Juristische Entscheidungen nach maschinellem Bauchgefühl

Jörn Erbguth ist Diplom-Informatiker, Diplom-Jurist, Legal Tech Berater und Mitglied im Vorstand des EDV-Gerichtstags.

Der Trend geht zu intelligenten Systemen, die allein anhand von Beispielen trainiert werden und trotzdem besser als Menschen entscheiden. Wir können uns damit scheinbar unsere Dogmatik und Regeln sparen. Wenn wir aber Systemen unsere Regeln nicht mitteilen, so werden sie sich auch nicht daran halten. Diskriminierung ist da noch eine der harmloseren Folgen.

Manche Bereiche entziehen sich klaren Regeln und wir müssen uns mit einer schwer strukturierbaren Kasuistik zurechtfinden. Da kommt uns die Nachricht gelegen, dass sogenannte Deep-Learning-Systeme – aus dem Bereich der Künstlichen Intelligenz (KI) – inzwischen so gut sind, dass sie häufiger richtig liegen als Menschen. Solche Systeme entscheiden in den USA und Großbritannien bereits über die Strafzumessung und darüber, wer auf Kaution freikommt. In Deutschland beschränkt sich ihr Einsatz bislang auf weniger kritische Bereiche: Versicherungen erkennen mit ihrer Hilfe Fälle, in denen ein Betrugsverdacht besteht; die Polizei in Baden-Württemberg und Bayern arbeitet mit einem System, genannt „PRECOBS", das prognostiziert, wo nach einer Straftat weitere Straftaten zu erwarten sind. Sollten wir das ausweiten und künftig zum Beispiel auch Entscheidungen zur Vorbeugehaft mit KI-Systemen tätigen oder diese auch nur unterstützen? Deep-Learning-Systeme sind künstliche neuronale Netze, deren Funktionsweise an die vernetzter Nervenzellen im Gehirn angelehnt ist. Sie arbeiten mit zwei Algorithmen:

Der erste ist ein häufig offen gelegter Lernalgorithmus. Dieser sorgt dafür, dass das System anhand von Beispielen „lernen" kann. Derselbe Lernalgorithmus kann dabei für sehr unterschiedliche Systeme oder auch Bereiche verwendet werden wie zum Beispiel die Bilderkennung und das Beurteilen juristischer Sachverhalte. Der Lernalgorithmus trainiert das künstliche neuronale Netz mithilfe einer großen Anzahl von Beispielfällen. Ausgehend von einem zufälligen Ausgangszustand werden die Verknüpfungen zwischen den Neuronen schrittweise so angepasst, dass das Ergebnis für die Trainingsfälle stimmt. Entscheidend ist danach, ob das System gut abstrahiert hat. Dies wird mit Test- und Validierungsfällen kontrolliert.

Heraus kommt dann der zweite, der eigentliche Algorithmus, den kein Mensch mehr verstehen kann – die berühmte „Black Box". Dieser Algorithmus hat mehr Ähnlichkeit mit menschlichem oder tierischem Erkennungs- oder Assoziierungsvermögen als mit einem logisch durchstrukturierten klassischen Algorithmus oder Computerprogramm. Diese Systeme lassen sich aber statistisch analysieren. Zunächst fiel auf, dass in den Trainingsdaten vorhandene Diskriminierungen und Stereotype reproduziert wurden. Dies ist auch nicht verwunderlich, geben wir dem System doch keinerlei Regelwerk mit. Es versucht die Trainingsfälle voneinander zu unterscheiden und weiß nicht, welche Kriterien dafür gut und welche eher schlecht geeignet oder unzulässig sind. Wenn Männer statistisch ein höheres Strafmaß erhalten, dann ist nicht unwahrscheinlich, dass das System eine Verbindung zum Geschlecht herstellt. Selbst wenn man dem System das Geschlecht vorenthält, so wird es eine Verbindung zwischen anderen, eher bei Männern vorhandenen Kriterien herstellen – auch wenn diese Kriterien nicht sachgerecht sind. Damit droht das System zu diskriminieren – denn es kann nicht unterscheiden, ob das höhere Strafmaß bei Männern das Resultat einer diskriminierenden Strafzumessung oder etwa einer bei Männern im Schnitt vorhandenen höheren Gewaltanwendung bei der Tatbegehung war.

Besonders kritisch sind Systeme, die Prognoseentscheidungen treffen. Die Wahrscheinlichkeit, dass jemand einen Kredit nicht zurückzahlen kann oder rückfällig wird, wird an Dingen festgemacht, die man in vielen Fällen den Betroffenen nicht vorwerfen kann. Familiär ungebunden zu sein etwa, zählt als Rückfallrisiko, obwohl daraus keinerlei Vorwurf abgeleitet werden kann. Auch ein statistisch einwandfrei arbeitendes System diskriminiert also. Damit ergibt sich die Frage, ob wir überhaupt Entscheidungen aufgrund von statistischen Prognosen treffen sollten. Wenn wir eine Prognose tätigen, dann beurteilen wir jemanden anhand des Verhaltens anderer Leute vor ihm. Wenn in der Vergangenheit Alleinstehende häufiger rückfällig wurden, dann nehmen wir dies als Maßstab für die Beurteilung einer Person, die keinerlei Einwirkungsmöglichkeit auf diese Vergangenheit hatte. Diese digitale Sippenhaft bringt die neuronalen Netze in Verruf, obwohl eigentlich das Instrument der Prognose problematisch ist. Allerdings erreichen diese Systeme statistisch sehr gute Ergebnisse. Doch können wir deshalb das Prinzip der individuellen Gerechtigkeit aufgeben?

Wenn von einem Offenlegen der Algorithmen gesprochen wird, macht dies im Zusammenhang mit neuronalen Netzen wenig Sinn. Diese Systeme arbeiten schlicht nicht nach Regeln, wie wir sie gewohnt sind. Sicher sollten die Trainingsbeispiele auf Ausgewogenheit und Abdeckung von Sonderfällen kontrolliert werden. Auch können einige Diskriminierungen festgestellt und durch zusätzliche Trainingsbeispiele „verlernt" werden. Doch selbst wenn die Diskriminierung nach Geschlecht, Rasse, Religion und Alter ausgeschlossen wurde, können immer noch Rothaarige, Bewohner eines bestimmten Stadtteils oder Alleinstehende diskriminiert werden.

Die Systeme lassen sich nicht nur statistisch untersuchen. Es lassen sich auch gezielt Fälle finden, in denen sich das System – ähnlich einer optischen Täuschung – überlisten lässt. So wurde ein Stopp-Schild wegen ein paar eher unscheinbarer Aufkleber von einem selbstfahrenden Auto als Tempo-Limit-Zeichen erkannt. Für menschliche Betrachter hatte das Schild jedoch keinerlei Ähnlichkeit mit dem Tempo-Limit-Schild und war weiterhin klar als Stopp-Schild zu erkennen. Das liegt daran, dass das System nicht nach unseren Differenzierungskriterien arbeitet. Vielmehr wählt es eher zufällig ein paar Kriterien aus, die statistisch bedeutsam sind. Auf eine Sozialprognose übertragen, könnte daher eine für den menschlichen Leser eher unbedeutende Umformulierung der Fallbeschreibung zu massiven Unterschieden in der Bewertung führen. Neuronale Systeme lassen sich also recht gut überlisten, wenn man Ihre Funktionsweise kennt. Auch eine Geheimhaltung der Systeme schließt diese Gefahr nicht verlässlich aus, sie reduziert zudem die Transparenz.

Sollten wir daher Deep Learning aus der Justiz verbannen? Ist manchmal eine schnelle „nur" statistisch gute Entscheidung nicht vielleicht besser als eine wirklich juristisch und individuell argumentierte Entscheidung, die Monate oder Jahre auf sich warten lässt? Können automatisierte Vorentscheidungen akzeptiert werden, wenn der Rechtsweg danach nach wie vor offensteht? Dies sind zentrale Fragen. Wir müssen uns jedenfalls im Klaren sein, dass diese Systeme zwar häufig recht haben, aber keinerlei Ahnung vom Recht haben. Sie arbeiten nicht nach juristischen oder anderen klaren Regeln, sondern bieten eine Art wohltrainiertes Bauchgefühl, ähnlich einem Judiz.