributes directly to simplifying the testing landscape of trustworthy AI in medicine.

2.1.2 Quality platform

In designing the AI quality platform, we draw on four different areas of related work: (1) initiatives for quality managementfrom other research areas that are both directly relevant and tangential to AI, e.g. data quality management [http://dx.doi.org/10.1145/3097983.3098021, https://doi.org/10.1145/3299869.3320210], (2) abstractions and data structures for the

knowledge representation [Patterson et al., 2018, https://doi.org/10.1145/3318464.3386146] and (3) overviews of real-world AI/ML management challenges, particularly in healthcare applications. Existing R&D approaches in each area are outlined below.

(1) Data quality - analogous to the desired AI quality - is a complex and multi-layered concept that encompasses dimensions such as data accuracy, completeness, consistency and timeliness. For each dimension, there is a series of abstract metrics. Each abstract measure in turn has a set of quantifiable metrics (analogous to e.g. accuracy, recall and ROC AUC score as specific metrics for the predictive performance of ML models). For a given dataset or application domain, each metric can be further specialized by a (static or dynamically evolving) threshold to define acceptable data quality. Examples of data quality management solutions include Tensorflow Data Validation which tests data against user-defined data schemas, Data linter [http://dx.doi.org/10.1145/3097983.3098021] which validates data against data lints - deviations from accepted practices of data analysis (analogous to code lints - deviations from best practices in software development), and the Deequ library for automating data quality checking at scale [, https://doi.org/10.1145/3299869.3320210] which proposes unit tests for data - a declarative specification of integration constraints. Schelter et al. also present functions for automated constraint suggestions based on data profiles (collected descriptive statistics on data attributes). However, this method requires the presence of reference data - a sample of the data population that is of acceptable quality - and is designed to generate suggestions that are validated by a domain expert. In Redyuk et al. (2021), one of our researchers proposed an approach for semi-automatic data quality assessment specifically tailored to dynamically changing data.

Even the best AI can only generate a limited amount of benefit from poor data. On the one hand, data quality specification as well as the detection of poor quality is therefore also part of AI quality and must be taken into account. On the other hand, concepts can also be extended to other aspects of AI.

(2) We look at existing solutions in the area of knowledge representation and management, both generic and ML/AI-specific. This is necessary to support auditors and developers of AI. A common justification is

1. "similar" AI was used "successfully" in a similar use case.

2. Individual partial decisions - e.g. the preprocessing of data - were based on successful preliminary work.

Making or reviewing such and similar decisions should not be based solely on the knowledge of individuals. This is an important aspect of the quality platform. Technical aspects that can be used for this and preliminary work are discussed in the next section.

The ontology language OWL [https://www.w3.org/2012/pdf/REC-owl2-primer-20121211.pdf] is a common tool in the semantic web for specifying entity relationships. The language formalism enables automated inferences that are used, for example, to validate ontological consistency or to derive new factual information. As an example of data science knowledge management, the ML Bazaar project [https://doi.org/10.1145/3318464.3386146] enables the treatment of various data science operators - preprocessing and modeling techniques - as "building block" primitives for the composition of complex data science workflows. Since the representation of operators includes the specification of interfaces for customizing a predictor or performing data transformations, this tool automatically handles the "glue code" between primitives originating from different libraries and frameworks. The IBM Data Science Ontology project [https://www.datascienceontology.org/] aims to catalog data science concepts and semantically annotate different frameworks and libraries of the Data Science Toolbox. However, maintaining the knowledge bases of these projects is associated with a considerable learning effort and manual overhead for the end user. (3) Challenges in the real world of ML/AI management, especially in healthcare applications.

It's often easy to publish a proof of concept for an AI application scientifically or in the news, but it's much harder to manage it in the real world, as for example for example in [https://ieeexplore.ieee.org/document/9825772, https://pubmed.ncbi.nlm.nih.gov/35873347/, https://ieeexplore.ieee.org/document/9920140, https://link.springer.com/chapter/10.1007/978-3-030-65854-0_4, https://dl.acm.org/doi/10.1145/3035918.3054782] is shown is shown. Also here see connections to the preliminary work of those responsible for Mission AI, such as their involvement in the AI4C standards. The user needs more support with the management and decision points and their documentation.

2.2 Previous work of the applicant

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ... & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. bmj, 368.

Mateen, B. A., Liley, J., Denniston, A. K., Holmes, C. C., & **Vollmer**, S. J. (2020). Improving the quality of machine learning in health applications and clinical research. Nature Machine Intelligence, 2(10), 554-556.

Sounderajah, V., Ashrafian, H., Golub, R. M., Shetty, S., De Fauw, J., Hooft, L., ... **Vollmer**, S., ... & Liu, X. (2021). Developing a reporting guideline for artificial intelligence-centered diagnostic test accuracy studies: the STARD-AI protocol. BMJ open, 11(6), e047709.

Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., Denniston, A. K., Ashrafian, H., ... **Vollmer**, S., ... & Yau, C. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The Lancet Digital Health, 2(10), e537-e548.

Dennis, J. M., McGovern, A. P., **Vollmer**, S. J., & Mateen, B. A. (2021). Improving survival of critical care patients with coronavirus disease 2019 in England: a national cohort study, March to June 2020. Critical care medicine, 49(2), 209.

Inkster, B., O'Brien, R., Selby, E., Joshi, S., Subramanian, V., Kadaba, M., Schroeder, K., Godson, S., Comley, K., **Vollmer**, S.J. and Mateen, B.A., 2020. Digital health management during and beyond the COVID-19 pandemic: opportunities, barriers, and recommendations. JMIR mental health, 7(7), p.e19246.

Sharma, R., **Redyuk**, S., Mukherjee, S., **Sipka**, A., **Vollmer**, S., & Selby, D. (2024). X Hacking: The Threat of Misguided AutoML. arXiv preprint, in peer-review

de Bie, K., Kishore, N., Rentsch, A., Rosado, P., & **Sipka**, A. (2020). Using AI to help healthcare professionals stay up-to-date with medical research. In AI for Social Good Workshop.

Redyuk, S., Schelter, S., Rukat, T., Markl, V., & Biessmann, F. (2019, July). Learning to validate the predictions of black box machine learning models on unseen data. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics (pp. 1-4).

Redyuk, S., Kaoudi, Z., Markl, V., & Schelter, S. (2021, March). Automating Data Quality Validation for Dynamic Data Ingestion. In EDBT (pp. 61-72).

Palacio, S., **Lucieri**, A., Munir, M., Ahmed, S., Hees, J., & **Dengel**, A. (2021). Xai handbook: towards a unified framework for explainable AI. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3766-3775).

DIN Spec - DIN (2023). Artificial intelligence - Life cycle processes and quality requirements - Part 3: Explainability; DIN SPEC 92001-3:2023-08

Lucieri, A., Bajwa, M. N., **Dengel**, A., & Ahmed, S. (2022). Explainable AI in medical diagnosis-successes and challenges. In Artificial intelligence in healthcare: Developments, examples and perspectives (pp. 727-754). Wiesbaden: Springer Fachmedien Wiesbaden.

Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., **Dengel**, A., & Ahmed, S. (2022). ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions. Computer Methods and Programs in Biomedicine, 215, 106620.

Lucieri, A., Bajwa, M. N., **Dengel**, A., & Ahmed, S. (2020, November). Explaining ai-based decision support systems using concept localization maps. In *International Conference on Neural Information Processing* (pp. 185-193). Cham: Springer International Publishing.

Lucieri, A., Bajwa, M. N., Braun, S. A., Malik, M. I., **Dengel**, A., & Ahmed, S. (2020, July). On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 international joint conference on neural networks (IJCNN)* (pp. 1-10). IEEE.

Zicari, R. V., Ahmed, S., Amann, J., Braun, S. A., Brodersen, J., Bruneault, F., Brusseau J., Campano, E., Coffee, M., **Dengel**, A., Düdder, B., Gallucci, A., Gilbert, T. K., Gottfrois, P., Goffi, E., Haase, C. B., Hagendorff, T., Hickman, E., Hildt, E., Holm, S., Kringen, P., Kühne, U., **Lucieri**, A., Madai, V. I., Moreno-Sánchez, P. A., Medlicott, O., Ozols, M., Schnebel, E., Spezzatti, A., Tithi, J. J., Umbrello, S., Vetter, D, Volland, H, Westerlund, M., & Wurth, R. (2021). Co-design of a trustworthy AI system in healthcare: deep learning based skin lesion classifier. *Frontiers in Human Dynamics*, *3*, 688152.

Mercier, D., **Lucieri**, A., Munir, M., **Dengel**, A., & Ahmed, S. (2021). Evaluating privacy-preserving machine learning in critical infrastructures: A case study on time-series classification. *IEEE Transactions on Industrial Informatics*, *18*(11), 7834-7842.

Lucieri, A., **Dengel**, A., & Ahmed, S. (2023). Translating theory into practice: assessing the privacy implications of concept-based explanations for biomedical AI. *Frontiers in Bioinformatics*, *3*.

Mercier, D., **Lucieri**, A., Munir, M., **Dengel**, A., & Ahmed, S. (2022). PPML-TSA: A modular privacy-preserving time series classification framework. *Software Impacts*, *12*, 100286.

Saifullah, S., Mercier, D., **Lucieri**, A., **Dengel**, A., & Ahmed, S. (2022). Privacy meets explainability: A comprehensive impact benchmark. *arXiv preprint arXiv:2211.04110*.

Lucieri, A., Schmeisser, F., Balada, C. P., Siddiqui, S. A., **Dengel**, A., & Ahmed, S. (2022, July). Revisiting the Shape-Bias of Deep Learning for Dermoscopic Skin Lesion Classification. In *Annual Conference on Medical Image Understanding and Analysis* (pp. 46-61). Cham: Springer International Publishing.

Muralidhara, S., **Lucieri**, A., **Dengel**, A., & Ahmed, S. (2022). Holistic multi-class classification & grading of diabetic foot ulcerations from plantar thermal images using deep learning. *Health Information Science and Systems*, *10*(1), 21.

2.3 Data and patent situation

The identification and use of any available public data is the subject of the work.

We are not currently aware of any patents that could have an impact on our planned work and subsequent utilization.

3 Detailed description of the work plan

3.1 Work package description incl. resource planning

The DFKI work planned in this sub-project description is to be understood as contributing to the overarching "Mission AI" goals and the cross-application test platform developed there. The cornerstones of the collaboration in Mission AI include

- Use test approaches:
  - The use cases of the DFKI are used to test the test approaches in the project, analogous to the implementation of WP3 of the Mission AI project plan.
  - The criteria and test procedures of Mission AI provide the basis for testing on the use cases, extended by the aspects for medical use cases as considered in the project.
- Integration of the platform:
  - DFKI's test platform for medical use cases and Mission AI's test platform represent a common point of contact, particularly with regard to user experience and story.
  - The software architectures of the platforms must be able to communicate.
  - The close integration of quality improvement tools, which can also be used as inspection tools, must be ensured.
  - All use cases are included in the testing of the test platform.
- Compatibility of the results:
  - The test tools developed by DFKI contribute to the test results and test statements of Mission AI - concretized and extended to the considered use cases, analogous to the implementation of WP3 of the Mission AI project plan.

The parallel processing of the DFKI subproject and the Mission AI requires regular coordination along the basic conceptual orientation of the Mission AI, especially in WP2 and WP3 of the Mission AI project plan.

3.1.1 Work package structure, main responsibilities

The following figure shows the work packages planned in the sub-project and their relationships.

The main result of the project will be the medical test platform to be developed in WP1. The quality platform to be developed in WP2 is an integral part of the medical test platform. The medical test platform itself is an integral part of the Mission AI platform.

The work on the test platform (WP1) and the quality platform (WP2) is being driven forward within the framework of 5 representative use cases.

Use case 1: Image-based diagnosis of skin cancer (AP3)

Use case 2: Tumor board decisions based on tabular data (WP4) Use case 3:

Spatio-temporal data in rescue services (WP5)

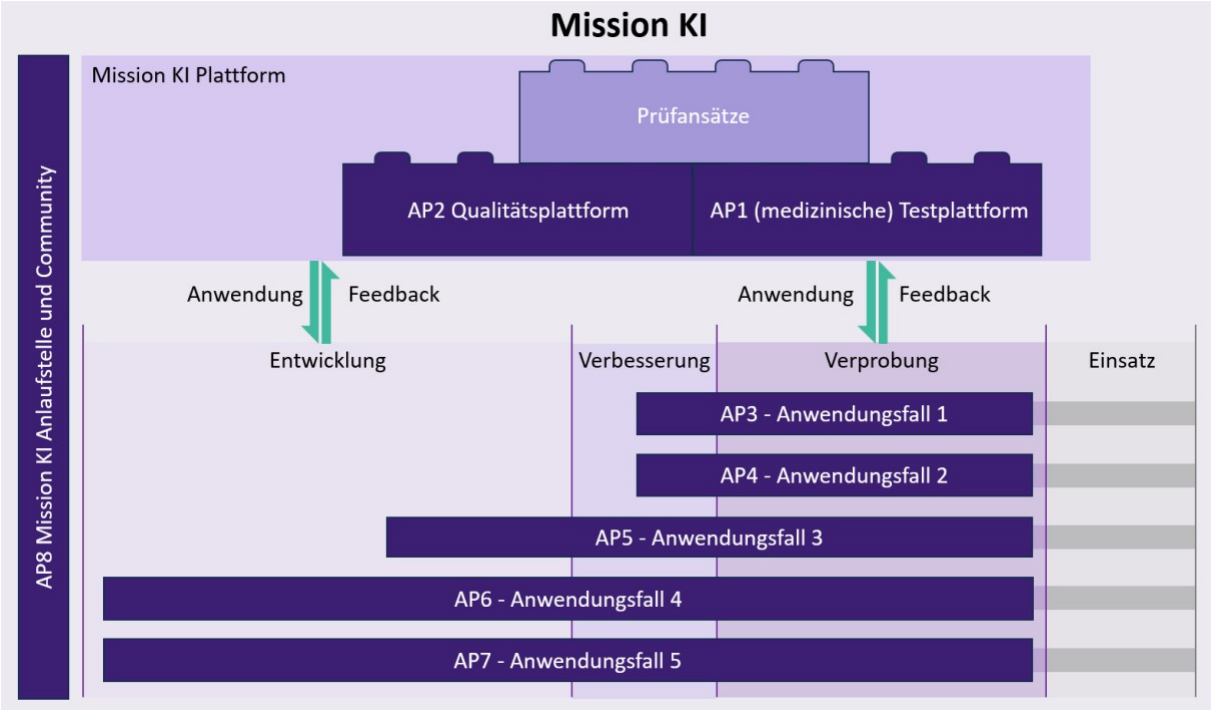
Use case 4: Text and survey data in psychiatric treatment and training (WP6)

Use case 5: Endoscopy videos for safe intubation (AP7)

The AI systems for implementing the use cases are in various stages of development at the start of the project.

- The systems for use cases 4 and 5 are being developed entirely within the project; in the case of use case 3, at least further development work is still taking place as part of the project. These three use cases will be used in particular to apply the prototype quality platform from WP2 and to improve it based on feedback from the applications.
- All five AI systems are to be tested using the test platform from WP1 in order to obtain feedback on the further development of the platform. The testing is preceded by a preparatory phase in which the focus on the development side is on ensuring that the testing is successful.

Finally, WP8 comprises the provision of the mission AI contact point and the corresponding community development.



All of the work described in this sub-project description will be carried out by DFKI. The work will be carried out in close coordination and using synergies with the other responsible parties of the Mission AI project with the aim of integration into the superordinate Mission AI test platform.

The parallel processing of the DFKI subproject and Mission AI requires regular coordination along the basic conceptual orientation of Mission AI, especially in WP2 and WP3 of the Mission AI project plan. Regular coordination with those responsible for Mission AI is planned for this purpose. The scheduling and frequency of the consultations is coordinated with those responsible for Mission AI.

3.1.2 Schedule and dependencies

		2024 Q2	Q3	Q4	2025 Q1	Q2	Q3	Q4
AP1	**Prüfplattform**							
1.1	Analyse der Anforderungen							
1.2	Analyse der relevanten Prüfplattformen							
1.3	Konzept und Prototyp einer Prüfplattform für vertrauenswürdige KI-Systeme in der Medizin							
1.4	Analyse, Kategorisierung und Bewertung bestehender Prüfwerkzeuge							
1.5	Anpassung und, falls erforderlich, Entwicklung neuer Prüfwerkzeuge für die Anwendungsfälle 1 und 2							
1.6	Vorbereitung und Durchführung prototypischer Prüfungen der Anwendungsfälle							
1.7	Bewertung und Gap-Analyse für die medizinische Prüfplattform							
1.8	Entwicklung von Handlungsempfehlungen für die medizinische Prüfplattform und deren Werkzeuge							
AP2	**Qualitätsplattform**							
2.1	Sprache der AI-Qualitätsanforderungen							
2.2	Wissensbasis (Knowledgebase)							
2.3	Automatisiertes Schlussfolgern							
2.4	Debugger							
AP3	**Anwendungsfall 1 - Bildbasierte Diagnose von Hautkrebs**							
3.1	Vorbereitung des KI-Systems für die prototypische Prüfung							
3.2	Anpassung und Entwicklung neuer Prüfwerkzeuge für die Anwendung							
3.3	Unterstützung der PoC Prüfung							
3.4	Evaluierung und Gap-Analyse des KI-Systems							
3.5	Nacharbeiten am KI-System zur Erfüllung der Prüfanforderungen und Nachprüfung							
3.6	Analyse des Prüfablaufs aus der Use Case Sicht und Handlungsempfehlungen für die medizinische Prüfplattform							
AP4	**Anwendungsfall 2 - Tumorboardentscheidungen anhand tabellarischer daten**							
4.1	Vorbereitung des KI-Systems für die prototypische Prüfung							
4.2	Anpassung und Entwicklung neuer Prüfwerkzeuge für die Anwendung							
4.3	Unterstützung der PoC Prüfung							
4.4	Evaluierung und Gap-Analyse des KI-Systems							
4.5	Nacharbeiten am KI-System zur Erfüllung der Prüfanforderungen und Nachprüfung							
4.6	Analyse des Prüfablaufs aus der Use Case Sicht und Handlungsempfehlungen für die medizinische Prüfplattform							
AP5	**Anwendungsfall 3 - Räumlich-zeitliche Daten im Rettungsdienst**							
5.1	Detaillierte Analyse des Anwendungsfalles bezüglich Testung und Risiko							
5.2	Nachfrageprognose							
5.3	Vorhersage über die verbleibende Ressourcennutzung							
5.4	Entscheidungsunterstützung, Identifizierung von Engpässen und verbesserte Gebietsabdeckung. Handlungsempfehlungen für die							
AP6	**Anwendungsfall 4 - Text- und Befragungsdaten in der psychiatrischen Behandlung und Ausbildung**							
6.1	Integration der Qualitäts- und Prüfplattform sowie Feedback an diese aus dem Anwendungsfall							
6.2	Einbeziehung der Nutzer:innen (Psychotherapeut:innen) als Schlüsselaspekt für die Entwicklung vertrauenswürdiger KI-Systeme							
6.3	Analyse der Transkripte							
6.4	Analyse der Ratingdaten therapeutischen Verhaltens							
6.5	Entwicklung eines erklärungsfähigen KI-Modells zur Verbesserung und Evaluation von klinischen Interventionen							
6.6	Nutzerstudie und Feedback							
AP7	**Anwendungsfall 5 - Endoskopie-Videos für eine sicherere Intubation**							
7.1	Integration der Qualitäts- und Prüfplattform sowie Feedback an diese aus dem Anwendungsfall							
7.2	3D-Rekonstruktion und Merkmalsextraktion aus Endoskopie-Videos							
7.3	Risikovorhersage für die Intubation							
7.4	Tests und Anforderungen für die Umsetzung							
7.5	Zukunftsplanung							
AP8	**Mission KI Anlaufstelle und Community**							
8.1	Sammeln und Analyse von Herausforderungen aus der Industrie							
8.2	Thematische Workshops zu vertrauenswürdigen Prinzipien							
8.3	Kuration von Showroom Exponaten							
8.4	Monatliche TrustifAI-Beratung							
8.5	Abschlussveranstaltung zur Präsentation der Projektergebnisse							

The dependencies can be found in the diagram in section 3.1.1.

3.1.3 Work package descriptions

AP1 Test platform

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	3 persons over the term, see section 3.2

Goals

The aim of WP1 is to design a test platform for trustworthy AI systems, especially in medicine. Medical AI systems must meet particularly high requirements in terms of trustworthiness. On the one hand, they must comply with existing legal regulations. However, due to the sensitive area of application with a direct impact on human well-being, it is also important for the acceptance of the systems by doctors and patients to fulfill further criteria of trustworthiness beyond the existing regulations. The basis for this is the verification of suitable test criteria for medical AI systems using a test platform. The work in WP1 is carried out in close cooperation with the other responsible parties in the Mission AI project and using synergies with the work there on a general, cross-application AI testing toolbox.

Expected results

The following results are expected from AP1:

- Identification of testing requirements for medical AI systems (WP1.1)
- Analysis of relevant test platforms (WP1.2)
- Design and prototype of a test platform for trustworthy AI systems in medicine (WP1.3)
- Analysis, categorization and evaluation of existing testing tools (WP1.4)
- Adapted and possibly new test tools for use cases 1 and 2 (WP1.5)
- Gap analysis for the medical testing platform based on use cases 1 and 2 (WP1.6 and WP1.7)
- Recommendations for the implementation of a final test platform (WP1.8)

Work

WP1.1 Analysis of requirements

In WP1.1, a classification of the various AI systems and components that can be used in medicine and healthcare is first created. This is intended to provide an overview of the breadth of application areas. The system will focus in particular on the AI technologies used and the role of trustworthiness in the practical use of medical systems.

Building on this, WP1.1 focuses on the fundamental analysis and precise understanding of the regulatory and normative landscape, which is particularly important for the use of medical AI systems as high-risk applications. Laws, standards and guidelines at local, national and international level are examined, including the GDPR, the AI Act, technical standards such as ISO42001:2023 and the German standardization roadmap for artificial intelligence. It is particularly important, for example, to identify and understand the terminology used by the various bodies in connection with trustworthy AI. Furthermore, it is analyzed whether there are criteria that have not yet been recorded that can contribute to the further acceptance of AI systems in medicine.

WP1.2 Analysis of the relevant test platforms

WP1.2 deals with the analysis and evaluation of existing testing platforms with regard to their suitability for testing AI systems in medicine. Both public and commercial platforms will be considered, including platforms that are still under development. The evaluation will consider various aspects such as adaptability to medical standards, generalizability to different medical use cases, interoperability with medical data and algorithms, and user-friendliness. In particular, however, the focus will be on the testing tools used, which include not only testing algorithms but also special synthetic and non-synthetic test data sets. Existing testing platforms and their tools will be cataloged in cooperation with the other persons responsible for the Mission AI project and deficits in existing testing environments for AI-based medical applications will be identified.

WP1.3 Concept and prototype of a test platform for trustworthy AI systems in medicine

With the help of the results of the previous work packages and in close cooperation with the other persons responsible for the Mission AI project, the prototype of a test platform for criteria of trustworthy AI in medical applications is being conceptualized in WP1.3. To this end, clearly formulated, yet generically applicable test conditions are first defined for the various dimensions of trustworthy AI (e.g. fairness, transparency, etc.). These abstract test conditions will be applicable to various types of algorithms, data and use cases through explicit instantiations in the form of test tools. The defined conditions do not claim to be complete, but can be expanded at any time as required. The clear structure of the test platform will make it possible to categorize test tools and thus easily identify the test coverage of different use cases and possible gaps. The aim is to ensure connectivity with the overarching overall platform as part of Mission AI.

WP1.4 Analysis, categorization and evaluation of existing testing tools

This work package initially concentrates on the state-of-the-art analysis of testing tools for AI algorithms that are already in commercial use or are still being researched, with a focus on medical applications. These testing tools will be catalogued and categorized in the context of medicine in collaboration with the other Mission AI project managers so that they can ultimately be incorporated and anchored in the Mission AI testing tool landscape. Based on the results, a method for the structured recording of testing tools will be developed, applied and incorporated into the concept of the testing platform, analogous to similar approaches (such as explainability fact sheets https://doi.org/10.1145/3351095.3372870) from the literature. The evaluation of the testing tools as part of a structured recording also relates to the applicability and transferability of existing methods to other medical applications and subsequently serves to formulate a task catalog for the further development of current testing tools, with the aim of ensuring sufficient coverage for a broad range of medical applications.

WP1.5 Adaptation and, if necessary, development of new test tools for use cases 1 and 2

Based on the task catalog from WP1.4 and with regard to the specific medical use cases 1 and 2 (WP3 and WP4), existing testing tools will be adapted and further developed. First, a requirements analysis for testing tools will be carried out with regard to the two use cases. As part of this analysis, the AI quality criteria relevant to the respective use cases and suitable for testing are identified and described. The aim of this work package is to provide customized testing tools that enable a precise and reliable evaluation of the AI systems in the selected medical use cases and at the same time, as far as possible, meet the established requirements.

WP1.6 Preparation and implementation of prototype tests of the use cases

The preparation for the prototype testing of the use cases includes the integration of the selected testing tools into a modular proof-of-concept (PoC) testing environment on the one hand and the adaptation and integration of the AI system in the work packages of the use cases on the other. Depending on the use case, prototype testing can also include the evaluation of quality criteria and test results through user studies in addition to the use of quantitative testing tools.

WP1.7 Evaluation and gap analysis for the medical testing platform

Once the PoC has been completed, this work package will evaluate the test results to assess the practicability and completeness of the Mission AI test platform based on our medical use cases. The work includes a detailed gap analysis in which the performance of the platform is critically scrutinized and compared with the original requirements and objectives. Identifying discrepancies and potential for improvement is crucial in order to further optimize and adapt the platform. This analytical work requires a deep understanding of both the technical and clinical aspects of medical AI applications to ensure that the platform provides a comprehensive and reliable testing environment.

WP1.8 Development of recommendations for action for the medical testing platform and its tools

Finally, precise recommendations for action are formulated based on the data collected and findings from the gap analysis. These recommendations should not only address identified shortcomings, but also show strategic ways in which the platform can be further developed to proactively take future requirements and developments in the field of medical AI into account. The development of these recommendations requires a holistic view of the platform, including technical, regulatory and market-specific aspects, in order to ensure a sustainable and future-proof solution.

WP2 Quality platform

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	3 persons over the term, see section 3.2

Goals

An AI quality platform is essential to enable trustworthy AI by facilitating difficult development processes, promoting transparency and enabling proactive risk minimization. The goal of WP 2 is therefore to develop such a quality platform that helps users to create efficient and effective data science pipelines and thus make an important contribution to the overall goal of Mission AI. A human-in-the-loop approach is intended to promote a development process that improves the quality and trustworthiness of AI systems. Human-in-the-loop aims to achieve what neither a human nor a machine can achieve alone. If a machine is unable to solve a problem, a human must step in and intervene. This process leads to a continuous feedback loop. Specifically in this context, WP2 focuses on the development of a common language for formulating AI quality requirements, jointly curating existing concepts for AI quality and application requirements in practice and aiming for integrability into the Mission AI platform. A continuously updated knowledge base for AI quality management will help in the design of tools for automated testing of AI applications and recommendations for preventive measures during the development process.

Expected results

The following results are expected from WP2:

- Design of a prototype of a standard terminology ("ubiquitous language") for the specification of AI requirements (WP2.1)
- Curating an expandable knowledge base of AI quality requirements and measures (WP2.2)
- Expansion of the knowledge base to include logical inference units for the automatic inference of requirements validations (WP2.4)
- The expected results 1-3 mentioned above together form a prototype of the AI Quality Platform
- Design of a prototype for an AI "debugger" to automatically analyze the AI development process and to recommend preventive measures during the AI development process in different application areas (WP2.4).
- Provide a mapping of AI quality requirements and corresponding tests of the AI test platform. (WP1.4 and WP2.1)

Work

<u>WP2.1: Language of AI quality requirements</u>

A "ubiquitous language" for specifying AI quality requirements and measures is a standardized terminology for stakeholders to precisely communicate domain-specific quality requirements, assumptions, use cases and test scenarios. The requirements language will be readable by both humans and machines to enable integration with a knowledge base (WP2.2) and tools for automated verification. The process of formulating such a language includes a
Literature analysis and consultations with experts. The result of WP2.1 is presented in WP2.2

as a schema for the knowledge database and in the implementation of the prototype of the test platform to specify common interfaces.

WP 2.2: Knowledge base (knowledgebase)

The knowledge base will build on WP2.1 to catalog curated knowledge about AI quality, existing metrics and quantifiable metrics, as well as models, algorithms, frameworks, and methods for AI development. In addition, it will map abstract concepts of AI quality via properties, requirements, and potential vulnerabilities to specific implementations of AI models. This unified data structure will enable efficient search, automated review and recommendation of tests for the AI testing platform. The evaluation of WP2.2 will be based on the accuracy, precision and scope of the stored information compared to publicly available ontologies and project use cases.

WP 2.3: Automated reasoning

The prototype language for AI quality requirements (WP2.1) in conjunction with information from the knowledge base (WP2.2) enables automated reasoning that can be used for rule checking, derivation of new facts and automatic execution of tasks. The latter can automatically select and execute tests for the AI quality requirements entered by the user to verify compliance with these requirements.

AP 2.4: Debugger

The debugger prototype illustrates the potential of the AI quality platform. It is intended as a human-in-the-loop utility for the (continuous) development of AI models. The debugger aims to identify potential vulnerabilities and suggest remediation strategies during (and not after) development. This will raise awareness of AI ethics guidelines among practitioners and, in conjunction with the AI testing platform, ensure their enforcement. WP2.5 will be measured qualitatively through stakeholder satisfaction surveys and quantitatively through A/B testing, where AI models developed with the debugger pass more tests in the AI testing platform than those developed without it.

AP3 Use case 1 - Image-based diagnosis of skin cancer

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	1.5 persons over the term, see section 3.2

Goals

Use case 1 is an AI system for detecting skin cancer on the basis of medical images. This system is already available at DFKI. The aim of WP3 is to test the test platform from WP1 with the help of the use case and to evaluate the applicability of the work results of the other people responsible for the Mission AI project. The AI system for detecting skin cancer is motivated and described below.

The rapid and reliable diagnosis of skin cancer, in particular the differentiation of malignant melanomas from melanocytic nevi, represents a significant challenge, as these lesions can be very similar. However, early identification of melanoma is crucial, as survival rates are drastically reduced if detection is delayed. The AI system used in this project is based on deep learning methods, which can be used for the automated evaluation of skin lesions based on dermatoscopic images. Since the classification of lesions is based on complex biomarkers that are difficult to interpret for non-experts, the system already has various multimodal explanatory components that are specially adapted to the diagnostic concepts used in medicine. This simplifies the communication and traceability of decisions and thus enables the system to be used, for example, to support the pre-selection of critical cases in general medical practices.

Expected results

The following results are expected from WP3:

- Detailed description of components of the AI system (WP3.1)
- New and adapted testing tools (WP3.2)
- Description of the test process for proof-of-concept (AP3.3)
- Evaluation of the trustworthiness of the AI system (WP3.4)
- Addressing deficits and review (WP3.5)
- Test report with recommendations (WP3.6)

Work

WP3.1: Preparation of the AI system for prototype testing

In WP3.1, the medical AI system is analyzed and categorized according to the test platform concept developed in WP1.3. The components of the AI system are described in detail. Potentially necessary additions to the system in terms of explainability, data protection and fairness are identified and addressed in order to work towards a positive outcome of the planned test. Furthermore, the actual audit execution is prepared technically and organizationally.

WP3.2: Adaptation and development of new testing tools for the application

In close cooperation with WP1, WP3.2 examines the applicability of existing testing tools to the AI system and identifies inadequate/missing tools.

are identified in advance. Based on this, application-specific instantiations of test tools are developed in collaboration with WP1.5. The structured recording of these tools is carried out using the test tool data sheets from WP1.4 for integration into the test platform and future use.

WP3.3: Support for PoC testing

In WP3.3, the PoC testing process coordinated by WP1.6 is supported by the stakeholders of the AI system. This includes the technical implementation of the test, the contribution of technical expertise on the mode of operation of the AI system and its components as well as the contribution of organizational knowledge about the practical application context of the system. This should ensure that the proof-of-concept test runs smoothly. In addition, WP3.3 also documents organizational aspects of the test execution and preparation from the perspective of the system provider.

WP3.4: Evaluation and gap analysis of the AI system

This work package comprises the detailed evaluation of the test results using the test tools. The degree of fulfillment of the defined test conditions is evaluated and open aspects and possible improvements are identified. The aim is to carry out a comprehensive assessment of the trustworthiness of the AI system and to create the basis for further optimization.

WP3.5 Reworking the AI system to fulfill the test requirements and verification

Based on the results of the gap analysis from WP3.4, the most important open aspects are addressed in WP3.5. Identified deficits in existing system components are rectified in order to improve the trustworthiness of the AI system under consideration. A subsequent evaluation and reporting on the achieved trustworthiness of the system rounds off this work package.

WP3.6 Analysis of the test procedure from the use case perspective and recommendations for action for the medical test platform

In WP3.6, the test procedure is analyzed from the perspective of the use case. Obstacles and deficits in both technical and organizational aspects during the test are assessed. The involvement of relevant stakeholders is evaluated and recommendations for action are developed for the testing process and the integration of technical components such as testing tools and assessment models. The findings from this work package will be incorporated into the recommendations for action from WP1.8 in order to further optimize the medical testing platform.

AP4 Use case 2 -Tumorboard decisions based on tabular data

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	1 person over the term, see section 3.2

Goals

Use case 2 is a system to support tumor board decisions in urology. This system is already being developed at DFKI as part of the KITTU funding project (https://kittu.org). The aim of WP4 is to test the test platform from WP1 with the help of the use case and to evaluate the applicability of the work results of the other project managers of the Mission AI project. The experience and knowledge gained will in turn be used to improve both the AI system itself and the test platform. The targeted AI system expected from KITTU is explained and described below.

A tumor board is a collaborative approach to making decisions about cancer treatment. It involves a group of physicians from different specialties who come together to review and discuss individual cancer cases based on a large amount of categorical, numerical and textual data from each patient. The AI system developed as part of KITTU takes all values into account to predict the optimal decision for the tumor board, as it is able to process much more data in less time than the expert does. The prediction of the system is to be presented together with an explanation to the tumor board to re-evaluate the case based on the automated decision. Therefore, the system uses a number of language models and other neural networks with explanatory components based on attention, feature permutation and other post-hoc approaches.

Expected results

The following results are expected from WP4:

- Detailed description of components of the AI system (WP4.1)
- New and adapted testing tools (AP4.2)
- Description of the test process for proof-of-concept (AP4.3)
- Evaluation of the trustworthiness of the AI system (WP4.4)
- Addressing the deficits, further evaluation (WP4.5)
- Test report with recommendations (WP4.6)

Work

<u>WP4.1: Preparation of the AI system for prototype testing</u>

In WP4.1, the medical AI system is analyzed and categorized according to the test platform concept developed in WP1.3. The components of the AI system are described in detail. Potentially necessary additions to the system in terms of explainability, data protection and fairness are identified and addressed in order to work towards a positive outcome of the planned test. Furthermore, the actual audit execution is prepared technically and organizationally.

WP4.2: Adaptation and development of new testing tools for the application

In close cooperation with WP1, the applicability of existing testing tools to the AI system is examined in WP4.2 and inadequate/missing tools are identified in advance. Based on this, application-specific instantiations of test tools are developed in collaboration with WP1.5. The structured recording of these tools is carried out using the test tool data sheets from WP1.4 for integration into the test platform and future use.

WP 4.3: Support for PoC testing

In WP4.3, the PoC testing process coordinated by WP1.6 is supported by the stakeholders of the AI system. This includes the technical implementation of the test, the contribution of technical expertise on the mode of operation of the AI system and its components as well as the contribution of organizational knowledge about the practical application context of the system. This should ensure that the proof-of-concept test runs smoothly. In addition, WP4.3 also documents organizational aspects of the test execution and preparation from the perspective of the system provider.

WP 4.4: Evaluation and gap analysis of the AI system

This work package comprises the detailed evaluation of the test results using the test tools. The degree of fulfillment of the defined test conditions is evaluated and open aspects and possible improvements are identified. The aim is to carry out a comprehensive assessment of the trustworthiness of the AI system and to create the basis for further optimization.

WP 4.5 Reworking the AI system to fulfill the test requirements and verification

Based on the results of the gap analysis from WP4.4, the most important open aspects are addressed in WP4.5. Identified deficits in existing system components are rectified in order to improve the trustworthiness of the AI system under consideration. A subsequent evaluation and reporting on the achieved trustworthiness of the system rounds off this work package.

WP 4.6 Analysis of the test procedure from the use case perspective and recommendations for action for the medical test platform

In WP4.6, the test procedure is analyzed from the perspective of the use case. Obstacles and deficits in both technical and organizational aspects during the test are assessed. The involvement of relevant stakeholders is evaluated and recommendations for action are developed for the testing process and the integration of technical components such as testing tools and assessment models. The findings from this work package will be incorporated into the recommendations for action from WP1.8 in order to further optimize the medical testing platform.

WP5 Use case 3 - Spatio-temporal data in rescue services

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	1 person over the term, see section 3.2
Partner:	DRK

Goals

The dynamic distribution of ambulances and rescue vehicles to existing stations for operational readiness is a complex optimization problem [1], the efficient solution of which can make the difference between life and death. The current status quo in the integrated control centers operated by the German Red Cross (DRK) in Rhineland-Palatinate includes neither forecasts about possible emergency calls nor about the imminent availability of rescue vehicles in operation. Instead, measures are only taken when operations actually arrive, vehicles from the area coverage are only dispatched when specific operations can no longer be attended to and vehicles are only taken into account in the dispatch proposal if they are in a status in which they can take over an operation; depending on the status, fixed deployment times are assumed, which, however, deviate greatly from the actual circumstances.

If the decision-makers for short to medium-term decisions in the control centers are to be supported by AI systems, quality criteria for these systems must be clearly defined and strictly adhered to during development, which is why this use case is ideal for testing the quality platform created in WP2. Furthermore, the aim of WP5 is to use the fully developed AI system to test the test platform from WP1 and to evaluate the applicability of the work results of the other project managers of the Mission AI project.

In this work package, we are working together with the German Red Cross as a project partner. In order to maximize the number of possible rescue service deployments, the limited resources available must be used optimally.

To do this, it is necessary to know how many emergency calls are to be expected where and how long the operations will take in order to be able to implement an early warning system for capacity utilization. Data-based decisions can support employees in the control centers in their decision-making and provide a good basis for comprehensible and justifiable operational-tactical decisions, for example when negotiating with cost bearers and political actors. In particular, there is currently no nationwide analysis of capacity utilization and shifting of emergency services across control center areas, which promises to further improve the rescue service, but cannot be implemented without AI support. Any use of AI in this context falls under the EU AI Act, as these are high-risk systems that are characterized by high requirements for monitoring by humans.

[1] M. Restrepo, S. G. Henderson, and H. Topaloglu, "Erlang loss models for the static deployment of ambulances," Health Care Manag Sci, vol. 12, pp. 67-79, Mar. 2009.

Expected results

The following results are expected from WP5:

- Detailed report on how the project was implemented within the framework of the EU AI ACT
  The criteria are based on the safety assessment and the verifiable criteria that must be met in order to permit actual use. Criteria lie in the

standard language from WP2 in order to be used in the quality platform. (WP5.1)

- Description of components of the AI system and applicability of the test and testing platform, development of new tools if necessary
- AI model (code, weights, documentation), which outputs the predicted ambulance demand using the factors mentioned in WP5.2 as inputs and complies with the criteria defined in WP5.1. (WP5.2)
- AI model (code, weights, documentation), which outputs the predicted remaining service life of a vehicle using the factors mentioned in WP5.3 as inputs and complies with the criteria defined in WP5.1. (WP5.3)
- AI system that uses the models from WP5.2 and 5.3 to calculate when and where a bottleneck is to be expected and calculates how available ambulances can be optimally deployed or whether reinforcements from non-resident ambulances need to be requested. The system also attempts to distribute missing ambulances fairly across different areas.
- Test report (WP5.4)

Work

WP5.1: Detailed analysis of the use case with regard to testing and risk

AI systems for ambulances and rescue vehicles could be classified as high-risk systems under the EU AI Act due to their role in emergency management and public safety. It is important that these systems take into account the strict requirements and obligations of the Act from the outset. In WP 5.1, these criteria are analyzed and system compatibility with the EU AI Act is checked. Furthermore, the criteria are transferred to the standard language and the quality platform from WP2 in order to ensure compliance during development. The use case will be used to test the test platform from WP1 and to check the applicability of the work results of the other responsible parties in the Mission AI project. During development, feedback from the use case to the two platforms should also enable dynamic development of these and the application systems.

WP5.2: Demand forecast

Improved forecasting of potential emergency calls and therefore the need for rescue resources is one of the cornerstones of efficient dispatching and reducing the workload on rescue stations. At present, such forecasts are not made, which means that measures such as alerting additional forces or calling in non-resident vehicles from areas with low capacity utilization can only be taken when actual operations occur. In this work package, predictive factors such as time of day, local distribution, hotspots, weather and events (sporting events, ...) are to be identified and used from historical deployment data [1] in order to forecast the local and temporal accumulation of future deployments. For this purpose, point process models such as Cox processes [2] are used and combined with new AI models. The development of the models and the processing of the data are checked using the knowledge base created in WP2 and the debugger based on the criteria formulated in WP5.1.

- [1] Li, C., Cui, K. Multivariate Hawkes processes with spatial covariates for analyzing spatio-temporal event data. Ann Inst Stat Math (2024). https://doi.org/10.1007/s10463-023-00894-2
- [2] Xenia Miscouridou, Samir Bhatt, George Mohler, Seth Flaxman, & Swapnil Mishra. (2022). Cox-Hawkes: Doubly stochastic spatio-temporal Poisson processes. https://arxiv.org/abs/2210.11844

WP5.3: Prediction of the remaining resource utilization

If a vehicle is on the road after an emergency call has been received, it is crucial for the control center to know when the vehicle will be free again for the next operation. This depends on factors such as (i) the type of operation, (ii) the clinical picture, (iii) the location and accessibility of the operation, (iv) whether other forces such as an emergency doctor need to be called in, (v) the location of the target hospital, (vi) the capacity utilization or handover time of the target hospital, (vii) the recovery time of the vehicle. (vii) the recovery time of the ambulance vehicle (e.g. disinfection in the case of infectious diseases), (viii) the distances to be covered and, last but not least, (ix) the traffic situation. In this work package, these factors will be used to make predictions about the remaining resource utilization from historical deployment data. We are working together with the ADAC for the historical traffic data. The predictions themselves are to be realized using time2event models, whereby the prediction is to be regularly updated with live information. Here we can rely on previous work [1, 2]. The development of the models and the processing of the data will be checked with the help of the knowledge base created in WP2 and the debugger using the criteria formulated in WP5.1.

- [1] Abbasi, A. F., Asim, M. N., Ahmed, S., Vollmer, S., & Dengel, A. (2024). Survival Prediction Landscape: An In-Depth Systematic Literature Review on Activities, Methods, Tools, Diseases, and Databases. Cold Spring Harbor Laboratory. https://doi.org/10.1101/2024.01.05.24300889
- [2] Sonabend, R., Bender, A., & Vollmer, S. (2022). Avoiding c-hacking when evaluating survival distribution predictions with discrimination measures. In Z. Lu (Ed.), Bioinformatics (Vol. 38, Issue 17, pp. 4178-4184). Oxford University Press (OUP). https://doi.org/10.1093/bioinformatics/btac451

WP5.4: Decision support, identification of bottlenecks and improved area coverage. Recommendations for action for the test platform.

The predictions obtained from WP5.2 and WP5.3 are intended to support short and medium-term decisions by the various dispatchers (head dispatchers, sub-dispatchers) and to identify possible future bottlenecks and the resulting improved area coverage. For example, if several vehicles are available in an area, vehicles could be actively dispatched to areas with an imminent shortage - due to an increase in demand recorded in real time, long active deployments or a shift in traffic volume - in order to occupy stations there or take up positions at other strategically favorable locations - even across control center areas. In contrast to WP5.2 and WP5.3, not only predictions are made here, but active support is provided for decisions, which is why the safety aspect is of great importance here. Fairness aspects should also come into play here: In the event of an undersupply of rescue resources, the aim should be to distribute errors evenly according to demographics. These fairness criteria should also be checked and guaranteed using the platform created in WP2. While it is outside the scope of the work package to develop a complete simulator, we plan to research the current state of research on simulators in detail. The system can then be tested using existing (traffic) simulators. The overall system will be tested using the test platform from WP1 and the findings from the process will be recorded as recommendations for action. In addition, a test report will be drawn up to verify the minimum standard co-developed by Acatech.

WP6 Use case 4 - Text and survey data in psychiatric treatment and training

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	0.8 Persons over the term, see section 3.2
Partner	University of Trier

Goals

Mental health and related interventions are determined and described by complex human conditions that are not always easy to quantify. Psychiatry and psychology rely on a mix of quantitative data (e.g. psychometric indicators, survey-based indicators, demographic indicators) and qualitative descriptions such as therapeutic notes and textual descriptions of the patient's condition. This creates a uniquely complex environment for the development of AI-driven solutions and tools that are trustworthy, reliable, useful, usable and non-harmful.

The aim of this work package is to present a development cycle for a high quality, trustworthy AI system that meets the needs of users and addresses the challenge of developing effective treatment plans for psychiatry by combining historical treatment plans (unstructured text documents) with psychometric, demographic and survey indicators. The system is intended to complement non-AI-based feedback from senior colleagues to support trainee psychiatrists in developing better treatment plans, with a focus on explainability, transparency, ease of use and end-user acceptance of the system. As the data contains highly sensitive information about individuals, we place particular emphasis on data privacy and the possibilities of developing such AI systems collaboratively while preserving the privacy of the individuals involved in the treatment process.

The aim of WP6 is, on the one hand, to test the quality platform created in WP2 and, on the other hand, to use the fully developed AI system to test the test platform from WP1 and to evaluate the applicability of the work results of the other persons responsible for the Mission AI project.

Expected results

The following results are expected from WP6:

- Description of components of the AI system and applicability of the test and testing platform, development of new tools if necessary
- Overview of NLP techniques; description of the testing process for proof of concept (WP6.3)
- Evaluation of the explainability of the AI system (WP6.4)

Work

<u>WP6.1: Integration of the quality and testing platform and feedback to it from the use case</u>

In WP6.1, an effective and dynamic integration and feedback between the development of AI applications, the quality platform (WP2) and the testing platform (WP1) will be developed. to ensure that all aspects of quality assurance and testing are embedded in the development process. The relevant criteria are translated into the standard language and

the quality platform from WP2 in order to ensure compliance during development. Furthermore, the use case will be used to test the test platform from WP1 and to check the applicability of the work results of the other project managers of the Mission AI project. During development, feedback from the use case to the two platforms should also enable dynamic development of these and the application systems.

WP6.2: Involvement of users (psychotherapists) as a key aspect for the development of trustworthy AI systems

As the main focus of this work package is on the usability, usefulness and acceptance of the system, we will first interview the primary users of the system - trainee psychotherapists - to understand their attitudes towards AI and AI-based solutions and to find out what would be the most important features of the system that would support their learning and practice. One focus of the content will be the evaluation of first clinical interventions of prospective psychotherapists.

WP6.3: Analysis of the transcripts

Before the transcripts can be analysed, data security must be checked and the anonymization of sensitive data ensured. Robust measures are implemented to ensure that the privacy of the source data is protected and that the models do not learn any sensitive correlations that could lead to the identification of individuals. This also serves to increase the trustworthiness of the system. The analysis of the transcripts themselves focuses on various natural language processing feature extraction and language modeling techniques that can be used for the analysis of transcripts in psychotherapy. Different feature extraction and language modeling methods are investigated in terms of their efficiency for the data at hand and their suitability for use in trustworthy AI use cases. The decisions are supported by the quality platform (WP2). Subsequently, a selected modeling technique will be used to determine whether an AI tool can be developed that meets the required criteria for trustworthiness.

WP6.4: Analysis of rating data of therapeutic behavior

Which indicators are best suited to describe therapists' initial interventions? A consistent basis is created by carefully reviewing and processing existing data/ratings on central constructs of basic psychotherapeutic behavior and intervention. This data can then be analyzed and provide information on whether there are indicators that can be used for partial explanation or prediction. This will also create a basis for the models to be developed further and an analysis of possible distortions in the data will be carried out.

WP6.5: Development of an explainable AI model for the improvement and evaluation of clinical interventions

Using a combination of features and models derived from existing data, an AI model will be developed that learns from transcripts of existing initial interventions in response to critical situations in videos to provide feedback on the current initial interventions based on new transcripts.
give. The main focus in the selection of techniques and models will be on their

The ability to explain the reasons for recommendations in a way that protects privacy and is understandable and trustworthy for prospective psychotherapists is a key challenge for the successful use of such systems in practice.

WP6.6: User study and feedback

Finally, a proof-of-concept user study will be conducted with prospective psychotherapists to investigate whether they could imagine using the system in clinical practice. Barriers to using the system will also be identified and feedback collected for further improvement. The aim is to promote the trustworthiness, usefulness and comprehensibility of the system. In addition to measuring these aspects, the quality of the treatment interventions will also be assessed on the basis of ratings by trained experts. They assess whether or not the collaboration between users and AI has led to a better quality treatment intervention. Particular attention will be paid to the effects of such a system on prospective therapists.

AP7 Use case 5 - Endoscopy videos for safer intubation

Start date: Month 1	
Runtime of the AP	Corresponds to project duration
Total personnel expenses	0.7 Persons over the term, see section 3.2
Partner:	Saarland University

Goals

In Germany, around 17 million anesthesias are performed every year (according to the German Society of Anesthesiology, 2020). Intubation, a crucial step in this process, is the insertion of a breathing tube through the mouth into the windpipe to enable artificial respiration during surgical procedures under general anesthesia. Despite the high standards, serious incidents are associated with significant risks even in healthy patients undergoing elective surgery [1]. In some patients (8-10 %), the airway is so obstructed that intubation is problematic or impossible [2]. This can lead to a life-threatening scenario, which is also very stressful for anaesthetists. All patients are therefore checked by anaesthetists during the pre-anaesthetic consultation for signs of obstacles to intubation or -difficulties. Currently, this assessment is based on external characteristics and clinical scoring systems [1,3,4]. This examination is important, but cannot rule out problems with absolute certainty. As a rule, the anesthesia department does not have high-resolution images of the airway situation available preoperatively in order to specifically and individually address the airway situation. However, the Ear, Nose and Throat Clinic at Saarland University Hospital has access to image data that is generated as part of the standardized examination of the larynx using endoscopes in daily clinical work. However, clinically validated predictors for a difficult airway based on these ENT endoscopy findings do not yet exist.

The aim is therefore to develop a trustworthy and reliable AI-based risk classification that uses routinely collected data to support intubation planning and improve patient safety. The success of such a project depends on collaboration between the departments of otorhinolaryngology and anesthesiology to enable effective sharing of the collected data for this secondary purpose.

The relevance of this use case lies in the prediction and evaluation of airway obstructions using AI. AI-supported identification of predictors for difficult intubation can not only significantly optimize intubation safety and thus patient care, this approach also offers the potential for more efficient use of available resources, a reduction in the workload of physicians and nursing staff due to unforeseen intubation difficulties and thus promotes a safe environment for patients and medical staff. In addition, it opens up new perspectives for medical research by providing deeper insights into the causes of difficult airway situations and thus contributes to the development of adapted treatment and intervention strategies. Due to the high relevance of this application, reliable AI-based risk prediction is equally important.

In the Department of Otorhinolaryngology and Head and Neck Surgery at Saarland University Hospital, around 200 airway findings are recorded and stored each month as part of the preoperative examination. As each finding is accompanied by a corresponding airway assessment by anesthesiology colleagues during intubation, we have very consistent data sets that can be used as training data for the development of an AI model.

The aim of WP7 is, on the one hand, to test the quality platform created in WP2 and, on the other hand, to use the completed AI system to test the test platform from WP1 and to evaluate the applicability of the work results of the other persons responsible for the Mission AI project.

- [1] Schiff, J.H.; Welker, A.; Fohr, B.; Henn-Beilharz, A.; Bothner, U.; Aken, H.V.; Schleppers, A.; Baldering, H.J.; Heinrichs, W. Major Incidents and Complications in Otherwise Healthy Patients Undergoing Elective Procedures: Results Based on 1.37 Million Anaesthetic Procedures. BJA: Br. J. Anaesth. 2014, 113, 109-121, doi:10.1093/bja/aeu094.
- [2] Piepho T, Kriege M, Byhahn C, Cavus E, Dörges V, Ilper H et al: S1 Guideline Airway Management 2023 Anästh Intensivmed 2024;65:69-96. DOI: 10.19224/ai2024.069
- [3] Samsoon GL, Young JR: Difficult tracheal intubation: a retrospective study. Anaesthesia 1987;42:487-490
- [4] Cormack RS, Lehane J: Difficult tracheal intubation in obstetrics. Anaesthesia 1984;39:1105-1111

Expected results

The following results are expected from WP7:

- Detailed description of components of the AI system (WP7.1)
- Risk score for intubation. New and adapted testing tools (AP7.2)
- Description of the testing process, evaluation of the trustworthiness of the system (WP7.3)
- Addressing the deficits, further evaluation (WP7.4)
- Stakeholder report with recommendations (WP7.5)

Work

WP7.1: Integration of the quality and testing platform and feedback to it from the use case

WP7.1 ensures effective and dynamic integration and feedback between the development of the reconstruction model (WP7.1) and the risk assessment model (WP7.2), the quality platform (WP2) and the testing platform (WP1) to ensure that all aspects of quality assurance and testing are embedded in the development. The relevant criteria are transferred to the standard language and the quality platform from WP2 in order to ensure compliance with them during development. Furthermore, the use case is used to test the test platform from WP1 and to check the applicability of the work results of the other people responsible for the Mission AI project. During development, feedback from the use case to the two platforms should also enable dynamic development of these and the application systems.

WP7.2: 3D reconstruction and feature extraction from endoscopy videos

In this sub-work package, existing experiences and methods for the partial reconstruction of 3D models of the trachea from endoscopy videos are adopted [1]-[6]. The reconstructed 3D model should enable the extraction of specific features for further learning tasks (see WP7.2). Two basic steps are necessary for this reconstruction. First, the search for dense correspondences between individual images in a video [2], [7]. Subsequently, the reconstructed model is optimized with the help of pattern models and physical rules [5], [6]. The challenges in analyzing endoscopy videos include tracheal deformation, optical distortions, reflections (wet surfaces) and homogeneous
Areas with low texture. Therefore, the existing methods must be extended to downstream

tasks and updated according to the requirements of the quality platform and adapted to the use case. The results of WP7.1 can be evaluated by comparing the reconstructed models with 3D from CT scans that are available for some patients.

- [1] Battrawy et al. "DeepLiDARFlow: A deep learning architecture for scene flow estimation using monocular camera and sparse LiDAR", IROS 2020.
- [2] Schuster et al. "SDC - Stacked Dilated Convolution: A unified descriptor network for dense matching tasks", CVPR 2019.
- [3] Schuster et al. "FlowFields++: Accurate Optical Flow Correspondences Meet Robust Interpolation". ICIP 2018.
- [4] Bailer et al. "Flow Fields: Dense Correspondence Fields for Highly Accurate Large Displacement Optical Flow Estimation", CVPR 2017.
- [5] Kairanda et al. "φ-SfT: Shape-from-Template with a Physics-Based Deformation Model", CVPR 2022.
- [6] Golyanik et al. "Intrinsic Dynamic Shape Prior for Fast, Sequential and Dense Non-Rigid Structure from Motion with Detection of Temporally-Disjoint Rigidity", 3DV 2020.
- [7] Liu et al. "Extremely Dense Point Correspondences Using a Learned Feature Descriptor", CVPR 2020.

WP7.3: Risk prediction for intubation

Using features from AP 7.1 or the direct endoscopy images and AI prediction models with deep learning, we are developing a risk score to correlate the laryngeal findings obtained during the ENT examination with the clinical anesthesiology scoring systems. Such a model can, for example, be a simple classifier that maps the recorded videos to the clinical scoring [1], [2]. In addition, other features can be identified that indicate a difficult airway and thus help to minimize the risk of intubation.

- [1] He et al. "Deep residual learning for image recognition". CVPR 2016.
- [2] Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv:2010.11929.

WP7.4: Tests and requirements for implementation

We want to measure the capability of the models and ensure that they can work with data that has different characteristics. For example, these variations may include different cameras with different technical specifications, operators with different endoscopy techniques or experience, and sensitivity to patients with different demographic and biological characteristics. Continued evaluation of AI explanations in collaboration with clinicians and patient advocates will help identify areas where the trained models can be improved, while respecting the limitations of endoscopy technique and patient comfort. Ideally, the test and quality platform and the continuous analysis of the models will be used to derive exact requirements for the endoscopy videos.

WP7.5: Planning for the future

Based on the AI quality platform (WP2) and the tests (WP1), we will identify possible further steps for operationalization that fall outside the scope of the previous sub-work packages. This could include alternative or additional sources for data collection and feedback or the development of new metrics for the quality of the sample data.

AP8 Mission AI contact point and community

Start date: Month 1	
Duration of the	WPCorresponds to project duration
Total personnel	expenses1 person over the term, see section 3.2

Goals

The overarching goal of WP8 is to establish and operate a Mission AI contact point in Kaiserslautern. This includes collecting industrial challenges related to AI applications through jointly organized events with the other responsible persons of the "Mission AI" project, conducting thematic workshops on trustworthy AI principles, curating potential exhibits, providing monthly consultation sessions on trustworthiness of AI and organizing a final event to present the project results. These activities aim to integrate trustworthiness into AI systems, address the challenges and problems, foster collaboration between different partners and present the results and success stories to a broad public.

Expected results

The following results are expected from AP8:

- Documented stakeholder feedback (WP8.1)
- Thematic workshops (WP8.2)
- Showroom exhibits (AP8.3)
- Drop-in sessions (AP8.4)
- Closing event (WP8.5)

Work

WP8.1: Collecting and analyzing challenges from the industry

In WP8.1, the challenges of the industry with regard to trustworthy AI systems are collected. This requires collaboration within the
Mission KI project includes researching potential interested parties and compiling corresponding contact lists, as well as making contact and systematically asking about challenges. The challenges are continuously collected and analyzed in an internal report and shared internally within Mission KI. The findings are incorporated into the work on the testing platform and the quality platform. They are also used to design and further develop the range of services, such as setting the topics for the planned workshops (see WP8.2).

WP8.2: Thematic workshops on trustworthy principles

A series of workshopswill be offered , whichaddress selected aspects of trustworthyAI. One possible topic would be the role of "human agency and control" - why this is important and how it can be practically implemented, while also discussing the challenges. The workshops will be attended by our use case partners, but other interested individuals, organizations, companies and Mission AI partners are also invited. The workshops, which will take place at t h e DFKI premises in Kaiserslautern, will be jointly organized by
conceptualized with the other responsible persons of the "Mission AI" project, content-wise

and filled with contributions. DFKI will involve regional and relevant stakeholders from the DFKI network and the surrounding area. In addition, the event will be moderated. The first workshop in the series will be organized as a representative kick-off workshop with high-ranking guests from politics.

WP8.3: Curation of showroom exhibits

The aim of WP8.4 is to identify, set up and curate possible exhibits for the showroom in "42" in Kaiserslautern. If possible, suggestions will also be made for exhibits at other locations.

WP 8.4: TrustifAI consulting

Industry representatives will have the opportunity to book consultation hours at the IQZ Kaiserslautern within the scope of available capacities. The consultation hours are intended to serve as an open exchange with and advice from the IQZ experts and include, among other things, the determination of the respective Challenges in the provision of trustworthy AI applications and their risk class and advice on AI testing.

WP8.5: Final event to present the project results

At the end of the project, we will organize an event where we will introduce the two platforms to various key stakeholders and interested parties and present the case studies. The aim of the event is to share the project results and tell success stories that show how AI systems can be made trustworthy for the benefit of organizations and end users and recipients.

3.2 Resource plan Personnel quantity structure, cost plan

The following table shows the number of full-time researchers (FTE) to be deployed for the project over its duration. The qualifications and resulting costs as well as other costs for the project can be found in the AZK forms and corresponding explanations.

Work package	FTE
AP1 test platform	3
WP2 Quality platform	3
AP3 Use case 1 - Image data for the diagnosis of primary cancer	1,5
WP4 Use case 2 - Tabular data for tumor board decisions	1
WP5 Use case 3 - Spatio-temporal data in rescue services	1
WP6 Use case 4 - Text and survey data in psychiatric treatment and training	0.8
AP7 Use case 5 - Endoscopy videos for safe intubation	0.7
AP8 Mission AI contact point and community	1
Total	**12**

3.3 Milestone planning

Due to the brevity of the project and the necessary flexibility in coordinating the work with the progress of the other responsible parties in the "Mission KI" project, no further overarching interim milestones are currently planned apart from the opening of the IQZ. The milestone plan with interim milestones during the course of the project will be detailed in close cooperation with Mission KI at the start of the project.

Milestone	Time	Contents
MS 1 (AP8)	July 2024	Opening of the Innovation and Quality Center (IQZ) in Kaiserslautern.
MS 2 (AP1)	End of project	Tried and tested PoC of the test platform for representative medical fields of application, recommendations for action for continuation
MS 3 (AP2)	End of project	Tested PoC of the quality platform for representative medical fields of application, recommendations for action for continuation
MS 4 (AP8)	End of project	Community and business offerings established and ready for consolidation

3.4 Data management plan

In our project, we are focusing on the development of testing and quality platforms. Concrete applications are decisive for the usefulness and connection and the further consortium. The data is provided by partners who have experience with data management and who carry this out for the respective application.

We therefore focus on identifying relevant existing data sources, the modalities for accessing this data and ensuring its compatibility with our analysis tools.

4 Utilization plan

4.1 Economic prospects of success

The TrustifAI project aims to create an innovative software platform for the development and verification of trustworthy AI systems in the healthcare sector. In view of the growing importance of AI in medicine and the increasing demands for ethical standards, transparency and reliability, TrustifAI offers considerable prospects of economic success.

The demand for trustworthy AI solutions in the healthcare sector is high and is expected to increase further. Hospitals, medical institutions and pharmaceutical companies are looking for reliable technologies to improve diagnoses, personalize treatments and optimize research and development processes. TrustifAI directly addresses this need and therefore offers significant market potential.

By positioning itself early on as a pioneer in the field of trustworthy AI for healthcare, DFKI is securing a decisive competitive advantage. The combination of highly specialized knowledge in the field of AI and healthcare as well as the focus on ethical aspects and user acceptance creates a unique value that is difficult for potential competitors to imitate.

4.2 Scientific and/or technical Prospects of success with time horizon

The TrustifAI project, which is being carried out by the German Research Center for Artificial Intelligence (DFKI), aims to make significant progress in the development and evaluation of trustworthy AI systems in the healthcare sector over a period of just under two years. The project aims to achieve both scientific and technical milestones within this tight timeframe.

Year 1 - Fundamentals and prototype development: In the first phase, TrustifAI focuses on the design and development of an initial version of the software platforms, which serves as a foundation for the development of trustworthy AI systems (Quality Platform) as well as the verification of the trustworthiness (Test Platform) of AI systems. In coordination with the other people responsible for the Mission AI project, the selected use cases are used to identify those quality features and test criteria that are particularly relevant for AI systems in the healthcare sector and that are accessible for improvements and automated checks.

Year 2 - Tests and quality optimization: In the second year, the focus is on validating and optimizing the developed platforms with the help of the selected use cases. This includes extensive testing in real application scenarios to ensure the practicability and effectiveness of the platforms. The aim is to present a viable concept for both platforms by the end of the project, which promotes and tests the criteria for trustworthiness and is suitable for widespread use in the healthcare sector.

Scientific and technical breakthroughs: Although the project is scheduled to run for less than two years, DFKI expects significant scientific and technical breakthroughs. These include the development of novel methods for establishing and evaluating the trustworthiness of AI systems in the healthcare sector.

4.3 Scientific and economic connectivity

TrustifAI lays the foundation for future research in the field of AI ethics and the development of trustworthy AI systems. The findings and methods developed during the project have been published and provide a valuable basis for academic institutions and research organizations to initiate further research and work. The results of TrustifAI will help provide the research community with much-needed data, tools and best practices that improve the quality, transparency and ethical alignment of AI systems.

By developing a platform that supports the development process of trustworthy AI systems, TrustifAI addresses a critical need within the healthcare sector and other industries that rely on AI solutions. The provision of such tools opens up new business opportunities for technology providers, consulting firms and AI developers. In addition, the project promotes the emergence of spin-offs and start-ups specializing in the implementation and adaptation of the TrustifAI platform, thus directly contributing to the economic dynamics in the field of AI technologies.

Although TrustifAI is primarily aimed at the healthcare sector, the principles of trustworthiness are universal and can be transferred to other areas in which AI systems are used. This opens up perspectives for the application of the project results in sectors such as financial services, transportation or education, where there are similar requirements for the security, transparency and ethical orientation of AI systems.

Promotion of norms and standards: Another aspect of connectivity concerns the development of norms and standards for trustworthy AI. TrustifAI can serve as a catalyst for the establishment of cross-industry standards in high-risk applications of AI, which not only improve their quality and safety, but also strengthen regulatory acceptance and consumer confidence.

4.4 Social added value

By ensuring the trustworthiness of AI systems in healthcare, TrustifAI contributes directly to improving the quality and safety of patient care. Trustworthy AI can help to make diagnostic procedures more precise, personalize treatment methods and ultimately increase the chances of recovery. This leads to a noticeable improvement in the quality of life for patients.

The central aim of TrustifAI is to strengthen trust in AI applications. By creating transparent, ethical and reliable AI systems, the project will make a significant contribution to increasing social acceptance of these technologies. Broader acceptance of AI can in turn promote innovation in various sectors and thus accelerate social progress.

TrustifAI sets new standards in terms of the ethical design and use of AI systems. By developing guidelines and best practices for AI ethics, the project makes an important contribution to the social debate on technology and morality. This promotes responsible behavior in dealing with AI and raises awareness of the social implications of technological developments.

4.5 Data-related utilization of results

Although the TrustifAI project does not generate its own data, but focuses on the development of findings, concepts and possibly software, the responsible handling of the information obtained through these processes is of central importance. The utilization of results therefore includes the following aspects:

TrustifAI aims to make the findings and concepts developed as part of the project accessible to the public. This is done through publications in scientific journals, presentations at conferences and workshops as well as through open access materials. The aim is to share the knowledge and approaches developed with the scientific community and industry in order to promote innovation and further research in the field of trustworthy AI.

If specific software solutions are developed as part of TrustifAI, a strategy for licensing and distributing these tools will be pursued. Particular emphasis is placed on open licenses to enable broad use and further development of the software by the research and development community. This promotes transparency and collaboration in the field of AI development.

5 Division of labor / cooperation with third parties

All work is carried out in close cooperation with acatech - National Academy of Science and Engineering. In addition, we are working with three partners - the German Red Cross Rhineland-Palatinate, Saarland University Hospital and the Psychiatric Clinic of Trier University.

6 Necessity of the grant

The realization of the TrustifAI project of the German Research Center for Artificial Intelligence (DFKI) is of crucial importance for the further development of trustworthy AI applications in the healthcare sector and beyond. Despite the high relevance and potential of AI in this area, trust in these technologies is still limited among users, as the latest survey results from Appinio make clear. This underlines the urgent need to establish standards and practices for the development and evaluation of AI systems that meet the criteria for trustworthiness.

The particular challenges of implementing trustworthy AI in healthcare require extensive research and development work that cannot be realized without adequate financial support. DFKI has the necessary expertise and capacity to lead and implement this ambitious project, but the investment required is substantial and cannot be met from its own resources. Funding through the grant is therefore essential.

The funding of TrustifAI will make an important contribution to strengthening Germany as a location for innovation in the field of AI research and application by laying the foundations for a globally leading standard in the development of trustworthy AI systems in a high-risk application area of AI ("AI made in Germany"). The investment in this project is an investment in the future of further application areas of trustworthy AI "Made in Germany" and thus in Germany's technological sovereignty and economic competitiveness.