Der Einsatz digitaler Systeme, insbesondere von Systemen Künstlicher Intelligenz (KI), in moralisch sensiblen Bereichen unseres Zusammenlebens muss stets ethisch-philosophisch beleuchtet werden. Dafür gibt es eine Vielzahl von Gründen, die tiefer liegen als ein allgemeiner Hinweis auf potenzielle oder gar Science-Fiction-Szenarien, die KI-Systeme nach dem *Terminator*-Modell als Gefahr für die Menschheit betrachten.[1] Im Folgenden werde ich zunächst erläutern, was unter KI beziehungsweise KI-Systemen zu verstehen ist. Im Anschluss werde ich dafür argumentieren, dass sich die Ethik der KI mit der Frage beschäftigt, wie die Mensch-Maschine-Interaktion ausgestaltet werden sollte. Damit verschiebt sich die Fragestellung auf den humanen Kontext der KI-Anwendung. Sie führt zuletzt auf den Menschen als unbedingte Wertquelle zurück, dem nicht nur als Objekt möglicher technisch-militärischer Operationen, sondern vor allem als Subjekt seiner Werturteile Würde zukommt.

Wert und Mensch in der Ethik der KI

Prof. Dr. Markus Gabriel

Was ist KI?

KI-Systeme sind Produkte der KI-Forschung, bei der es sich um einen Forschungsbereich der Informatik handelt, der sich seit einigen Jahrzehnten rasant weiterentwickelt. In diesem Forschungsgebiet werden Methoden des maschinellen Lernens erforscht, deren Ziel es ist, menschliche Wahrnehmungs-, Denk- und Handlungsmuster zu modellieren und die Modelle vom Menschen unabhängig, das heißt autonom entscheiden, agieren und beurteilen zu lassen. Dabei wird es meistens vermieden, anzugeben, worin Intelligenz eigentlich besteht und was eine Intelligenz als künstlich (im Unterschied zu natürlich beziehungsweise animalisch) auszeichnet.

Allerdings ist das ein potenziell fatales Problem. Wenn ein Forschungsgebiet keine explizite oder implizite hinreichend klare Vorstellung davon hat, worin sein zu modellierendes Zielsystem besteht, verfügt es über keine hinreichend klar definierten Erfolgskriterien. Das zeigt sich in der KI-Forschung in problematischen Auswüchsen eines KI-Hypes, der sich über Ideologie und Mythologie verbreitet, was inzwischen gut untersucht ist.[2]

Die erste Aufgabe jeder Philosophie der KI ist daher eine Klärung der relevanten Begriffe. Erst davon lässt sich der Fokus einer Ethik der KI ableiten. Im Allgemeinen kann man unter „Intelligenz“ als Zielsystem der KI-Forschung das Vermögen (paradigmatisch eines Lebewesens) verstehen, ein gegebenes Problem in einem endlichen Zeitraum zu lösen. Der Zeitraum unserer Problemlösungen ist evolutionär, das heißt durch Parameter der Reproduktion individueller Organismen, von Populationen und ganzen Spezies gegeben. Für einzelne Aufgaben, die sich als Probleme verstehen lassen, ist der Zeitraum

enger zu fassen. Doch alles, was wir als Lebewesen tun, geschieht auch im Horizont unseres Überlebens. Ein System S_1 ist intelligenter als ein anderes System S_2, wenn es dasselbe Problem schneller löst. Intelligenz in diesem Sinne ist also ein Maß der Problemlösungseffizienz. KI-Systeme sind besser im Schach, in Go, oder im Durchsuchen großer Datensätze im Hinblick auf bestimmte Muster als Menschen, weil sie schneller sind als wir.

Eine Intelligenz ist dabei „künstlich", wenn sie nicht durch biologische Reproduktion entstanden ist. Wir Menschen sind ebenso wie die meisten anderen (nicht von uns produzierten) Lebewesen nicht künstlich. Das bedeutet, dass unser Organismus aus Zellen besteht, die nach erfolgter Befruchtung einer Eizelle durch Zellteilung entstanden sind. Intelligenz entsteht im Vorgang der Ontogenese, das heißt der Entwicklung eines individuellen Menschen im Mutterleib. Wann genau, wissen wir heute nicht, auch weil unklar ist, wie genau Bewusstsein und Intelligenz verbunden sind und woran sich Bewusstsein beziehungsweise Intelligenz neurobiologisch festmachen lassen.

Künstliche Intelligenzen sind jedenfalls Systeme, die im humanen Kontext unserer Verwendung von KI-Systemen industrielle Produkte unserer nicht biologischen Herstellung sind. Diese Systeme erscheinen im Anwendungskontext als intelligent, weil wir sie einsetzen, um unsere Probleme zu lösen.

An sich sind die Systeme wohlgemerkt nicht intelligent. Anders als Lebewesen sind die internen Vorgänge eines KI-Systems nämlich nicht an der Reproduktion der Grundlagen des Systems interessiert. KI-Systeme bauen (jedenfalls bisher) keine Hardware und designen keine Siliziumchips, Halbleiter und Elektrizitätswerke, um ihren Fortbestand zu sichern. Vielmehr sind KI-Systeme in ihrer Fortexistenz vollständig vom humanen Kontext abhängig, in dem sie gebaut und verwendet werden. Abstrahiert man von dieser Tatsache und schaut sich ausschließlich die Funktionsweise eines einzelnen KI-Systems an, scheint es, als ob man es mit einer höheren Intelligenz zu tun hat. Doch dies ist eine Illusion. Denn KI-Systeme können zwar in der Anwendung Problemlösungen generieren, die alles übersteigen, was Menschen bisher leisten. Doch die Problemlösungen, die sie liefern, finden ihren Sinn und Unsinn nur innerhalb des humanen Kontextes. Streicht man diesen aus der Gesamtbetrachtung, handelt es sich bei

Im Zeitalter einer nur vermeintlich zu Ende gegangenen Geschichte kommt es darauf an, diesen Gedanken eines moralischen Auftrags der Menschheit wieder in Erinnerung zu rufen, um die irrige Vorstellung eines nackten Überlebenskampfes abzuwehren.

KI-Systemen letztlich nur um bestimmte Computer, in denen Vorgänge ablaufen. Anders ist es bei Menschen. Menschen (wie auch andere Lebewesen) sind intrinsisch intelligent und nicht nur dadurch, dass wir innerhalb des humanen Kontextes zum Einsatz kommen.

Daraus folgt nicht, dass KI-Systeme nicht intelligent sind, sondern nur, dass ihre Intelligenz nicht autonom, sondern eine Funktion ihrer Einbettung in den humanen Kontext ist.

Ethik der KI

KI ist eine Soziotechnologie. Sie ereignet sich an der Mensch-Maschine-Schnittstelle und ist damit stets Ausdruck einer Mensch-Maschine-Interaktion. KI ist nicht autonom, sondern relational. KI ist auch nicht an sich sozial. Denn etwas ist nur dann sozial, wenn es dasjenige, was es tut, im Licht einer Vorstellung davon tut, was andere tun und was man selbst tun soll. Soziale Systeme sind normativ, was voraussetzt, dass sie sich selbst Regeln geben, mit denen sie auf Regeln reagieren, denen andere folgen. Es reicht nicht aus, dass ihnen Regeln vorgeschrieben werden.

Sozialität lässt sich dabei ebenso wie Intelligenz teilweise modellieren. Sie ist in fortgeschrittene KI-Systeme bereits über Datensätze eingebaut, die Vorstellungen enthalten, die sich Menschen voneinander machen. Es bleibt aber dabei, dass ein Modell einer Gesellschaft keine Gesellschaft ist. Allerdings sind Gesellschaftsmodelle und damit auch soziale Netzwerke gesellschaftlich wirksam. Und das ist der eigentliche Grund, warum wir einer Ethik der KI bedürfen.

Thema der KI-Ethik ist der Einsatz von KI-Systemen innerhalb des humanen Kontextes. Welche Muster eine KI identifizieren soll und wie wir mit den Ergebnissen umgehen, die sie uns liefert, ist das relevante Problemfeld.

Im Kontext digitaler Waffensysteme gilt, dass KI-Systeme, die zur Aufklärung dienen, die also im englischsprachigen Sinne Teil der *Intelligence* sind, wünschenswert sind. Ethisch sinnvoll, das heißt gut, ist ihr Einsatz nur unter den Bedingungen, unter denen

Kriegsführung beziehungsweise der allgemeine Einsatz militärischer Mittel gut ist. Damit unterscheidet sich die allgemeine Ethik der KI im militärischen Feld nicht wesentlich von der allgemeinen Kriegsethik. Denn KI-Systeme sind zwar Waffen oder Instrumente einer besonderen Art, keineswegs aber ein Quantensprung der Menschheitsentwicklung. Gerade im Zeitalter einer in voller Wucht zurückgekehrten Geopolitik (die freilich nie verschwunden war, sondern nur im Wunschdenken weggedacht wurde) gilt, dass die militärische Struktur einer Gesellschaft ihrer Selbsterhaltung dient. Da Verteidigung ethisch unproblematisch ist, ist es auch Verteidigung durch Einsatz intelligenter Systeme, sofern dieser Einsatz in den humanen Kontext eingebettet bleibt.

Sollten eines Tages von KI angetriebene Waffensysteme weitgehend ohne Einsatz von Soldatinnen und Soldaten eine Schlacht oder einen Krieg entscheiden können, der ethisch vertretbar ist, wäre dies wünschenswert. Allerdings setzt dies voraus, dass die Entwicklung solcher Systeme durch staatliche und private Akteure, die die allgemeinen Richtlinien ethisch vertretbarer Kriegsführung akzeptieren, vertretbar ist. Das gilt nur eingeschränkt, weil die Entwicklung einer Hochrisikotechnologie (man denke nur an die Atomkraft in all ihren Dimensionen) nur ethisch vertretbar und wünschenswert ist, wenn ihr Einsatz ethisch eingehegt ist. Hocheffiziente KI-Systeme dürfen deswegen nicht in die Hände von Parteien geraten, die sich über jede Kriegsethik hinwegsetzen.

Damit treffen wir auf dem Feld der KI-Ethik auf ein bekanntes Paradoxon. Nur wenn der Werterahmen einer Gesellschaft und ihrer Teilsysteme in die richtige Richtung zeigt, sind Entwicklung und Einsatz von Hochrisikotechnologien vertretbar und wünschenswert. Die gesellschaftlichen Wertvorstellungen und die allgemeinen humanen Werte, um die es in der Ethik geht, werden wirkungslos, wenn das Böse überhandnimmt, das heißt der unbedingte, grausame Wille der Zerstörung eines faktischen Feindes oder einer dehumanisierten Menschengruppe.

Dagegen hilft es nicht, die irrige Vorstellung zu mobilisieren, wir könnten KI-Systeme programmieren, die von sich aus moralische Ideen entwickeln, die unseren vielleicht sogar überlegen sind. Denn der Problemlösungsraum ethischen Denkens hängt damit

zusammen, dass wir verletzliche endliche Lebewesen sind, die sich in andere einfühlen können. Die entsprechenden Vermögen des empathischen Miterlebens mit anderen sind nach heutigem Kenntnisstand nicht modellierbar, weil sie nicht in den Bereich derjenigen Probleme gehören, die sich schnell lösen lassen.

Individueller und kollektiver moralischer Fortschritt, das heißt ein Zugewinn an ethischer Erkenntnis, setzt biografische und historische Entwicklungsprozesse voraus, die weit über die Vorstellung hinausgehen, die Ethik befasse sich mit individuellen, ahistorischen Entscheidungssituationen. Selbst in einem Gefecht in einem heißen, konventionellen Krieg geht es keineswegs nur darum, die eigene Truppe zusammenzuhalten und dem Gegner einen möglichst großen Schaden zuzufügen. Die umfangreiche historische, biografische und künstlerische Literatur insbesondere zu den beiden Weltkriegen belegt dies eindrücklich. Dies betrifft das Problemfeld einer Truppenmoral, das von den altgriechischen Historikern, die den Angriffskrieg der Perser auf die numerisch unterlegenen Griechen mit Staunen beschrieben haben, bis in die gegenwärtige Situation einer moralisch im Recht befindlichen Verteidigungsarmee der Ukraine reicht. Kurzum: Das David-gegen-Goliath-Motiv weist darauf hin, dass Kampfhandlungen zwischen Menschen auch in ihrer Effizienz daran gebunden bleiben, dass historisch verankerte Vorstellungen der Gerechtigkeit gegeneinander antreten.

Im Zeitalter einer nur vermeintlich zu Ende gegangenen Geschichte kommt es darauf an, diesen Gedanken eines moralischen Auftrags der Menschheit wieder in Erinnerung zu rufen, um die irrige Vorstellung eines nackten Überlebenskampfes abzuwehren, in dem sich das Recht des Stärkeren durchsetzt.[3]

Wert und Mensch

Die allgemeine Ethik ist eine Teildisziplin der Philosophie. Sie untersucht das Bestehen moralischer Tatsachen. Hierbei ist eine „moralische Tatsache“ lediglich insofern eine wahre Antwort auf eine sinnvoll gestellte Frage bezüglich dessen, was jemand tun beziehungsweise unterlassen soll, als sie oder er ein Mensch ist. Ein

Wir müssen unter allen Umständen vermeiden, unsere Ethik im Allgemeinen an nichtmenschliche Akteure zu delegieren.

einfaches Beispiel soll das illustrieren. Wer vor die Wahl gestellt ist, ein Kleinkind vor dem Ertrinken in einem flachen Gewässer zu retten, ohne dabei selbst irgendeine Gefahr für sein eigenes Leben einzugehen, ist moralisch verpflichtet, das Kind zu retten. Die Frage: „Soll ich das Kind retten?", ist deswegen eindeutig mit „Ja" zu beantworten und zwar ganz und gar unabhängig davon, wer man selbst und wer das Kind ist. Jede und jeder soll in dieser Situation das Kind retten.

Menschen sind die paradigmatischen Subjekte und Objekte ethischer Überlegungen. Es geht uns in der Ethik primär um uns als Menschen und darum, was wir einander schulden.[4] Die Tier-, Technik- und Umweltethik als besondere Anwendungsgebiete der allgemeinen Ethik folgen ohne Umschweife aus der anthropologischen Begründung der allgemeinen Ethik. Menschen selbst sind Tiere, also gilt eine Tierethik, die auch auf andere Lebewesen auszuweiten ist, sofern sie mit uns zusammenleben und sofern es in unserer Macht liegt, die Biodiversität zu achten, ohne dadurch Menschen schweren Schaden zuzufügen.

Die Umweltethik folgt daraus, dass Menschen als Tiere in eine Umwelt eingebettet sind, die sie gestalten. Umwelt ist nichts Gegebenes, sie ist keineswegs reine Natur, die es auch ohne uns gäbe. Sie ist vielmehr stets unsere Umwelt, also etwas, was an Menschen gekoppelt ist, die keineswegs dieselbe Umwelt wie andere Lebewesen haben. Es gibt nicht die eine Umwelt, die sich alle Tiere teilen. Die Umwelt ist keine bloße Ansammlung von Ressourcen zur Reproduktion von Lebewesen. Dies wäre eine umweltethisch nicht vertretbare These.

Die Technikethik, zu der die Ethik der KI zählt, widmet sich der Mensch-Maschine-Interaktion und damit der Frage, unter welchen Bedingungen wir neue Techniken entwickeln und einsetzen dürfen, die Menschen massiven Schaden zufügen können.

Im Umfeld der KI-Ethik hat sich die irrige und ethisch gefährliche Vorstellung verbreitet, wir könnten KI-Systeme entwickeln, die moralische Urteile fällen, die womöglich sogar besser als unsere eigenen menschlichen Urteile sind. Die damit verbundenen Argumentationen übersehen allerdings, dass KI-Systeme zwar in der Tat stringenter und kohärenter urteilen können als Menschen.[5] Aber

sie blockieren genau dadurch auch moralischen Fortschritt im Sinne einer fundamental neuen Einsicht in das moralisch Richtige und Falsche, die damit zusammenhängt, dass wir unsere gesellschaftlich wirksamen Werturteile und Vorurteile verändern. Diese Veränderung unterliegt keinem Algorithmus und keiner allgemeinen Handlungsanweisung – sie wird damit auch weder von einem kategorischen Imperativ noch von utilitaristischen Kalkülen oder einem anderen Ethiktypus (wie der Tugendethik) gesteuert.

Insbesondere übersieht die Vorstellung, wir könnten moralische Maschinen bauen, die uns womöglich ethisch sogar überlegen sind, dass unsere Wertfindung innerhalb des humanen Kontextes immer von Wertvorstellungen abhängt, die hochgradig diversifiziert sind und die auch davon abhängen, welche soziale Stellung ein menschliches Individuum hat. Das ist einer der guten Gründe des derzeitigen Diversitätsdiskurses, sofern er darauf abzielt, die Vielfalt menschlichen Erlebens als ethisches Thema ernst zu nehmen.

Für unseren Kontext bedeutet das konkret, dass das Thema des maschinellen Bias und der Diskriminierungsstrukturen, die bereits in Datensätzen hinterlegt sind und die einer KI als Grundlage präsentiert werden, ins Zentrum einer KI-Ethik gehört. Wenn etwa KI-Systeme im Schlachtfeld zum Einsatz kommen, die gezielt Menschen einer diskriminierten Gruppe angreifen und andere verschonen, so verstärkt das gegebene moralische Schieflagen.

Natürlich kann man an dieser Stelle einwenden, dass die KI-Systeme damit zumindest nicht zwingend schlechter, sondern allenfalls so schlecht wie Menschen urteilen, deren Urteilspraxis von moralisch verwerflichen Voreinstellungen geprägt sind. Doch die Verstärkung stereotyper Vorurteile gegen historisch vernachlässigte Gruppen zu vermeiden und umgekehrt eine tolerante, offene demokratische Gesellschaft zu erzeugen, haben wir eine Vielzahl ethischer Praktiken der Nachsicht, des Verzeihens, der Debatte, des Minderheitenschutzes, des Widerstands, der Opposition, des Mitleidens und so weiter entwickelt, die ihrem Wesen nach nicht algorithmisch und damit auch nicht programmierbar sind.

Die Ethik bleibt, als Erkenntnisform und damit als Teildisziplin der wissenschaftlichen (nicht weltanschaulichen) Philosophie, ebenso wie

unsere alltäglichen und professionellen Wertvorstellungen auf unsere individuellen und kollektiven, menschlichen und damit verkörperten Standpunkte angewiesen. Daraus folgt keinerlei Kulturrelativismus. Denn das Ziel der Ethik ist die Erkenntnis und die Anerkennung universal gültiger Werte, die in der menschlichen Lebensform begründet sind. Doch die Erkenntnis des Universalen setzt voraus, dass wir die individuellen Kontexte der Entscheidungsfindung und der konkreten Lebensumstände von Menschen berücksichtigen, was kulturell fundierte Wissensformen voraussetzt, über die KI-Systeme nach heutigem Stand in keiner Weise verfügen. Dazu müssten wir ihnen geistes-, kultur- und sozialwissenschaftliche Erkenntnisse sowie eine subjektive Erlebensperspektive einprogrammieren, die sich allesamt niemals in digitale Datensätze übersetzen lassen. Denn digitale Datensätze sind bestenfalls Modelle eben jener menschlichen Urteils- und Lebenspraktiken, sodass wir uns im Kreis drehen, wenn wir versuchen, digitalen Systemen unsere Ethik beizubringen, indem wir sie mit Daten füttern, die höchstens Wertvorstellungen abbilden, über deren Richtigkeit gesellschaftlich zu diskutieren bleibt.

Für die Kriegsethik und damit auch für die Bundeswehr ergibt sich daraus, dass die Ausbildung in allgemeiner Ethik sowie in Kriegsethik ihrerseits eine Bildung der Mitglieder voraussetzt. Diese Bildung zielt darauf ab, eine Haltung zu erzeugen, deren Ziel es ist und bleibt, das moralisch Richtige in komplexen Handlungssituationen zu identifizieren. Gleichzeitig darf zu keinem Zeitpunkt am Rahmen der Menschenrechte und damit der Menschenwürde gerüttelt werden. Und all dies impliziert, dass wir es unter allen Umständen vermeiden müssen, unsere Ethik im Allgemeinen an nicht menschliche Akteure zu delegieren. Vielmehr gilt es, die KI-Systeme zu optimieren, um im digitalen Wettrüsten nicht unterlegen zu sein, ohne dabei zum Opfer der auch militärisch nicht sinnvollen Illusion zu werden, wir könnten die Soldatinnen und Soldaten der Zukunft durch ungleich klügere, effizientere und gar moralisch gerüstete Roboter ersetzen.

1 In diesem Sinne irren die Analysen von Nick Bostrom auf eine paradigmatische Weise, weil er futuristische KI-Systeme beschreibt, ohne die Wirklichkeit von KI in Rechnung zu stellen, die uns bereits vor ethisch-philosophische Herausforderungen stellt. Vgl. Nick Bostrom (2016): Superintelligenz. Szenarien einer kommenden Revolution. Berlin; dagegen: Markus Gabriel (2018): Der Sinn des Denkens. Berlin.

2 Vgl. etwa Stephen Cave/Kanta Dihal/Sarah Dillon (Hrsg.) (2020): AI Narratives. A History of Imaginative Thinking. Oxford; und Stephen Cave/Kanta Dihal (2022): Imagining AI. How the World Sees Intelligent Machines. Oxford. Vgl. auch Markus Gabriel (2020): Fiktionen. Berlin, Paragrafen 15–17.

3 Vgl. dazu in Auseinandersetzung mit Francis Fukuyamas berühmter, aber oftmals im Detail nicht genau zur Kenntnis genommener These vom „Ende der Geschichte“ bei Markus Gabriel (2020): Moralischer Fortschritt in dunklen Zeiten. Universale Werte für das 21. Jahrhundert. Berlin; sowie ders. (2022): Der Mensch als Tier. Warum wir trotzdem nicht in die Natur passen. Berlin.

4 Wie die berühmte Formulierung Thomas M. Scanlons (2000) lautet in: What We Owe to Each Other. Cambridge, MA.

5 Vgl. dazu insbesondere im Hinblick auf rechtliche und ökonomische Werturteile: Daniel Kahneman/Olivier Sibony/Cass R. Sunstein (2021): Noise. Was unsere Entscheidungen verzerrt – und wie wir sie verbessern können. München.