Klaus-Jürgen Grün
3.
Werte und KI
Die Wertung der Werte durch KI
Müssen KI-Systeme Werte haben? Von Menschen sagt man, sie „haben“ Werte. Werte erscheinen dabei als ein Besitzstand. Können KI-System etwas „besitzen“? Und wie würde sich das für ein solches System anfühlen, wenn es etwas besitzt? Vom Standpunkt der Maschine aus betrachtet, verändert sich auch unsere Auffassung von Werten. Aber das haben Werte schon immer getan.
Unsere heutige Auffassung von Werten ist eine recht junge Erscheinung. Aus den Studien zur antiken Philosophie spricht eine Vorstellung von Werten, die ihren Ursprung in der Wahrnehmung von Nützlichem hatte.

Was für den Zusammenhang zwischen Ethik und KI gilt, wird in diesem Kapitel auf das System der Werte angewandt. Wenn ein *Lexikon der Werte*[1] das Wort „Achtsamkeit“ aufführt, dann verliert sich vor dem Hintergrund der Intelligenzverkörperung in Maschinen seine Bedeutung als Wert. Warum sollen wir beispielsweise einem Roboter Achtsamkeit entgegenbringen? Das kann sich nur auf einen schonenden Umgang mit Ressourcen beziehen. Und für eine solche Forderung bedarf es keiner moralischen oder ethischen Aufladung. Es genügen die am Nutzen orientieren Interessen. Denn es verursacht Kosten, wann man einen undichten Roboter im Regen stehen lässt. Außerdem erfordern unsere ethisch-moralischen Empfindungen Reziprozität. Nur weil wir alle Menschen für „vernunftbegabt“ halten, fordern wir wechselseitig Vernunftgründe - was auch immer darunter verstanden werden mag - für unser Handeln. Von einem geistig behinderten Menschen zu verlangen, dass er sich an einem vorgegebenen Maßstab von Vernünftigkeit orientiere, könnte am Ende auf körperliche Nötigung hinauslaufen. Gleiches gilt für KI-Systeme.
Die Betrachtung unserer Wertschätzung für Werte gewinnt neue Konturen, wenn wir sie vor dem Hintergrund des Umgangs mit KI-Systemen betrachten.
Werte haben keinen eingebauten Wert, keine Bedeutung an sich, so wie Wegweiser auch keinen ihnen eigenen Weg beschreiten. Was Werte sind, erzeugt sich im Gebrauch des Wortes in bestimmten Zusammenhängen. Welche Werte einer Maschine einprogrammiert werden, hängt davon ab, welche Werte der Programmierer hat.

Wenn wir uns die Geschichte der Entwicklung eines Bewusstseins für Werte anschauen, dann erkennen wir, dass sie aus dem inneren Bedürfnis des Menschen entstehen, etwas wertzuschätzen. So sind in der Antike Werte gebunden an die ökonomischen Interessen der Menschen. Um den Wert aus der Willkürlichkeit zu befreien, versuchten Philosophen stets einen „natürlichen“, also objektiven Wert von Dingen zu erfassen. In der spätrömischen Stoa schreibt Seneca den natürlichen Dingen einen Wert zu, den unnatürlichen dagegen einen Unwert. Seine Wertvorstellung fügt sich ein in die Forderung nach einem Leben im Einklang mit der Natur.
Diese Vorstellung passt zu der in den in der Philosophie bis nach dem Mittelalter gebräuchlichen Tugendethik. Tugenden können wir als handlungsleitende Werte auffassen. Zumeist werden seit Aristoteles vier Kardinaltugenden angegeben, also Primärtugenden, die als „Dreh- und Angelpunkt“ (cardo) gelten. Regelmäßig werden hier genannt: Tapferkeit, Mäßigung, Besonnenheit, Gerechtigkeit. In dem, was Tugenden verlangen, wird man tüchtig, indem man es tut. Wer also ein liebenswürdiger Mensch sein will, der muss sich in Liebenswürdigkeit üben; wer ein toleranter Mensch sein will, muss sich in Toleranz üben und wer ein besonnener Mensch sein will, der muss sich in Besonnenheit üben.

---

[1] Eine Liste aller Werte will die *Enzyklopädie der Wertvorstellungen* sein (https://www.wertesysteme.de/alle-werte-definitionen/). Die Seite bietet die komplette Liste aller im Wörterbuch definierten Werte-Begriffe (129) als Übersicht sowie die wichtigsten Wertesysteme und Synonyme.

In den Ausführungen zur Tugendethik steht ebenfalls eine objektive Werthaftigkeit der jeweiligen Tugend im Vordergrund. Autoren denken sich die Tugend der Gerechtigkeit deswegen als Tugend, weil Gerechtigkeit ein objektiver Wert sei. Diese Auffassung spiegelt auch das althochdeutsche Wort „Werd“ wider. Es bezeichnet den „Preis“ oder die „Kaufsumme“. Darin enthalten ist die Bedeutung von „Geltung“ und „(Wert-)Schätzung“. So soll das Wertsein und das Werthaben einer Sache benannt werden. Schließlich spüren wir mit dem deutschen Wort „Geld“ noch einen Nachhall der Forderung nach absoluter „Geltung“. Aber mit dem Blick auf das Wort „Geld“ offenbart sich auch das bislang uneinlösbare Versprechen absoluter Werte. Natürlich sagen Moralisten und Ethiker, dass der Wert des Geldes nicht vergleichbar sei mit dem objektiven Wert der Werte. Sie haben Recht, denn wenn aus den allseits vertrauten Umgangsformen mit Geld kein objektiver Wert des Geldes definierbar ist, dann wird es für die vagen Ideen ethisch-moralischer Werte noch weniger gelingen.

Nachdem die Moralphilosophie Immanuel Kants ein letztes Mal mit einem großangelegten Entwurf den absoluten Wert moralischer Geltung durch die Formulierung des kategorischen Imperativs - Handle stets so, dass die Maxime deines Handelns jederzeit und überall zur Grundlage der allgemeinen Gesetzgebung herangezogen werden kann - einklagen wollte, wächst unaufhaltsam der Zweifel am Wert der Werte. Allen voran diagnostizierte Friedrich Nietzsche am Ende des 19. Jahrhundert die „Umwertung aller Werte“. Er dachte an ein dynamisches, auf Steigerung durch Kampf und Überwindung ausgerichtetes Prinzip und nannte es den „Willen zur Macht“. „Diese Welt ist der Wille zur Macht – und nichts außerdem! Und auch ihr selber seid dieser Wille zur Macht – und nichts außerdem.“[2] Er sah die Welt sich gestalten durch Wille und Gegenwille bei Menschen und in der gesamten Natur. Damit ist jeder Gedanke an einen übernatürlichen Wert ausgeschaltet, auch wenn Nationalsozialisten die Idee eines Willen zur Macht zur übernatürlichen Ideologie eines Rassenwahns umstilisierten.
Es ist vor diesem Hintergrund schwer vorstellbar, dass ein KI-System einen „Willen zur Macht“ entfalte, noch auch, dass er ein Wertebewusstsein ausbilde, das seinem funktionierenden Handeln eine Richtung geben könnte und verschiedene individuelle Willen emotional synchronisiere. Zudem müssen wir aus der Erfahrung mit KI-Systemen erkennen, dass Daten stets nur auf Datenförmiges wirken und auch nur Daten generieren. Das erinnert uns daran, dass auch Werte nur auf Wertendes wirken und von diesem ausgehen. Nur Menschen und andere Tiere bilden eine Wertschätzung aus. Und nur Menschen bilden daraus Worte als leitende Begriffe. Werte gehen also von Menschen aus und wirken auf Menschen zurück. Maschinen und KI-Systeme erweisen sich dabei als zwischengeschaltetes Instrument, mit deren Hilfe Menschen ihre Werte „besser“ verwirklicht werden sollen. Da wir aber als Menschen bislang nicht herausgefunden haben, wie wir die Mitglieder der Mafia und anderer ähnlicher Organisationen davon überzeugen können, dass ihre Werte keine Werte seien, müssen wir damit rechnen, dass niemand auf Dauer werde verhindern können, dass KI-Systeme wie andere Maschinen in der Vergangenheit auch zur Durchsetzung des „Willens zur Macht“ in der Zukunft ausgenutzt werden. Vielleicht ist es ratsam, bei jeder Entwicklung neuer Systeme auch die Bauanleitung zu liefern, wie das System im Bedarfsfall gestört werden kann. Aber der Glaube an das Gute rechnet zu wenig mit der Notwendigkeit, unerwünschten Gebrauch erfolgreich stören zu können.
Die Datenförmigkeit der Algorithmen eines KI-Systems offenbart eine weitere Kränkung des abendländischen philosophischen Bewusstseins. Haben Philosophen und Theologen mit ihren metaphysischen Werten – dem Wahren, Guten und Schönen – den Rang objektiver und universaler Geltung versprochen, so hält die Datenförmigkeit der Welt, was Philosophen stets nur versprochen haben. Alles, womit Menschen zu tun haben, was

[2] Friedrich Nietzsche, *Aus dem Nachlaß der Achtzigerjahre*, in: Werke in drei Bänden, München 1954, Band 3, S. 918.

ihnen in den Sinn kommt und wovon sie sich einen Nutzen erwarten, lässt sich in Datenform bringen. Weder Religion noch Ethik konnten dieses Versprechen einlösen. Daten erweisen sich mehr und mehr als die allgemeine Formel, auf die sich alle Dinge bringen lassen – und zwar unabhängig davon, ob uns das passt oder nicht. Die Erfindung des binären Zahlencodes und dessen Bedeutung für die universale Codierung der Welt geht auf Gottfried Wilhelm Leibniz (1646-1716) zurück. Er identifizierte die 1 mit Gott, dem höchsten Welt, und die 0 mit dem Nichts, dem niedrigsten Wert. Dazwischen ereigne sich Alles: „unitas in multitudine – Einheit in der Vielheit“.