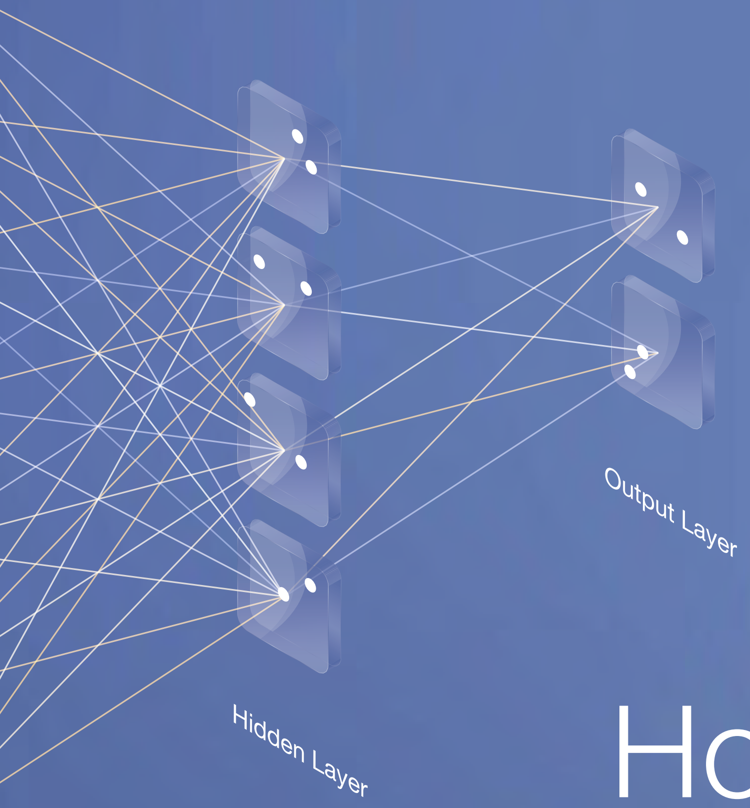
Link
www.aim-lab.io



How Medical AI Can Become **Trustworthy**

AI systems in the health sector should be ethically above reproach and as trustworthy as possible. A research group working with computer scientist Prof. Daniel Rückert is developing methods by which privacy can be maintained with AI applications – with mathematical certainty.

Gesamter Artikel (PDF, DE): www.tum.de/faszination-forschung

Wie medizinische KI vertrauenswürdig wird

D

KI-Systeme in der Medizin müssen vertrauenswürdig sein. Sie sollten wie eine menschliche Ärztin zuverlässig und fair agieren und die Privatsphäre von Patienten achten. Das Forschungsteam um Prof. Daniel Rückert untersucht, wie die Trainingsdaten von Patientinnen sicher geschützt werden können und eine „Privatsphäre wahrende KI“ möglich ist. Das Team hat gezeigt, dass Differential Privacy mathematische Garantien für die Privatsphäre gibt – und sie weder durch aktuelle noch durch zukünftige Angriffe unterminiert werden kann. Diese Garantien sind umfassend und unabhängig vom Stand der Technik. □

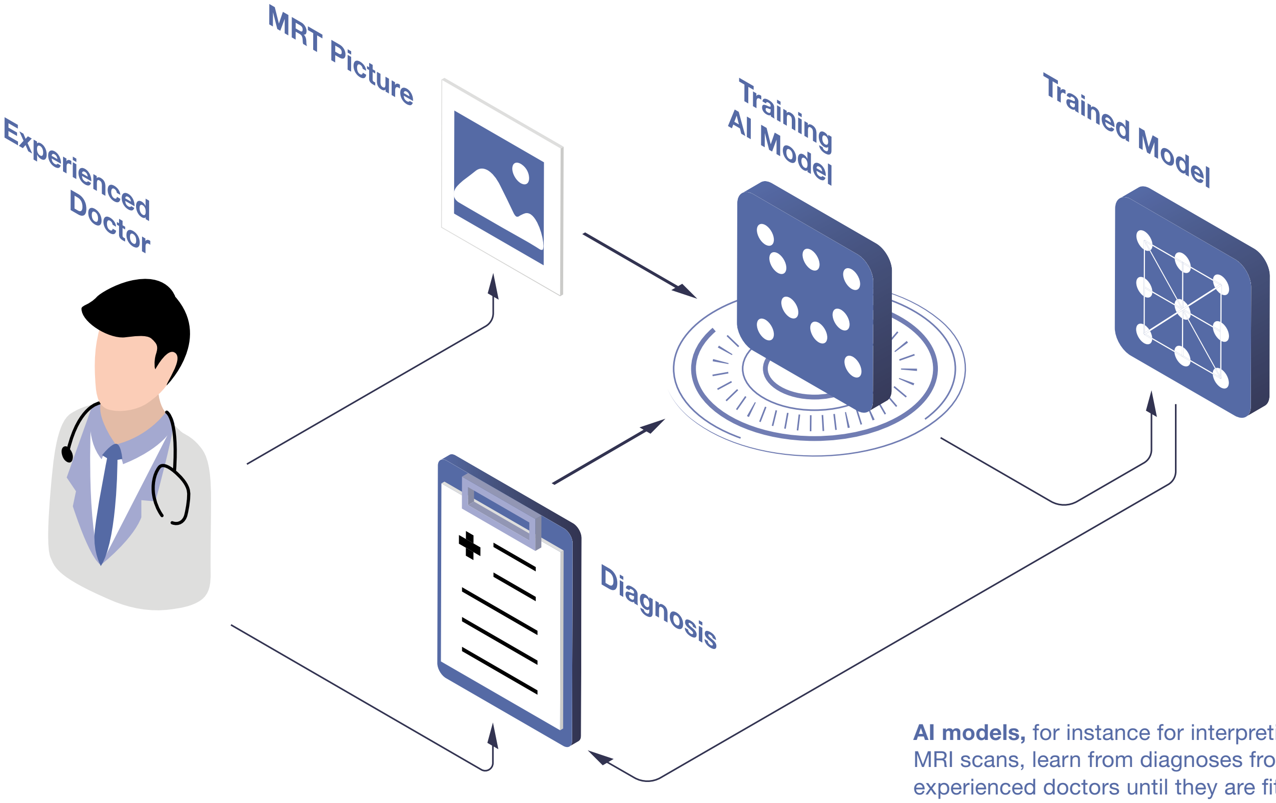
Graphics: ediundsepp

> *“The requirements for AI systems are high. They should handle patients’ personal data with care and not store any identifiable information.”*
>
> *Daniel Rückert*

Prof. Daniel Rückert

has been active as the Alexander von Humboldt Professor for AI in Medicine at TUM since 2020. He is also a professor at Imperial College London. Rückert studied computer science at TU Berlin (1993) and then earned his doctorate at Imperial College followed by a postdoc at King’s College London. In 1999, he became Assistant Professor at Imperial College. He has held the Chair for Visual Information Processing at Imperial College since 2005 and was also the Dean there from 2016 to 2020.
At TUM, Prof. Rückert heads up the Center for Digital Medicine and Health. Daniel Rückert works in the field of artificial intelligence (AI) and machine learning and their applications in medicine. In his research, he focuses on the development of innovative algorithms for acquiring, analyzing and interpreting images and, with respect to AI, on the extraction of clinical information from medical images – particularly for computer-aided diagnosis and prognosis.

Artificial intelligence (AI) is in the process of changing medicine with intelligent systems. Most AI applications are based on models for machine learning. These models are trained to recognize certain patterns on the basis of patient data. The more data are incorporated into the training, the more accurate the diagnoses and prognoses become.
In medicine, such AI systems are now supporting doctors very successfully in the diagnosis and treatment of illnesses, the analysis of X-ray images and many other medical fields besides. But the rapid development in this area also raises questions of a fundamental nature: Are the AI systems as reliable as a human doctor? Can medical users trust them? And are the patient data used to train the model handled with care?



AI models, for instance for interpreting MRI scans, learn from diagnoses from experienced doctors until they are fit to detect diseases on their own.

Picture credit: Juli Eberle; Graphics: ediundsepp (source: D. Rückert)

Computer scientist Daniel Rückert from TUM is working on making automatic systems as trustworthy as a human doctor – an essential factor for the acceptance of such programs: “In medicine, we have two groups of people with whom an AI system interacts,” Daniel Rückert explains. “Doctors and clinicians make up one group, and patients the other. Both groups have very high requirements in terms of the quality of the decision-making processes.” AI systems also have to meet these requirements. For example, they need to handle patients’ personal data with care and not store any identifiable information – in other words, safeguard their privacy. They should be fair and treat men the same as women, for example. And they should state how certain their decisions are. Because, just like a human doctor, an AI system will be able to make some diagnoses with 99 percent certainty but others perhaps with only 80 percent. And the system has to communicate such figures as transparently as possible.

“Generally speaking, there are many definitions and categorization approaches for trustworthy AI,” says Dr. Georgios Kaissis from Rückert’s team. The consensus is that intelligent systems in medicine should act in the same way, in the widest sense, as a responsible doctor. “Trustworthy AI must be compatible with human values,” says Kaissis. “The output from such systems should not conflict with basic human values such as fairness or the protection of data.” ▷

The data protection dilemma

Among other things, Daniel Rückert and his research groups are focusing on the topics of fairness and transparency – with the emphasis currently on privacy-preserving AI. Assistant Professor Georgios Kaissis is leading the research group on this topic. The question the radiologist and computer scientist is grappling with is this: How can you train AI models with patient data without enabling such data to be reconstructed from the models?

The relevance of this question must not be underestimated. Fundamentally, patient data such as MRI scans, for example, are essential for training AI models. However, these patient data are problematic for two reasons. Firstly, such data are not available in medicine in the same quantity as for non-medical AI applications – where millions if not billions of training datasets are often used. Here, one has to make do with fewer – which can limit the reliability of the models and diagnoses.

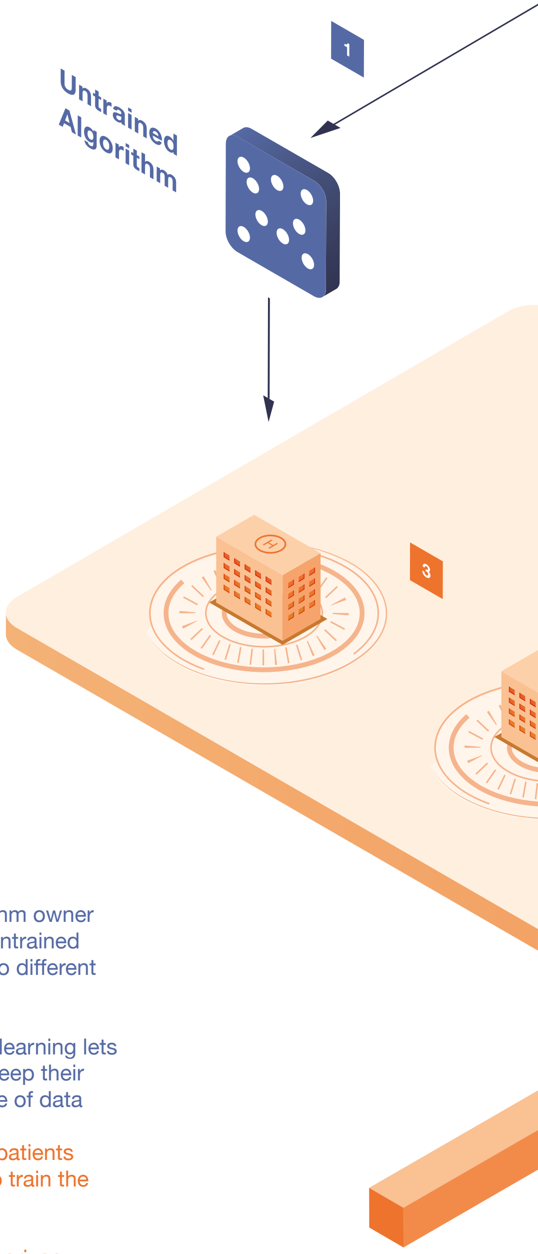
Secondly, the medical data used for training purposes are highly sensitive and very much in need of protection. After all, illness is a private affair – as a matter of principle, medical practitioners must not disclose such data without the consent of the data subjects, not even to train a computer system that may be able to save lives in the future.

Both challenges – not enough data and highly sensitive data – can be solved by reliable privacy protection. Anonymization and pseudonymization have largely established themselves as techniques for providing adequate protection for such data. In the case of anonymization, the names or identifying information are completely removed from the dataset. Bob Dylan's "Greatest Hits" album can be anonymized by deleting the name, with the result that the dataset now only contains the entry "Greatest Hits". With pseudonymization, the name "Bob Dylan" is replaced by a different name such as "Bob Marley".

The snag is that anonymization and pseudonymization are now no longer secure. The ways and means to attack AI models have become so powerful that even very well anonymized data can be re-identified with relative ease.

▷

Rückert and his team rely on differential privacy and federated learning to keep health data used to train AI models private. Differential privacy adds a calibrated statistical noise to protect sensitive data. In federated learning, the AI model is successively sent to individual hospitals instead of sending sensitive data to a central server. This means that control of the data remains in the respective hospital.



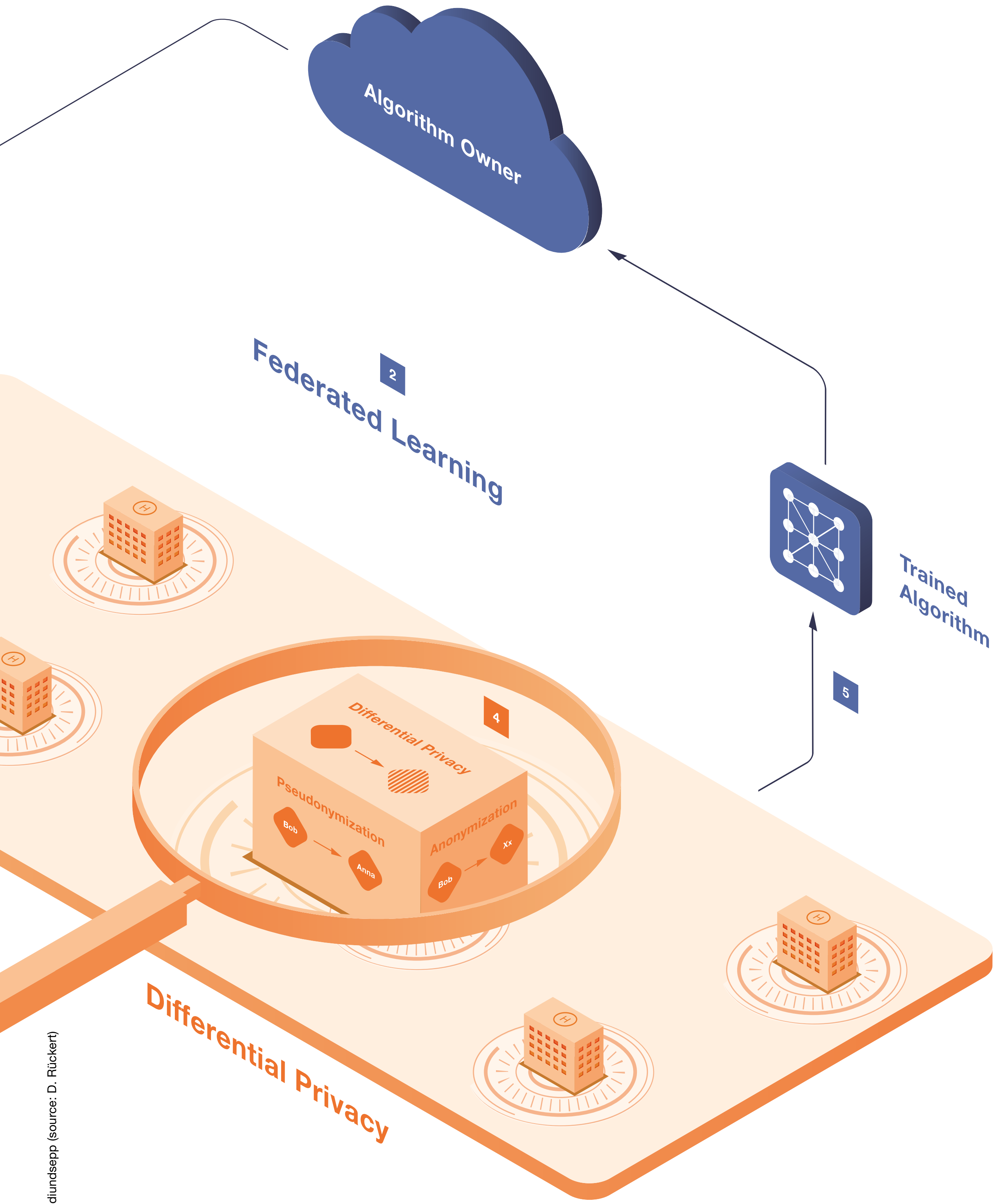
1 The algorithm owner sends an untrained algorithm to different hospitals

2 Federated learning lets hospitals keep their governance of data

3 Data from patients are used to train the algorithm

4 Differential privacy guarantees data security now and in the future, independent of the state of technology

5 The trained algorithm is then sent back to the algorithm owner



Graphics: ediundsepp (source: D. Rückert)

Rückert and his team recently showed that data that flow into AI models are effectively protected by differential privacy.

“The mere removal of the name is completely meaningless for the latest hacking techniques,” Georgios Kaissis explains. “We were able to show multiple times in our work that patient data can be reconstructed from the models if you use them for training purposes without any additional protective measures.” For example, Kaissis and his staff succeeded in completely reconstructing patients’ X-rays from the models – a disaster for data protection.

Nevertheless, anonymization and pseudonymization continue to be used in practice. “That is down to the discrepancy between the state of research results and the legal framework,” Kaissis says. “Legally, anonymized data are still not considered personal data and are therefore legally permissible. However, research shows that anonymization is not secure.” The legal framework would therefore need to be amended.

Besides protecting sensitive data, any AI that safeguards privacy can also solve the problem of insufficient quantities of data – even if only indirectly. AI systems that preserve privacy are trustworthy for users and data providers alike and therefore have a highly motivating effect on patients, with the result that they approve the use of their data. More training data then become available, making the models more reliable and robust.

Interdisciplinary research: Center for Digital Medicine and Health (ZDMG)

Prof. Daniel Rückert heads up the Center for Digital Medicine and Health (ZDMG), for which TUM has received 43 million euros from the federal government and from the Free State of Bavaria. The intention is for researchers in medicine, computer science and mathematics to develop new approaches together in the fields of data science and artificial intelligence and drive their clinical application. Thanks to the targeted inclusion of expertise from the natural sciences and engineering disciplines, the development of innovative methods and technologies in the fields of AI and data science will be made usable for various medical applications at the new interdisciplinary research center.

Mathematical guarantees

Daniel Rückert and his team working with Georgios Kaissis are using a technique by the name of differential privacy that overcomes the limitations and lack of security of anonymization and pseudonymization. Differential privacy is essentially based on the fact that when training the AI systems, “calibrated statistical noise” – i.e. random noise – is added to the data. The whole method is mathematically complex but the result is that the privacy of individual patients is guaranteed.

The major benefit of differential privacy is that, in contrast to traditional techniques, this method offers a mathematical guarantee that it cannot be undermined by either current or future attacks. While an empirical guarantee only ensures that a current attack will be repelled, it cannot be ruled out that a future attack might evade this guarantee. A mathematical or formal guarantee, on the other hand, is a guarantee that privacy can never be circumvented, either now or in the future. This formal guarantee is significantly stronger than a merely empirical one – it is comprehensive and does not depend on the state of technology. “If I want to convince a data protection officer from Klinikum rechts der Isar to allow me to use such methods, it’s naturally much more appealing to them if I can tell them that I can mathematically guarantee that it will never be possible to re-identify the patient from such data,” says Rückert.

▷

PD Dr. med. Georgios Kaissis, MHBA

is the leader of a working group at the Institute of Artificial Intelligence and Computer Science in Medicine and Senior Physician at the Institute for Radiology at TUM as well as the leader of a working group at the Helmholtz Center in Munich. He researches in the field of privacy-preserving, trustworthy artificial intelligence, particularly on the subject of differential privacy as well as on applications in the fields of medicine and biomedical imaging.

Picture credit: Juli Eberle, Andreas Heddergott

The "holy trinity" – algorithmic privacy

Three methods have established themselves for the protection of sensitive data – under the heading of "algorithmic privacy".

Federated learning

With federated learning, the data are not brought to the algorithms but the algorithms to the data. The model to be trained is moved to the hospital, trained in the hospital using the data available there and then returned to be further trained with data from a different hospital. The advantage here is that the data never have to be released from the custody of the hospital. The disadvantage is that hackers might be able to simply copy patient data from the training algorithm and smuggle them out.

Cryptographic methods

Cryptographic methods encrypt systems and primarily protect the algorithms – i.e. the model weights, for example. Model weights are the learnable parameters in a machine learning model that control its behavior and capabilities. Cryptographic methods are useful when sending out models. This means they cannot be used if they end up in the wrong hands.

Differential privacy

Differential privacy is seen as the gold standard of data protection and was developed at the start of the 2000s. With differential privacy, mathematical noise – i.e. false data – is added to the data. In this process, the characteristic features of individual datasets are changed as a result of the algorithm, or "spurious" datasets are added, which are included in the evaluation.

All three methods are used in AI. Prof. Rückert's team is primarily backing differential privacy but also combining it with federated learning.

But differential privacy offers further benefits. For example, the method allows models to be trained with a "privacy budget". This privacy budget works in a similar way to a purchase in which a certain amount of money can be spent. Applied to data protection, this means that if you have exhausted the privacy budget as a result of several iterations (computation sequences) with private data, the system will not permit any further interaction with this dataset – it is quite simply blocked.

"For example, with the privacy budget, every participating institution (or even every patient) can define a quantitative volume of privacy that they would like to expend for training this model," Rückert explains. "This budget is correlated with the risk of datasets being re-identified. The higher the budget becomes, the higher the risk that my data can be reconstructed."

Rückert's team recently examined whether that can be put into practice. To do so, a dataset with patients' X-ray images was used to train algorithms. The test was successful: The team succeeded in reliably analyzing X-rays with the algorithms trained in hospital, and in showing that they are protected from external attack. "We demonstrated in an article published in the journal 'Nature Machine Intelligence' that it can actually work in a case study," the researcher stated. ■

Klaus Manhart

Picture credit: Juli Eberle