3.5 Ein KI-TÜV für Europa

Ferdinand Müller

Interdisziplinäres Projektteam: (von links) Ferdinand Müller, Martin Schüßler und Elsa Kirchner

Der Jurist Ferdinand Müller (Forschungsgruppe 16 „Verlagerungen in der Normsetzung) und der Informatiker Martin Schüßler (Forschungsgruppe 20 „Kritikalität KI-basierter Systeme) haben in einem interdisziplinären Team gemeinsam mit der Biologin und Informatikerin Elsa Kirchner vom Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) einen Vorschlag für die Risikobewertung von algorithmischen Entscheidungssystemen erarbeitet.

Künstliche Intelligenz in Form von voll- oder teilautomatisierten algorithmischen Entscheidungssystemen kommt in immer mehr Anwendungsbereichen zum Einsatz. Solche Systeme werden etwa bei der Kreditwürdigkeitsprüfung durch die SCHUFA eingesetzt, beim Hochfrequenzhandel an der Börse, bei der Vorauswahl von Bewerbungsschreiben im Personalmanagement, bei selbstfahrenden Fahrzeugen oder bei der Auswertung medizinischer Bilddaten in Bereichen wie Pränatalmedizin oder Krebserkennung.

Eine Besonderheit von algorithmischen Entscheidungssystemen (AES) liegt darin begründet, dass durch die Entkoppelung von menschlicher Mithilfe die Bearbeitung auch äußerst großer Datenmengen in relativ kurzer Zeit möglich wird. Dabei werden unter Umständen auch sehr komplexe Modelle eingesetzt, die es schwierig bis unmöglich machen, Resultate im Nachhinein nachzuvollziehen.

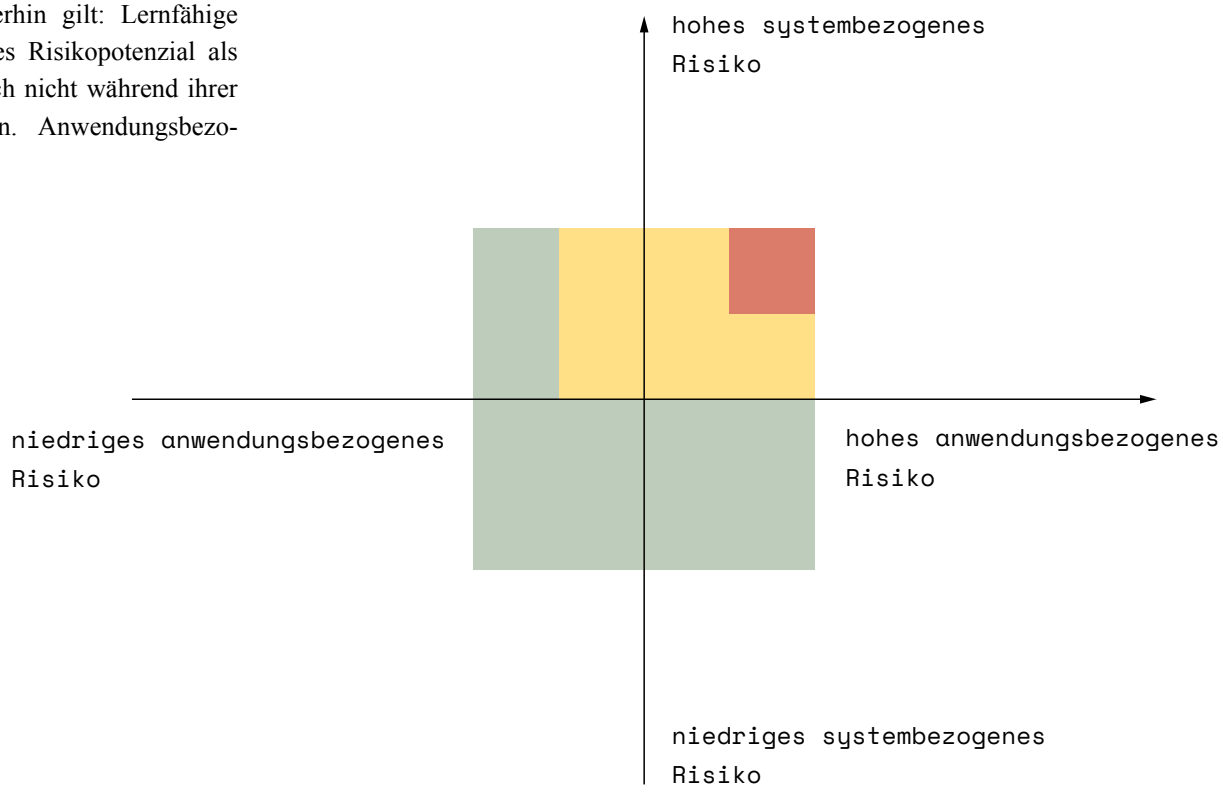
Viele Staaten überlegen aktuell, neue Gesetze für AES zu schaffen. Auf EU-Ebene hat die europäische Kommission zuletzt Mitte Februar 2020 das „Weißbuch zur Künstlichen Intelligenz“ vorgelegt, in das vorangegangene Empfehlungen wie etwa das Sachverständigengutachten der High Level Expert Group on Artificial Intelligence eingeflossen sind. Von einer Lösung in Form einer konkreten Regulierung jedoch ist das Weißbuch noch weit entfernt.

Was wohl alle bisher vorgestellten Gutachten und Strategien zur Regulierung von AES eint, ist das Streben nach einer strukturierten Kritikalitäts- oder Risikobewertung. Große Beachtung hat diesbezüglich ein Modell erfahren, welches die durch die Bundesregierung eingesetzte Datenethikkommission in ihrem 2019 veröffentlichen Gutachten präsentiert hat. Das Modell basiert auf einer klassischen Risikobewertung. Auf der einen Seite wird die Schwere eines möglichen Schadens betrachtet, den eine spezifische Technologie verursachen kann, auf der anderen Seite die Wahrscheinlichkeit des Schadenseintrittes. In der Gesamtschau ergibt sich unter Berücksichtigung der beiden Faktoren Schwere und Eintrittswahrscheinlichkeit die Einstufung einer Technologie auf einer Risikoleiter oder -pyramide. Je nach Einstufung sind regulatorische Folgemaßnahmen notwendig. Diese können zum Beispiel ein spezielles Zulassungsverfahren sein (für Anwendungen mit erheblichem Schädigungspotenzial) oder Verpflichtungen zur Selbstkontrolle (für Anwendungen mit geringerem Schädigungspotenzial).

Wir schlagen als Alternative zu dieser klassischen Risikobewertung ein verändertes Verfahren vor, das unserer Meinung nach den Besonderheiten von AES besser Rechnung trägt. Die Vielzahl der Anwendungsbereiche und -formen von AES erschwert eine einheitliche Einordnung auf einer einzigen Skala. Unseres Erachtens nach ist daher eine Matrix besser für eine Risikobewertung geeignet als eine einstufige Pyramide. Das Matrix-Modell, das wir als Alternative vorschlagen, fokussiert sich auf die qualitative Bewertung der Risiken. Gleichwohl soll die Matrix die Ableitung konkreter Handlungsmaßnahmen ermöglichen.

Anstelle von „Schwere“ und „Eintrittswahrscheinlichkeit“ betrachtet die Matrix einerseits „systembezogene Risiken“, die sich aus der eingesetzten AES-Technologie ergeben, und andererseits „anwendungsbezogene Risiken“, die aus dem spezifischen Einsatz der Technologie resultieren.

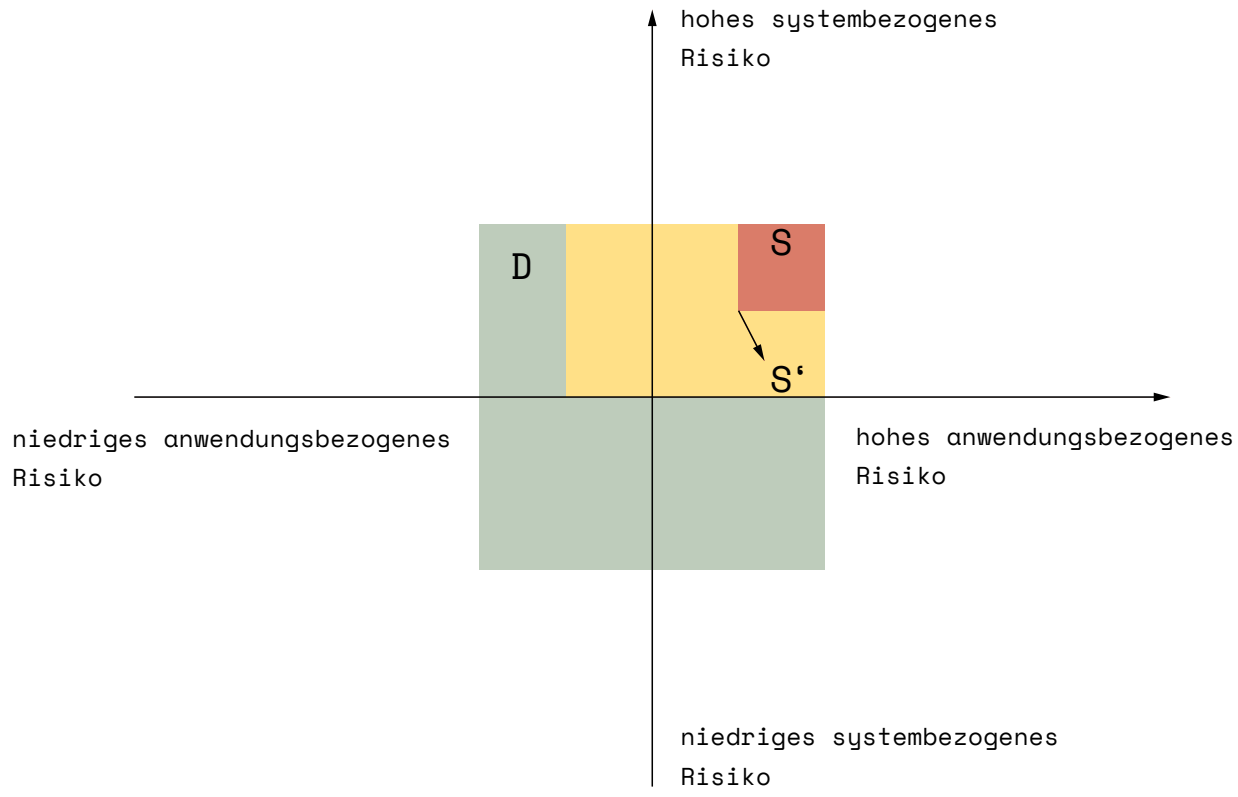
Systembezogene Risiken sind bedingt durch den verwendeten Algorithmus, das Model oder die Trainingsdaten, auf die ein AES-System aufbaut. Hier kann es etwa zu systematischen Verzerrungen des Ergebnisses kommen (auch als „biased AI“ bezeichnet), das durch eine lückenhafte oder kurzsichtige Auswahl der entscheidungsrelevanten Parameter entstehen können. Ein anderes Problem ist die Intransparenz von manchen AES-Systemen, deren Resultate nur schwer oder gar nicht durch Menschen nachvollzogen und somit korrigiert werden können. Weiterhin gilt: Lernfähige AES haben ein höheres Risikopotenzial als solche Systeme, die sich nicht während ihrer Verwendung verändern. Anwendungsbezogene Risiken, auf der anderen Seite, ergeben sich durch das spezifische Einsatzfeld. Bei der Verwendung von AES zur Prognose der Rückfallwahrscheinlichkeit von Straftätern werden andere Rechtsgüter tangiert als bei AES im Hochfrequenzhandel an der Börse oder bei der Auswertung medizinischer Bilddaten.



Indem man die system- und die anwendungsbezogenen Risiken miteinander ins Verhältnis setzt, kann man die Gefahren des Einsatzes von AES ermitteln. Technologien, die im grünen Bereich der Matrix verortet werden, benötigen eine vergleichsweise geringe Regulierung; Technologien im gelben bis roten Bereich eine anspruchsvolle Regulierung. Ein Beispiel dafür wäre die Einführung einer Betreiberhaftung, verbunden mit einer Pflichtversicherung.

Zwei Beispiele veranschaulichen, wie die Matrix funktioniert:

Beispiel D:
Eine Dating-App ist aufgrund der Vielzahl an verwendeten Parametern und Entscheidungsebenen intransparent, das heißt, ihre Ergebnisse sind nur schwer oder gar nicht durch Menschen interpretierbar. Dies steht für ein hohes systembezogenes Risiko. Gleichzeitig ist das anwendungsbezogene Risiko relativ gering. Anwender einer Dating-App haben weder körperliche noch finanzielle Schäden zu befürchten. Sie haben sich freiwillig zur Nutzung der App im Rahmen eines Vertragsverhältnisses entschieden. Trotz des hohen systembezogenen Risikos wird solch eine Dating-App deshalb im Resultat im grünen Bereich verortet.



Eine Dating-App (D) und ein AES zur Berechnung der Rückfallwahrscheinlichkeit von Straftätern (S) in der Risikobewertung.

Beispiel S:

In den Vereinigten Staaten werden seit einiger Zeit zur Berechnung der Rückfallwahrscheinlichkeit von mutmaßlichen Straftäter*innen algorithmische Entscheidungssysteme durch Richter*innen eingesetzt. Diese Systeme sollen Entscheidungsvorschläge zu Fragen der Festsetzung einer Kaution oder Freilassung auf Bewährung generieren. Diese Anwendung betrifft damit also indirekt die persönliche Freiheit des Betroffenen. Außerdem kann sich der Einzelne einer solchen Anwendung nicht entziehen, da sie von staatlicher Seite aus eingesetzt wird. Dementsprechend ist bereits im Vorfeld ein hohes anwendungsbezogenes Risiko gegeben.

Gleichzeitig, so zeigen Untersuchungen, besteht ein hohes systembezogenes Risiko in Form möglicher Verzerrungen in den Resultaten. Die Software COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) beispielsweise setzt zur Berechnung der Rückfallwahrscheinlichkeit 137 Parameter ein. Enthalten sind statische Faktoren wie die Schulbildung des Angeklagten, dessen Nachbarschaft oder die Länge des Vorstrafenregisters. Wissenschaftler*innen konnten jedoch 2016 zeigen, dass der Algorithmus zur Verschärfung von Ungleichheiten führt. Bei People of Color nahm das System überdurchschnittlich häufig fälschlicherweise eine zu hohe Rückfallwahrscheinlichkeit an. Auf diese Weise werden bestehende Ungleichheiten durch den Einsatz der AES-Technologie verstärkt. Eine Entschärfung des Problems wurde von einer anderen Gruppe von Forscher*innen vorgeschlagen. Dieser Gruppe gelang es, mit Hilfe der Reduktion von eingesetzten Parametern die Verzerrung zu verringern und die Nachvollziehbarkeit zu verbessern. Wenn man diese Strategie befolgt, könnte man theoretisch die systembezogenen Risiken einer Software wie COMPAS so weit verringern, dass die Anwendung insgesamt vom roten in den gelben Bereich (von S auf S‘) hinaufgestuft werden kann. Das Beispiel zeigt somit, wie die Matrix nicht nur die Einordnung einer Technologie als grün, gelb oder rot erleichtert, sondern zugleich den Blick auf praktische Maßnahmen lenkt, die zum Zwecke der Risikominderung unternommen werden könnten. ´