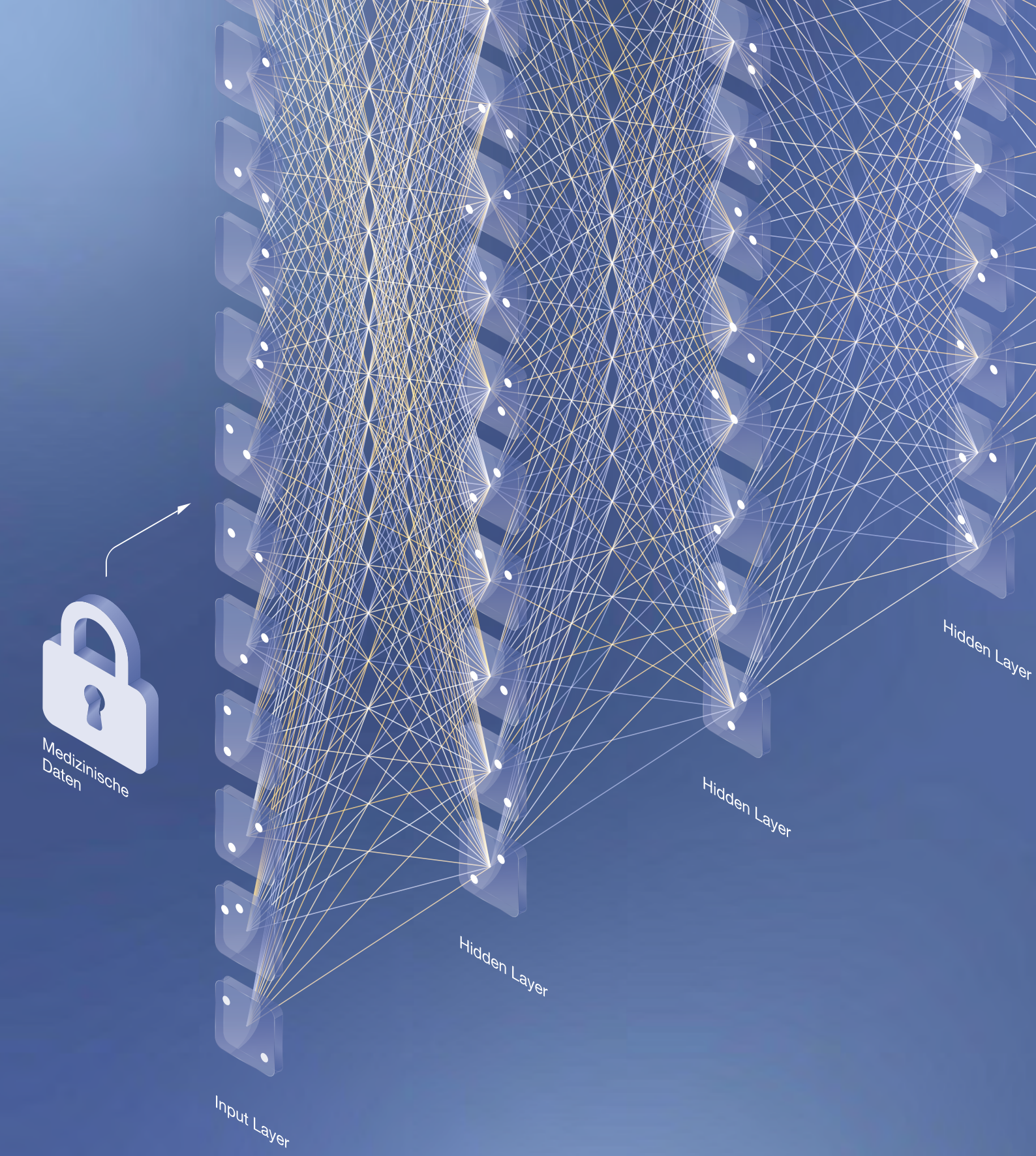
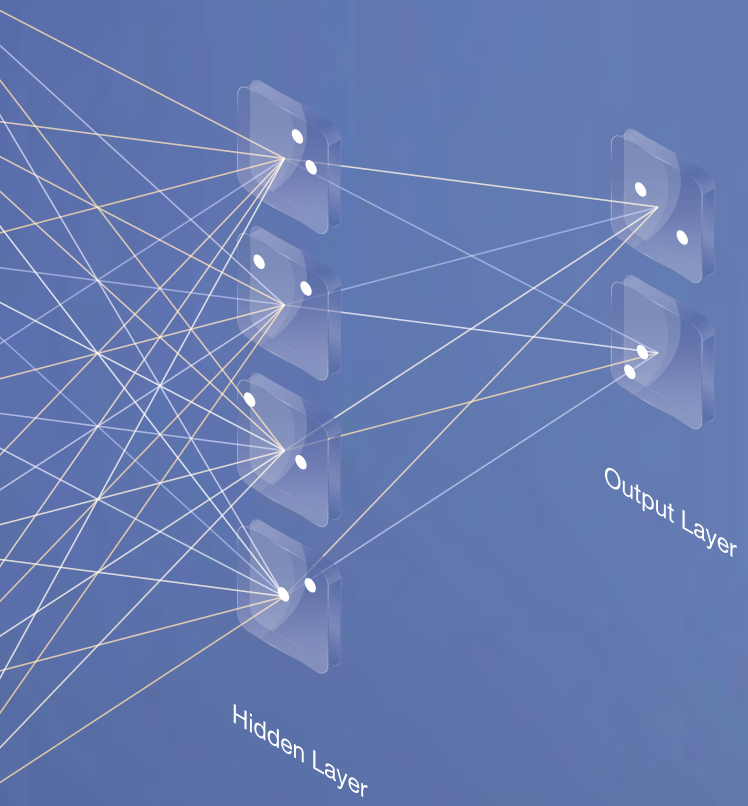


Link
www.aim-lab.io



Wie medizinische KI **vertrauenswürdig** wird

KI-Systeme im Gesundheitswesen sollten ethisch einwandfrei und möglichst vertrauenswürdig sein. Eine Forschungsgruppe um den Informatiker Prof. Daniel Rückert entwickelt Verfahren, mit denen die Privatsphäre bei KI-Anwendungen gewahrt werden kann – mit mathematischer Garantie.

Full Article (PDF, EN): www.tum.de/faszination-forschung

How Medical AI Can Become Trustworthy

E

AI systems in medicine have to be trustworthy. They should act reliably and fairly just like a human doctor and respect patients' privacy. The research team working with Prof. Daniel Rückert is examining how patients' training data can be safely protected and how "privacy-respecting AI" can be achieved. The team has shown that differential privacy offers mathematical guarantees of privacy – that cannot be undermined by either current or future attacks. These guarantees are comprehensive and not dependent on technical progress. □

Grafiken: ediundsepp

„Die Anforderungen an KI-Systeme sind hoch. Sie sollen mit den persönlichen Daten von Patienten sorgfältig umgehen und keine identifizierbaren Informationen abspeichern.“

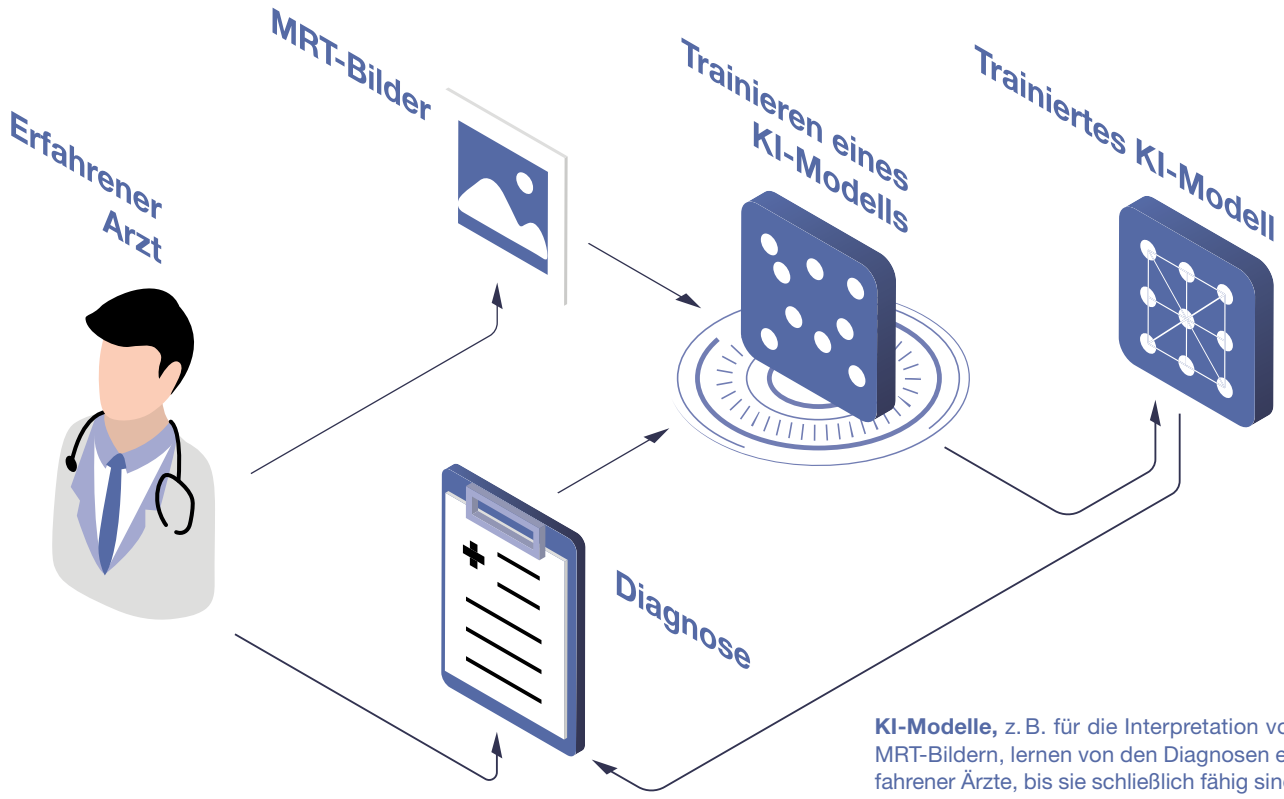
Daniel Rückert

Prof. Daniel Rückert

ist seit 2020 an der TUM Alexander von Humboldt-Professor für KI in der Medizin. Zusätzlich ist er Professor am Imperial College London. Er studierte Informatik an der TU Berlin (1993) und ging danach zur Promotion an das Imperial College, gefolgt von einem Post-Doc am King's College London. 1999 wurde er Assistant Professor am Imperial College. Seit 2005 hat er den Lehrstuhl für Visual Information Processing am Imperial College inne, wo er auch von 2016 bis 2020 als Dekan tätig war. An der TUM leitet Daniel Rückert zudem das Zentrum für Digitale Medizin und Gesundheit. Er arbeitet auf dem Gebiet der Künstlichen Intelligenz (KI) und des Maschinellen Lernens und deren Anwendungen in der Medizin. Seine Forschungsschwerpunkte liegen in den Bereichen Entwicklung von innovativen Algorithmen zur Bilderakquisition, Bildanalyse und Bildinterpretation und im Bereich KI zur Extraktion klinischer Informationen aus medizinischen Bildern – insbesondere zur computergestützten Diagnose und Prognose.

Künstliche Intelligenz (KI) verändert mit intelligenten Systemen gerade die Medizin. Die meisten KI-Anwendungen basieren auf Modellen für Maschinelles Lernen. Diese werden anhand von Patientendaten trainiert, um bestimmte Muster zu erkennen. Je mehr dieser Daten in das Training mit einfließen, umso genauer sind Diagnosen und Prognosen.

In der Medizin unterstützen solche KI-Systeme Ärztinnen und Ärzte inzwischen sehr erfolgreich bei der Diagnostik und Behandlung von Krankheiten, der Analyse von Röntgenbildern und in vielen anderen medizinischen Gebieten. Doch die rasante Entwicklung in diesem Bereich wirft auch Fragen grundsätzlicher Art auf: Sind die KI-Systeme ebenso verlässlich wie ein menschlicher Arzt? Können ihnen medizinische Anwender vertrauen? Und werden die für das Modelltraining genutzten Patientendaten sorgsam behandelt?



Bildnachweis: Juli Eberle; Grafiken: ediundsepp (Quelle: D. Rückert)

KI-Modelle, z. B. für die Interpretation von MRT-Bildern, lernen von den Diagnosen erfahrener Ärzte, bis sie schließlich fähig sind, Krankheiten selbstständig zu erkennen.

Der Informatiker Daniel Rückert von der TUM arbeitet daran, dass automatische Systeme ähnlich vertrauenswürdig sind wie ein menschlicher Arzt – für die Akzeptanz der Programme ein unerlässlicher Faktor: „Wir haben in der Medizin zwei Gruppen von Menschen, mit denen ein KI-System interagiert", sagt Daniel Rückert. „Die eine Gruppe sind Ärzte und Kliniker und die andere die Patienten. Beide Gruppen haben sehr hohe Anforderungen an die Qualität der Entscheidungsprozesse."

Diese Anforderungen sollten auch KI-Systeme erfüllen: Sie sollten beispielsweise mit den persönlichen Daten von Patienten sorgfältig umgehen und keine identifizierbaren Informationen abspeichern – also die Privatsphäre wahren. Sie sollten fair sein und beispielsweise Männer nicht anders als Frauen behandeln. Und sie sollten angeben, wie sicher ihre Entscheidungen sind. Denn wie ein menschlicher Arzt wird auch eine KI manche Diagnosen zwar mit 99 Prozent Sicherheit stellen können, andere aber vielleicht nur mit 80 Prozent. Und das muss das System möglichst transparent kommunizieren.

„Generell gibt es viele Definitionen und Kategorisierungsansätze für vertrauenswürdige KI", sagt Dr. Georgios Kaissis aus dem Team von Prof. Rückert. Der Konsens dabei ist, dass intelligente Systeme in der Medizin im weitesten Sinn ähnlich agieren sollten wie eine verantwortungsbewusste Ärztin oder ein verantwortungsbewusster Arzt. „Eine vertrauensvolle KI muss mit menschlichen Werteinstellungen vereinbar sein", sagt Kaissis. „Der Output solcher Systeme sollte menschlichen Grundwerten – wie etwa Fairness oder Schutz von Daten – nicht widersprechen." ▷

Dilemma Datenschutz

Daniel Rückert hat mit seinen Forschungsgruppen unter anderem die Themen Fairness und Transparenz im Fokus – und als aktuellen Schwerpunkt die privatsphärewahrende KI. Privatdozent Georgios Kaissis leitet die Forschungsgruppe zu dem Thema. Den Radiologen und Informatiker beschäftigt die Frage: Wie kann man KI-Modelle mit den Daten von Patientinnen und Patienten trainieren, ohne dass diese Daten wieder aus den Modellen rekonstruiert werden können?

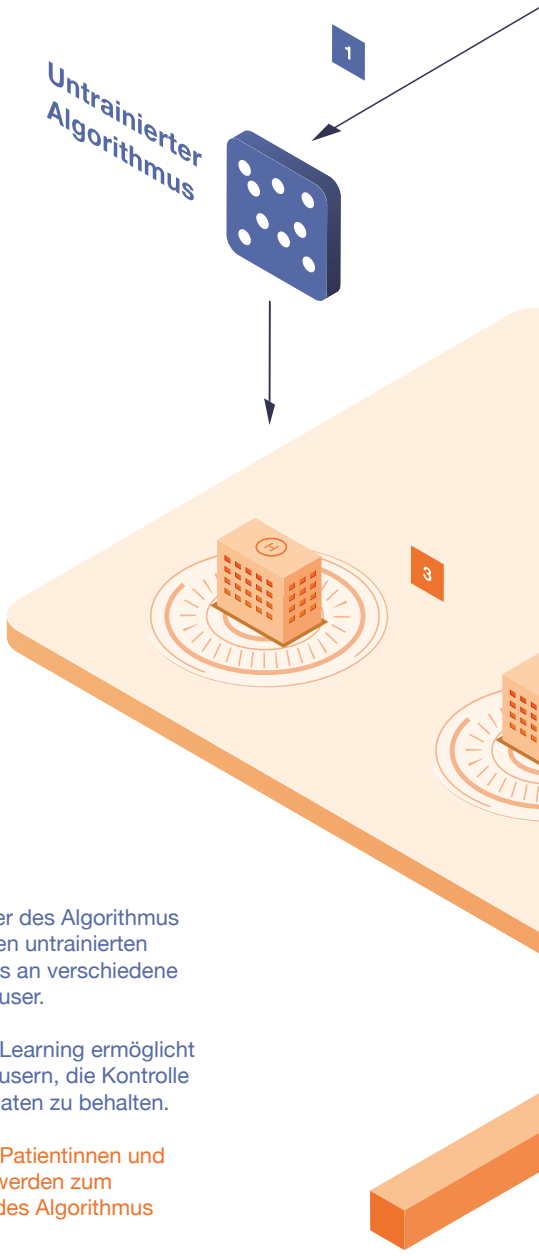
Die Relevanz dieser Frage darf nicht unterschätzt werden. Grundsätzlich sind Patientendaten, wie beispielsweise MRT-Bilder, unerlässlich für das Training der KI-Modelle. Diese Patientendaten sind aber aus zwei Gründen problematisch: Zum einen stehen diese Daten in der Medizin nicht in der Häufigkeit zur Verfügung wie bei nichtmedizinischen KI-Anwendungen – wo oft Millionen oder gar Milliarden von Trainings-Datensätzen genutzt werden. Man muss sich hier mit weniger begnügen – was die Verlässlichkeit der Modelle und Diagnosen einschränken kann.

Zum anderen sind die für das Training verwendeten Gesundheitsdaten hochsensibel und äußerst schützenswert. Krankheit ist schließlich Privatsache – Mediziner dürfen solche Daten prinzipiell nicht ohne Zustimmung der Betroffenen aus der Hand geben, auch nicht, um damit ein Computersystem zu trainieren, das künftig Leben retten kann.

Beide Herausforderungen – zu wenig Daten und sehr sensible Daten – lassen sich durch zuverlässigen Privatsphärenschutz lösen. Weitgehend etabliert als Verfahren, solche Daten hinreichend zu schützen, haben sich Anonymisierung und Pseudonymisierung. Bei der Anonymisierung werden die Namen oder identifizierenden Informationen komplett aus dem Datensatz entfernt. Die CD „Bob Dylan“ „Greatest Hits“ kann durch Löschung des Namens anonymisiert werden, so dass der Datensatz nur mehr den Eintrag „Greatest Hits“ enthält. Bei der Pseudonymisierung wird der Name „Bob Dylan“ durch einen anderen Namen ersetzt wie „Bob Marley“.

Der Haken an der Sache: Anonymisierung und Pseudonymisierung sind inzwischen nicht mehr sicher. Die Angriffsmöglichkeiten gegen die KI-Modelle sind so mächtig ▷

Rückert und sein Team setzen auf Differential Privacy und Federated Learning, um Gesundheitsdaten, die zum Training von KI-Modellen verwendet werden, sicher zu schützen. Bei der Differential Privacy wird ein kalibriertes statistisches Rauschen hinzugefügt, um sensible Daten zu schützen. Beim Federated Learning wird das KI-Modell sukzessive an einzelne Krankenhäuser geschickt, anstatt sensible Daten an einen zentralen Server zu senden. Dadurch bleibt die Kontrolle über die Daten in der jeweiligen Klinik.



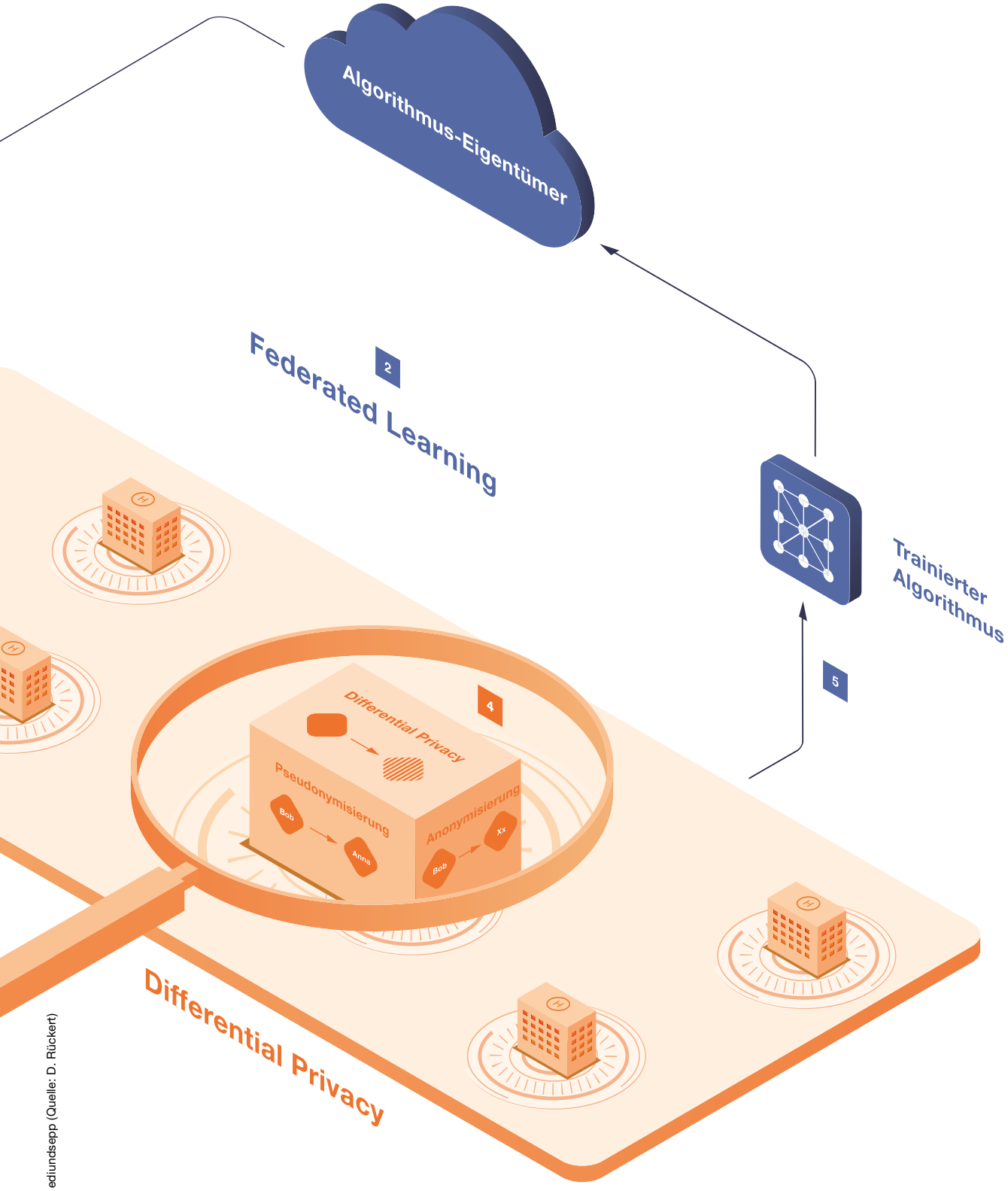
1 Der Besitzer des Algorithmus sendet einen untrainierten Algorithmus an verschiedene Krankenhäuser.

2 Federated Learning ermöglicht Krankenhäusern, die Kontrolle über ihre Daten zu behalten.

3 Daten von Patientinnen und Patienten werden zum Trainieren des Algorithmus genutzt.

4 Differential Privacy garantiert die Datensicherheit jetzt und in Zukunft unabhängig vom Stand der Technik.

5 Der trainierte Algorithmus wird dann zum Algorithmus-Eigentümer zurück geschickt.



Grafiken: ediundsepp (Quelle: D. Rückert)

Rückert und sein Team haben kürzlich gezeigt, dass Daten, die in KI-Modelle einfließen, durch Differential Privacy effektiv geschützt sind

geworden, dass selbst sehr gut anonymisierte Daten relativ einfach zu re-identifizieren sind. „Die bloße Entfernung des Namens ist für neuartige Angriffsmethoden völlig belanglos“, erklärt Georgios Kaissis. „Wir konnten in unseren Arbeiten mehrfach zeigen, dass Patientendaten wieder aus den Modellen heraus rekonstruiert werden können, wenn man diese ohne zusätzliche Schutzmaßnahmen trainiert.“ So ist es Kaissis mit seinen Mitarbeitern beispielsweise gelungen, Röntgenbilder von Patientinnen und Patienten wieder aus den Modellen komplett zu rekonstruieren – ein Desaster für den Datenschutz.

Dennoch werden in der Praxis Anonymisierung und Pseudonymisierung weiter genutzt. „Das liegt an der Diskrepanz zwischen dem Stand der Forschungsergebnisse und dem rechtlichen Rahmen“, sagt Kaissis. „Juristisch gelten anonymisierte Daten nach wir vor als nicht personenbezogen und sind deshalb rechtlich zulässig. Die Forschung zeigt allerdings, dass Anonymisierung nicht sicher ist.“ Erforderlich wäre deshalb eine Novelle des rechtlichen Rahmens.

Neben dem Schutz der sensiblen Daten kann eine KI, die die Privatsphäre wahrt, auch das Problem der zu geringen Datenmengen lösen – wenn auch nur indirekt: Eine KI, die die Privatsphäre wahrt, ist nämlich für Anwender und Datengeber vertrauenswürdig und wirkt damit stark motivierend auf Patienten, dass sie ihre Daten zur Nutzung freigeben. Es stehen dann mehr Trainingsdaten zur Verfügung, was die Modelle zuverlässiger und robuster macht.

Interdisziplinäre Forschung: Zentrum für Digitale Medizin und Gesundheit (ZDMG)

Prof. Daniel Rückert leitet das Zentrum für Digitale Medizin und Gesundheit (ZDMG), für das die TUM 43 Millionen Euro vom Bund und vom Freistaat Bayern erhielt. Forschende der Medizin, Informatik und Mathematik sollen dort gemeinsam neue Ansätze in den Bereichen Data Science und Künstliche Intelligenz entwickeln und deren klinische Anwendung vorantreiben. Durch die gezielte Einbindung natur- und ingenieurwissenschaftlicher Kompetenzen soll am neuen interdisziplinären Forschungszentrum die Entwicklung innovativer Methoden und Technologien in den Bereichen KI und Data Science für verschiedene medizinische Anwendungsbereiche nutzbar gemacht werden.

Mathematische Garantien

Daniel Rückert und sein Team um Georgios Kaissis nutzen mit Differential Privacy ein Verfahren, das die Limitierungen und Unsicherheiten von Anonymisierung und Pseudonymisierung hinter sich lässt. Im Wesentlichen beruht Differential Privacy darauf, dass beim Training der KI-Systeme den Daten „kalibriertes statistisches Rauschen" – also zufälliges Rauschen – hinzugefügt wird. Das Ganze ist mathematisch komplex, führt aber dazu, dass die Privatsphäre von einzelnen Patientinnen und Patienten gewährleistet ist.

Der große Pluspunkt von Differential Privacy: Die Methode gibt – anders als herkömmliche Verfahren – mathematische Garantien, dass sie weder durch aktuelle noch durch zukünftige Angriffe unterminiert werden kann. Während eine empirische Garantie nur sicherstellt, dass ein aktueller Angriff abgewehrt wird, ist es nicht ausgeschlossen, dass ein zukünftiger Angriff diese Garantie umgeht.

Eine mathematische oder formale Garantie ist hingegen eine Garantie, die weder jetzt noch in der Zukunft jemals umgangen werden kann. Diese formale Garantie ist deutlich stärker als eine bloße empirische – sie ist umfassend und unabhängig vom Stand der Technik. „Wenn ich den Datenschützer vom Klinikum rechts der Isar davon überzeugen will, dass er mir erlaubt, solche Verfahren einzusetzen, dann ist es natürlich für diesen sehr viel attraktiver, wenn ich ihm sagen kann: Ich kann mathematisch garantieren, dass man daraus den Patienten nie re-identifizieren kann," sagt Rückert. ▷

PD Dr. med. Georgios Kaissis, MHBA

ist Arbeitsgruppenleiter am Institut für Künstliche Intelligenz und Informatik in der Medizin und Oberarzt am Institut für Radiologie der TUM sowie Arbeitsgruppenleiter am Helmholtz-Zentrum München. Er forscht im Bereich der privatsphärewahrenden und vertrauenswürdigen Künstlichen Intelligenz, insbesondere zum Thema „Differential Privacy" sowie zu Anwendungen im Bereich der Medizin und der biomedizinischen Bildgebung.

Die „heilige Dreifaltigkeit“ – Algorithmic Privacy

Zum Schutz sensibler Daten haben sich – unter der Rubrik „Algorithmic Privacy“ (algorithmische Privatsphärenwahrung) – drei Verfahren etabliert

Verteiltes Lernen (Federated Learning)

Beim verteilten Lernen werden die Daten nicht zu den Algorithmen gebracht, sondern die Algorithmen zu den Daten. Das zu trainierende Modell wird in die Klinik verlegt, mit den dortigen Daten in der Klinik trainiert, das Modell wird dann zurückgeschickt und weiter mit Daten an einer anderen Klinik trainiert. Der Vorteil ist, dass die Daten nie aus der Obhut der Klinik herausgegeben werden müssen. Der Nachteil ist, dass Hacker mit dem trainierenden Algorithmus Patientendaten einfach kopieren und diese nach außen schmuggeln könnten.

Kryptographische Verfahren

Kryptographische Verfahren verschlüsseln Systeme und schützen vor allem die Algorithmen – also zum Beispiel die Modellgewichte. Modellgewichte sind die lernbaren Parameter in einem maschinellen Lernmodell, die dessen Verhalten und Fähigkeiten steuern. Kryptographische Verfahren sind nützlich beim Versenden von Modellen. So können sie, wenn sie in falsche Hände gelangen, nicht genutzt werden.

Differential Privacy

Differential Privacy gilt als Goldstandard des Datenschutzes und wurde Anfang der 2000er Jahre entwickelt. Bei Differential Privacy wird den Daten mathematisches Rauschen – das sind falsche Daten – hinzugefügt. Dabei werden aufgrund eines Algorithmus die Merkmalsausprägungen einzelner Datensätze verändert oder „unechte“ Datensätze hinzugefügt, welche in die Auswertung miteinbezogen werden.

Die drei Verfahren werden in der KI genutzt. Das Team von Prof. Rückert setzt vor allem auf Differential Privacy, kombiniert es aber mit verteiltem Lernen.

Differential Privacy hat aber noch weitere Vorteile. So erlaubt es die Methode, Modelle mit einem „Privatsphären-Budget“ zu trainieren. Dieses Privatsphären-Budget funktioniert analog wie ein Einkauf, bei dem ein bestimmter Betrag Geld ausgegeben werden kann. Übertragen auf den Datenschutz heißt das: Wenn man durch mehrere Iterationen (Rechendurchgänge) mit privaten Daten das Privatsphären-Budget aufgebraucht hat, dann verbietet das System, dass man weiter mit diesem Datensatz interagiert – er wird einfach gesperrt. „Mit dem Privatsphären-Budget kann zum Beispiel (jeder Patient oder) jede teilnehmende Institution eine quantitative Menge an Privatsphäre festlegen, die sie gerne für das Training dieses Modells aufwenden möchte“, erklärt Rückert. „Dieses Budget korreliert mit dem Risiko einer Re-Identifikation von Datensätzen. Je höher das Budget wird, desto höher wird das Risiko, dass meine Daten wieder heraus rekonstruiert werden können.“ Ob das auch in der Praxis umsetzbar ist, hat Rückerts Team kürzlich untersucht. Dazu wurde ein Datensatz mit Röntgenbildern von Patientinnen und Patienten verwendet, um Algorithmen damit zu trainieren. Der Test war erfolgreich: Es gelang mit den im Krankenhaus trainierten Algorithmen, Röntgenbilder verlässlich zu analysieren und zu zeigen, dass sie vor Angriffen von außen geschützt sind. „Wir haben das im Journal „Nature Machine Intelligence“ in einer Veröffentlichung gezeigt, dass es ganz konkret in einer Fallstudie funktionieren kann“, so der Forscher. ■ *Klaus Manhart*

Bildnachweis: Juli Eberle