SIMPLY RATIONAL

Der EU KI Act – Eine Gelegenheit!

Dr. Niklas Keller, Prof. Dr. Florian Artinger, Dr. Malte Petersen & Prof. Dr. Gerd Gigerenzer

Simply Rational – Ein Max-Planck Spin-Off

Es gibt einen guten Grund, warum die EU den Einsatz von künstlicher Intelligenz (KI) regulieren will. In den letzten Jahren wurden Fälle aufgedeckt, in denen Systeme der künstlichen Intelligenz zum Nachteil von Bürgern und Unternehmen katastrophal versagt haben. In seiner jetzigen Form konfrontiert die EU-KI-Verordnung die Unternehmen mit erheblicher regulatorischer Unsicherheit. Erstens in Bezug auf das System zur Risikoklassifizierung und die Frage, in welche Kategorie die KI-Systeme eines Unternehmens fallen werden. Zweitens in Bezug auf die sich daraus ergebenden regulatorischen Anforderungen, die an diese Systeme gestellt werden. Die Forderung nach mehr Transparenz für Hochrisikosysteme in der EU-KI-Verordnung bietet aber auch eine Chance. Wie sich herausstellt, haben Forschungen an Max-Planck-Instituten und darüber hinaus gezeigt, dass einfache und transparente Algorithmen, die auf den Prinzipien der menschlichen Informationsverarbeitung beruhen ("psychologische KI"), selbst die modernsten und komplexesten "Black-Box"-Algorithmen des maschinellen Lernens oft übertreffen können. Durch den Einsatz transparenterer KI können Unternehmen die regulatorische Unsicherheit im Zusammenhang mit dem Abschnitt zur Risikoklassifizierung der KI-Verordnung erheblich verringern. Der Aufwand für viele der regulatorischen Anforderungen kann mit Hilfe transparenter Algorithmen stark reduziert werden.

Der Druck steigt. Das Bundeskartellamt hat kürzlich die 50 größten Versandhändler und alle größeren Kreditauskunfteien aufgefordert, das Innenleben ihrer Algorithmen offenzulegen. Auch wenn noch nicht klar ist welche Form die neue EU-KI-Verordnung letztendlich annehmen wird, fordern die Behörden bereits jetzt mehr Transparenz ein, testen Grenzen aus, schaffen Präzedenzfälle und probieren verschiedene Möglichkeiten der Regulierung von KI für eine Zukunft aus, in der die KI-Verordnung vollständig eingeführt sein wird. Den Kopf in den Sand zu stecken, ist nicht länger eine Option. Die Zeit, sich auf die neue regulatorische Realität vorzubereiten, ist gekommen.

Die KI-Verordnung kam nicht aus dem Nichts

Böse Zungen mögen behaupten, dass es sich bei der EU-KI-Verordnung lediglich um einen weiteren Machtrausch der EU-Bürokraten handelt, doch man sollte nicht vergessen, dass die KI-Verordnung einer zunehmenden Zahl von KI-Einsätzen folgt, die sich im Nachhinein als potenziell katastrophal für das Wohlergehen von Hunderttausenden von Bürgern erwiesen haben. Zu den bekanntesten Fällen gehören:

> Gerichtliche Entscheidungen: Der Algorithmus "Correctional Offender Management Profiling for Alternative Sanctions (COMPAS)" hat in den USA Millionen von Kautions- und anderen Gerichtsentscheidungen unterstützt. Eine Studie hat jedoch gezeigt, dass COMPAS unter Verwendung von über 130 Variablen die Rückfälligkeit nicht besser vorhersagt als Menschen ohne juristischen Hintergrund, die nur 7 Variablen verwenden. Darüber hinaus ist es aufgrund der Black-Box-Natur von COMPAS schwierig festzustellen, ob Minderheiten diskriminiert werden.
>
> Gesichtserkennung: Die Bilderkennungssoftware von Google Photos stufte zwei Afroamerikaner als "Gorillas" ein. Auch die Gesichtserkennungs-KI von Amazon hatte eine weitaus höhere Genauigkeit für Männer (3 % Fehlerquote) als für Frauen (19 % Fehlerquote) und die niedrigste Genauigkeit für schwarze Frauen (39 % Fehlerquote).
>
> Onkologie: IBMs hochmoderne KI "Watson" sollte die Onkologie revolutionieren: die Versorgung von Krebspatienten verbessern, die besten Behandlungen empfehlen und die Suche nach neuen Arzneimitteln völlig umgestalten. Aber die Versprechen der IBM-Marketingabteilung konnten nie eingehalten werden und gefährdeten das Leben von Patienten. “Watson” wurde im Jahr 2021 im Wesentlichen in Einzelteilen verkauft, einschließlich der Patientendaten. Dies geschah, nachdem es bereits in Versuchen zur Unterstützung der Versorgung tausender Patienten eingesetzt worden war.

Dies ist nur eine kleine Auswahl von Fällen mit hohem Bekanntheitsgrad. Insbesondere in der Medizin, einem Sektor, der aufgrund seiner Größe die Hauptlast der KI-Initiativen trägt, tauchen fast täglich Berichte über "hochentwickelte" KI-Systeme auf, die bei den ihnen zugedachten Aufgaben kläglich versagen. Zusammengenommen haben diese Systeme bereits in Millionen von Fällen Entscheidungen im Gesundheitswesen "unterstützt" - mit unvorhersehbaren Folgen für Patienten und Ärzte gleichermaßen. Die Regulierungsbehörden sind daher zu Recht besorgt über die Leistung von KI-Algorithmen in sensiblen Bereichen wie dem Gesundheitswesen, dem Recht oder dem Finanzwesen.

Die große Unbekannte: das Risikolevel einer KI

Die Art und Weise, wie die KI-Akte Risiken kategorisiert (siehe Abbildung 1) ist verbesserungsbedürftig. Während der Abschnitt über die Risikoklassifizierung im Vergleich zu den ersten Entwürfen deutlich verbessert wurde, ist es immer noch der Fall, dass manchmal bestimmte Technologien für sich allein kategorisiert werden, ohne dass ihr Ziel oder Zweck erwähnt wird (z.B. werden KI-Chatbots immer noch als niedriges Risiko eingestuft). Manchmal werden Ziele für sich allein kategorisiert, ohne dass die zugrunde liegende Technologie erwähnt wird (z. B. alle Formen des Social Scoring). Manchmal werden Ziele und Technologien in einen Topf geworfen (z. B. biometrische Echtzeit-Fernerkennungssysteme, die in öffentlichen Räumen für die Strafverfolgung eingesetzt werden). Und die meisten dieser Kategorien schließen sich nicht gegenseitig aus. Ist ein (nicht-manipulativer) KI-Chatbot, der Menschen, die mit ihm interagieren, dazu bringt häufiger Selbstmord zu begehen, ein System mit minimalem Risiko? Ist eine eBay- oder Uber-Bewertung, mit der die Vertrauenswürdigkeit eines Verkäufers oder Fahrers überprüft wird, nicht auch eine Form des Social Scoring, auch wenn sie das Risiko für die Nutzer dieser Dienste verringert?

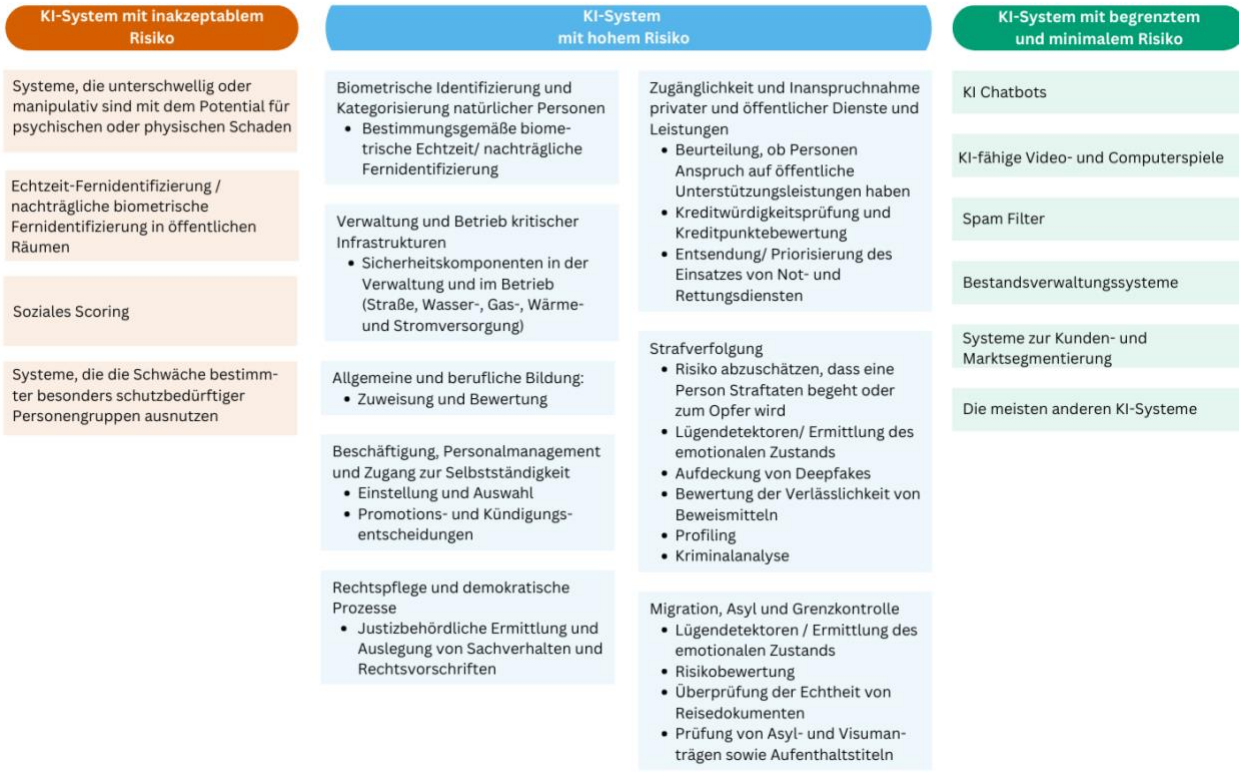

Abbildung 1. Die drei Risikokategorien der EU-KI-Verordnung und die entsprechende Kategorisierung von KI-Technologien bzw. deren Anwendungsbereiche.

Darüber hinaus wurde die Liste der Hochrisikotechnologien im Laufe der verschiedenen Versionen der KI-Verordnung immer wieder erweitert und die Kriterien, die die Verordnung für die Aufnahme weiterer Sektoren in diese Liste vorschlägt, bleiben erheblichen Interpretationsmöglichkeiten ausgesetzt. Sie könnten in der Tat eines der Hauptschlachtfelder der Lobbyarbeit nach der Einführung der Verordnung darstellen. Dies spiegelt sich auch in der

Bewertung der Verordnung durch den Rat der Europäischen Union vom 6. Dezember 2022 wider, in der die anhaltende Unsicherheit hinsichtlich der Definition von KI hervorgehoben wird und insbesondere Änderungen am Risikoklassifizierungssystem vorgeschlagen werden (sowohl an der Liste der Hochrisikotechnologien als auch an der Art und Weise, wie diese definiert werden). Bis zur endgültigen Überarbeitung (die in jedem Fall weiterhin durch die Aufnahme weiterer Sektoren in die Liste der Hochrisikotechnologien geändert werden kann), sind weiterhin erhebliche Änderungen zu erwarten. Infolgedessen wird die Unsicherheit darüber, in welche Risikokategorie das Produkt oder die Dienstleistung eines Unternehmens fallen wird, auf absehbare Zeit groß bleiben.

Die Illusion der Komplexität

Wenn wir uns die Gründe für die KI-Verordnung ansehen, sind "Vertrauen und Transparenz" sicherlich die von Brüssel am häufigsten genannten. Trotz des anfänglichen Hypes um die KI haben die vielen dokumentierten Fehler der KI verständlicherweise das Vertrauen untergraben. Die Notwendigkeit einer stärkeren Aufsicht ist die Folge. Wo kein Vertrauen herrscht, muss es zumindest Transparenz geben. Aber warum ist so viel von der KI, die wir heute einsetzen, so intransparent? Ein Grund ist die in Wirtschaft, Politik und Gesellschaft weit verbreitete "Komplexitätsillusion".

Die Annahme, dass mehr Informationen und ausgefeilte, komplexe Modelle mit größerer Flexibilität (d. h. mehr freien Parametern) automatisch bessere Ergebnisse liefern als einfachere Modelle, ist so tief verwurzelt, dass sie selten diskutiert und fast nie in Frage gestellt wird. Für bestimmte Aufgaben können simplere Ansätze verwendet werden, aber nur, weil der Aufwand, die zusätzlichen Informationen oder die benötigte Rechenleistung, die für die Anwendung eines komplexen Modells erforderlich sind, überwiegt die Vorteile einer höheren Leistung. Dies wird als "Aufwand-Genauigkeit-Abwägung" bezeichnet. Je mehr Aufwand man betreibt, desto besser ist die Genauigkeit (mit abnehmendem Ertrag). Diese Annahme führt auch direkt zu der (falschen) Behauptung, dass die Verwendung leicht verständlicher, einfacher und transparenter Algorithmen immer zu einer geringeren Vorhersagegenauigkeit führt. Dies hat die Defense Advanced Research Projects Agency (DARPA), eine militärische Forschungsagentur und einer der größten Forschungsförderer der Vereinigten Staaten, 2016 dazu veranlasst diese Grafik zu veröffentlichen:

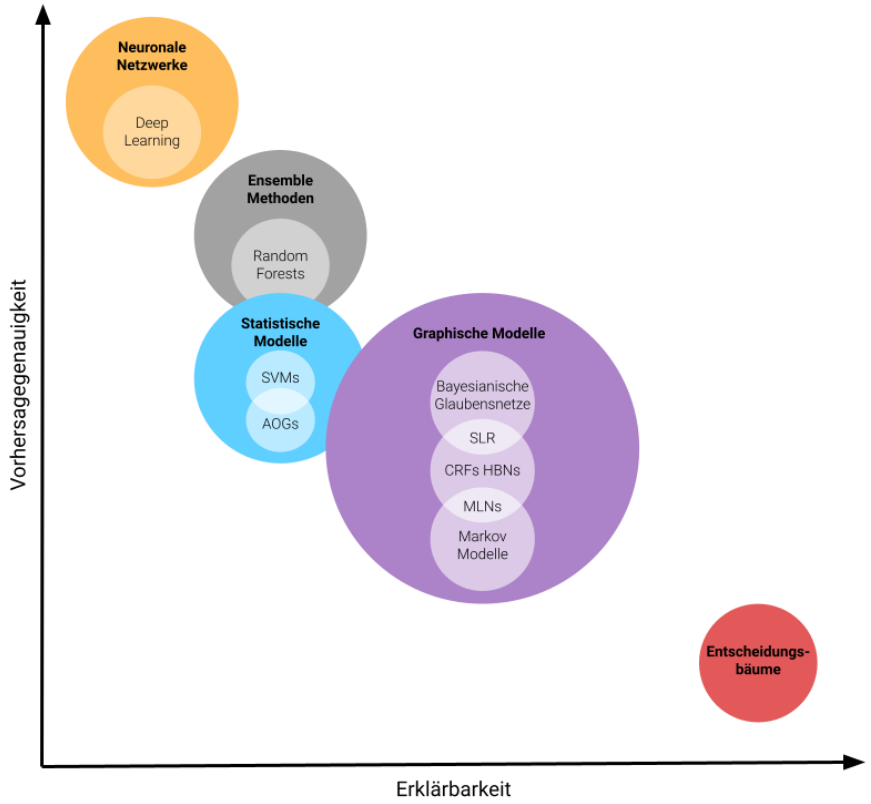

Abbildung 2 (DARPA, 2016). Der Kompromiss zwischen Aufwand und Genauigkeit, der sich in den Kompromiss zwischen Erklärbarkeit und Genauigkeit umwandelt. Die (fehlerhafte) Annahme ist hier, dass Modelle mit geringerer Komplexität (z. B. Entscheidungsbäume unten rechts) eine höhere Erklärbarkeit, aber eine geringere Vorhersagegenauigkeit aufweisen.

In der oberen linken Ecke befindet sich der Höhepunkt der Komplexität der modernen künstlichen Intelligenz: das neuronale Netz, in dem Deep Learning stattfindet. Schade, dass niemand die innere Funktionsweise solcher Modelle wirklich versteht, wenn sie einmal trainiert wurden. Nicht einmal die Ingenieure, die diese Modelle entwickelt haben! Als Google Photos Afroamerikaner als Gorillas einstufte, gab es keine Codezeile, die die Google-Ingenieure austauschen oder löschen konnten. Die einzige Möglichkeit das Problem zu lösen, bestand darin, die gesamte Kategorie "Menschenaffen" aus den Auswahlmöglichkeiten zu entfernen.

In der rechten unteren Ecke: der einfache Entscheidungsbaum - leicht zu verstehen und zu erklären, aber angeblich mit wenig Leistung. Verurteilt uns die EU-KI-Verordnung nun dazu, die schlechtesten Technologien für die wichtigsten Aspekte unseres Lebens, unserer Wirtschaft und unserer Gesellschaft einzusetzen?

Transparente Algorithmen können leistungsstärker sein als Black-Boxes

Glücklicherweise nicht. Es hat sich herausgestellt, dass der Kompromiss zwischen Aufwand und Genauigkeit im Allgemeinen nicht zutrifft. Die Vorstellung, dass komplexere, ausgefeiltere Modelle immer besser sind als einfachere, ist nachweislich falsch. Modelle, die einfacher und transparenter sind, weniger Daten verwenden und diese Daten auf unkompliziertere Weise kombinieren, können oft hoch entwickelte und komplexe KI-"Black-Box"-Modelle übertreffen. Weniger kann mehr sein! Hier sind drei veranschaulichende Beispiele.

Gerichtliche Entscheidungen: Im Anschluss an die COMPAS-Studie haben unsere Kollegen von Microsoft Research und der Stanford University einen transparenten Algorithmus entwickelt, der auf Prinzipien der Psychologie und der Kognitionswissenschaft basiert (was wir als "psychologische KI" bezeichnen) und ihn mit der Leistung eines komplexen KI-Modells (eines so genannten Random Forest) verglichen. Beiden Methoden standen sämtliche Daten zur Verfügung. Der Random Forest integrierte schließlich alle 64 verfügbaren Variablen auf sehr intransparente Weise. Der transparente Algorithmus verwendete nur zwei Variablen (das Alter und die Anzahl der Fälle, in denen der Angeklagte nicht vor Gericht erschienen war) und integrierte diese in eine transparente Bewertungsmethode. Trotz dieser enormen Unterschiede bei der Verwendung der Daten und der Komplexität des Algorithmus schnitten beide Algorithmen in etwa gleich gut ab (und waren beide deutlich besser als die Richter; siehe Abbildung 3).

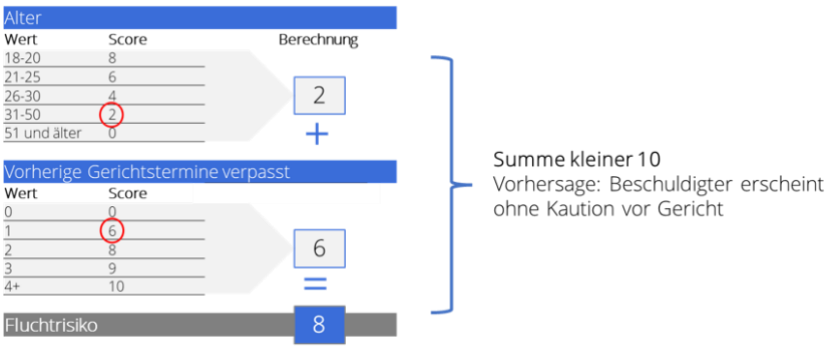

Abbildung 3.1. Ein einfacher, transparenter Algorithmus zur Vorhersage der Rückfälligkeit. Der Algorithmus berücksichtigt nur zwei Faktoren (Alter und Anzahl der versäumten Gerichtstermine) und integriert diese mithilfe einer einfachen Bewertungsmethode.

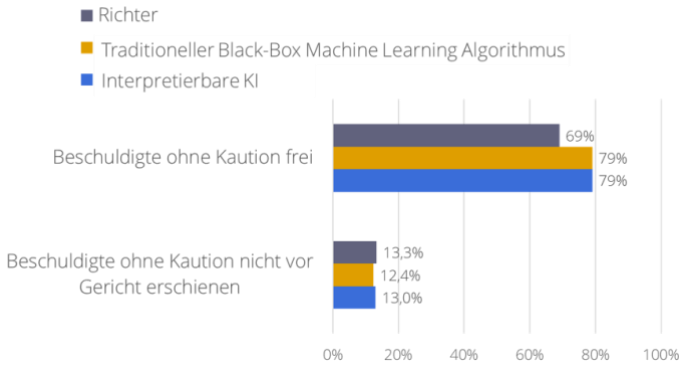

Abbildung 3.2. Die Leistung des einfachen, transparenten Algorithmus ist besser als die von erfahrenen Richtern und gleichauf mit der eines komplexen Random Forest.

Vorhersage von Pandemien: Im Jahr 2008 sollte Google Flu Trends die Vorhersage und Verfolgung von Grippepandemien revolutionieren. Die Idee war einfach: Anstelle des langwierigen Prozesses, bei dem das U.S. Centre for Disease Control Daten über Grippediagnosen von Ärzten sammelt, sollte die KI von Google Suchbegriffe für Grippe und Grippesymptome analysieren, um vorherzusagen, ob sich die Grippe in einem bestimmten Gebiet ausbreitet. Zunächst wurden 50 Millionen Suchbegriffe und über 100 Millionen verschiedene Vorhersagemodelle ausgewertet. Die Ingenieure wählten dann 45 (geheime) Suchbegriffe aus, aber nach einer Reihe von Fehlschlägen bei der Vorhersage von Grippepandemien (Schweinegrippe, MERS, SARS 1) erhöhten die Google-Ingenieure die Zahl auf etwa 160 Suchbegriffe. 2015 wurde das Projekt auf Eis gelegt. Anstelle von mehr Komplexität verfolgte ein Team, dem auch Forscher von Simply Rational angehörten, den gegenteiligen Ansatz: Was ist, wenn es bei großer Unsicherheit am besten ist, die Vergangenheit zu ignorieren und sich nur auf die jüngsten Datenpunkte zu verlassen? Das von uns entwickelte Modell verwendete einen einzigen "intelligenten" Datenpunkt, nämlich die Anzahl der grippebedingten Arztbesuche in der aktuellen Woche, und sagte dieselben Besuche für die nächste Woche voraus. In den acht Jahren, in denen wir die Grippe vorhersagten (2007-2015), war die Fehlerquote dieser intelligenten Regel nur etwa *halb so hoch* (0,2 % gegenüber 0,38 %) wie bei Google Flu Trends. Weniger ist mehr.

Analyse des Kundenstamms: Um die Effizienz und den Erfolg eines Unternehmens zu verbessern, ist es wichtig zu wissen, welche Kunden angesprochen werden sollen (z. B. für Marketingkampagnen) und welche der bestehenden Kunden in Zukunft Käufe tätigen und welche nicht. Bereits 2008 haben zwei ehemalige Kollegen gezeigt, dass eine einfache Management-Faustregel ein mathematisches "Optimierungsmodell" aus dem Marketingbereich (Pareto/NBD-Modell) über drei Märkte hinweg schlagen kann (siehe Abb. 4, links). Das einfache Modell, die so

genannte "Hiatus-Heuristik", wird von Vertriebsleitern in den unterschiedlichsten Branchen verwendet: "Wenn ein Kunde seit X Monaten keinen Kauf getätigt hat, stufen Sie ihn als inaktiv ein, ansonsten als aktiv." Wir haben uns kürzlich zusammengetan, um diese Analyse auf 61 (Einzelhandels- und Nicht-Einzelhandels-) Datensätze auszuweiten, wobei wir logistische Regressionen und wiederum Random Forests in unsere Vergleiche einbezogen haben. Auch hier konnte sich der transparente heuristische Algorithmus nicht nur gegen diese weitaus komplexeren Konkurrenten behaupten, sondern übertraf sie in all diesen Datensätzen und Anwendungen sogar erheblich (siehe Abb. 4, rechts).

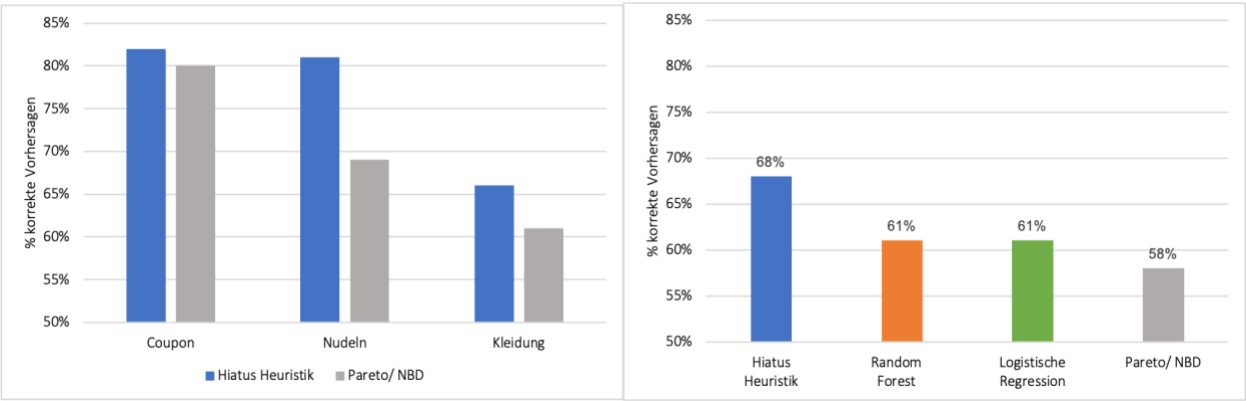

Abbildung 4. Vorhersage der Kundenaktivität. Eine transparente Heuristik schlägt nicht nur ein komplexes mathematisches Modell aus dem Marketing (linke Seite: Hiatus-Heuristik: blau; Pareto/NBD: grau), sondern auch statistische und moderne maschinelle Lernmethoden über 61 Anwendungen hinweg (rechte Seite: Hiatus Heuristik: blau; Random Forest: rot; logistische Regression: grün; Pareto/NBD: grau).

Transparenz: In einer unsicheren Welt kann weniger mehr sein

Diese Beispiele sind nur die Spitze des Eisbergs von Studien und Anwendungen, die gezeigt haben, dass einfache Modelle bessere Vorhersagen machen können, *selbst wenn viele Daten verfügbar sind*. Der Grund dafür ist der so genannte "Bias-Variance-Trade-off". Jeder Vorhersagealgorithmus kann zwei Arten von Fehlern produzieren: *Bias*, d. h. eine systematische Abweichung des geschätzten Durchschnittswerts vom wahren Wert, und *Varianz*, d. h. eine Streuung um den geschätzten Durchschnittswert. Wenn ein Algorithmus zu einfach ist, läuft er Gefahr, wichtige Faktoren zu „übersehen“ und wird systematisch und vorhersehbar daneben liegen – was zu einem hohen Bias führt. Ist ein Algorithmus jedoch zu komplex, besteht die Gefahr, dass er irrelevante Faktoren in seine Vorhersagen einbezieht, was den Schätzfehler erhöht und zu einer hohen Varianz führt. Bei vielen komplexen Algorithmen, die auf reale Probleme angewandt werden, kann der Vorhersagefehler aufgrund der Varianz weitaus größer sein als der Fehler aufgrund des Bias.

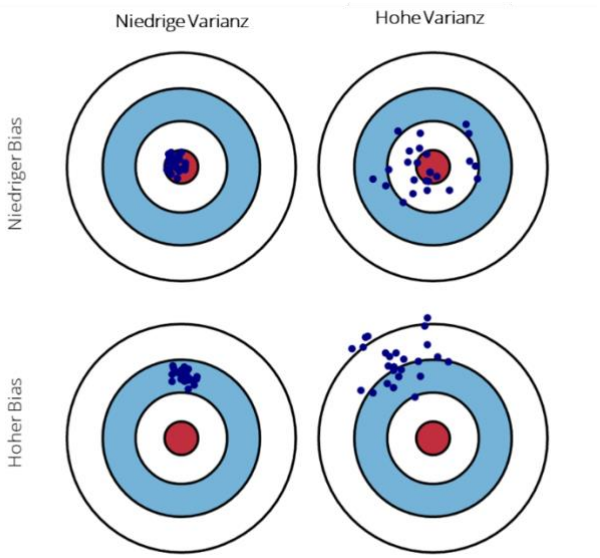

Abbildung 5. Ein Bias (die beiden unteren Ziele zeigen hohen Bias) ist eine systematische Abweichung des geschätzten Durchschnittswerts vom wahren Wert (dem Bullauge) - denken Sie an ein Gewehr mit einem schiefen Lauf, das das Ziel immer um denselben vorhersehbaren Grad verfehlt. Die Varianz (die beiden rechten Zielscheiben zeigen eine hohe Varianz) ist die Streuung der einzelnen Werte (der Einschusslöcher) um ihren Durchschnitt - denken Sie an eine abgesägte Schrotflinte, die eine größere Streuung hat als eine mit einem längeren Lauf.

Bias und Varianz sind eng mit dem Umfeld verbunden, in dem sie angewendet werden. Wenn eine Umgebung stabil ist und davon ausgegangen werden kann, dass die Vergangenheit der Zukunft gleicht, kann ein komplexer, ausgeklügelter KI-Algorithmus, der alle verfügbaren Daten nutzt, immer kleinere Nuancen herauskitzeln, um zu immer besseren Vorhersagen zu gelangen. Ist das Umfeld jedoch nicht stabil, stellen ihre vielen freien Parameter "Angriffspunkte" für Unsicherheiten dar, bei denen Schätzfehler zu einer erhöhten Variabilität und Schwankungen von Vorhersagen und schließlich zu einem Leistungsabfall führen. Im Gegensatz dazu hat ein einfacher, transparenter Algorithmus, der sich bei seinen Vorhersagen nur auf einige wenige robuste Faktoren stützt, weniger Angriffspunkte. Selbst wenn eine einfache psychologische KI und eine komplexe Black-Box-KI dieselben drei robusten Faktoren für ihre Vorhersagen verwenden, die komplexe KI aber zusätzlich 30 weitere Faktoren für ihre Vorhersagen nutzt, ist es wahrscheinlich, dass die einfache KI, die weniger Faktoren verwendet, die komplexe KI übertrifft, da sie nicht durch nutzlose Informationen in die Irre geführt wird.

Was bedeutet die EU-KI Verordnung für die Wirtschaft? Eine Gelegenheit!

Die Erkenntnis, dass die Transparenz von Algorithmen nicht im Widerspruch zu ihrer Vorhersagegenauigkeit stehen muss, kann sich erheblich darauf auswirken, wie man an die KI-Verordnung und die damit verbundene regulatorische Unsicherheit herangeht. Eine große

Unsicherheit betrifft den Abschnitt über die Risikostratifizierung, d.h. es ist für die meisten Unternehmen, die KI zur Steuerung ihres Geschäfts einsetzen, schwer zu sagen, welche ihrer Algorithmen in die Kategorie mit hohem Risiko eingeordnet werden. Weniger ungewiss ist jedoch, was die KI-Verordnung von den Unternehmen verlangen wird, um zu zeigen, dass sie diese Risiken effektiv handhaben können. Dazu gehören derzeit:

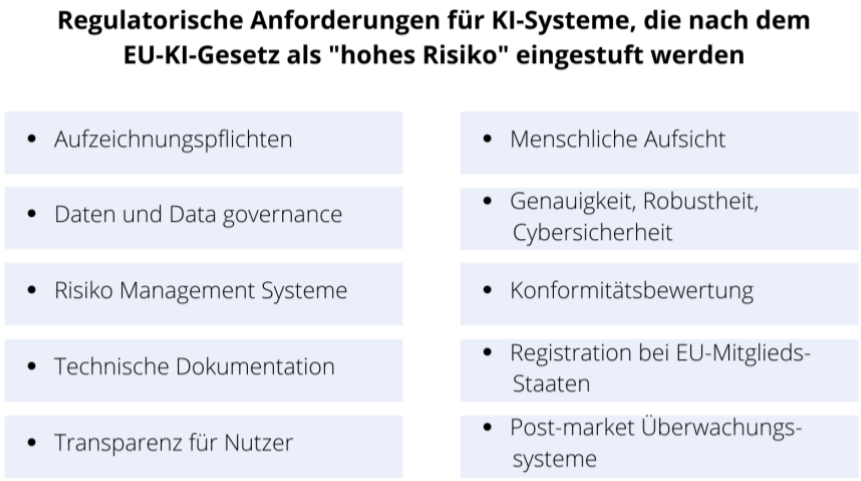

Abbildung 6. Erforderliche Regulierungsmaßnahmen für hochriskante KI. Viele der Anforderungen lassen sich sehr viel leichter erfüllen, wenn transparente Algorithmen anstelle von Black-Boxes eingesetzt werden.

In der endgültigen Fassung der KI-Verordnung können ein oder zwei weitere Maßnahmen hinzukommen, entfallen, oder sie können auf eine bestimmte Weise umstrukturiert werden. Aber im Großen und Ganzen wird es keine großen Überraschungen in Bezug auf die Systeme und Prozesse geben, die die Aufsichtsbehörden von den Unternehmen verlangen werden, um die Einhaltung der KI-Verordnung zu gewährleisten. Ein Blick auf jede dieser Anforderungen zeigt, dass transparente psychologische KI viele Aspekte erleichtert, sowohl die Bewertung vor als auch die Überwachung nach der Markteinführung. So kann auch die Kosten-Nutzen-Rechnung für viele der oben genannten regulatorischen Anforderungen positiv beeinflusst werden. Aufzeichnungen, Datenverwaltung und -management sowie die technische Dokumentation sind einfacher. Risiken können proaktiver verwaltet werden, und Risikomanager oder Compliance-Beauftragte brauchen keinen IT-Abschluss, um zu verstehen, was vor sich geht, einschließlich der Erkennung von Risiken durch Voreingenommenheit und andere ethische oder moralische Verstöße. Menschliche Aufsicht und Transparenz sind der Standard und begleitende Konformitätsbewertungen sind für alle Beteiligten weniger aufwändig. Gleichzeitig wird, wie bereits erwähnt, die Genauigkeit einfacher Modelle in vielen Fällen gleich oder sogar höher sein als bei den aktuellen, komplexen Lösungen. Und ihre Robustheit, d. h. ihre

Anwendbarkeit auf neue Fälle außerhalb derer, auf die der Algorithmus trainiert wurde, wird wahrscheinlich höher sein.

Psychologische KI-Algorithmen erfüllen nicht nur die meisten Anforderungen, die die KI-Verordnung wahrscheinlich stellt, sondern haben auch Vorteile aus geschäftlicher und leistungsbezogener Sicht. Sie ermöglichen eine einfachere Integration mit menschlichem Wissen und Fachkenntnissen. Dies führt zu einer besseren Gesamtleistung des Systems. Ein Beispiel ist das Schachspiel, bei dem die beste KI die besten Menschen schlägt, aber wenn beide zusammenarbeiten, können sie die beste KI schlagen. In vielen Situationen können transparente psychologische KI-Algorithmen auch kostengünstiger und einfacher implementiert, gewartet und überwacht werden. Wann immer es wichtig ist, den Menschen einzubeziehen, können sowohl der Nutzen als auch der Output eines Systems, das eine psychologische KI einsetzt, leichter kommuniziert, gerechtfertigt oder verteidigt werden. Dies hängt auch mit einem weiteren großen Vorteil der Transparenz zusammen: In allen Bereichen wie Versicherung, Kreditwürdigkeitsprüfung, Gesundheit, Rechtsprechung usw., in denen sich eine Verhaltensänderung des Kunden positiv auf ein Unternehmen auswirken kann, dient eine psychologische KI, die auf Faktoren basiert, die der Kunde aktiv beeinflussen kann, als Orientierungshilfe für dieses Verhalten in einem Prozess der wirklich gemeinsamen Entscheidungsfindung zwischen Unternehmensvertretern und bestehenden oder potenziellen Kunden. Wenn es darum geht, welche Art von KI Ihr Unternehmen einsetzen sollte, stellen Sie sich diese drei Fragen:

> Ist der komplexe Black-Box-KI-Algorithmus besser als die transparente und intuitive psychologische KI?
>
> → Wenn nein, verwenden Sie die psychologische KI.
>
> Wenn ja: Schlägt sich dieser Leistungsvorteil in einer besseren Systemleistung nieder (d. h. in der Leistung von Mensch und KI zusammen)?
>
> → Wenn nein, verwenden Sie psychologische KI.
>
> Wenn ja: Ist dieser Leistungsvorteil groß genug, um ihn gegen die aufwändigere und kostspieligere Einhaltung der regulatorischen Anforderungen für Black-Box-KI der KI- Verordnung aufzuwiegen?
>
> → Wenn nein, verwenden Sie psychologische KI

Nur wenn Sie alle Fragen eindeutig mit JA beantworten können, sollte Ihr Unternehmen Black-Box-KI-Modelle in Betracht ziehen.

Die KI-Verordnung: Eine Chance für den Fortschritt

Die Komplexitätsillusion (siehe Abbildung 2), die unhinterfragte und weit verbreitete Annahme, dass komplexe Algorithmen, die mehr Daten verwenden, immer besser sind als einfachere, ist eine der Hauptursachen für die zunehmende Intransparenz von KI-Tools und dem daraus

resultierenden Mangel an Vertrauen, den wir heute erleben. Wir können es besser machen. Wir wissen, dass komplexe, intransparente KI-Modelle in stabilen Umgebungen besser funktionieren und in instabilen Umgebungen nicht so gut. Und wir wissen, dass die transparenten psychologischen KI-Modelle, die in den letzten 20+ Jahren entwickelt wurden, in instabilen, dynamischen Umgebungen besser funktionieren. Vor diesem Hintergrund stellen wir die folgenden Forderungen für die KI-Verordnung auf:

> Wir schlagen vor, dass die KI-Verordnung Bestimmungen enthält, die einfache und transparente Algorithmen als Maßstab festlegen. Nur wenn ein komplexer Algorithmus eindeutige und sinnvolle Leistungsvorteile gegenüber einer einfachen Methode nachweisen kann, sollte er eingesetzt werden. Argumente, die sich auf nationale Sicherheit oder Geschäftsgeheimnisse stützen, sollten von einem unabhängigen EU-Fachgremium eingehend geprüft werden.
>
> Jeder Algorithmus, dessen Ergebnis das Potenzial hat, die Lebenssituation eines Menschen direkt zu beeinflussen, muss transparent sein, und zwar nicht nur für die Regulierungsbehörden, sondern auch für die Öffentlichkeit. Eine solche Transparenz wird vorzugsweise dadurch erreicht, dass von vornherein transparente psychologische KI-Modelle verwendet werden, anstatt durch Post-hoc-Methoden (wie Shapley-Werte oder LIME), die ihrerseits nur eine Annäherung an die zugrunde liegenden Black-Box-Algorithmen darstellen. Das Gleiche gilt für Simulatoren für intransparente Modelle, die nur einen Teil der Faktoren der zugrundeliegenden Black-Box-Modelle verwenden.
>
> Argumente, dass Endnutzer und andere Interessengruppen nicht in der Lage sind, eine solche Transparenz zu verstehen, müssen empirisch belegt werden. Selbst wenn nachgewiesen werden kann, dass die meisten Beteiligten die volle Transparenz nicht verstehen können, sollten für Beteiligte mit unterschiedlichen technischen Kenntnissen verschiedene Transparenzstufen eingesetzt werden.

Wir haben uns unser Leben schwieriger gemacht, als es eigentlich sein müsste. Es gibt viele Situationen, in denen einfache, transparente Algorithmen den intransparenten Black-Box-Lösungen, die derzeit von Unternehmen eingesetzt werden, überlegen sind. Vor diesem Hintergrund kann die KI-Verordnung der EU als eine Gelegenheit für Unternehmen gesehen werden, ihre Verwendung von „Black-Boxes“ für die Entscheidungsfindung zu überdenken.