Responsible AI

Transparenz, Bias, und Verantwortung in der KI

von Bernd „Benno“ Blumoser & Dr. Ulli Waltinger

Das Gebiet der Künstlichen Intelligenz mit ihren vielfältigen Disziplinen im Bereich der Wahrnehmung, des Lernens, der Logik und der Sprachverarbeitung hat in den letzten zehn Jahren signifikante Fortschritte in ihrer Anwendung gemacht. Ausschlaggebend für diesen Fortschritt waren: die Quantität von verfügbaren Daten, der Anstieg der Rechenleistung, die Verfügbarkeit von freien Software-Entwicklungsumgebungen und die Weiterentwicklung neuer Algorithmen und Architekturen aus dem Umfeld des Maschinellen Lernens. Systeme aus dem Bereich der Künstlichen Intelligenz und ihre algorithmischen Entscheidungsmuster beeinflussen immer mehr Elemente des täglichen Lebens: Die Interaktion im Haushalt mittels Heimassistenten, die Relevanz von Such- und Werbe-Angeboten, die mobile Fahr- und Verkehrsführung, die Diagnose in der Medizin oder auch die Vergabe des persönlichen Kreditrahmens.

“It’s not a human move. I’ve never seen a human play this move.” (Fan Hui zum 37. Zug des Go-Spiels zwischen Lee Sedol und AlphaGo, 2016).

Die herausragenden Erfolge vor allem im Bereich des überwachten Maschinellen Lernens und hier insbesondere der Aspekte des Repräsentations-Lernens und der tiefen Neuronalen Netze, dem sogenannten Deep Learning, haben signifikanten Einfluss nicht nur in der akademischen Welt, wie beim herausragenden Meilenstein im Go-Spiel zwischen Lee Sedol und AlphaGo, sondern zeigen auch in der industriellen Anwendung ihren Mehrwert.

Algorithmen machen unser Leben effizienter, unterstützen zunehmend unsere Entscheidungsprozess, teilweise übernehmen sie diese. Von der Spracherkennung und Übersetzung, der visuellen Qualitätsinspektion, der dynamischen Preisermittlung, der autonomen Parameter-Optimierung für Datenzentren, der verbesserten Adaptivität von Robotik-Steuerungen und KI-optimierten Lieferketten bis hin zur vorausschauenden Wartung in der Produktion.

Im Kontrast zum Industrieumfeld KI (B2B), welche schon immer den Fokus auf Effizienz- und Produktivitätssteigerung setzte, fokussiert sich der Konsumentenbereich (B2C) vor allem auf Aspekte der Prädiktion von Verhaltensmustern und Optimierung des Aufmerksamkeitshorizonts der Kunden. Von der Platzierung von Werbeanzeigen über das automatisierte Filtern von News-Artikeln bis hin zur KI-gestützten Bilderprüfung.

Die Stärke - und gleichzeitig die Gefahr - bei der Anwendung von Deep Learning zur Vorhersage oder Klassifizierung neuer Situationen besteht darin, dass die Qualität des Ergebnisses stark von der Größe, Ausgewogenheit und Reinheit der Inputdaten abhängt. Die Übertragbarkeit des Modells auf eine neue Fragestellung ist demnach nur sehr schwer einzuschätzen. Während in der Statistik das Konzept der Signifikanz die Aussagekraft jeder Analyse in jedem ihrer Schritte sehr kritisch hinterfragt, fehlt in der KI ein derart durchgestochenes Maß für die Güte eines Algorithmus‘. Validiert werden kann in der Regel nur die Qualität des Ergebnisses auf Basis der entsprechenden Trainingsdaten. Ob hier allerdings innerhalb vielschichtiger Neuronaler Netzwerke Korrelationen auftreten, die trotz großer Daten nur zufällig ein sinnvolles Muster ergeben (vielleicht weil sich Fehler gegenseitig aufheben oder gerade in diesem Anwendungsfall nicht zu relevanten Verzerrungen geführt haben) und welche Eingangsdaten verwendbar sind, um richtige Ergebnisse zu bringen, bleibt hier in der Regel verborgen. Dementsprechend kann ein solcher Algorithmus aufgrund falscher Annahmen im Lernprozess, wie z. B. einer geringen Vielfalt von Datenquellen, fehlerhafter Robustheit in wechselnden Anwendungsdomänen oder falschen Annahmen im Modellierungsprozess zu unausgewogenen Ergebnissen führen wie z. B. bei der geschlechterdiskriminierenden Vergabe von Kreditrahmen.

Dr. Ulli Waltinger ist Leiter der Forschergruppe Machine Intelligence bei Siemens Corporate Technology und Technologischer Leiter des Siemens AI Lab

Bernd „Benno“ Blumoser ist Innovationsleiter des Siemens AI Lab

“One of the first things taught: […] correlation is not causation. It is also one of the first things forgotten.” (Thomas Sowell, Stanford, 1930)

Gerade die verschachtelte nicht-lineare Struktur moderner Deep Learning-Systeme lässt den Anwendern aber auch KI-Experten nur schwer transparent erscheinen, welche Informationen oder Merkmale für eine Entscheidung vom System herangezogen worden sind. Welche Informationen eventuell nur einer zufälligen Korrelation geschuldet sind, aber möglicherweise keine signifikante Kausalität zulassen, ist nur schwer einzuschätzen. Daher wird diese Art von KI-System häufig als „Black Box System“ bezeichnet. Fehlerhafte Ergebnisse beeinträchtigen dann direkt die Wirtschaftlichkeit. Sobald aber auf Basis der Empfehlungen eines KI-Systems Entscheidungen getroffen werden, die wesentliche, vielleicht sogar existentielle Auswirkungen auf einzelne Personen oder Personengruppen haben, sind natürlich die Anforderungen an solch einen nicht-diskriminierenden Algorithmus ungleich höher und die Fehlertoleranz deutlich geringer. Wie also kann sichergestellt werden, dass ein System auch in einem realen Umfeld erwartbar funktioniert und die hohen Anforderungen an nicht-diskriminierende Ergebnisse erfüllt? Wie

kann das Risiko, von einer neuronalen „Black Box“ schlechte, unfaire und instabile Ergebnisse geliefert zu bekommen, reduziert werden?

“Professional responsibility [...] is not to discover the laws of the universe, but act responsibly in the world by transforming existing situations into more preferred ones.” (Herb Simon, 1996)

In einer ersten Näherung spielen natürlich die sich derzeit schnell weiterentwickelnde Regulierungen eine große Rolle dabei, die Risiken, die in Design und Nutzung der KI entstehen können, einzugrenzen. Neben existierenden Produkthaftungsgesetzen, der DSGVO oder Sicherheitsbestimmungen konkretisieren allerdings viele Institutionen und Unternehmen ihren Ansatz in Chartas oder Regelwerken, die den Besonderheiten der KI gerecht werden sollen. Bei Siemens nutzen wir einen Satz von sieben „mitigation principles“ (s. Abbildung unten), von denen wir glauben, dass sie dabei helfen, die unbestrittenen Vorteile der KI in einem verantwortlichen Rahmen nutzbar zu machen.

Neben sinnvollen, auf KI zugeschnittenen Regeln spielt als zweiter großer Hebel bei der Risikominimierung das ganzheitliche Einbeziehen verschiedenartiger Perspektiven eine große Rolle. Denn die Tatsache, dass wir in einer volatilen, von Unsicherheit geprägten Welt leben, in der wir alle blinden Flecken in unserer Wahrnehmung und Urteilskraft haben und damit oft in unbewusster Voreingenommenheit entscheiden, macht es erforderlich, diese menschliche Schwäche auch im Bereich der Künstlichen Intelligenz aus verschiedenen Perspektiven zu erkennen und zu korrigieren.

Diversität als elementarer Baustein im KI Lebenszyklus: Die Daten, mit denen KI-Systeme trainiert werden, müssen die Bandbreite ihrer späteren Anwendungsfälle abdecken, um zu gültigen Ergebnissen zu kommen. Aber unterschiedliche Perspektiven müssen auch im Forschungs-, Entwicklungs-, und Anwendungsbereich der KI einbezogen werden. Von Endnutzern über Domänenexperten bis zur Software-Entwicklung – wir müssen Vielfalt in all ihren Dimensionen wie Geschlecht, sozialer- und ethnischer Herkunft als gesellschaftliches Potenzial wertschätzen und integrieren. Diversität ist somit nicht nur der Motor für herausragende Innovationsleistungen, sondern auch elementar für die Minderung von Voreingenommenheit und Bias in der Künstlichen Intelligenz.

„Algorithmen machen unser Leben effizienter, unterstützen zunehmend unsere Entscheidungsprozess, teilweise übernehmen sie diese.“

Die Umsetzung geschieht dabei in kurzfristigen Innovationsformaten wie Hackathons, Bootcamps oder Innovation Sprints, aber auch in dedizierten Orten für Co-Creation, die den Anspruch erfüllen sollen, eine Plattform für unterschiedliche Perspektiven zu sein.

Ein eleganter und gerade im europäischen Kontext sehr relevanter Weg, um die Widersprüche zwischen Potenzial und Gefahr von KI auflösen zu können, stellen neue Technologien dar, welche einen ganzheitlichen Blick auf die Einflussfaktoren von KI-Anwendungen in ihren Lebenszyklen ermöglichen: Datengenerierung und -auswahl, Algorithmen-Auswahl und Erklärbarkeit, Genauigkeit und Laufzeit, aber auch Bereitstellung, Aktualisierung und Monitoring der Applikationen.

Hierzu gibt es bereits viele relevante technologische Bausteine, die dabei unterstützen können, implizite Vorurteile in den Daten und damit auch den KI-gestützten Empfehlungen aufzudecken und korrigierbar zu machen. Darüber hinaus tragen sie dazu bei, die hohen Anforderungen an den Schutz der persönlichen Daten zu erfüllen, ohne durch die höhere Komplexität in Prozessen und Produkten einen Wettbewerbsnachteil zu haben, und verbessern schließlich auch die Robustheit der KI-Systeme, was zugleich deren Fehleranfälligkeit minimiert und die ökonomisch attraktive Skalierbarkeit auf weitere Applikationsfelder erleichtert. Aktuell relevante Technologien sind:

Explainable AI ist ein Feld, das die Interpretierbarkeit von Black-Box-Entscheidungen in der KI adressiert. Hierbei wird in Erklärbarkeit vor (z.B. Daten, Eingangsmerkmale), während (z.B. Model-Architekturen, Relevanzmerkmale) und nach (z.B. Test- und Ziel-Referenzen) der Modellierung unterschieden. In der Industrie werden diese Methoden u.a. in Kombination mit Black-Box-Ansätzen verwendet, um die Genauigkeit der KI-Algorithmen erklärbar zu machen. Dies hilft den Prozess für Kunden verständlicher zu machen, aber auch system-interne Verzerrungen zu veranschaulichen.

Active Learning ist ein weiteres aufstrebendes Feld in der KI, das nicht nur den Prozess der KI mit wenig gelabelten Daten beschleunigen kann, sondern auch das „Human-in-the-Loop-Paradigma“ innerhalb der KI prägt. In der industriellen Anwendung erlaubt dieser Ansatz Rückmeldung von Domänen-Experten in den KI-Trainingszyklus zu integrieren, d.h. ▶

Let’s create a future-oriented society together with Responsible Industrial Artificial Intelligence

01 **Shape sustainable development**
Increase our positive economic, societal and environmental impact and thus contribute to achieving the Sustainable Development Goals

02 **Foster inclusiveness & shared benefit**
Ensure diversity, fairness and inclusiveness by co-creating value for all stakeholders in a multidisciplinary approach

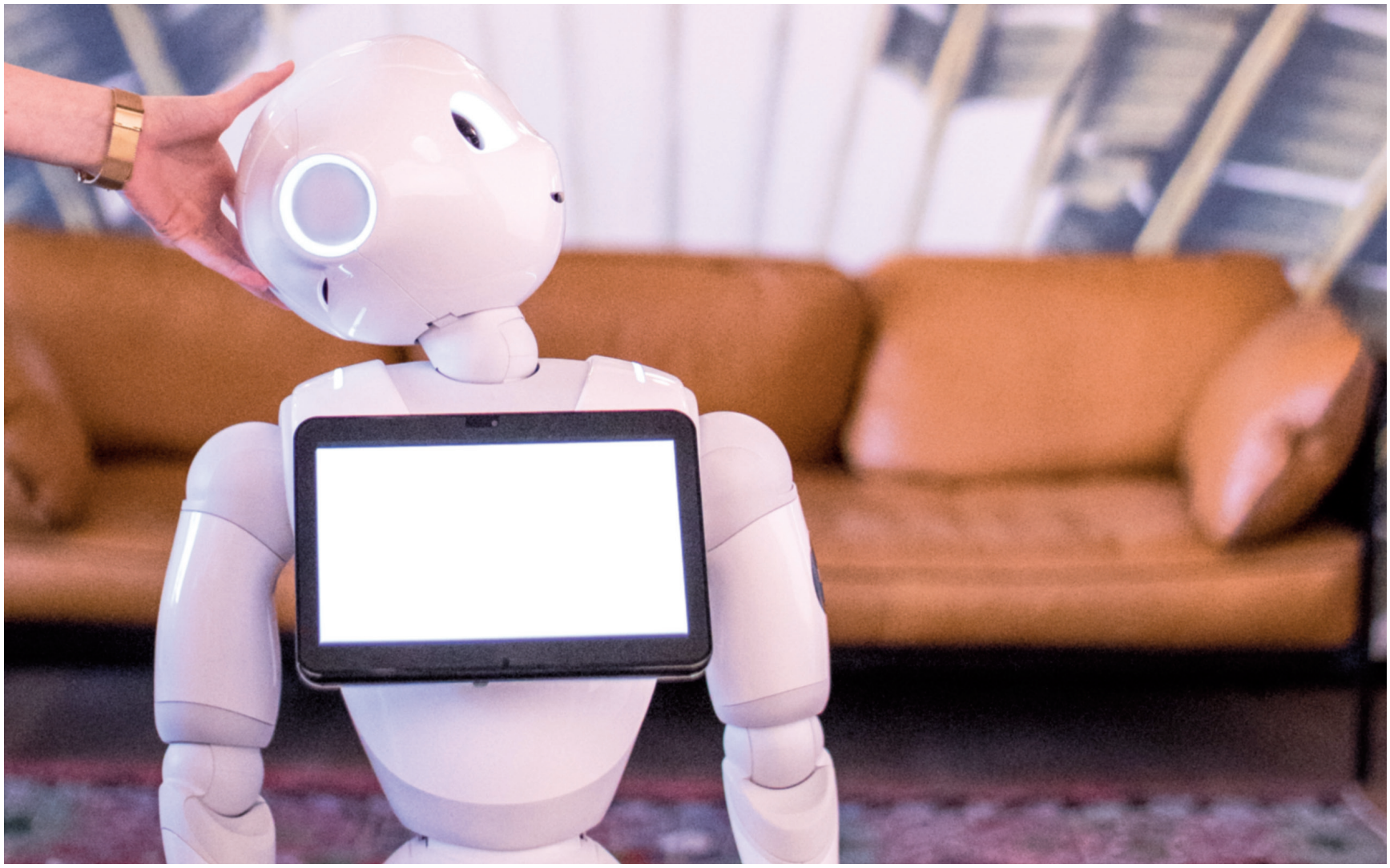
03 **Safeguard human oversight**
The design of AI systems should always convey the objectives clearly defined humans

04 **Guarantee data governance & privacy**
Protect fundamental rights of partners, respecting their right to the protection and governance of personal and non personal data

05 **Ensure system security & safety**
Apply honest, credible, holistic rules and concepts as standards for security and safety

06 **Endorse explainability**
Create awareness, trust and acceptance by explaining the rationale of AI solutions whilst safeguarding intellectual property

07 **Promote accountability & liability**
Make policies and processes clear and accessible to guide stakeholders to take responsibility



Auch Pepper, der Junior Concierge im Siemens AI Lab, braucht zuweilen ein wenig human control, um wieder in die Spur zu kommen.

das System durch menschliches Nutzungsverhalten und Domänen-Wissen kontinuierlich und effizient zu verbessern.

Trustworthy AI zielt auf das Gebiet der Vertrauenswürdigkeit und Robustheit von Algorithmen ab, welche dem Nutzer Feedback über Fehler, Robustheit oder Inkonsistenzen in allen Phasen der KI geben. Ziel ist es, dass KI-Anwendungen in die Lage versetzt werden, einen möglichen Domänen-Wechsel und dementsprechende adaptive Unsicherheit zu erkennen und rückzumelden.

Federated Learning ist ein verteilter Ansatz des maschinellen Lernens, der das Modelltraining auf großen Datenmengen dezentral-verteilter Edge-Geräte ermöglicht. Die Grundidee dahinter ist, den Code zu den Daten, anstatt die Daten an den Code zu schicken, und adressiert hier die grundlegenden Aspekte der Privatsphäre, Eigentum und Speicherort der Daten

Differential Privacy ist ein mathematisches Verfahren zur Anonymisierung von Datensätzen über die Verwendung von Metadaten, wodurch die Privatsphäre des Einzelnen gewahrt werden kann. Ein Algorithmus analysiert dabei einen Datensatz und berechnet Statistiken darüber (z. B. den Mittelwert, die Varianz, oder den Median). Er wird als differenziell privat bezeichnet, wenn man anhand der Ausgabe nicht erkennen kann, ob die Daten einer Person im ursprünglichen Datensatz enthalten waren oder nicht.

„Diversität ist nicht nur der Motor für herausragende Innovationsleistungen, sondern auch elementar für die Reduzierung von Voreingenommenheit und Bias in der KI.“

Edge AI erlaubt nicht nur die Echtzeitverarbeitung von Daten, welche auf einem Hardware-Gerät („Edge“) eingesammelt werden, sondern erlaubt, dass die KI als Treuhänder agiert, da die Datengenerierung und -auswertung innerhalb der Edge durchgeführt wird. D.h., es können Daten verarbeitet und Entscheidungen dezentral getroffen werden, ohne dass eine Verbindung zu einem zentralen System (z. B. Cloud) bestehen muss.

Ein verantwortungsvoller Umgang mit der Künstlichen Intelligenz erfordert jedoch nicht nur ein technisches und institutionelles Steuern und Überwachen des KI Prozesses hinsichtlich Bias, Fairness, Transparenz, Verantwortlichkeit und Erklärbarkeit, sondern auch eine kontinuierliche Weiterqualifizierung der Entwickler, Anwender und Entscheider. Diese müssen die Vorteile, Gefahren und Verfahren zur Risikominderung von KI-Methoden und Applikationen in Trainings kennen und einschätzen lernen.

“Trust is not necessarily about transparency but about interaction.“ (Ulli Waltinger, 2019)

Vertrauen und Sicherheit sind die wichtigsten Imperative für Mensch, Prozess und Produkte im gesamten Lebenszyklus der KI. Die Vorteile aber auch Konsequenzen von KI entfalten sich immer noch und werden weiterhin die Gesellschaft und Wirtschaft grundlegend verändern. Daher ist es umso wichtiger, den technologischen und kulturellen Wandel gemeinsam und verantwortlich zu gestalten.