From Diagnosis to Therapy: AI in Medical Imaging

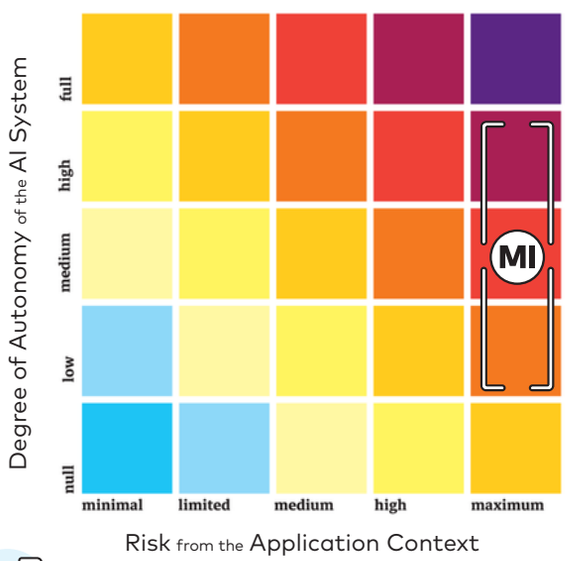
Opportunities and Risks of AI-based Diagnostic and Therapeutic Support in Medicine

Executive Summary

In the healthcare system, the advanced capabilities of AI systems to recognize and classify images are used for diagnosis and treatment. For example, image recognition is used to evaluate retinal scans in ophthalmology, to analyze digitized tissue sections in pathology, and to detect tumors such as breast cancer. In the present ALAIT Risk Radar, the use of this type of AI application in medical imaging is classified as having a relatively high overall risk, as indicated by its placement in the red area of the graphic. However, healthcare providers can reduce this risk by implementing the risk assessment and reduction recommendations contained in this dossier (see p. 4).

In detail, the classification of AI applications in the ALAIT risk radar takes two dimensions into account:
1) The risk arising from the application context: Here, there is a „maximum risk" (level 5), as faulty systems and applications can have life-threatening consequences for affected individuals. Additionally, according to the EU AI Act Annex I, medical devices are generally classified as high-risk. However, until the European AI Regulation comes fully into force in mid-2027, AI-based medical devices are not yet required to meet specific quality requirements. This lack of transparency and clarity currently makes it even more difficult for healthcare providers to assess the performance of these tools and thus makes their use particularly risky.
2) The second dimension assesses **the degree of autonomy of AI applications in image-based diagnostic and therapeutic support.** This varies depending on the level of human control over specific applications. If AI systems largely automate diagnoses and/or therapy suggestions, which are only reviewed by medical professionals in defined cases, the degree of autonomy is considered high („Human in Control," e.g., if, within the framework of AI-supported breast cancer screening, only unclear or critical cases are evaluated by medical professionals). While this is technically feasible, such a high degree of automation is not possible in the Austrian context due to the final decision being made by humans, e.g., physicians during diagnosis. The degree of autonomy is low when AI systems serve as supporting tools for diagnosis and/or therapy planning („Human-in-the-Loop" or „Human-on-the-Loop"), for example, when medical professionals have case-specific image analyses performed.

ALAIT risk radar for image-based diagnostic and therapeutic support



MI AI applications in Medical Imaging

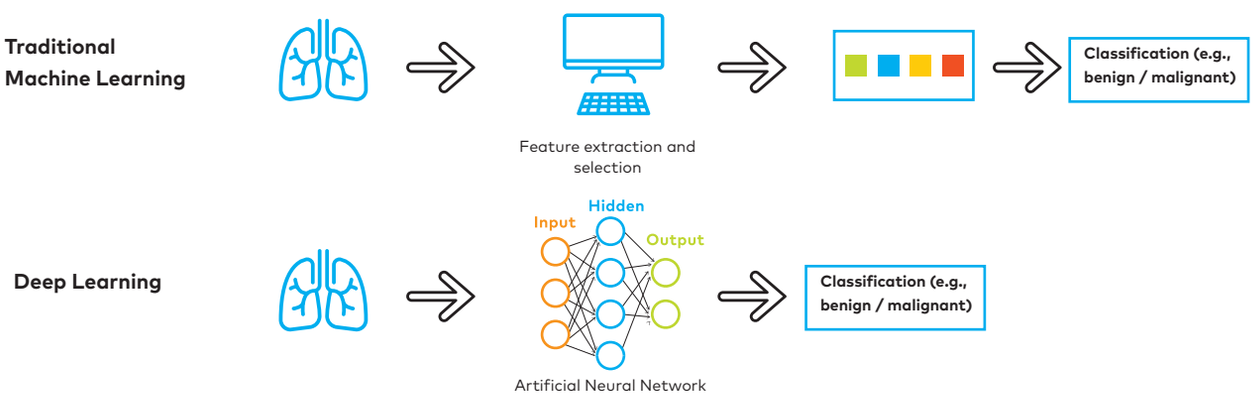
The ALAIT Risk Radar is a scientifically developed risk analysis tool for artificial intelligence (AI) that classifies AI applications contextually and according to their degree of technical autonomy, thus making the risk level visible to users at a glance. The principle is: the higher the deployment risk arising from the application context and the greater the degree of autonomy of the AI system with regard to decisions, the riskier the deployment is classified. An extended bracket indicates a range in the risk classification. A low degree of autonomy in an AI system does not mean that one can sit back and relax. It requires a strong role from the people using it. (For details on the stage model, see pp. 7-8).

**ALAIT is a research project initiated by the Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI) to develop a novel form of technology assessment for AI technologies using scientific methods.*

Computer Vision: The Road to Machine Sight

Modern medical diagnostics and therapy increasingly rely on AI. Particularly in the field of image-based diagnostic and therapeutic support, significant technologi-cal advances have been made in recent decades.[1] Current scientific studies suggest that the use of AI-based tools for image and video analysis in radiology, patholo-gy, dermatology, or surgery could lead to more accurate, and potentially even earlier, diagnoses or more individualized treatment approaches, with the expectation of more precise diagnoses.[2,3,4]

Technically speaking, applications in this area belong to the field of computer vision; that subfield of AI which deals with recognizing, understanding, and interpreting images and videos – much like humans do with their eyes and brain. AI systems used in image-based diagnostic and therapeutic support process image data in various steps, as schematically illustrated in the following graphic:[5,6,7]



In the automated analysis of images/videos, a primary distinction is made between traditional machine learning and deep learning. In traditional machine learning, relevant features (e.g., shape, color, texture) are manually extracted from the raw data—that is, by humans—and provided to the algorithm. This feature selection is usually rule-based and requires expert knowledge. Afterward, an algorithm can take over the actual learning task, for example, the classification „tumor/no tumor" within the framework of *supervised learning*. Here, the model trains with pre-labeled data where the correct answer is known (e.g., „tumor/no tumor"). In contrast, *unsupervised learning* works with unlabeled data: The model independently searches for patterns or groups, for example, by grouping similar cell types into clusters, without prior instruction. Because traditional machine learning clearly identifies which features led to the prediction, these systems are usually easier to explain and decisions are more comprehensible.

Deep learning, on the other hand, is based on arti-ficial neural networks and learns relevant features and patterns directly and automatically from the raw data – without requiring explicit human defi-nition beforehand. This can offer efficiency advan-tages, especially with unstructured data such as images. Deep learning can also be used in a super-vised manner when the raw data is already labe-led (e.g., classification with pre-labeled images) or unsupervised when the raw data is unlabeled (e.g., automatic clustering). In any case, deep learning requires large amounts of data and significant computing power. The decision-making processes of these models are often difficult to understand, which is why they are referred to as *black-box systems*: It is usually unclear why the model made a particular decision (e.g., classifi-cation), which limits explainability.

Opportunities

The application of AI in image-based diagnostic and therapeutic support offers great opportunities for patients, healthcare professionals, and the healthcare system, and is expected to have positive effects on these groups.2,8,9,10 For patients, the focus is primarily on improved medical services, while for healthcare professionals and administrators, more efficient and accelerated processes are expected.

Opportunities for Patients

- **Early detection of diseases at often earlier stages than with traditional methods** (e.g., tumors and diabetic retinopathy). Furthermore, the number of missed breast cancers was reduced by 26% through AI assistance.
- **Personalized diagnostics,** e.g., support in predicting the individual course of the disease.

Opportunities for Medical Personnel

- **Accelerating diagnosis** through rapid image and video analysis and prioritizing urgent cases with systems like CHIEF, which have an accuracy of approximately 94%.[12]
- **Creating time resources within the workforce.**

Opportunities at the Administrative Level

- **Reducing the workload of medical staff** (radiologists, pathologists, etc.) by up to 62% through the automated analysis of large volumes of imaging data.[13]

Challenges and Risks

Despite the opportunities mentioned above, there are a number of detailed challenges that need to be overcome for a trustworthy and secure integration of image-based AI diagnostic and therapeutic systems into medicine:

1. Accuracy and reliability of the results:

The performance of AI imaging systems in clinical practice is not always guaranteed. This means there is a risk of false results. This is partly because the accuracy achieved in pre-market tests can only be realized if the patient population in the respective field of application matches the dataset used to train the AI. If patient cha-racteristics, such as gender, age, ethnicity, etc., differ significantly from those in the training data, concerning conditions indicative of disease may go undetected, or healthy conditions may be incorrectly classified as cause for concern. To avoid this data shift and the resulting false results, specific tests within the respective organization are necessary before deployment. [11]

2. Bias:

Even a suitable, large, and high-quality dataset does not guarantee that AI is free from *systematic bias* and performs equally well for everyone. In general, it has been shown that AI systems in medical imaging very often produce incorrect results for individuals who are underrepresented in medical data, belong to risk groups who use the healthcare system less frequently, or for people who have no access to the system. For example, these might be women or children, migrants, or people with darker skin. However, in principle, anyone can be affected by bias.[8]

3. Transparency:

AI systems are often not very transparent due to the machine learning methods they employ (see *supervised and unsupervised learning*) and are therefore also referred to as *black-box systems*.[3] While manufacturers of AI systems must comply with transparency requirements and operators must ensure that users, such as doctors and healthcare professionals, possess the necessary AI literacy to use the AI, this is often insufficient to fully understand the output of AI systems. This lack of transparency stems not only from the AI's learning methodology but also from trade secrets and intellectual property protection (i.e., a conscious decision by the manufacturers) and the GDPR (confidentiality of training data).

4. Privacy and Security:

Protecting patient *privacy* is always paramount and must be guaranteed at all times when using AI systems. To date, many AI systems operate in the cloud (and not on the organization's own servers – on-premises), which increases vulnerability to data breaches and security attacks.[1,2,11]

Recommendations for the practical application of AI in medical imaging:

Six key points

Despite technological leaps and significant opportunities for patients and the healthcare system, healthcare providers require specific AI expertise and accompanying risk management measures to use AI applications for image analysis appropriately and responsibly. It is essential that such AI systems are tested before and du-ring deployment and that accompanying measures are implemented within the organization. The following six points are recommended to ensure the safety and ac-ceptance of all involved:

1. Careful selection of the AI application and demand for transparency from the provider:
AI systems in medical imaging are not all-rounders, but are trained on very narrowly defined, specific tasks. Therefore, essential information must be verified before deployment (see points a-d). This allows for initial conclusions to be drawn as to whether the AI system is fundamentally applicable in the intended field. This information will be easier to obtain from mid-2027 onwards, as the EU AI Act, with its transparency obligations for medical devices, will then be fully applicable.

a) The Basic Funtionality (Intended Purpose):
- What is the system optimised for?
- Which technologies are used?
- What unput data does the application work with?

b) The Training and Test Data used
including a description of the patient populations from which they are taken (since these should be as similar as possible to the patient populations in the intended area of application)
- What standards and metrics were used to test the population, and what were the results?

c) Known limitations, e.g., regarding target groups for whom the application does not work reliably

d) Approval as a medical device in the EU (if it is a system from, for example, the USA) and any additional standards that are met (will be particularly important in the future).

2. Assessment of performance under real-world conditions (AI model performance clinical performance):
The performance of AI systems can vary considerably in clinical practice. For this reason, testing AI systems in a real-world environment using input data from actual patient populations (in hospitals, laboratories or other healthcare facilities) is an important prerequisite for being able to assess their performance and the risk of false positives more accurately. Such tests should be carried out.

3. Ensure that the final decision is taken by people with the relevant expertise:
Although AI provides valuable support, it must be controlled and monitored by humans at all times due to technical imperfections (see 'black-box' systems) and legal requirements (see the EU AI Act). Doctors must continue to bear responsibility for diagnoses and treatment decisions. At the same time, this entails an internal organisational mandate to draw up clear guidelines on the use of AI (organisational AI guidelines) and to provide training and further education for (medical) staff (keyword: AI literacy), so that AI results can be critically scrutinised and interpreted (see next point). It is worth noting here that a system has recently been developed to help healthcare professionals correctly contextualise and interpret AI results. MONET increases the transparency of results from certain AI applications for decision-makers through, amongst other things, data verification, model validation and model interpretation.[4]

4. Building specific skills among users (AI literacy):

The successful integration of AI into image-based diagnostic and therapeutic support requires the training and continuing professional development of medical staff to provide a general understanding of the capabilities and limitations of AI systems, their strengths and weaknesses, and the correct interpretation of AI-assisted diagnoses. Staff must be able to make informed decisions and feel confident in their ability to challenge the AI application in the event of incorrect results.

5. Informing patients about the use of AI:

The acceptance and safety of AI-assisted image analysis depend crucially on its transparency and continuous monitoring (see next point). Patients must be informed when AI is used in diagnosis or the interpretation of findings. This can be done through explanatory consultations, written information or specific consent forms. Furthermore, it must be clearly communicated that the ultimate responsibility lies with healthcare professionals.

6. Ongoing supervision and monitoring of the AI system:

Continuous monitoring of AI performance is essential to ensure quality in the long term and to minimise risks on an ongoing basis. This involves the regular analysis of incorrect decisions, adapting algorithms and models to new findings through continuous training with new datasets, and implementing a reporting system for medical staff to identify and resolve problems at an early stage.

Important Terms

Black-box systems: This term refers to AI systems whose internal decision-making processes are neither transparent nor comprehensible to humans. This means that only inputs and outputs can be observed, without understanding the precise processing that occurs in between.

Data protection in AI: The entirety of practices and concerns related to the ethical collection, storage and use of personal data by artificial intelligence systems.

EU AI Act (European Union Act on Artificial Intelligence): The European regulation on artificial intelligence (AI) – the first ever comprehensive regulation on AI issued by a major regulatory authority. It focuses in particular on high-risk AI systems.

AI autonomy: The ability of an AI system to achieve a set of objectives, amidst a range of uncertainties in its environment, independently and without external inter-vention.

AI accuracy: Refers to an AI system's ability to make correct predictions or decisions. It is an important measure of its performance and is crucial for determining its effectiveness and reliability.

Systematic bias: Bias is the systematic difference in the treatment of certain objects, individuals or groups compared with others. Treatment refers to any kind of action, including perception, observation, representation, prediction or decision-making.

Transparency: This means that the functionality, decision-making processes and areas of application of an AI system are comprehensible, explainable and openly ac-cessible – for developers, users and other stakeholders.

Supervised learning: A machine learning method in which the training data is pre-labeled and the system learns the pattern between the image and the label. The task of such an AI system is to find a relationship that maps each input of the training set (the data) to an output (the label).

Unsupervised learning: A subfield of machine learning in which an algorithm recognizes patterns, structures, or relationships in unlabeled data without being told what is right or wrong.

Explanation of the ALAIT Risk Radar's Tiered Model

The ALAIT Risk Radar illustrates the relationship between operational risk within the scope of application and the decision-making autonomy of an AI. Lower autonomy and lower application risk are indicated by cooler colours (blue and yellow), whilst higher autono-my and application risk are indicated by warmer co-lours (orange to violet). Ideally, AI applications that are classified in the highest risk categories on both axes should only be deployed after very careful consideration or avoided altogether. AI applications that are placed in the highest category on only one axis (high appli-cation risk and low autonomy, or vice versa) represent an overall medium level of risk. The risk categories are based on the EU AI Act, in particular Article 6 and An-nex III, which deal with high-risk areas of AI application.

Autonomy

Level 1: No Autonomy

AI is a passive tool; humans make all the decisions and take the necessary action.

Example: Diagnostic systems that display raw medical data or analyse the data (without making recommendations)

Recommended use cases: scenarios involving high risks or where ethical decisions are of great importance (e.g. medical diagnostics, the justice system).

Level 2: Low Level of Autonomy (Human-in-the-Loop)

The AI provides recommendations or options, but users remain responsible for selecting and approving actions (human-in-the-loop).

Example: AI suggests optimal routes for logistics or recommendation systems in e-commerce.

Recommended use cases: tasks of moderate complexity involving moderate risks (e.g. supply chain optimisation).

Level 3: Mid-level Autonomy (Human-on-the-Loop)

The AI provides recommendations or options, but users remain responsible for selecting and approving actions

Example: AI-supported manufacturing processes in which the system controls the machines, but operators intervene in the event of anomalies.

Recommended use cases: Scenarios where continuous human involvement is not required, but critical risks do require human monitoring (e.g. industrial automation, monitoring of financial transactions).

Level 4: High Degree of Autonomy (Human in Control)

The AI system operates largely autonomously, but allows users to override it themselves in order to avoid undesirable results.

Example: Autonomous vehicles

Recommended use cases: Low- to medium-risk environments (e.g. logistics, basic traffic management).

Level 5: Full autonomy with minimal supervision

The AI system operates independently and requires minimal or no human intervention.
Human involvement is limited to long-term oversight (audits).

Examples: Autonomous agricultural machinery, AI for electricity grid distribution, underground trains, airport shuttle trains

Recommended use cases: Environments with low safety or ethical risks and high reliability of the AI system (e.g. repetitive tasks in controlled environments).

Scope of Application Risk

Stage 1: Minimal Risk

The AI system has no impact on the user or the decision-making process.

Examples: Filters, NPCs in computer games, recommendationalgorithms without serious consequences (DeepL, other translation tools).

Criteria: No direct impact on the high-risk areas of the EU AI Act.

Stage 2: Limited Risk

AI systems that interact with users but do not make high-risk decisions. The risk increases when there is a lack of transparency regarding the involvement of AI.

Examples: AI systems that interact with users but do not make high-risk decisions. The risk increases when there is a lack of transparency regarding the use of AI.

Criteria: Areas not included in the 'high-risk' list set out in the EU AI Act.

Stage 3: Medium Risk

Artificial intelligence systems do not have any particular impact on individuals; rather, their effects are felt at a collective or societal level.

Examples: Generative AI such as ChatGPT and other systems that can indirectly influence the environment in which they are used and may ultimately lead to major changes in society and people's lives.

Criteria: AI systems that are available for public use and have the potential to influence and, in the long term, change existing practices.

Stage 4: High Risk

Any algorithm used in what the EU AI Act defines as 'high-risk areas' or which has a direct impact on the life and limb of individuals.

Examples: Medicine, biometrics, critical infrastructure, education and vocational training, employment, access to services, law enforcement, migration.

Criteria: Classification as a 'high-risk sector' under the EU AI Act is only applicable if the rules on transparency and data quality are complied with.

Stage 5: Extreme Risk

Any algorithm used in what the EU AI Act defines as 'high-risk areas'.

Examples: Medicine, biometrics, critical infrastructure, education and vocational training, employment, access to services, law enforcement, migration.

Criteria: Classification as a 'high-risk area' if the rules on transparency and data quality are NOT complied with.

Endnotes

1 Barragán-Montero, A., Javaid, U., Valdés, G., Nguyen, D., Desbordes, P., Macq, B., Willems, S., Vandewinckele, L., Holmström, M., Löfman, F., Michiels, S., Souris, K., Sterpin, E., & Lee, J. A. (2021). Artificial intelligence and machine learning for medical imaging: A technology review. Physica Medica, 83, 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

2 Herington, J., McCradden, M. D., Creel, K., Boellaard, R., Jones, E. C., Jha, A. K., Rahmim, A., Scott, P. J. H., Sunderland, J. J., Wahl,R. L., Zuehlsdorff, S., & Saboury, B. (2023). Ethical Considerations for Artificial Intelligence in Medical Imaging: Data Collection, Development, and Evaluation. Journal of Nuclear Medicine, 64(12), 1848–1854. https://doi.org/10.2967/jnumed.123.266080

3 Castiglioni, I., Rundo, L., Codari, M., Di Leo, G., Salvatore, C., Interlenghi, M., Gallivanone, F., Cozzi, A., D'Amico, N. C., & Sardanelli, F. (2021). AI applications to medical images: From machine learning to deep learning. Physica Medica, 83, 9–24. https://doi.org/10.1016/j.ejmp.2021.02.006

4 Kim, C., Gadgil, S. U., DeGrave, A. J., Omiye, J. A., Cai, Z. R., Daneshjou, R., & Lee, S.-I. (2024). Transparent medical image AI via an image–text foundation model grounded in medical literature. Nature Medicine, 30(4), 1154–1165. https://doi.org/10.1038/s41591-024- 02887-x

5 Boesch, G. (2023, January 20). What is Computer Vision? The Complete Guide for 2025. Viso.Ai. https://viso.ai/computer-vision/what-is-computer-vision/

6 Szeliski, R. (2022). Computer Vision: Algorithms and Applications. Springer Nature.

7 Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. Com- putational Intelligence and Neuroscience, 2018(1), 7068349. https://doi.org/10.1155/2018/7068349

8 Leimüller, G., & Wazir, R. (2025). Die Sache mit dem Bias in der Künstlichen Intelligenz.

9 Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., & Pattichis, C. S. (2020). AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE Journal of Biomedical and Health Informatics, 24(7), 1837–1857. IEEE Journal of Biomedical and Health Informatics. https://doi.org/10.1109/JBHI.2020.2991043

10 Khalifa, M., & Albadawy, M. (2024). AI in diagnostic imaging: Revolutionising accuracy and efficiency. Computer Methods and Pro- grams in Biomedicine Update, 5, 100146. https://doi.org/10.1016/j.cmpbup.2024.100146

11 Tang, X. (2020). The role of artificial intelligence in medical imaging research. BJR|Open, 2(1), 20190031. https://doi.org/10.1259/bjro.20190031

12 Wang, X., Zhao, J., Marostica, E. et al. (2024) A pathology foundation model for cancer diagnosis and prognosis prediction. Nature 634, 970–978 . https://doi.org/10.1038/s41586-024-07894-z

13 Chen, M., Wang, Y., Wang, Q. et al. (2024). Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation. https://doi.org/10.1038/s41746-024-01328-w

About the ALAIT Project

The Austrian Lab for AI Trust (ALAIT) is a research and development project initiated by the Austrian Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI) to build trust through knowledge in the field of artificial intelligence (AI). The ALAIT project aims to empower interested individuals and key societal groups to use AI technologies responsibly and to establish ethical and high-quality standards for the application of AI.

The project is led by **winnovation** (Gertraud Leimüller and Lena Müller-Kress) and implemented by a consortium including **leiwand.ai** (Rania Wazir and Silvia Wasserbacher-Schwarzer), **TU Vienna** (Sabine Köszegi and Ilya Faynleyb) and **Austria Press Agency – APA** (Verena Krawarik and Sophia Marecek).

The ALAIT dossiers are available on the project homepage: https://science.apa.at/project/alait/

The contents of this dossier reflect the current state of the art and have been carefully compiled according to scientific criteria. However, they do not constitute legally binding information or advice.

Imprint

Media owner and publisher:
winnovation consulting gmbh
Linke Wienzeile 124/5
1060 Vienna

Published in 2025, funded by the Austrian Federal Ministry for Innovation, Mobility and Infrastructure.