Vorteile der KI nutzen, ohne Akzeptanz zu verlieren

Einordung von überraschendem Verhalten und Hinweise für dessen Beherrschung

Stand: 06.05 2022

Einleitung

Der Einsatz von Künstlicher Intelligenz (KI) im industriellen Umfeld ist ein weiterer Schritt und Treiber, Geschäftsmodelle neu zu denken, sowie Optimierungspotentiale in Prozessen zu heben. Potentiale, die durch ihre Komplexität und ihre schiere Menge an zu bewertenden oder sich dynamisch ändernden Informationen, nicht durch den Menschen bewältigt werden können.

Industrielle technische Systeme können – so der Gedanke – durch KI funktional erweitert, sowie autonomer und resilienter gegen äußere Einflüsse gestaltet werden. Aktuelle Herausforderungen in Optimierungsprozessen zur Minimierung des Energie- oder Ressourcenverbrauchs sind beispielhafte Aufgaben in denen KI-basierte Algorithmen Wirkung zeigen.

Der Vorteil der KI, Maschinen und Anlagen während der Laufzeit an ihren Nutzungsprozess zielorientiert anzupassen, kann erst dann ausgespielt werden, wenn die beteiligten Menschen diese Veränderungen annehmen und akzeptieren. Für den Menschen ist eine Maschine eine „Black Box“, die sich im Wesentlichen immerzu gleich verhält und wenig bis keine Überraschungen im Maschinen- oder Produktionsablauf vorhält. Mit KI „im Bauch“ kann sich diese Situation deutlich verändern, gerade dann, wenn die durch KI beeinflussten Prozesse eine direkte Interaktion mit den Menschen aufweisen. Dies können Änderungen in der Prozessreihenfolge, Änderungen von technischen Parametern (wie z.B. Beschleunigung) oder Reaktionen auf diese Interaktion (wie z.B. Feedback auf Maschinenbedienung), ungewohnte Geräusche oder auch optische Verhaltensänderungen sein. Die Einordnung von solchen Änderungen des Maschinenverhaltens und wie der Mensch darauf reagiert – wie etwa unangenehm berührt zu sein oder gar dieses Verhalten ganz abzuweisen - ist Kern des nachfolgenden Beitrags.

Veränderungen akzeptieren und mit Überraschtheit umgehen

Der Nutzen von KI-Technologien bildet sich vielfach in Funktionen ab, die für den Anwender oft nicht sichtbare, aber für den Produktionsprozess spürbare Nutzenpotentiale erschließen. Diese Optimierungen in Prozessen von Maschinen- und Anlagenfunktionen bzw. Service- oder Supportfunktionen unterstützen den Menschen bei seiner täglichen Arbeit. Werden KI-Funktionen in der Mensch-Maschine-Interaktion spürbar, kann dieses für den Menschen zu überraschenden Situationen führen, in denen er das veränderte Verhalten einer Maschine trotz Überraschtheit dennoch akzeptiert oder aber eben nicht mehr akzeptiert.

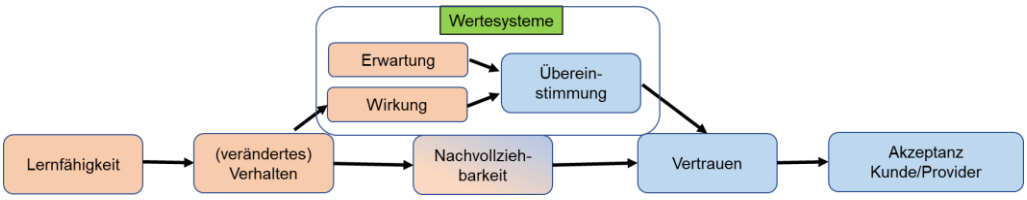

Bild 1: Erwartung und Wirkung bei der Akzeptanzgenerierung in KI-basierten Systemen

In einfachsten Fällen werden in der Lernphase die anzuwendenden Algorithmen ertüchtigt, nachvollziehbare Handlungen mit einem erwarteten Ergebnis durchzuführen. Führen die

Lernphase und Lernfähigkeit des KI-basierten Systems bei der Ausführung zu einem veränderten Verhalten, welches nachvollziehbar oder auch nicht nachvollziehbar sein kann, so kommt es beim Einsatz von KI-Technologien darauf an, ob die Erwartung des Kunden mit der entstandenen Wirkung des KI-System übereinstimmt. Nur dann wird Vertrauen und Akzeptanz zu generieren sein.

Akzeptanz setzt sich dabei in Summe aus verschiedensten Komponenten zusammen, wie dies in Bild 1 veranschaulicht ist. Die Übereinstimmung zwischen der spürbaren und erwarteten Veränderung ist hierbei ein wesentlicher Faktor. Dieser Faktor ist auch an die Veränderbarkeit und Lernfähigkeit der Maschine und Anlage über KI-Algorithmen (in orange dargestellt) und des erlebten Wertesystems des Menschen (in blau dargestellt) gekoppelt.

Ist die Veränderung nachvollziehbar und entsprechend nachhaltig, um Vertrauen in die Maschinenfunktion aufbauen zu können, erhöht sich die Akzeptanz und die Überraschtheit bei sich veränderndem Verhalten nimmt ab.

Der Erfolg eines Maschinenherstellers hängt am Ende von der Akzeptanz des Nutzers ab, der neben der technischen Funktionsbeschreibung der Maschine oder Anlage auch die Beschreibung der durch KI gestalteten technischen Möglichkeiten erkennen und bewerten muss.

Als Indikator dafür ist bei der Verwendung von KI in den Maschinen und Anlagen der Grad an Überraschtheit zu sehen, die den Zusammenhang zwischen vertrautem und übereinstimmendem Verhalten und der akzeptierten Verhaltensänderungen beschreibt.

Überraschtheit und Akzeptanz messbar machen

Bei der Anwendung von KI-Systemen ist die Vertrauenswürdigkeit des Systems für die Kundenakzeptanz von wesentlicher Bedeutung. Das Vertrauen in Empfehlungen, Entscheidungen und das Verhalten von (technischen) Systemen unter dem Einsatz von KI stellt aktuell sowohl eine technische, organisatorische wie normative Herausforderung dar, welche noch nicht ausreichend adressiert wird. Aus diesem Grund ist beim Einsatz von KI auf die Rolle des Menschen und seiner Interaktion mit dem KI-System sowie eine geeignete Integration von Mensch und Maschine zu achten.

Bild 2: Ablauf für die Erstellung des Orientierungsrahmens

Mit dem Ansatz, die Akzeptanz messbar und in ihrer Konsequenz bewertbar zu machen, wird ein Prozess vorgeschlagen, der neben deterministisch technischen Parametern auch nicht deterministische Betrachtungsformen enthält (siehe auch Bild 2).

Technisch organisatorische Auslegung

Für die technisch organisatorische Auslegung kann auf bekannte Mechanismen wie z.B. den Risikografen der funktionale Sicherheitstechnik zurückgegriffen werden. Er dient zur Bewertung des Sicherheits-Integritätslevels (s. auch IEC 61508), der das Schadensausmaß, die Gefährdungsexposition, die Schadensvermeidungsmöglichkeit und die Eintrittswahrscheinlichkeit miteinander kombiniert, um das Risiko für einen Schadenseintritt abzuschätzen.

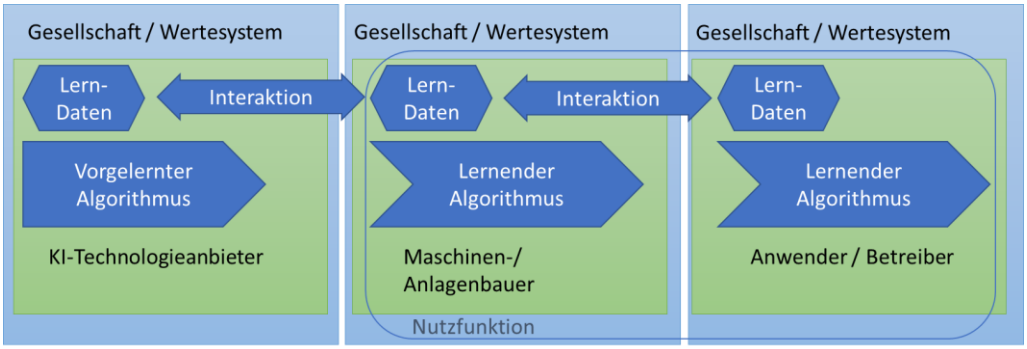

Bild 3: Technische Parameter und Systemgrenzen

In der Analogie dazu werden bei KI und Akzeptanz die Eingangsparameter durch den Systemaufbau der verwendeten KI unter Betrachtung deren genutzter Systemgrenzen bestimmt. Ausschlaggebend sind dabei die Phasen und Systemgrenzen des Algorithmus in der Vorlern-Phase beim KI-Hersteller und der Lerneffekte bei der Erstellung der Maschine bzw. Anlage und der weiteren dazu gelernten Inhalte im tatsächlichen Betrieb des Systems. Betrachtet werden neben dem KI-Algorithmus die verwendeten Daten und ihr Ursprung bezüglich der Datenerhebung und Qualifizierung. Dies schließt auch die Betrachtung von Rechtsräumen mit ein. Die Art und die Hintergründe der Erstellung des KI-Algorithmus ist für das Vertrauen in ein maschinelles System und damit für die Akzeptanz des Systems beim Betreiber eine wesentliche Entscheidungsbasis.

Auslegung durch nicht deterministische Parameter

Zur weiteren Abschätzung der Akzeptanz wird auf nicht deterministische Parameter zugegriffen, die den Grad der Überraschtheit abschätzen lassen. Um diese Sichtweise zu erfassen, spielen vier Ebenen für die Risikowahrnehmung wichtige Rollen:

- Auf der ersten Ebene wird die individuelle Informationsverarbeitung beschrieben: Jeder Mensch baut durch seine Wahrnehmung der Umgebung und den Umgang mit Dingen Erfahrungswissen auf und entwickelt so einen „individuellen gesunden Menschenverstand“. Mit dessen Hilfe erfasst und bewältigt er dann neue Situationen, gerade auch dann, wenn er dabei mit unvollständigen Informationen und Wissen gegebenenfalls sogar noch unter Zeitdruck zu annehmbaren Entscheidungen kommen muss. Diese

Fähigkeit hilft ihm auch bei Überraschungen, wenn sich eine Situation der Erfahrung entzieht oder ihr widerspricht und Heuristiken der Informationsverarbeitung scheitern.

- Die zweite Ebene wird durch kognitiv-affektive Faktoren gebildet: Wenn das situative Erleben den individuellen oder kollektiven Überzeugungssystemen zuwiderläuft, dann formt dies ein „Überrascht-Sein“. Da hierbei individueller bzw. kollektiver Glaube, in Form von Zuneigung oder Stigmata, ins Spiel kommen, ist die Relevanz dieser Form der Überraschung tendenziell sehr hoch.
- Die beiden letzten Ebenen beziehen sich auf den sozialen und politisch-institutionellen Rahmen sowie auf den kulturellen Hintergrund. Darin spiegeln sich einerseits die Werte und Regeln und andererseits die Grundüberzeugungen eines Kollektivs als Identität wider.

Diese vier Kontext-Ebenen können als Faktorisierung der Überraschung herangezogen werden. Die Komplexität der Zusammenhänge beim Einsatz KI-basierter autonomer Systeme erfordert neben der Diskussion dieser verschiedenen Dimensionen aber eine handhabungsfähige Methodik, um die Akzeptanz bzgl. der Überraschtheit abzuschätzen.

Mit Überraschung bezeichnet man das Erleben unvorhergesehener Situationen. Als Analogie kann die Methode für die Beurteilung von Lawinengefahren für „Skibergsteiger“ betrachtet werden:

Bei Skitouren heißt die zu vermeidende Überraschung Lawine. Früher war die Beurteilung der Lawinengefahr eine Aufgabe, welche die Schneedecke auf den Hängen anhand von verschiedenen Faktoren bewertete. Diese klassische Analyse war aus mehreren Gründen fehlerbehaftet. In der Schneedecke laufen sehr komplexe Vorgänge ab, die eine verlässliche Vorhersage zur Lawinengefahr kaum zulassen. Zusätzlich standen der Mangel an fundiertem Wissen und die subjektive Wahrnehmung der Faktoren durch den Menschen einer sicheren Vorhersage der tatsächlichen Lawinengefahr entgegen.

Der Schweizer Lawinenforscher Werner Munter erkannte dies und entwickelte mit der 3x3 Reduktionsmethode eine auf Statistiken und Erfahrungswerten beruhende datengetriebene Entscheidungsstrategie. Sie erleichtert und vereinfacht den Umgang mit der komplexen und im Einzelnen nicht mit der notwendigen Exaktheit erfassbaren Zusammenhänge der Lawinenbildung im winterlichen, alpinen Gebirge. Dabei zeigt sich, dass in der praktischen Anwendung bereits die Beurteilung von wenigen Risikofaktoren ausreicht, um eine gute Aussage über das Risiko zu machen.

Die Munter-Methode gehört zu den probabilistischen Methoden, bei denen nicht auf Detailfragen eingegangen wird, sondern Erfahrung, die wiederum in Wahrscheinlichkeiten ausgedrückt wird.

Charakteristisch ist, dass die auf Erfahrung beruhende Methode nicht allgemeingültig ist, sondern nur auf den Raum zutrifft, in dem die Erfahrung gemacht wird. Sie gilt für die Alpen. Sie ist nicht 1:1 auf die Anden, den Himalaya oder den nordamerikanischen Kontinent übertragbar.

Das aufgeführte Beispiel aus der Welt der Skitouren zeigt, dass es für eine Risikoabschätzung oft genügt, wenige greifbare Kriterien im Sinne einer Heuristik zu bewerten, um Gefahren ausreichend vermeiden zu können. Für die Bewertung des Grades an Überraschtheit wird ebenfalls ein heuristischer Ansatz gewählt. Dazu werden vier essenzielle Faktoren ausgewählt, mit denen es auf einfache Art gelingt, diesen Grad auszudrücken. Überraschtheit kann dabei erst dann entstehen, wenn ein Ereignis eintritt, mit dem nicht gerechnet wurde.

Auf die Maschinen- und Anlagenwelt bezogen, handelt es sich bei diesem Ereignis um eine Änderung des Verhaltens einer Maschine bzw. Anlage. Dies setzt die technische Fähigkeit eines Systems voraus, sich in seiner Auswirkung auf den Menschen und seine Umgebung verändern zu können. Der erste Faktor beschreibt genau diese **Verhaltensänderung** in Prozent, wobei eine kleine Prozentzahl eine marginale Verhaltensänderung und eine große Prozentzahl eine gravierende Verhaltensänderung ausdrückt. Die Verhaltensänderung kann auch als Abweichung vom erwarteten Verhalten ausgedrückt werden.

Die Überraschtheit wird generell durch eine solche Verhaltensänderung der Maschine bzw. Anlage ausgelöst; wie überrascht auf diese aber letztendlich reagiert wird, hängt von mindestens drei weiteren Erfahrungswerten des Individuums ab. Diese Erfahrungswerte beziehen sich auf die Ratio (**Erklärbarkeit**), die Kultur (Rolle des **Individuums**) und das Milieu (**Risikosensibilität**) desjenigen Menschen, der die Verhaltensänderung feststellt.

Der Ratio-Erfahrungswert **Erklärbarkeit** gibt an, ob für das abweichende Verhalten eine plausible Erklärung gefunden werden kann oder nicht. Wenn das Ergebnis nachvollzogen werden kann, ist es nicht mehr sehr überraschend, dass es zu diesem Ergebnis gekommen ist. Natürlich hängt die Nachvollziehbarkeit eines Ergebnisses wieder von den Fähigkeiten des Individuums ab, wie z.B. dem Expertenstatus, der Ausbildung, der Erfahrung, dem Wissen und weiteren erkenntnisbildenden Faktoren.

Der nächste Erfahrungswert **Individualität** hängt mit der Kultur des Individuums zusammen und hier insbesondere, wie das individuelle Verhalten relativ zum kollektiven Verhalten der Gesellschaft ausgeprägt ist. Wenn die Kultur eher das Kollektiv in den Mittelpunkt stellt, ist ein bestimmtes Maß an Fremdbestimmtheit nichts Ungewöhnliches. Dementsprechend ist aufoktroyiertes Verhalten eher akzeptiert. Auch das abweichende Verhalten einer Maschine wird dann eher geduldet. Steht hingegen das Individuum im Mittelpunkt, wird die Abweichung des Verhaltens als Einschränkung der persönlichen Entfaltung und damit störend wahrgenommen. Dementsprechend ist das abweichende Verhalten einer Maschine dann ebenfalls störend und auffällig.

Der dritte Erfahrungswert **Risikosensibilität** hat mit der Stabilität der Umgebung/des Milieus eines Individuums zu tun. Wenn ein Individuum in einer instabilen Umgebung „aufgewachsen“ ist, sind Änderungen vom „Normalzustand“ an der Tagesordnung und damit nicht besonders überraschend. Dementsprechend wird auch abweichendes Verhalten von Maschinen und Anlagen toleriert. Ist das Individuum in einer sehr stabilen (behüteten) Umgebung aufgewachsen, wird eine Verhaltensänderung u.U. regelrecht als Bedrohung empfunden.

Die vier Faktoren **Verhaltensänderung**, **Erklärbarkeit**, **Individualität** und **Risikosensibilität**, die jeweils Werte zwischen 0 und 1 annehmen können, spannen eine Fläche auf, deren Flächeninhalt den Grad der Überraschtheit ausdrückt, wie in Bild 4 dargestellt.

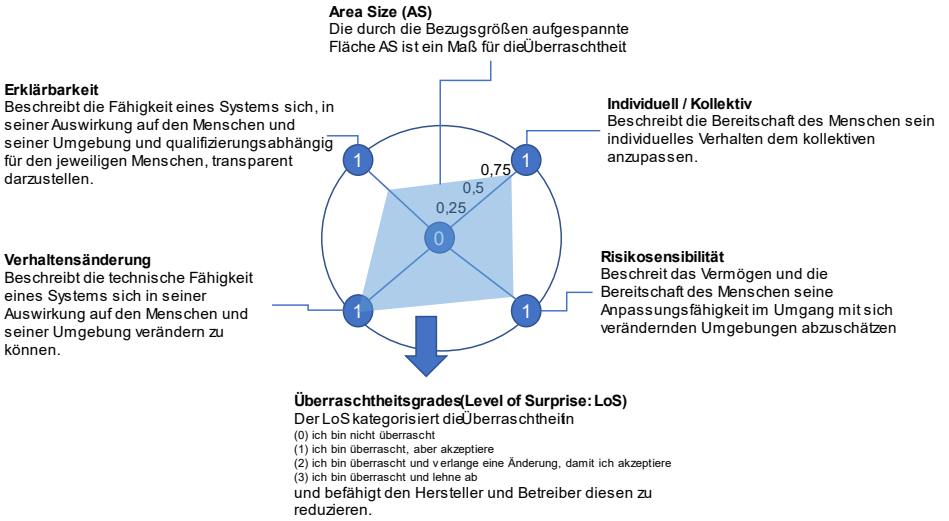

Bild 4:Verfahren zur Ermittlung des Überraschtheitsgrades (Level of Surprise: LoS)

Die Größe des Flächeninhalts (Area Size; AS) wird in vier Kategorien unterschieden, denen jeweils ein Überraschtheitsgrad (Level of Surprise: LoS) von 0 bis 3 zugeordnet wird:

LoS 0 (0% < AS < x): ich bin nicht überrascht

LoS 1 (x < AS < y): ich bin überrascht, aber akzeptiere

LoS 2 (y < AS < z): ich bin überrascht und verlange eine Änderung, damit ich akzeptiere

LoS 3 (z < AS < 100%): ich bin überrascht und lehne ab

Abhängig vom errechneten Grad der Überraschtheit (LoS) werden in der weiteren Diskussion der Arbeitsgruppe 2 „Technologie- und Anwendungsszenarien“ der Plattform Industrie 4.0 diese Überraschtheitsgrade eingeteilt und entsprechende Handlungsempfehlungen abgeleitet. Diese sollen dem Ersteller oder Nutzer von Maschinen und Anlagen, in denen KI verwendet wird, aufzeigen, welche Maßnahmen er bei der jeweiligen Verwendung der Maschine oder Anlage bezogen auf den Anlagenstandort, den Nutzer und die eingesetzten KI-Technologien zu berücksichtigen hat. Weiterhin geben die technischen Kriterien Aufschluss über die Einbindung von KI-Technologien in die Konstruktion und Organisation von Maschinen und Anlagen, um diese in unterschiedlichen Regionen, mit vielfältigen Menschen und/oder vernetzten Maschinen anwendergerecht betreiben zu können.

Zusammenfassung

Durch eine methodische Herangehensweise und den Ansatz, Überraschtheit als Maß für die Akzeptanz von KI-basierter-Systeme zu verwenden, sollen Hersteller und Anwender in die Lage versetzt werden, die Funktions- und Wirkungsweise von Maschinen und Anlagen in Verbindung mit dem Menschen und seiner Umgebung zu organisieren, zu bewerten und in ihre Entscheidungen über den gesamten Lebenszyklus mit einzubeziehen.

Autoren
Diese Publikation ist ein Ergebnis der Arbeitsgruppe „Technologie- und Anwendungsszenarien" der Plattform Industrie 4.0.

Dieses Papier entstand unter der Mitarbeit von Johannes Kalhoff, Johannes Diemer, Alexander Schließmann, Stefan Elmer, Gerd Bachmann.

Kontakt: Geschäftsstelle Plattform Industrie 4.0, Bülowstraße 78, 10783 Berlin

geschäftsstelle@plattform-i40.de