Prof. Dr. Jörg Kopecz:

KÜNSTLICHE INTELLIGENZEN WERDEN AUTONOM UND ZUNEHMEND UNUNTERSCHEIDBAR VON MENSCHEN -UND DAMIT POTENZIELL FÜR IHR HANDELN VERANTWORTLICH[1]

Dies ist ein stark gekürzter Beitrag eines längeren Artikels, der unter dem Titel: „Moralische Maschinen. Zur ethischen Ununterscheidbarkeit von Mensch und Maschine“ demnächst bei „Springer Nature“ erscheinen wird, und den ich gemeinsam mit meinem Kollegen U. Dettmann verfasst habe.

ABSTRACT

Je mehr sich Künstliche Intelligenzen an menschliche Fähigkeiten annähern, desto stärker treten ethische Fragen in den Vordergrund. Diese sind vor allem mit Begriffen wie Autonomie, Verantwortung für Handeln, aber auch der Frage, ob KIs „Personen sein“ können und damit verbunden auch ob sie haftbar bzw. moralisch verantwortlich für ihr Handeln sein können. In diesem Artikel nähern wir uns von zwei Seiten diesem Thema: zum einen betrachten wir die wachsenden Fähigkeiten von KIs bzw. ihr beobachtbares Verhalten und zu anderen betrachten wir die Kombination Mensch-Maschine am Beispiel Body Hacking bzw. Enhancement, sowie die möglichen Konsequenzen für Menschen. Aus der philosophische Betrachtung kommend, ergänzen wir dies durch systemtheoretische Betrachtungen aus dem Kontext der nichtlinearen Systeme und kommen zu dem Schluss, dass es keine prinzipielle Gründe dafür gibt, dass hinreichend komplexe künstliche Intelligenz nicht Träger moralischer Rechte und Pflichten sein könnten und damit auch moralische Verantwortung tragen müssen.

SCHLÜSSELWÖRTER:

Neuronale Netze, Künstliche Intelligenz, Turing Test, Personsein, Body Hacking, Nichtlineare Systemtheorie, Neuromorphic Computing, Robotik, Digitale Ethik

Künstliche Intelligenz und der Begriff der Person

KI boomt und verfügt über immer mehr Eigenschaften und Fähigkeiten, die bisher allein dem Menschen vorbehalten waren. 2017 hat das Europäische Parlament deshalb veranlasst, eine Entschließung mit Empfehlungen an die Kommission zu zivilrechtlichen Regelungen im Bereich Robotik zu verabschieden. In dieser Entschließung hat das Europäische Parlament erste Vorschläge zu rechtlichen Regelungen über Roboter und die künstliche Intelligenz gemacht. Besonders der Vorschlag des Parlaments, langfristig die Einführung eines eigenen Rechtsstatus für Roboter als elektronische Personen zu erwägen, hat für enormen Diskussionsbedarf gesorgt.

Wir werden uns im Folgenden vor allem mit der Frage beschäftigen, inwiefern der Begriff der Person zentral für die Frage ist, ob Systeme der KI in naher oder ferner Zukunft als elektronische Personen moralische Rechte, Pflichten und moralische Verantwortung besitzen können und welche Auswirkungen dies auf die zur Zeit noch geltende moralische Unterscheidbarkeit von Mensch und Maschine haben kann. Das Personsein spielt in unserer Kultur eine bedeutende Rolle. Der Mensch begreift sich seit jeher als von anderen Lebensformen verschieden, woraus sich spezifische normative Grundwerte wie etwa die Menschenwürde und der Personenstatus ableiten.

Verbunden mit dem Personsein sind zentrale Eigenschaften, durch die sich der Mensch von anderen uns bekannten Lebensformen abgrenzt. Diese Eigenschaften, die mit dem Personsein verbunden sind, bilden zugleich die Grundlage dafür, dass Menschen sich und anderen als Personen einen besonderen ethischen und rechtlichen Status zuschreiben. Diese Eigenschaften und Fähigkeiten werden immer auch als Begründung dafür angeführt, weshalb der Mensch als autonome Person Respekt verdient und es den Begriff der unverletzlichen Menschenwürde gibt. Für Systeme der KI ist deshalb zunächst zu fragen, ob diese die für das Personsein charakteristischen Eigenschaften und Fähigkeiten in hinreichendem Maße besitzen (können) und ab wann sie als Personen zu gelten haben.

Aufgrund welcher Eigenschaften und Fähigkeiten gehört eine Entität zur Klasse der Personen (und hat damit moralische Rechte und Pflichten)? Diesbezüglich gibt es eine Reihe von Vorschlägen, welche Bedingungen notwendig und hinreichend dafür

sein sollen, Entitäten als Personen zu klassifizieren und damit zu ethisch relevanten Trägern von Rechten und Pflichten zu machen.

Die Analyse dieser Kriterien ist also wesentlich für die Frage, ob und unter welchen Bedingungen künstliche Systeme Träger ebendieser Rechte und Pflichten sein können.

Als Kandidaten für die genannten Bedingungen werden unter anderem genannt:

- Praktische Urteilskraft,
- Intelligenz
- Rationalität
- Lernfähigkeit
- Bewusstsein, Selbstbewusstsein, Selbsterkenntnis
- Die Seele und das Ich
- Freier Wille, Autonomie
- Wissen um die eigene zeitlich ausgedehnte Existenz
- Verständnis für die evaluativen und normativen Aspekte der Wirklichkeit
- Einsichtsfähigkeit
- Emotionen
- Menschenwürde als Folge von Personalität

Auffällig ist, dass all die Besonderheiten, die Personen auszeichnen sollen, keineswegs klar, sondern ihrerseits erklärungsbedürftig sind. Was meint „freier Wille“, „Selbstbewusstsein“ oder „praktische Urteilskraft?“ Sind wir überhaupt vernünftige und freie Wesen in dem noch recht vagen Sinne, der uns vorschwebt? Auf welche Weise unterscheidet sich unser Gedächtnis und unser Erleben von Zeit von dem anderer Lebewesen? Worin besteht das Bewusstsein, und sind wir tatsächlich die einzigen, die so etwas haben? Dabei sollte man aber nicht vergessen, dass auch Menschen nicht immer alle der hier genannten Kriterien erfüllen und wir dennoch nicht zögern, sie als Personen zu behandeln. Insofern liegt etwas Eigentümliches darin, dass wir einerseits eine beträchtliche Anzahl von Bedingungen formulieren können, die ein Wesen erfüllen muss, um als Person gelten zu können, andererseits aber von uns und anderen gar nicht verlangen, dass wir über alle diese Eigenschaften verfügen. Jemand wird von uns als eine Person wahrgenommen, auch ohne

Anspielungen, Metaphern oder Witze zu verstehen, ohne Einsicht oder praktische Urteilskraft und ohne als besonders vernünftig zu gelten. Auch Menschen, die aufgrund einer Erkrankung oder Schädigung des Gehirns keine Emotionen empfinden können, bleiben selbstverständlich Personen mit Rechten und Pflichten.

Bedeutet das, so die naheliegende Frage, dass wir bereits hinreichend komplexen künstlichen Systemen Personenstatus zusprechen müssen, weil KI-Experten es für wahrscheinlich halten, dass diese Systeme in absehbarer Zukunft über die genannten Fähigkeiten und Eigenschaften verfügen werden oder verfügen könnten? Die oben genannten Bedingungen für die Zuschreibung eines personalen Status sind nun in einer Weise miteinander verknüpft, dass Ergebnisse in einem Thema, z.B. in der Physikalismusdebatte, relevant sind für Argumente in anderen Themen.

Bieten die Erfolge bei der Entwicklung intelligenter Maschinen oder Programme vielleicht einen Hinweis darauf, dass der Physikalismus wahr ist? Wenn dem so wäre, warum sollten dann künstliche Systeme nicht prinzipiell auch Personen sein können und damit Träger normativer Prädikate? Wenn schon eine programmierte Maschine geistbegabt ist, wie kann man dann noch behaupten, das Mentale sei eine eigene Substanz? Natürlich muss dazu erst einmal geklärt sein, wie es um die Entwicklung intelligenter Maschinen bestellt ist und ob der dort verwendete Intelligenzbegriff überhaupt einer ist, der demjenigen entspricht, den wir dem menschlichen Geist zusprechen. Von den Antworten auf diese Fragen ist es dann abhängig, wie wir die Frage, ob künstliche Systeme Träger moralischer Rechte und Pflichten sein können, beantworten.

KI, Enhancement und personale Identität

Dem Begriff der personalen Identität kommt sowohl in der praktischen Philosophie wie auch in der Metaphysik eine zentrale Bedeutung zu. In der Metaphysik wird personale Identität behandelt unter der Fragestellung „Welche Bedingungen müssen erfüllt sein, um von einer Entität X zu einem Zeitpunkt t und von einer Entität Y zu einem von t unterschiedenen Zeitpunkt t‘, sagen zu können, es handele sich um ein und dieselbe Person?“ Die Antwort auf diese Frage hat unmittelbare Auswirkungen auf die Ausgangsfrage, ob künstliche Systeme Träger moralischer Rechte und Pflichten sein können. Die Überschrift, unter der diese Thematik verhandelt wird lautet

„Enhancement“: es gibt für KI zwei Entwicklungsrichtungen: zum einen können wir uns mit künstlichen Systeme verbinden und damit „künstlicher“ werden und zum anderen können künstliche Systeme (mit uns verbunden) immer natürlicher werden. Im Zentrum der Debatte um Enhancement und Neuroenhancement steht die Frage, ob diese mit den „Ursprungssystemen“ in einer irgendwie gearteten Identitätsrelation stehen. Sollte dies so sein, dann ist schwer begründbar warum künstlichen Systemen nicht – unter bestimmten eher technischen Bedingungen – dieselben normativen Prädikate zugeschrieben werden können als „menschlichen Systemen“. Bereits heute gibt es viele Möglichkeiten den Menschen durch Techniken des Enhancement und Neuroenhancement zu optimieren, zu verändern und Teile des Körpers und des Gehirns durch künstliche Implantate zu ersetzen. Frühe konkrete Konzepte hierzu findet man z.B. bei sog. Gehirn-Computer-Schnittstellen (Brain-Computer-Interfaces, BCIs).

Ein weiterer, bisher vernachlässigter Aspekt des Enhancements, betrifft die Plastizität des Gehirns. Wie verändern sich menschliche Eigenschaften, wenn die Plastizität und Anpassungsfähigkeit des Gehirns durch fremde, angekoppelte Systeme ein biologisch/evolutiv entstandenes Wesen wie den Menschen (oder Tiere) ergänzt bzw. verändert? Die Plastizität des Gehirns bewirkt, dass Implantate auf neuronaler Basis strukturelle Veränderungen in den Verarbeitungsstrukturen des Gehirns erzeugen. Da Teile des Cortex nach dem Prinzip funktionaler Karten arbeiten, die Sensor- oder Aktorflächen funktional ergänzt repräsentieren, verändern sich diese Strukturen, wenn wie auch immer geartete Eingangssignale dort ankommen. So können Menschen Farben hören oder fühlen, wenn die richtigen Hirnareale angesprochen werden. Darüber hinaus verändern diese Manipulationen auch unser gesamtes Denken: Wir benutzen Begriffe wie „begreifen“, „unerhört“ oder „einsehen“ für kognitive Begriffe, die jedoch offensichtlich aus unseren aktorischen oder sensorischen Fähigkeiten abgeleitet sind. Wenn wir geänderte sensorische Fähigkeiten haben, werden sich möglicherweise auch unsere Begriffe und damit unsere Art zu denken verändern. Wenn sich die Verknüpfung nicht nur auf die Sensorik oder Aktorik bezieht, sondern auch auf intellektuelle Fähigkeiten, wie das bei der KI der Fall ist, lässt sich dies weiter extrapolieren und man kann vermuten, dass dies auch Auswirkungen auf das ethische Selbstverständnis des Menschen und auf das Verständnis von Autonomie

hat. Umgekehrt gilt dies auch für KI Systeme: es erhält weitere Fähigkeiten, die es zur Problemlösung nutzen kann und die seine Autonomie befördern.

Was folgt nun aber aus den neuen Möglichkeiten des Enhancements und Neuroenhancements für unsere Fragestellung?
Was würde passieren, wenn wir eines Tages den Menschen mit technischen Mitteln komplett verändern und verbessern, ihn also in ein künstliches System umwandeln könnten? Wenn wir Menschen, unter bestimmten Einschränkungen, moralische Verantwortlichkeit und damit moralische Rechte und Pflichten zuschreiben, ab welchem Zeitpunkt würden wir sagen, dass dies nun nicht mehr möglich sei, da das „menschliche System" „zu künstlich" sei?. Dieselbe Überlegung betrifft auch eine weitere essentielle Eigenschaft des Menschen, nämlich das Bewusstsein. Können wir diese Stelle nicht benennen, müssen wir erklären warum ein künstliches System mit Bewusstsein nicht zumindest denkbar ist. Mit der Verbindung von Mensch und Maschine verschwimmen die Grenzen beider Systeme und wir können keine definierte prinzipielle Grenze mehr hinsichtlich des autonomem Denkens oder Handelns zwischen Mensch und Maschine festlegen. Die feste Grenze zwischen beiden Systemen löst sich auf.

Die systemtheoretische Perspektive von KI und der Begriff der „Autonomie"

Der Begriff ‚Künstliche' Intelligenz weckt unmittelbar die Frage danach, was ‚natürliche' Intelligenz sei. Die Wissenschaft gibt hier vielfältige Antworten, die zunächst nichts mit den in diesem Aufsatz zentralen Begriffen „Person" oder „Autonomie" zu tun haben. Die einfachste Definition ist der oft zitierte Satz von Boring (1923) „Intelligenz ist das was Intelligenztests messen".

Wir reden von Künstlicher Intelligenz, wenn bei Computersystemen Eigenschaften und Fähigkeiten beobachtbar sind, die menschlichen Eigenschaften und Fähigkeiten gleichen. Dabei lassen sich generell zwei Klassen von Systemen unterscheiden: die der regelbasierten Expertensysteme, deren Regelwerke oder Grammatiken in speziell formulierten Sprachen abgebildet werden. Diese Systeme werden z.B. in der Medizindiagnostik oder -medikation intensiv eingesetzt und sind überall dort erfolgreich, wo kontrollierbare Umgebungen oder Randbedingungen klare Problemformulierungen zulassen. Das resultierende Systemverhalten ist komplett

vorhersehbar und in allen Verästelungen abbildbar. Die Verantwortung für die Systemleistung (z.B. Haftung) liegt beim Erzeuger. So sind auch ethische Fragen eindeutig dem Produktverantwortlichen zuordenbar. In den 90er Jahren wurden Ergänzungen in Form sog. Fuzzy-Technik vorgenommen, damit Expertensysteme auch mit unscharfen Eingangsdaten funktionieren konnten, ohne jedoch den festen Logikrahmen zu sprengen. Ein anderer Entwurf sind die sog. künstlichen neuronalen Netze, die nicht explizit alle Lösungswege einprogrammiert haben, sondern anhand von Beispielen eine Approximation zu einer gestellten Aufgabe erzeugen. Zurzeit erleben die Neuronalen Netze eine Renaissance, die zunächst, kommend aus der Kybernetik der 50er und 60er Jahre, in den 80er und 90er Jahren des letzten Jahrhunderts einen ersten Boom hatten. Die praktischen Anwendungen (z.B. Gesichtserkennung, Prognoseverfahren) scheiterten jedoch an der mangelnden Rechnerleistung der damaligen Computer, und so kam es zum oft zitierten „Neuronalen Winter“, der erst vor wenigen Jahren von einem beispiellosen Boom abgelöst wurde. Die Grundlagen für diese Idee des „Lernens“ wurde bereits 1949 gelegt, als der Psychologe Donald O. Hebb erste Lernregeln aufstellte, mit denen biologische neuronale Systeme modelliert werden sollten. “Wenn ein Axon der Zelle A Zelle B erregt und wiederholt und dauerhaft zur Erzeugung von Aktionspotentialen in Zelle B beiträgt, so resultiert dies in Wachstumsprozessen oder metabolischen Veränderungen in einer oder in beiden Zellen, die bewirken, dass die Effizienz von Zelle A in Bezug auf die Erzeugung eines Aktionspotentials in B größer wird." D.h. in einem Netzwerk verstärken sich Verbindungen zwischen Rechenknoten immer dann, wenn diese zu einem möglichst „guten“ Gesamtergebnis beitragen. „Gut“ war dabei als Approximationsgüte an eine Zielfunktion definiert und der Grad der Anpassung der gewichteten Verknüpfungen der Knoten wurde durch eine Differenzfunktion abgebildet. In einem solchen programmierten System ist die Gesamtleistung des Systems nicht mehr explizit programmiert, sondern ist abhängig vom Lernverfahren und von den gezeigten Trainingsdaten. Die Frage, ob das so erzeugte emergente Verhalten als „autonom“ zu bezeichnen ist, beantworten die folgenden Ausführungen.

Nichtlineare Dynamik als Sprache neuronaler Netze: Neben den eben skizzierten Expertensystemen, die regelbasiert und damit strenger Logik folgend arbeiten, ist die Situation bei sog. neuronalen Netzen völlig anders: Ein Neuronales Netz ist formal eine Superposition verschieden gerichteter Erzeugendenfunktionen, die keine Basis

im mathematischen Sinne darstellen. Dennoch spannen diese (i.d.R. Sigmoid- oder Gauss- ähnliche) Funktionen einen hochdimensionalen Funktionenraum auf, in dem nichtlineare Approximationen bzw. Diskriminierungen verschiedener Muster möglich sind.

Generell wird zwischen überwachten und unüberwachten Systemen unterschieden. Die Optimierung dieses Funktionenraumes erfolgt in einem Fall durch sog. „Lernen“ von Zielfunktionen, indem die Wichtungen (und ggf. weitere Funktionsparameter) in den Superpositionen der Funktionen optimiert werden. Je nach Verfahren unterscheidet man einfache Feedforward Netze von komplexeren, z.B. lokalen Netzen oder rückgekoppelten (rekurrenten) Netzen mit eigener Aktivierungsdynamik. In einfachen Netzen ist die Zeit lediglich ein Parameter; komplexere Systeme bilden Dynamiken ab und haben daher Zeit als Variable in den Gleichungen. Während erstere einfach in einen (oder mehrere) Fixpunkt(e) hineinlaufen, können rekurrente dynamische Netze Attraktordynamiken entwickeln, wie sie aus zahlreichen nichtlinearen gekoppelten Systemen wie Aktivator-Inhibitorsysteme bekannt sind und wiederholt zur Modellierung biologischer oder evolutiver Prozesse genutzt wurden. Die damit verbundenen Probleme sind bekannt: selten lassen sich diese geschlossen lösen, sondern es lassen sich nur Sonderfälle betrachten, Stabilitätsnachweise in der Nähe von Fixpunkten z.B. durch Störungsrechnung erbringen oder die Art der Attraktordynamik in der Nähe der Lösungen eruieren. Die Komplexitätstheorie und auch die Chaostheorie haben hier ausführliche Betrachtungen dazu ermöglicht. In der Regel sind diese Systeme nur numerischen Methoden zugänglich oder nicht polynomial lösbar. Transferiert man diese Betrachtungen auf KI mit KNN, so ist nun eine der damit verbundenen Fragen, ob es KNN geben kann, die zwar nichtlineare, hochdimensionale und gekoppelte Dynamiken abbilden, jedoch

1. prinzipiell deterministisch sind, oder
2. nicht deterministisch sind.

Der erste Fall lässt sich nochmals unterteilen in solche Fälle, in denen der Determinismus praktikabel nachweisbar ist und in solche, die zwar deterministisch sind, die jedoch z.B. aufgrund der hohen Dimensionalität oder Zahl der möglichen Lösungen in praktischen Fällen nicht deterministisch erscheinen oder als solche nicht beobachtbar sind. Dies trifft u.a. auf sog. np-vollständige Probleme zu, da deren mögliche Lösungskombinationen exponentiell mit dem Parameter p ansteigen. Im

letzteren Fall mag ein solches System wie nichtdeterministisch erscheinen, z.B. indem bereits minimale Änderungen im Eingangsraum, die eigentlich identische Signale darstellen sollten, durch Halbleiterrauschen, Rundungsfehler im Algorithmus o.ä. zu anderen Ergebnissen im Lösungsraum führen. Mathematisch können Ursachen hierfür Bifurkationen sein, oder unterschiedliche Typen von Attraktoren. Es wurde bereits gezeigt, dass solche deterministischen Systeme äquivalent sind zu Müller Turing Maschinen bzgl. ihres Lösungsvermögens. Systeme haben eine besondere Robustheit, indem z.B. sogenannte Graceful Degradation dafür sorgt, dass minimale Abweichungen nicht zu einem völligen Versagen des Systems - im Extremfall bzgl. der Gesamtfunktion, in der Praxis bzgl. der Approximationsgüte - führt. Das damit verbundene notwendige Wohlverhalten ist sowohl in biologischen wie auch in künstlichen Systemen wünschenswert, führt jedoch ebenso in das bekannte Stabilitäts-Flexibilitätsdilemma, das für alle Systeme gilt: bis zu welchen Variationen im Eingangsraum sollte ein System gleiche oder ähnliche Lösungen zeigen und ab wann sollte entweder eine andere stabile Lösung angesteuert werden, bzw. ein Umschalten der Dynamik in ein anderes Muster erfolgen. Graceful Degradation analog zu biologischen Systemen wird in der Regel versucht, durch sog. Neuromorphic Computing nachzubilden: Die Struktur auf der gerechnet wird, ist dem biologischen Vorbild nachgebildet und möglichst problemadäquat gewählt. Dies erzeugt auch eine strukturelle Robustheit gegen Störungen, da ähnliche Eingangsinformationen zu ähnlichen Orten im Funktionen- oder Vektorraum führen, in dem die Abbildung geschieht oder die Dynamik abläuft.

Fazit:
Es gibt eine Klasse neuronaler Netze, die aufgrund der oben beschriebenen Eigenschaften die Voraussetzungen für Autonomie besitzen, die über eine reine komplexe Abbildung hinaus geht und eigene Freiheitsgrade besitzen kann, zugleich jedoch eine definierte Problemklasse gut genug beherrscht, um auch das skizzierte Wohlverhalten zu zeigen: Mit den zur Zeit üblicherweise betrachteten Feed Forward überwacht trainierten neuronalen Netzen ist ein nichtdeterministisches Verhalten nicht zu erwarten. Wohl aber können diese Systeme, wenn z.B. in autonomen Fahrzeugen eingesetzt, autonom erscheinen, da aufgrund der hohen Dimensionalität und der Varianz der Eingangssituationen nicht alle in der Praxis eintretenden Situationen explizit programmiert, trainiert oder getestet werden konnten. Nichtdeterministische

Systeme haben dagegen das prinzipielle Potenzial, Autonomie zu zeigen, was nicht heißt, dass alle nichtdeterministischen Systeme sinnvolle Autonomie zeigen. Denn Autonomie setzt auch ein problemangepasstes Verhalten voraus, sonst ist es lediglich chaotisch und unvorhersagbar. Mit den oben beschriebenen Eigenschaften, kann KNN also - unter bestimmten Bedingungen - echte Autonomie zugeschrieben werden. Und so stellt sich auch hier die Frage, in welchem Sinne sich künstliche Systeme essentiell vom Menschen unterscheiden, wenn wesentliche Eigenschaften, die für den Status einer Entität als Person relevant sind, bereits heute durch künstliche Systeme realisiert werden können. Die KI-Forschung nähert sich auf jeden Fall mit großen Schritten der moralischen Ununterscheidbarkeit von Mensch und Maschine. Damit stellt sich die Frage, wie wir damit umgehen wollen. In den weiteren Entwicklungen wird es langfristig entscheidend sein, welche Moral und welche ethischen Werte wir auch im Kontext von KI leben werden.

Literatur zum Thema -auch die hier verwendete- gerne auf Anfrage beim Autor.