Rechtsperson Roboter – Philosophische Grundlagen für den rechtlichen Umgang mit künstlicher Intelligenz[1]

Dr. iur. Jonathan Erhardt, B.Phil. und Prof. Dr. iur. et lic. phil. Martino Mona, Universität Bern

A. Einführung

Technische und wissenschaftliche Fortschritte stellen Rechtsordnungen regelmässig vor neue Herausforderungen.[2] Selbst ein oberflächlicher Blick auf die letzten Jahrzehnte offenbart zahlreiche Beispiele. Die Rechtswissenschaft, Rechtsprechung und Rechtsetzung mussten sich mit neuen Phänomenen wie Cyberkriminalität, der DNA-Analyse, der CRISPS/Cas Methode zur Genveränderung, Waffen, die sich mit 3D-Drucker erstellen lassen, erschwinglichen Drohnen, Fortschritten in der Überwachung von Telekommunikation, selbstfahrenden Autos und der Präimplantationsdiagnostik auseinandersetzen. Bei all diesen neuen Technologien galt es, einen mit der bisherigen Rechtsprechung und Rechtsetzung konsistenten Umgang mit diesen Phänomenen zu entwickeln. Bei einigen ist dieser Prozess noch nicht abgeschlossen und es wurde noch kein befriedigender rechtlicher Umgang gefunden.

Die vermutlich grösste solche Herausforderung steht uns noch bevor: Wie soll die Rechtsordnung mit fortgeschrittenen künstlichen Intelligenzen und Robotern umgehen, mit intelligenten Akteuren, die nicht durch Evolution entstanden sind, sondern durch Menschen erschaffen wurden? Bisher hat der Mensch in allen Rechtsordnungen einen Sonderstatus genossen. Viele Rechte und Pflichten waren dem Menschen vorbehalten. So

---

1 Wir bedanken uns bei den Teilnehmerinnen und Teilnehmern am Law & Robots Workshop vom 22. April 2016 an der Universität Basel für konstruktive Anregungen. Vielen Dank vor allem an Dr. Tizian Troxler für seine ausführliche Rückmeldung und an Dr. Kristin Boosfeld für ihre überaus wertvolle Unterstützung. Jonathan Erhardt wurde für diesen Aufsatz von der Stiftung für Effektiven Altruismus finanziert.

2 Siehe etwa *M. Schulte*, Technische Innovation und Recht: Antrieb oder Hemmnis, Heidelberg 1997, für eine Untersuchung des Zusammenspiels zwischen technischer Innovation und Recht.

konnten sich üblicherweise[3] nur Menschen strafbar machen, Verträge schliessen, Prozesse führen und sich auf verfassungsrechtlich garantierte Grundrechte berufen.[4] In Zukunft könnten uns fortschrittliche künstliche Intelligenzen und Roboter diese rechtliche Sonderposition streitig machen. Bisher hatten künstliche Intelligenzen und Roboter bloss eingeschränkte Fähigkeiten in einem bestimmten Anwendungsgebiet. Diese berühren den besonderen rechtlichen Status der Menschen nicht. Zukünftig könnten aber künstliche Intelligenzen entstehen, die dem Menschen in allen Aufgaben und Tätigkeiten mindestens ebenbürtig sein werden, oder ihn sogar weit überflügeln. Bei solchen künstlichen Akteuren stellt sich die Frage, ob sie erstmals die Anthropozentrizität heutiger Rechtsordnungen zu erschüttern vermögen.

Der beeindruckende technische Fortschritt in den letzten Jahren legt die Vermutung nahe, dass eine solche künstliche Intelligenz in absehbarer Zeit Wirklichkeit werden könnte. Gemäss einer Umfrage nehmen die renommiertesten Experten im Bereich der künstlichen Intelligenz im Median mit 90% Wahrscheinlichkeit an, dass 2070 eine künstliche Intelligenz existieren wird, die Menschen in den meisten Aufgaben mindestens ebenbürtig sein wird.[5] Und aufgrund des Singularitätsarguments scheint es wahrscheinlich, dass es von einer solchen Intelligenz zu einer dem Menschen in allen Bereichen deutlich überlegenen künstlichen Intelligenz nicht annähernd so lange dauern wird, wie von der Erfindung des Computers bis zur Entwicklung der ersten menschenähnlichen künstlichen Intelligenz.[6]

---

3 Ausnahmen sind namentlich Strafprozesse gegen Tiere (siehe etwa *P. Dinzelbacher*, Das fremde Mittelalter: Gottesurteil und Tierprozess, Essen 2006) oder Fälle von Unternehmensstrafbarkeit (siehe etwa *K. Seelmann*, Personalität und Zurechnung von der Aufklärung bis zur Philosophie des Idealismus, in: M. Heer et al. (Hrsg), Toujours agité, jamais abattu – Festschrift für Hans Wiprächtiger, Basel 2001, S. 575 ff.).

4 Es existieren aber interessante Argumente dafür, auch Primaten durch gewisse Grundrechte zu schützen, vgl. *R. Fasel/A. Mannino/T. Baumann/C. Blattner*, Grundrechte für Primaten. Positionspapier von Sentience Politics 2016, S. 1 ff., abrufbar unter https://ea-stiftung.org/positionspapiere/ (zuletzt abgerufen am 12.09.2016); vgl. grundlegend auch *S. Stucki*, Grundrechte für Tiere, Baden-Baden 2016.

5 *N. Bostrom*, Superintelligence, Paths, Dangers, Strategies, Oxford 2014, S. 19.

6 In vereinfachter Form lautet das Singularitätsargument: Wenn es einmal eine den Menschen in allen Aufgaben mindestens ebenbürtige künstliche Intelligenz gibt, dann wird diese vermutlich eine noch bessere künstliche Intelligenz programmieren können, die dann eine noch bessere künstliche Intelligenz programmieren kann, und

Einige wirtschaftliche Indikatoren sprechen dafür, dass die Einschätzung der Experten auch von Unternehmen geteilt wird: In letzter Zeit haben mehrere grosse Unternehmen wie Google und IBM grosse Investitionen im Bereich der künstlichen Intelligenz getätigt. So wurde beispielsweise die Firma DeepMind 2014 von Google für 400 Millionen Dollar gekauft.[7] Insgesamt haben sich die Investitionen in diesem Bereich in den letzten fünf Jahren verdreifacht und 2013 ein Gesamtvolumen von 581 Millionen Dollar erreicht.[8] Diese Indikatoren legen nahe, dass in absehbarer Zeit, vermutlich noch innerhalb dieses Jahrhunderts, eine übermenschliche künstliche Intelligenz entwickelt wird.[9]

I. Neue und vertraute Probleme der künstlichen Intelligenz

Roboter und künstliche Intelligenz stellen uns aus zwei Gründen vor eine aussergewöhnliche rechtliche Herausforderung. Zum einen kombinieren sie viele Elemente, die den rechtlichen Umgang mit bisherigen technologischen Innovationen schwierig gemacht haben. So haben beispielsweise schon frühere technologische Errungenschaften die Frage nach den ethischen und rechtlichen Grenzen invasiver Überwachungspraktiken aufgeworfen, aber Algorithmen im Bereich der künstlichen Intelligenz machen neue flächendeckende Überwachungsprogramme deutlich mächtiger, und dementsprechend werden die damit verbundenen rechtlichen Probleme

---

so weiter. So wird die Entwicklung mit jeder Generation schneller. Für eine ausführliche Untersuchung des Arguments siehe *D. Chalmers*, The Singularity: A Philosophical Analysis, Journal of Consciousness Studies 17 (2010), S. 7 ff. Für einen Überblick über verschiedene Szenarien zum Übergang von menschenähnlicher zu übermenschlicher künstlicher Intelligenz siehe *Bostrom*, Superintelligence (Fn. 5), Kapitel 4.

7 Für einen Überblick über einige Akquisitionen vgl. *S. Betschon*, Ende der Bescheidenheit, Neue Zürcher Zeitung vom 30.1.2014, sowie *C. Eliasmith*, On the Eve of Artificial Minds, in: T. Metzinger/J.M. Windt (Hrsg.), Open MIND, Frankfurt am Main 2015, S. 1 ff. (S. 2 f.).

8 Vgl. dazu https://www.cbinsights.com/blog/artificial-intelligence-venture-capital (zuletzt abgerufen am 12.09.2016).

9 Eine ausführliche Analyse der Aussichten auf eine solche künstliche Intelligenz liegt jenseits der Möglichkeiten dieses Aufsatzes. Siehe dazu *Bostrom*, Superintelligence (Fn. 5), Kapitel 2. Zum gegenwärtigen Stand der Forschung vgl. *K. Mnih et al.*, Playing Atari with Deep Reinforcement Learning, 2013, abrufbar unter https://www.cs.toronto.edu/~vmnih/docs/dqn.pdf (zuletzt abgerufen am 12.09.2016).

drängender. Zum anderen stellt uns der wissenschaftliche Fortschritt im Bereich der künstlichen Intelligenz aber vor eine Reihe gänzlich neuartiger Probleme. Diese Probleme hängen mit dem umstrittenen Status künstlicher Intelligenzen als Rechtspersonen zusammen.

Erstere können wir die einfachen rechtlichen Probleme der künstlichen Intelligenz nennen.[10] In der Praxis sind diese Probleme zwar oft alles andere als einfach zu lösen, aber wenigstens sind wir mit ihnen aufgrund von früheren technischen Errungenschaften schon vertraut und kennen die richtige Methode, um sie zu lösen. Beispiele für einfache rechtliche Probleme der künstlichen Intelligenz sind die Problematik flächendeckender Überwachung der Kommunikation dank künstlicher Intelligenz (Beispiele dafür sind das US-amerikanische PRISM-Überwachungsprogramm und das indische NETRA), die Frage, ob künstliche Intelligenz patentiert werden darf, die Problematik der rechtlichen Zulassung selbstfahrender Autos[11] sowie die haftungsrechtliche Einordnung des Gefahrenpotentials der Entwicklung künstlicher Intelligenz.[12]

Die schwierigen rechtlichen Probleme der künstlichen Intelligenz gehen insofern über die einfachen Probleme hinaus, als dass sie mit dem Status künstlicher Intelligenzen als *Rechtspersonen* zusammenhängen: Können künstliche Intelligenzen als selbständige Vertragsparteien auftreten bzw. Verträge schliessen?[13] Können künstliche Intelligenzen für ihr Handeln

10 In Anlehnung an *D. Chalmers*, Facing Up to the Problem of Consciousness, Journal of Consciousness Studies 2, 3 (1995), S. 200 ff. *Chalmers* unterscheidet zwischen den einfachen Problemen und dem schwierigen Problem in der Philosophie des Geistes. Die einfachen Probleme betreffen die Erklärung des Gehirns und seiner Funktionen, das schwierige Problem besteht in der Erklärung von Bewusstsein selbst.

11 Vgl. *A. Sharma*, Driving the Future: The Legal Implications of Autonomous Vehicles, 2012, abrufbar unter: http://law.scu.edu/hightech/autonomousvehiclecon frecap2012 (zuletzt abgerufen am 12.09.2016).

12 *Bostrom*, Superintelligence (Fn. 5), Kapitel 8–15. Für eine Übersicht über die einfachen juristischen Probleme der künstlichen Intelligenz siehe *M.F. Müller*, Roboter und Recht – Eine Einführung, Aktuelle Juristische Praxis 5 (2014), S. 595 ff. und *R. Calo*, Roboters in American Law, University of Washington School of Law Research Paper 2016-04.

13 Vgl. *T. Allen/R. Widdison*, Can Computers Make Contracts?, Harvard Journal of Law & Technology 9, 1 (1996), S. 25 ff.; *E. Abdel/R. Dahiyat*, Intelligent Agents and Contracts: Is a Conceptual Rethink Imperative?, Artificial Intelligence and Law 15 (2007), S. 375 ff.

(strafrechtlich) verantwortlich sein? Haben künstliche Intelligenzen einen Anspruch auf Grundrechte?

Nach der Klärung einiger relevanter Konzepte werden wir uns im zweiten Teil dieses Aufsatzes mit dem Kernproblem all dieser schwierigen Probleme der künstlichen Intelligenz auseinandersetzen: Können schon existierende oder zukünftige künstliche Intelligenzen Rechtspersonen sein? Die Beantwortung dieser Frage soll die Grundlage liefern für die Verteidigung der These, dass zumindest zukünftige künstliche Intelligenzen für ihr Handeln strafrechtlich verantwortlich sein können. Bei diesen Fragen berücksichtigen wir nicht nur den heutigen Stand der Technik, sondern stellen diese Fragen auch für zukünftige künstliche Intelligenzen. Deshalb werden unsere Einschätzungen notwendigerweise auch eine spekulative Komponente enthalten. Es sollte aber betont werden, dass sich die spekulativen Teile nur auf die zukünftige Entwicklung im Gebiet der künstlichen Intelligenz beziehen, nicht auf die Bedingungen, die erfüllt sein müssen, damit ein intelligenter Agent Rechtspersönlichkeit hat. Deshalb kann die folgende Diskussion auch dann von Nutzen sein, wenn sich die technische Entwicklung anders abspielen wird, als wir das erwarten. Der Fokus unserer Untersuchung liegt dabei auf den philosophischen Grundlagen dieser Fragen. Vielen naheliegenden juristischen Folgefragen werden wir in diesem Aufsatz daher nicht nachgehen.

II. Terminologie und Grundlagen der künstlichen Intelligenz

Manchmal leiden juristische Diskussionen zu künstlicher Intelligenz an unzureichend präziser Terminologie. Alltagssprachliche Begriffe wie „Intelligenz“ oder „Autonomie“ sind meistens vage, manchmal mehrdeutig und bringen oft Assoziationen mit sich, die für die Klärung der Probleme hinderlich sind. Es lohnt sich deshalb, einige Kernkonzepte explizit zu definieren, um Missverständnisse und falsche Assoziationen zu vermeiden.

Unter einer künstlichen Intelligenz verstehen wir einen nicht durch Evolution entstandenen, sondern künstlich erschaffenen, intelligenten Akteur oder Agent. Wir können einen sehr breiten Agentenbegriff verwenden, wonach jedes System, dessen Handlungen sich durch Zuschreibung von Zielen und Überzeugungen systematisch erklären und prognostizieren lassen, ein Agent ist. Diese Charakterisierung orientiert sich an Daniel Dennetts „Intentional Stance“: „The intentional stance is the strategy of interpreting the behavior of an entity (person, animal, artifact, whatever)

by treating it as if it were a rational agent who governed its ‚choice' of ‚action' by a ‚consideration' of its ‚beliefs' and ‚desires'."[14] John McCarthy, der Erfinder der Programmiersprache Lisp, hat einen ähnlichen Ansatz gewählt: „To ascribe certain beliefs, knowledge, free will, intentions, consciousness, abilities or wants to a machine or computer program [...] is useful when the ascription helps us understand the structure of the machine, its past or future behavior, or how to repair or improve it."[15]

Nach diesem Kriterium sind Menschen, Säugetiere, Insekten und vielleicht sogar bereits gewisse existierende Computersysteme Agenten. Bei all diesen Wesen und Objekten ist es möglich, ihre Handlungen durch Kombinationen von Zielen und Überzeugungen systematisch zu beschreiben. McCarthy geht sogar davon aus, dass sehr einfache Systeme Agenten in unserem Sinne sind: „Machines as simple as thermostats can be said to have beliefs, and having beliefs seems to be a characteristic of most machines capable of problem solving performance."[16] Dem widerspricht David Caverley: „If the complexity of AI behavior did not exceed that of a thermostat, then it is not likely that anyone would be convinced that AIs really possess intentional states – that they really believe things or know things."[17] Wir müssen hier nicht eine genaue untere Grenze für die Anwendung des Begriffs „Agent" festlegen, da die Komplexität rechtlich interessanter Computerprogramme weit über diejenige eines Thermostats hinausgeht.

Unter Intelligenz verstehen wir die Fähigkeit eines Akteurs, seine Ziele in einer grossen Vielfalt an unbekannten Umgebungen erreichen zu können.[18] Dieses Verständnis fängt einige Elemente ein, die oft mit Intelligenz bezeichnet werden, etwa Mittel-Zweck-Rationalität,[19] Anpassung an neu-

14 *D. Dennett*, Intentional Stance, in: R.A. Wilson/F.C. Keil (Hrsg.), The MIT Encyclopedia of the Cognitive Sciences, Cambridge (Mass.) 2001, S. 412.

15 *J. McCarthy*, Ascribing Mental Qualities to Machines, Philosophical Perspectives in Artificial Intelligences 1979, Introduction, abrufbar unter: http://www-formal.stanford.edu/jmc/ascribing/ascribing.html (zuletzt abgerufen am 21.11.2016).

16 *McCarthy*, Qualities (Fn. 15), Introduction.

17 *D.J. Calverley*, Legal Rights for Machines: Some Fundamental Concepts, in: M. Anderson/S.L. Anderson (Hrsg.), Machine Ethics, Cambridge 2011, S. 213 ff. (S. 221).

18 *S. Legg/M. Hutter*, Universal Intelligence: A Definition of Machine Intelligence, Minds and Machines 17, 4 (2007), S. 391 ff. (S. 415).

19 *Bostrom*, Superintelligence (Fn. 5), S. 107.

artige Bedingungen und die Fähigkeit, neuartige Probleme zu lösen.[20] Dabei können wir zwischen genereller Intelligenz und domänenspezifischer Intelligenz unterscheiden.[21] Erstere zeichnet sich durch die Fähigkeit aus, Ziele in besonders vielen verschiedenen, möglicherweise sogar allen denkbaren Umgebungen so gut wie möglich erreichen zu können. Eine domänenspezifische Intelligenz hingegen kann dies nur in einer sehr limitierten Anzahl Umgebungen. Ein Beispiel für eine generelle Intelligenz ist ein gewöhnlicher Mensch, der in verschiedensten Umgebungen die bestehenden Probleme lösen kann, um seine Ziele zu erreichen. Beim Schachcomputer Hydra oder Google's Go-Algorithmus AlphaGo handelt es sich um domänenspezifische Intelligenzen, die für eine sehr spezifische Umgebung programmiert wurden und nur dort ihre Ziele erreichen können.[22]

Da wir künstliche Intelligenz als Agent oder Akteur bestimmt haben, können wir ihr Ziele und Wünsche zuschreiben. Dieser Gedanke kann weiter präzisiert werden: Wir können die Ziele und Wünsche einer künstlichen Intelligenz in Form einer Nutzenfunktion beschreiben, die Weltzustände nach ihrer Erwünschtheit ordnet. Die Nutzenfunktion einer künstlichen Intelligenz muss dabei keineswegs menschenähnlich sein. Die Nutzenfunktion eines Schachcomputers kann beispielsweise alle Schachkonfigurationen, in denen der Computer gewinnt, als maximal erwünscht, alle Konfigurationen, in denen er unentschieden spielt, als mittelmässig erwünscht und alle Konfigurationen, in denen er verliert, als minimal erwünscht qualifizieren. Anders als bei natürlichen Agenten besteht bei künstlichen Agenten oft die Möglichkeit, als Programmierer die Nutzenfunktion direkt zu bestimmen.[23] So wurde künstliche Intelligenz beispiels-

---

20 *H.-M. Süss*, Intelligenz, in: G. Strube et al. (Hrsg.), Wörterbuch der Kognitionswissenschaften, Stuttgart 1996, S. 280.

21 *Fasel/Mannino/Baumann/Blattner*, Grundrechte (Fn. 4), S. 7.

22 Gute aber nicht unfehlbare Evidenz für generelle Intelligenz ist die Fähigkeit, den Turing Test zu bestehen, siehe dazu *A.M. Turing*, Computing Machinery and Intelligence, 1950, abrufbar unter http://orium.pw/paper/turingai.pdf (zuletzt abgerufen am 12.09.2016) und *J.H. Moor*, The Turing Test: The Elusive Standard of Artificial Intelligence, Dordrecht 2003.

23 Wie stark auf die Nutzenfunktion Einfluss genommen werden kann, hängt von der Art der künstlichen Intelligenz ab. Bei Gehirnsimulationen wie dem Blue Brain Project (http://bluebrain.epfl.ch, zuletzt abgerufen am 12.09.2016) ist diese Möglichkeit vermutlich eingeschränkt. Bei „Good Old-Fashioned Artificial Intelligence", bei der künstliche Intelligenz als Programme zur Symbolmanipulation direkt programmiert wird, *Bostrom*, Superintelligence (Fn. 5), S. 7, ist die Möglichkeit der Einflussnahme sehr ausgeprägt. Wie künstliche Intelligenzen am besten

weise programmiert, um Schachspiele zu gewinnen, Spammails zu löschen und Fahrzeuge unfallfrei und unter Beachtung der Verkehrsregeln an den Zielort zu manövrieren. Die Fähigkeit, die Nutzenfunktion zu programmieren, führt nicht unbedingt dazu, dass eine künstliche Intelligenz kontrollierbar oder berechenbar ist. Ein sehr intelligenter künstlicher Agent könnte seine Ziele auf für uns sehr überraschende Art und Weise verfolgen. Der Teslagründer Elon Musk hat dies mit folgendem Beispiel illustriert: Eine künstliche Intelligenz, die Spam-mails verhindern soll, könnte bemerken, dass die einfachste Art Spam-mails permanent zu verhindern, die Zerstörung der Menschheit ist.[24] Bei Menschen ist die Nutzenfunktion teilweise biologisch bestimmt und kann nur innerhalb biologischer Grenzen durch Erziehung und andere Lebenserfahrungen beeinflusst werden.

Als Akteure können künstliche Intelligenzen auch Überzeugungen über die Welt bilden und diese nutzen, um ihre Ziele und Wünsche zu erreichen. Mit anderen Worten: Sie können Gründe abwägen und Entscheidungen treffen. Auf den ersten Blick scheint es vielleicht merkwürdig zu sagen, ein Schachprogramm oder AlphaGo habe *Überzeugungen*. Und es gibt tatsächlich wichtige Unterschiede in der Art und Weise, wie Menschen und solche Algorithmen sich in der Welt orientieren und Daten über die Welt abspeichern. Es existieren aber mehr Gemeinsamkeiten, als man zunächst annehmen könnte.

In der Entscheidungstheorie, einem Forschungsgebiet im Schnittbereich zwischen Philosophie und angewandter Wahrscheinlichkeitstheorie, werden Überzeugungen von Personen häufig als subjektive Wahrscheinlichkeiten charakterisiert, dass ein bestimmter Sachverhalt wahr ist.[25] Dieses Verständnis wird manchmal mit dem Satz „credences are degrees of belief" eingefangen und passt zu der verbreiteten Annahme, dass Überzeugungen in verschiedenen „Graden" oder „Intensitäten" vorkommen. So bin ich beispielsweise sehr sicher, dass ich existiere, und dementsprechend habe ich eine starke Überzeugung, dass ich existiere, während ich weit

---

mit Nutzenfunktionen ausgestattet werden können, wird als „Value-Loading Problem" bezeichnet, vgl. *Bostrom*, Superintelligence (Fn. 5), S. 185 ff.

24 Vgl. http://www.businessinsider.com/elon-musk-robots-could-delete-humans-like-spam-2014-10?IR=T (zuletzt abgerufen am 12.09.2016). Für eine ausführliche Diskussion des sogenannten Kontrollproblems siehe *Bostrom*, Superintelligence (Fn. 5), Kapitel 9.

25 *D.H. Mellor*, Probability: A Philosophical Introduction, London 2004, S. 66 ff.

weniger sicher bin, dass ich heute noch diesen Abschnitt fertig schreibe, und deshalb nur eine moderat starke Überzeugung habe, dies zu tun. In der Bayes'schen Entscheidtheorie wird die Stärke oder Intensität einer Überzeugung als subjektive Wahrscheinlichkeit aufgefasst und zur Grundlage der rationalen Entscheidungsfindung gemacht.[26]

Wenn wir Überzeugungen bloss als Wahrscheinlichkeitszuordnungen zu Sachverhalten verstehen, scheint es schon viel plausibler, dass gewisse künstliche Intelligenzen Überzeugungen haben und darauf basierend Entscheidungen treffen können. Denn viele künstliche Intelligenzen speichern Daten der Form „Sachverhalt X ist mit einer Wahrscheinlichkeit von Y wahr" ab. So ordnen beispielsweise Bayes'sche Spamfilter dem Sachverhalt, dass eine bestimmte Email Spam ist, eine berechnete Wahrscheinlichkeit zu.[27] Diese Wahrscheinlichkeit wird aufgrund von Textmerkmalen der Email gebildet und aktualisiert, ähnlich wie ein Mensch dies beim Lesen einer Email machen würde. Des Weiteren fällen künstliche Intelligenzen aufgrund ihrer Überzeugungen auch Entscheidungen. Ein Spamfilter verschiebt Emails in den Spamfilter sobald eine bestimmte Wahrscheinlichkeitsschwelle überschritten ist, und AlphaGo wählt zwischen verschiedenen Handlungsalternativen im Spiel diejenige mit dem höchsten Erwartungswert.

Die Erwägungen im vorherigen Abschnitt helfen uns auch bei der Frage, ob künstliche Intelligenzen sich Wissen aneignen können. Der Wissensbegriff ist zwar in der philosophischen Literatur notorisch umstritten, wir können aber die traditionelle philosophische Analyse von Wissen als wahrer, gerechtfertigter Überzeugung mindestens als erste Annäherung an die richtige Analyse verwenden. Wir haben dafür argumentiert, dass künstliche Intelligenzen Überzeugungen haben können. Sie können also auch Wissen haben, sofern mindestens eine ihrer Überzeugungen wahr und gerechtfertigt ist.

Menschliche und viele nicht-menschliche Tiere haben Bewusstsein. Die korrekte Zuschreibung vieler mentaler Zustände und Eigenschaften erfordert beim Zuschreibungsobjekt die eine oder andere Form von Bewusst-

---

26 *E.J. Horvitz/J.S. Breese/M. Henrion*, Decision Theory in Expert Systems and Artificial Intelligence, International Journal of Approximate Reasoning 2, 3 (1998), S. 247 ff.

27 *P. Graham*, A Plan for Spam, 2002, abrufbar unter: http://www.paulgraham.com/spam.html (zuletzt abgerufen am 12.09.2016).

sein.[28] Ein Objekt hat dann Bewusstsein, wenn es sich irgendwie anfühlt, dieses Objekt zu sein.[29] Können auch künstliche Intelligenzen über Bewusstsein verfügen? Es ist nützlich, noch etwas präziser zu werden und folgende zwei Fragen separat zu diskutieren:[30] Können künstliche Intelligenzen überhaupt je Bewusstsein entwickeln? Und falls ja, welche Typen von künstlichen Intelligenzen werden Bewusstsein haben? Intuitiv würden viele Menschen beide Fragen negativ beantworten, Computer sind schliesslich nur deterministische Datenverarbeitungsmaschinen. David Chalmers schreibt zu diesem Einwand:

> „It is easy to think of a computer as simply an input-output device, with nothing in between except for some formal mathematical manipulations. This way of looking at things, however, leaves out the key fact that there are rich causal dynamics inside a computer, just as there are in the brain. Indeed, in an ordinary computer that implements a neuron-by-neuron simulation of my brain, there will be real causation going on between voltages in various circuits, precisely mirroring patterns of causation between the neurons. [...] It is the causal patterns among these circuits, just as it is the causal patterns among the neurons in the brain, that are responsible for any conscious experience that arises. "[31]

Die erste Frage wird dementsprechend von Chalmers und auch von den meisten Experten positiv beantwortet.[32] Grund dafür ist die intuitive Plausibilität des von Chalmers in der Textstelle angesprochenen Prinzips der organisatorischen Invarianz. Nach diesem Prinzip haben zwei Objekte dieselben Bewusstseinszustände, wenn sie dieselbe feingliedrige organisatorische und kausale Struktur teilen.[33] Wenn also zwei Gehirne dieselbe orga-

---

28 *D. Chalmers*, The Conscious Mind: In Search of a Fundamental Theory, Oxford 1996, S. 11 f. und S. 16 ff.

29 *T. Nagel*, What is it like to be a bat?, The Philosophical Review 83, 4 (1974), S. 435 ff.

30 *Fasel/Mannino/Baumann/Blattner*, Grundrechte (Fn. 4), S. 10.

31 *Chalmers*, Conscious Mind (Fn. 28), S. 321.

32 *H.P. Moravec*, Mind Children: The Future of Robot and Human Intelligence, Cambridge (Mass.) 1988; *D. Chalmers*, Absent Qualia, Fading Qualia, Dancing Qualia, in: T. Metzinger (Hrsg.), Conscious Experience, Paderborn 1995, S. 309 ff., abrufbar unter: http://consc.net/papers/qualia.html (zuletzt abgerufen am 15.09.2016); *T. Metzinger*, Der Ego Tunnel. Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsethik, München 2014. Siehe jedoch *J.R. Searle*, Minds, Brains, and Programs, Behavioral and Brain Sciences 3 (1980), S. 417 ff. für eine prominente Gegenstimme.

33 *Chalmers*, Absent Qualia (Fn. 32).

nisatorische und kausale Struktur aufweisen, so teilen sie auch dieselben Bewusstseinszustände. Aus diesem Prinzip folgt, dass mindestens neuromorphe künstliche Intelligenzen, also solche, die auf denselben organisatorischen Strukturen wie das Gehirn basieren, Bewusstsein haben können.[34] Bei nicht neuromorphen künstlichen Intelligenzen ist der Expertenkonsens weniger ausgeprägt. Es ist beispielsweise umstritten, ob reine Simulationen Bewusstsein haben können; dies wird von Chalmers[35] und Bostrom[36] bejaht und von Koch,[37] Searle[38] und Tononi[39] verneint.[40]

Weitere Dimensionen menschlichen Handelns sind Autonomie bzw. Willensfreiheit. Auch hier gibt es eine Vielzahl philosophischer Positionen. Wir möchten mit zwei etablierten Gruppen von Willensfreiheitskonzepten arbeiten, nämlich den inkompatibilistischen und den kompatibilistischen Konzeptionen. Gemäss ersterer ist Willensfreiheit nicht mit Determinismus vereinbar: Wenn das Handeln eines Akteurs vollständig durch die Naturgesetze und die Anfangsbedingungen des Universums bestimmt sind, so hat er keine Willensfreiheit.[41] Nach dieser Konzeption ha-

---

34 *Chalmers*, Absent Qualia (Fn. 32), nennt einige der stärksten weiteren Argumente für die Bewusstseinsfähigkeit fortschrittlicher künstlicher Intelligenzen. Gemäss dem „Fading Qualia"-Argument könnte bei einem menschlichen Gehirn ein Neuron nach dem anderen durch einen Silikonchip mit demselben kausalen Profil ersetzt werden, bis am Schluss das ganze Gehirn aus Silikonchips besteht. Es ist aber unplausibel, dass bei diesem Prozess das Bewusstsein entweder bei einem bestimmten Neuron plötzlich verschwindet, oder langsam verblasst, bis es ganz weg ist. Also hat auch das Silikonhirn, d.h. eine Form von künstlicher Intelligenz, Bewusstsein.

35 *Chalmers*, Absent Qualia (Fn. 32).

36 *N. Bostrom*, Are We Living in a Computer Simulation?, The Philosophical Quarterly, 53, 211 (2003), S. 243 ff.

37 *C. Koch*, What it Will Take for Computers to Be Conscious, MIT Technology Review 2014, abrufbar unter: http://www.technologyreview.com/news/531146/ what-it-will-take-for-computers-to-be-conscious (zuletzt abgerufen am 12.09.2016).

38 *Searle*, Minds (Fn. 32).

39 *G. Tononi*, Integrated Information Theory. Scholarpedia 10, 1 (2015), 4164, abrufbar unter: http://scholarpedia.org/article/Integrated_Information_Theory (zuletzt abgerufen am 12.09.2016).

40 Bewusstsein ist eine Voraussetzung für Leidensfähigkeit. Die Vereinigung „People for the Ethical Treatment of Reinforcement Learners" (PETRL) spricht sich dafür aus, dass künstliche Intelligenzen, sofern sie leidensfähig sind, die gleiche moralische Berücksichtigung erhalten sollten wie „biologische Intelligenzen": http://petrl.org (zuletzt abgerufen am 12.09.2016).

41 *P. van Inwagen*, The Incompatibility of Free Will and Determinism, Philosophical Studies 27, 3 (1975), S. 185 ff.

ben künstliche Intelligenzen voraussichtlich keine Willensfreiheit, denn ihre Algorithmen funktionieren nach streng deterministischen Gesetzen. Es ist aber zu betonen, dass auch umstritten ist, ob Menschen nach dieser Konzeption überhaupt frei sind. Denn es gibt interessante Argumente dafür, dass das Universum entweder deterministisch ist,[42] oder dass allfälliger Indeterminismus sich auf der makroskopischen und entscheidungsrelevanten Struktur des Gehirns nicht auswirkt.[43]

Kompatibilistische Konzeptionen von Willensfreiheit gehen davon aus, dass auch determinierte Akteure frei sein können. Eine der populärsten kompatibilistischen Theorien ist der Gründe-Kompatibilismus, der unter anderem von Dennett und Nozick[44] vertreten wird. Grob besagt der Gründe-Kompatibilismus, dass ein Akteur dann frei ist, wenn er auf angemessene Weise auf Gründe für und gegen eine Handlung reagieren kann. Er ist dann unfrei, wenn er nicht auf rationale Weise auf Gründe eingehen kann, z.B. weil er unter Zwangsstörungen leidet.[45] Nach diesem Verständnis scheint es für künstliche Intelligenzen problemlos möglich, frei zu sein. Sie müssen nur mit einem genügend guten Entscheidungsalgorithmus versehen worden sein und entsprechend rational auf ihre Überzeugungen reagieren können.

Eine künstliche Intelligenz kann in einen Roboter, d.h. in einen mechanischen Körper implementiert werden, oder diesen aus Distanz kontrollieren. Dies ist aber nicht zwingend, viele domänenspezifische künstliche Intelligenzen wie beispielsweise künstliche Intelligenzen in Computerspielen interagieren nie mit Robotern, sondern laufen ausschliesslich auf gewöhnlichen Computern. Wir werden im Verlauf der weiteren Diskussion

---

42 *L.E. Szabó*, Is Quantum Mechanics Compatible with a Deterministic Universe? Two Interpretations of Quantum Probabilities, Foundations of Physics Letters 8, 5 (1995), S. 417 ff.; *L. Vaidman*, Many-Worlds Interpretation of Quantum Mechanics, in: E.N. Zalta (Hrsg.), The Stanford Encyclopedia of Philosophy, 2016, abrufbar unter: http://plato.stanford.edu/archives/spr2016/entries/qm-manyworlds (zuletzt abgerufen am 12.09.2016).

43 *P.G. Clarke*, Neuroscience, Quantum Indeterminism and the Cartesian Soul, Brain and Cognition 84, 1 (2014), S. 109 ff.

44 *D.C. Dennett*, Elbow Room: The Varieties of Free Will Worth Wanting, Cambridge (Mass.) 2015; *R. Nozick*, Philosophical Explanations, Cambridge (Mass.) 1981.

45 *M. McKenna/D.J. Coates*, Compatibilism, in: E.N. Zalta (Hrsg.), The Stanford Encyclopedia of Philosophy, 2015, abrufbar unter: http://plato.stanford.edu/entries/compatibilism (zuletzt abgerufen am 12.09.2016).

auf die hier eingeführten Konzepte zurückgreifen, um die Brücke zu spezifisch rechtlichen Begriffen zu schlagen, die in einem ersten Schritt für die Rechtspersönlichkeit und in einem zweiten Schritt für strafrechtliche Verantwortlichkeit künstlicher Intelligenzen eine Rolle spielen.

B. Künstliche Intelligenzen als Personen und Rechtspersonen

Im Folgenden wenden wir uns mehreren schwierigen Problemen der künstlichen Intelligenz zu. Die Problemauswahl ist nicht zufällig, sondern bereitet die Beantwortung der zentralen Frage dieses Aufsatzes vor: Können existierende oder zukünftige künstliche Intelligenzen Rechtspersonen sein, die man sodann in einem weiteren Schritt auch strafrechtlich zur Verantwortung ziehen könnte?

I. Künstliche Intelligenzen als Personen

Das Konzept der Person ist, anders als der Begriff der Rechtspersönlichkeit, kein rechtsdogmatisches, sondern ein alltagssprachliches oder allenfalls philosophisches Konzept. Deshalb ist die Klassifizierung einer künstlichen Intelligenz als Person auch nicht eine notwendige Voraussetzung für deren rechtlichen Status. Personen sind aber immerhin die traditionellen Subjekte des Rechts und des Strafrechts im Besonderen[46] und deshalb ist es interessant zu fragen, ob, und wenn ja, unter welchen Voraussetzungen künstliche Intelligenzen als Personen gelten dürfen. Und wie wir sehen werden, gibt es Überlappungen zwischen den beiden Konzepten, so dass es sich anbietet, zuerst diese Frage zu beantworten.

Wie Harry Frankfurt feststellt, kann der Begriff Person verwendet werden, um bloss Spezieszugehörigkeit auszudrücken: „There is a sense in which the word ‘person’ is merely the singular form of ‘people’ and in which both terms connote no more than membership in a certain biological species.“[47] Nach diesem Personenbegriff würden künstliche Intelligenzen als Angehörige einer anderen Spezies von vornherein ausgeschlossen.

---

46 *S. Gless/T. Weigend*, Intelligente Agenten und das Strafrecht, Zeitschrift für die gesamte Strafrechtswissenschaft, 126, 3 (2014), S. 561 ff. (S. 568).

47 *H. Frankfurt*, Freedom of Will and the Concept of a Person, The Journal of Philosophy 68, 1 (1971), S. 5 ff. (S. 6).

Dieser Personenbegriff ist aber in den meisten Kontexten wenig interessant und soll uns hier nicht weiter beschäftigen.

Stattdessen wollen wir mit einem Personenkonzept arbeiten, das stets im Zentrum des philosophischen Interesses stand. Dieses Konzept endet nicht automatisch an der Speziesgrenze: „Our concept of ourselves as persons is not to be understood, therefore, as a concept of attributes that are necessarily species-specific. It is conceptually possible that members of novel or even of familiar nonhuman species should be persons; and it is also conceptually possible that some members of the human species are not persons"[48] Die richtige Analyse dieses Konzepts ist philosophisch umstritten und es hat sich kein klarer Konsens herauskristallisiert. Daniel Dennett bemerkt dazu: „One might well hope that such an important concept, applied and denied so confidently, would have clearly formulatable necessary and sufficient conditions for ascription, but if it does, we have not yet discovered them."[49] Statt hier einen weiteren Beitrag zur Diskussion über die richtigen Anwendungsbedingungen des Konzepts zu liefern, ist es sinnvoll, kursorisch einige Vorschläge zu erwägen.

Harry Frankfurts Begriffsanalyse[50] gehört zu den bekanntesten zeitgenössischen Vorschlägen. Nach Frankfurt unterscheiden sich Personen von Nichtpersonen unter anderem durch die Struktur ihres Willens. Auch Nichtpersonen haben Wünsche und Überzeugungen und treffen darauf aufbauend Entscheidungen. Charakteristisch für Personen ist aber, dass sie auch „Wünsche zweiter Stufe" bilden können. Sie können nicht nur etwas wünschen, sondern sie können sich auch wünschen, sich etwas zu wünschen. Personen können sich beispielsweise wünschen, sich zu wünschen, gerne täglich Sport zu machen, ohne dass sie *tatsächlich* gerne täglich Sport machen. Nach Frankfurt sind also alle Personen Agenten in unserem Sinne, da sie Wünsche und Überzeugungen haben. Aber nicht alle Agenten sind Personen, denn nicht alle Agenten verfügen auch über Wünsche zweiter Stufe. Könnten künstliche Intelligenzen nach diesem Konzept in die Kategorie der Personen fallen?

Nach diesem Vorschlag gibt es mindestens eine Art künstliche Intelligenz, deren Mitglieder höchstwahrscheinlich als Personen zu gelten haben, nämlich Simulationen menschlicher Gehirne (sog. „whole brain emu-

48 *Frankfurt*, Freedom (Fn. 47), S. 6.

49 *D. Dennett*, Conditions of Personhood, in: R.A. Oksenberg (Hrsg.), The Identities of Person, Berkeley 1976, S. 175 ff. (S. 175).

50 *Frankfurt*, Freedom (Fn. 47).

lations"),[51] wie sie durch das Blue Brain Project der EPFL ermöglicht werden sollen.[52] Am Beginn des Prozesses einer Gehirnsimulation steht das Scannen eines existierenden Gehirns. Bei der Simulation eines *menschlichen* Gehirns, wird ein solches gescannt.[53] Da die Simulation auf einem existierenden Gehirn basiert, wird sie auch wesentliche, bei einem perfekten Scan sogar alle psychologischen Merkmale des ursprünglichen Gehirns teilen. Wenn also der Inhaber des ursprünglichen Gehirns die für Personenstatus relevante Willensstruktur aufweist, dann verfügt auch die Simulation über sie. Damit ist der Personenstatus für menschliche Gehirnsimulationen wohl zu bejahen.[54] Diese Klassifizierung hat wichtige ethische Konsequenzen. Beispielsweise sollen mit dem Blue Brain Project psychische Krankheiten wie Depressionen oder Autismus untersucht werden. Es wäre aber ethisch höchst problematisch, eine Gehirnsimulation mit Personenstatus in depressive Zustände zu versetzen, um Depressionen zu studieren. Thomas Metzinger warnt davor, dass bewusste Computerprogramme wie sie im Rahmen des Blue Brain Project entwickelt werden, für Forschungszwecke missbraucht werden könnten und möglicherweise als „Bürger zweiter Klasse" ohne Rechte gelten werden.[55]

Bei anderen Typen künstlicher Intelligenz ist es schwieriger festzustellen, ob sie Wünsche oder Ziele zweiter Stufe haben können. Denkbar wäre dies bei einer künstlichen Intelligenz, die aufgrund ihrer finalen Ziele in Form der Nutzenfunktion von Zeit zu Zeit instrumentelle Ziele als Heuristiken bildet, beispielsweise um Rechenleistung zu sparen. Wir können uns eine künstliche Intelligenz vorstellen, die für das Brettspiel Go programmiert wurde. Sie hat als finales Ziel, das Spiel zu gewinnen. Ab und zu

---

51 Vgl. *Bostrom*, Superintelligence (Fn. 5), S. 30 ff.

52 Für eine ausführliche Darstellung der Ziele und des technischen Hintergrunds des Projekts vgl. "The Human Brain Project – A Report to the European Commission", abrufbar unter: https://www.humanbrainproject.eu/documents/10180/ 17648/ TheHBPReport_LR.pdf/18e5747e-10af-4bec-9806-d03aead57655 (zuletzt abgerufen am 12.09.2016).

53 Einen Überblick über die wichtigsten technischen Aspekte eines Gehirnscans findet man in *Bostrom*, Superintelligence (Fn. 5), S. 30 ff.

54 Man könnte einer Gehirnsimulation allenfalls den Personenstatus verweigern, indem man dafür argumentiert, dass blosse Simulationen kein Bewusstsein haben, und Bewusstsein eine notwendige Bedingung für die relevante Art von Wünschen und Überzeugungen sei. Wie im ersten Teil erwähnt, gibt es einige Experten, die Gehirnsimulationen ein Bewusstsein absprechen.

55 *Metzinger*, Ego Tunnel (Fn. 32).

bildet sie instrumentelle Ziele wie „dominiere die untere Hälfte des Spielbrettes“, die dann so lange verfolgt werden, bis neue Ziele gesetzt werden. Die instrumentellen Ziele werden also manchmal aufgrund der finalen Ziele revidiert.[56] Es scheint durchaus angemessen, dies als Fall zu beschreiben, bei dem die künstliche Intelligenz ihre instrumentellen Ziele erster Stufe aufgrund ihrer Wünsche zweiter Stufe angepasst hat.

Für Juristen als technische Laien ist es wohl oft schwierig zu bestimmen, ob eine künstliche Intelligenz die nötige Willensstruktur aufweist. Deshalb ist es sinnvoll, dass diese Frage von Fall zu Fall von Experten beantwortet wird. Existierende künstliche Intelligenzen mögen häufig noch nicht über Wünsche zweiter Stufe verfügen, aber es scheint keine Gründe zu geben, warum zukünftige künstliche Intelligenzen nicht die relevante Willensstruktur besitzen könnten.

Als zweites Personenkonzept wählen wir John Lockes, das zu den historisch einflussreichsten Analysen des Begriffs gehört. Er beschreibt sein Verständnis des Konzepts an zwei Stellen in seinem *Essay Concerning Human Understanding*:

> „[W]e must consider what person stands for; which, I think, is a thinking intelligent being that has reasons and reflection, and can consider itself as itself, the same thinking thing, in different times and places“.[57]

> „[Person] is a forensic term [...] and so it belongs to intelligent agents capable of a law, and happiness, and misery. This personality extends itself beyond present existence to what is past, only by consciousness, whereby it becomes concerned and accountable.“[58]

Die beiden Analysen unterscheiden sich geringfügig: Nur die zweite erfordert, dass eine Person Glück und Leid empfinden kann und fähig ist, Recht zu verstehen und zu befolgen. Im ersten Teil des Aufsatzes haben wir dafür argumentiert, dass künstliche Intelligenzen Überzeugungen haben, Gründe abwägen und Entscheide treffen können. Lockes Intelligenz-

---

56 *C. Metz*, The Sadness and Beauty of Watching Google’s AI Play Go, Wired vom 03.11.2016, abrufbar unter: http://www.wired.com/2016/03/sadness-beauty-watching-googles-ai-play-go (zuletzt abgerufen am 12.09.2016), beschreibt einen Spielzug der Google KI AlphaGo, der als Wechsel der instrumentellen Ziele verstanden werden kann: „AlphaGo’s move didn’t seem to connect with what had come before. In essence, the machine was abandoning a group of stones on the lower half of the board to make a play in a different area.“

57 *J. Locke*, Essay Concerning Human Understanding, XXVII, 11.

58 *Locke*, Essay (Fn. 57), XXVII, 28.

kriterium scheint damit auf den ersten Blick zwar erfüllt, aber die meisten existierenden künstlichen Intelligenzen sind nur *domänenspezifisch* intelligent. Sie können nur in einem sehr eng definierten Problembereich Gründe abwägen, Überzeugungen bilden und Entscheidungen treffen. Es scheint durchaus plausibel, dass Locke in seinen Textstellen *generelle* Intelligenz fordert, also die Fähigkeit, in vielen oder sogar allen Problembereichen durch Überlegen seine Ziele verfolgen zu können. Die Suche nach *genereller* künstlicher Intelligenz gehört zu den wichtigsten Zielen der KI-Forschung und hat insbesondere in Form von rekurrenten neuronalen Netzen, die viel flexibler Daten verarbeiten können als bisherige neuronale Netze,[59] beeindruckende Fortschritte gemacht. So hat beispielsweise ein einziger DeepMind Algorithmus selbständig gelernt, 49 verschiedene Atari-Spiele rein aufgrund von Pixeldaten auf menschenähnlichem Niveau zu spielen.[60] Dieser Algorithmus ist aber immer noch weit davon entfernt, die Generalität menschlicher Intelligenz zu erreichen. Es scheint daher, dass existierende künstliche Intelligenzen wohl eher noch nicht die Generalität erreicht haben, um Lockes Intelligenzkriterium zu erfüllen, dies aber zukünftig durchaus möglich sein könnte.

Locke fordert zusätzlich, dass Personen einen Bezug zu sich selbst darstellen können, und sich selbst als zeitüberdauernde Objekte verstehen. Mit anderen Worten: Personen haben Überzeugungen über sich selbst, über ihre Vergangenheit und Zukunft. Für viele praktische Anwendungsbereiche künstlicher Intelligenzen sind Überzeugungen dieser Art nicht nötig und wurden deshalb vom Entwickler nicht programmiert. Aber bei gewissen künstlichen Intelligenzen in Computerspielen (sogenannten „Bots“) ist es nötig, dass die künstliche Intelligenz die von ihr kontrollierte Spielfigur, ihr „ich“, von den Spielfiguren anderer Spieler unterscheiden kann. Das kann als sehr rudimentäre Version eines Selbstverständnisses verstanden werden. Bots in existierenden Spielen müssen keine Erinnerungen an vergangene Ereignisse speichern, es scheint aber nicht unwahrscheinlich, dass Spieleentwickler früher oder später Bots entwickeln werden, die von vergangen Erlebnissen lernen und so eine grössere Herausforderung für den Spieler darstellen. So hat ein Team der Universität Texas eine künstliche Intelligenz für das Spiel „Unreal Tournament“ entwickelt,

---

59 *A. Karpathy*, The Unreasonable Effectiveness of Recurrent Neural Networks, 2015, abrufbar unter: http://karpathy.github.io/2015/05/21/rnn-effectiveness (zuletzt abgerufen am 12.09.2016).

60 *Mnih et al.,* Playing (Fn. 9).

die teilweise auf neuronalen Netzwerken basiert und anhand ihrer Erfahrung lernt und sich verbessert.[61]

Können künstliche Intelligenzen fähig sein, Recht und Gesetze zu verstehen und zu befolgen? Dieses Kriterium scheint zunächst sehr anspruchsvoll. Wir denken, dass dieser Schein trügt. Als Beispiel soll erneut AlphaGo dienen. Der Algorithmus kennt die Spielregeln von Go, er wählt seinen nächsten Zug aus der Menge aller von den Spielregeln erlaubten Züge. Die Spielregeln sind eine Form von *Gesetz* oder *Recht*: Nur eine Teilmenge aller logisch möglichen Züge sind auch regelkonform. AlphaGos Gesetzesverständnis ist zwar weit von dem entfernt, was Locke fordert – unter anderem bezieht es sich auf das falsche Gesetz, nämlich auf die Spielregeln von Go. Es gibt aber einen ersten Hinweis darauf, wie das erforderliche Gesetzesverständnis in einer künstlichen Intelligenz aussehen könnte. Beispielsweise könnte AlphaGo mit einem CheatingNetwork als sechste Komponente[62] versehen werden. Dieses bestimmt für jeden Zug, ob jetzt eine gute Gelegenheit wäre, einen nicht regelkonformen Zug zu vollziehen, um die eigene Situation zu verbessern. Das CheatingNetwork könnte anhand von Interaktionen mit menschlichen Opponenten trainiert werden und herausfinden, bei welchen Spielkonstellationen sie am wenigsten aufmerksam sind und deshalb Regelverstösse mit einer kleineren Wahrscheinlichkeit bemerken. Damit hätte AlphaGo die Wahl, gegen die Regeln zu verstossen. AlphaGo könnte auch bemerken, dass Regelverstösse des Gegners wahrscheinlicher werden, wenn der Gegner einen Regelverstoss von AlphaGo bemerkt. Oder es könnten Strafen für Verstösse eingeführt werden, die AlphaGo in den Erwartungswert der Züge einbeziehen muss. So könnte Schritt für Schritt die Komplexität des Akteurs erweitert werden, bis Lockes Erfordernis erfüllt ist.

Ist es möglich, dass künstliche Intelligenzen – gegenwärtige oder zukünftige – Freude und Leid erleben können? Zum Teil wird vertreten, dass schon existierende „reinforcement-learning" Algorithmen, wie sie beispielsweise bei Google DeepMind zum Einsatz kommen, Freude und Leid

61 Artificially Intelligent Game Bots Pass the Turing Test on Turing's Centenary, The University of Texas at Austin News vom 26.09.2016, abrufbar unter: http://news.utexas.edu/2012/09/26 (zuletzt abgerufen am 12.09.2016).

62 Eine hilfreiche Übersicht über die AlphaGo Architektur findet man auf Patrick Mineaults Blog: https://xcorr.net/2016/02/03/5-easy-pieces-how-deepmind-mastered-go (zuletzt abgerufen am 12.09.2016).

empfinden können.[63] Der Philosoph Thomas Metzinger schlägt anspruchsvollere Kriterien vor, die von existierenden Algorithmen wohl noch nicht erfüllt werden.[64] Metzinger warnt aber davor, dass zukünftigen künstlichen Intelligenzen im Rahmen von Experimenten schweres Leid zugefügt werden könnte.[65] Es scheint also denkbar, dass zukünftige künstliche Intelligenzen alle von Locke geforderten notwendigen und hinreichenden Bedingungen erfüllen könnten, und dementsprechend Personen sein werden.

Zuletzt soll der Vorschlag von Daniel Dennetts untersucht werden, einem Philosophen der sich schon häufiger mit künstlicher Intelligenz auseinandergesetzt hat. Sein Vorschlag vereint Traditionen verschiedenster Philosophen und kann deshalb als „ökumenischer Vorschlag" bezeichnet werden. Er schlägt folgende notwendige (und möglicherweise gemeinsam hinreichende) Bedingungen vor:[66]

(1) Personen sind rationale Wesen.
(2) Sie haben Bewusstsein, Wünsche und Überzeugungen.
(3) Ob ein Objekt eine Person ist, hängt davon ab, welche Haltung wir ihm gegenüber einnehmen.
(4) Das Objekt muss unsere Haltung ihm gegenüber auch uns gegenüber einnehmen können.
(5) Eine Person muss fähig sein, verbal zu kommunizieren.
(6) Eine Person muss sich ihrer selbst bewusst sein.

Das erste Kriterium spielt eine wichtige Rolle in den Theorien von Rawls, Kant, Aristoteles und – wie wir gesehen haben – Locke. Es gibt wiederum verschiedene Verständnisse von Rationalität, aber ein Minimalstandard, der wohl alle gängigen Definitionen erfüllt, ist, dass ein rationaler Agent auf Gründe angemessen reagieren muss. Wir haben schon gesehen, dass künstliche Intelligenzen Überzeugungen haben können, die in gewissen Kontexten Gründe für eine Handlung darstellen, und dass sie auf diese Gründe im Rahmen ihrer Entscheidungstheorie reagieren können. Je nach

63 *M. Daswani/J. Leike*, A Definition of Happiness for Reinforcement Learning Agents, 2015, abrufbar unter: http://arxiv.org/abs/1505.04497 (zuletzt abgerufen am 12.09.2016).
64 *T. Metzinger*, Two Principles for Robot Ethics, in: E. Hilgendorf/J.-P. Günther (Hrsg.), Robotik und Gesetzgebung, Baden-Baden 2012, S. 263 ff.
65 *Metzinger*, Ego Tunnel (Fn. 32).
66 *Dennett*, Personhood (Fn. 49).

der Menge an Gründen, auf die eine künstliche Intelligenz anspricht, und der Qualität ihrer Entscheidungsfindungsroutine kann sie mehr oder weniger rational sein. Angesichts der teilweise ausgeprägten Irrationalität von Menschen[67] – zweifellos Personen – sollten die Anforderungen hier aber nicht allzu hoch gestellt werden. Wie wir bei Locke gesehen haben, könnte auch ein bestimmter Grad an Generalität der Rationalität gefordert werden – schon existierende domänenspezifische künstliche Intelligenzen sind in engen Bereichen perfekt rational, sie sind aber in anderen Problembereichen hilflos. Auch hier ist deshalb denkbar, dass Rationalität aufgrund der domänenspezifischen Natur existierender Algorithmen noch nicht im geforderten Sinne erfüllt ist, wohl aber durch zukünftige Algorithmen erfüllbar sein wird.

Das zweite Kriterium scheint etwas problematischer, ist aber vermutlich ebenfalls erfüllbar. Künstliche Intelligenzen haben Wünsche in Form einer Nutzenfunktion und wie wir dargelegt haben, können sie auch Überzeugungen haben. Am schwierigsten ist die Frage, ob sie über Bewusstsein verfügen können. Es gibt einen soliden Expertenkonsens darüber, dass künstliche Intelligenzen grundsätzlich Bewusstsein haben können.[68] Umstrittener ist, welche Formen und Architekturen künstlicher Intelligenz Bewusstsein haben können. Künstliche Intelligenzen können also Bewusstsein haben, aber es kann im Einzelfall schwierig sein zu beantworten, ob eine spezifische künstliche Intelligenz Bewusstsein hat.[69] Wir halten es für plausibel, dass jede künstliche Intelligenz Bewusstsein hat, deren Verhalten mit einem genügend komplexen Netz aus Wünschen und Überzeugungen prognostiziert und erklärt werden kann.

Dennett geht davon aus, dass das dritte Kriterium die Grundlage für die ersten beiden Kriterien ist. Im terminologischen Teil haben wir stipuliert, dass der Begriff „Agent" auf Objekte anwendbar ist, wenn wir ihnen Wünsche und Überzeugungen zuschreiben können, mit denen sich ihr Verhalten erklären und prognostizieren lässt. Wir haben uns dabei an Daniel Dennetts „intentional stance" orientiert. Dennett geht davon aus, dass die „intentional stance" die im dritten Kriterium geforderte Haltung ist. Und

67 *T. Gilovich/D. Griffin/D. Kahneman*, Heuristics and biases: The psychology of intuitive judgment, Cambridge 2002.

68 *Moravec*, Mind Children (Fn. 32); *Chalmers*, Consciousness (Fn. 10); *Metzinger*, Ego Tunnel (Fn. 32).

69 Vgl. dazu auch *L.B. Solum*, Legal Personhood for Artificial Intelligences, North Carolina Law Review 70 (1992), S. 1231 ff.

er geht auch davon aus, dass sogar relativ einfache Computersysteme, nämlich Schachcomputer, diese Bedingung erfüllen: „[W]e can even use the procedure to predict the behavior of some machines. For instance, it is a good, indeed the only good strategy to adopt against a good chess-playing computer. By *assuming* the computer has certain beliefs (or information) and desires (or preference functions) dealing with the chess game in progress, I can calculate – under auspicious circumstances – the computer's most likely next move, *provided I assume the computer deals rationally with these beliefs and desires*."[70] Schon existierende künstliche Intelligenzen erfüllen also das dritte Kriterium.

Das vierte Kriterium setzt die Anforderungen höher. Es fordert, dass ein Objekt, dem gegenüber wir die „intentional stance" einnehmen, diese auch uns gegenüber einnehmen kann. Dennett nennt solche Systeme „second-order intentional systems": „Let us define a *second-order intentional system* as one to which we ascribe not only simple beliefs, desires and other intentions, but beliefs, desires, and other intentions *about* beliefs, desires, and other intentions."[71] Es scheint unwahrscheinlich, dass existierende künstliche Intelligenzen dieses Kriterium schon erfüllen – selbst im Tierreich schreibt Dennett diese Fähigkeit nur höheren Säugetieren zu. Es gibt aber Bemühungen, künstliche Intelligenzen zu nutzen, um Emotionen bei Menschen zu erkennen.[72] Das wäre ein erster Schritt, um einem System psychologische Prädikate zuzuordnen. Es ist deshalb wahrscheinlich, dass zukünftig auch weitere solche Prädikate, darunter solche die sich auf Wünsche und Überzeugungen beziehen, von künstlichen Intelligenzen zugeschrieben und zur Verhaltenserklärung und Prognose verwendet werden.

Das fünfte Kriterium erfordert verbale Kommunikation. Die Forschung hat im sog. „natural language processing", dem Verarbeiten natürlicher Sprachen durch künstliche Intelligenzen, beeindruckende Fortschritte gemacht. Beispielsweise können schon öffentlich zugängliche neuronale Netze, die auf gewöhnlichen Heimcomputer laufen, beeindruckende Textbausteine selbständig generieren. Andrej Karpathy hat beispielsweise ein auf git-hub verfügbares neuronales Netz mit dem Shakespeare Corpus trai-

70 *Dennett*, Personhood (Fn. 49), S. 179.
71 *Dennett*, Personhood (Fn. 49), S. 181.
72 *K. Dai/J. Leike/J. MacAuslan*, Recognizing Emotion in Speech Using Neural Networks, Telehealth and Assistive Technologies (2008), S. 31 ff.

niert und selbständig folgende Textbausteine im Shakespeare-Stil generieren lassen:[73]

> Second Lord:
> They would be ruled after this chamber, and
> my fair nues begun out of the fact, to be conveyed,
> Whose noble souls I'll have the heart of the wars.
>
> Clown:
> Come, sir, I will make did behold your worship.
>
> VIOLA:
> I'll drink it.

Zwar sind die Sätze teilweise unsinnig, sie zeigen aber auf beeindruckende Weise, wie selbst gewöhnliche neuronale Netze auf normaler Hardware wichtige Elemente der Syntax und des Stils einer Textauswahl lernen können. Intuitiv näher an semantischem Verständnis ist Googles Algorithmus zur Generierung von Textbeschreibungen für Bilder. So hat der Algorithmus beispielsweise richtigerweise ein Foto, auf dem einige Pizzascheiben auf einem Herd zu sehen sind, beschrieben als: „Two pizzas sitting on top of a stove top oven."[74]

Diese Beispiele erreichen noch nicht die von Dennett gesetzte Messlatte. Unter Rückgriff auf Paul Grice[75] verlangt Dennett, dass der Akteur Sprache auch zum Zweck der Kommunikation verwendet und damit eine bestimmte Reaktion hervorrufen will. Der Google Algorithmus verwendet Sprache wohl kaum zu diesem Zweck. Er könnte aber eine Komponente eines Systems sein, das die generierten Sätze zu diesem Zweck verwendet. Deshalb scheint es, dass wir echter verbaler Kommunikation künstlicher Intelligenzen schon erstaunlich nahegekommen sind.

Für das sechste Kriterium können wir auf die Ausführungen zu Lockes Konzeption verweisen. Abschliessend lässt sich sagen, dass existierende künstliche Intelligenzen wohl unter keiner Konzeption Personen sind. Unsere Untersuchung der Kriterien hat aber gezeigt, dass der Weg zur Erfül-

---

73 http://karpathy.github.io/2015/05/21/rnn-effectiveness/ (zuletzt abgerufen am 12.09.2016).

74 https://research.googleblog.com/2014/11/a-picture-is-worth-thousand-coherent.html (zuletzt abgerufen am 21.11.2016).

75 *H.P. Grice*, Meaning, The Philosophical Review 66, 3 (1957), S. 377 ff.; *H.P. Grice*, Utterer's meaning and intention, The Philosophical Review 78, 2 (1969), S. 147 ff.

lung dieser Kriterien nicht mehr weit ist: anders als es noch vor einem Jahrzehnt schien und teilweise auch heute noch behauptet wird. Ebenfalls bemerkenswert ist, dass viele der Kriterien der verschiedenen Personenbegriffe durch existierende Systeme schon „teilweise" erfüllt werden – oft wurden schon einige nötige Kernelemente programmiert und diese müssen nur noch mit teilweise existierenden, teilweise noch zu entwickelnden Komponenten kombiniert werden, um die Kriterien vollständig zu erfüllen.

II. Künstliche Intelligenzen als Rechtspersönlichkeiten

Nur Rechtspersonen sind Inhaber von Rechten und Pflichten.[76] Im Strafrecht zeigt sich diese Maxime unter anderem darin, dass Art. 9 StGB den persönlichen Geltungsbereich des StGB implizit auf Rechtspersonen beschränkt, indem alle Ausnahmen vom persönlichen Geltungsbereich stets auf „Personen" bezogen sind. Auch werden Föten, Menschen die noch keine Rechtspersonen sind,[77] nicht wie Rechtspersonen durch den Katalog der Tötungsdelikte geschützt, sondern durch die im Rahmen der Fristenregelung ergänzten Artikel zum Schwangerschaftsabbruch. Für uns besonders wichtig ist, dass nur Rechtspersonen strafrechtlich verantwortlich sein können.[78]

Anders als der Begriff der Person ist der Begriff der Rechtsperson ein rechtsdogmatischer Begriff, dessen Bedeutung und Extension von der Rechtswissenschaft bestimmt werden muss.[79] Damit kann er grundsätzlich auch neuen Umständen, wie beispielsweise technischen Fortschritten in der KI-Forschung, und neuen Kenntnissen, z.B. über die kognitiven Fähigkeiten höherer Säugetiere, Rechnung tragen und in seinem Anwendungsbereich erweitert werden. Können künstliche Intelligenzen über Rechts-

76 Vgl. Art. 11 ZGB.

77 *A. Büchler/M. Frei*, Der Lebensbeginn aus juristischer Sicht – unter besonderer Berücksichtigung der Problematik des Schwangerschaftsabbruchs, Jusletter 29.08.2011.

78 *M. Hildebrandt*, Criminal Liability and 'Smart' Environments, in: R.A. Duff/S. Green (Hrsg.), Philosophical Foundations of Criminal Law, Oxford 2011, S. 507 ff. (S. 511).

79 *D. Reuter*, Rechtsfähigkeit und Rechtspersönlichkeit: Rechtstheoretische und rechtspraktische Anmerkungen zu einem grossen Thema, Archiv für die civilistische Praxis 207 (2007), S. 673 ff. (S. 674).

persönlichkeit verfügen? Erforderlich ist dafür, dass Kandidaten für Rechtspersönlichkeit die notwendigen und gemeinsam hinreichenden Bedingungen des Konzepts erfüllen.[80] In seiner Untersuchung zum Begriff der Rechtspersönlichkeit identifiziert Dieter Reuter folgende Anforderungen:[81]

- Die Fähigkeit zum Eigeninteresse
- Das Rechtssubjekt muss über eine eigene Identität verfügen, um am Rechtsverkehr teilnehmen zu können

Zum ersten Kriterium schreibt Reuter: „Das Haben von Rechten und Pflichten macht nur für denjenigen Sinn, der ein Eigeninteresse hat und notfalls mit Hilfe anderer verfolgen kann."[82] Mit Eigeninteresse können nicht „egoistische Interessen" gemeint sein, denn zweifellos reichen altruistische Interessen, um diese Bedingung zu erfüllen. Sonst würden auch viele Menschen dieses Kriterium nicht erfüllen. Vielmehr zählt *jede Art* von subjektiven Interessen des Akteurs. Damit reichen auch die in Form der Nutzenfunktion verankerten Wünsche und Ziele künstlicher Intelligenzen.

Gemäss dem zweiten Kriterium muss eine Rechtsperson über eine eigene Identität verfügen; wer am Rechtsverkehr teilnimmt und Inhaber von Rechten und Pflichten werden soll, muss eindeutig feststehen. Die dogmatische Debatte dazu bezieht sich praktisch ausschliesslich auf die Frage, welche Anforderungen dabei an Gesellschaften als juristische Personen zu stellen sind.[83] Bei natürlichen Personen wird diese Frage selbstverständlich positiv beantwortet, und andere Kandidaten für Rechtspersönlichkeiten werden in der zeitgenössischen Rechtswissenschaft nur selten disku-

80 *Hildebrandt*, Criminal Liability (Fn. 78), S. 511, schreibt dazu: „Criminal liability of a smart device does not imply that smart entities are equivalent to natural persons, but rather suggests that good reasons can be given to assign a measure of responsibility to non-human persons." Sie betont damit, dass einer künstlichen Intelligenz Rechtspersönlichkeit zuschreiben, nicht heisst, sie als äquivalent zu einer natürlichen Person zu sehen. Es heisst bloss, dass beide – trotz ihrer Verschiedenheit – die Anwendungsbedingungen des Konzepts erfüllen und deshalb in dessen Extension fallen.

81 *Reuter*, Rechtsfähigkeit (Fn. 79), S. 680 ff.

82 *Reuter*, Rechtsfähigkeit (Fn. 79), S. 680.

83 Vgl. *Reuter*, Rechtsfähigkeit (Fn. 79), S. 681 ff. mit weiteren Verweisen.

tiert.[84] Mit der Frage, ob künstliche Intelligenzen eine genügend klar ausgeprägte Identität haben, betreten wir also Neuland.

Zunächst ist festzustellen, dass Algorithmen – wie auch Menschen – physikalisch realisiert und damit auch lokalisiert sind. Sie befinden sich an einem oder mehreren *Orten* und existieren auf einem physikalischen Computersystem, so wie auch der menschliche Geist auf einem physikalischen (bzw. biologischen) System realisiert ist.[85] Zudem können sie – wie Menschen oder Gesellschaften – einen Namen haben: Der Go-spielende Google-Algorithmus heisst „AlphaGo". Was die Frage im Vergleich zu Menschen und sogar Gesellschaften schwieriger macht, ist, dass künstliche Intelligenzen als Algorithmen einfach *kopiert* und damit *vervielfältigt* werden können. Google könnte sehr einfach zahlreiche Instanzen von AlphaGo kreieren. Der Ökonom Robin Hanson geht davon aus, dass dereinst zahllose Kopien von Gehirnsimulationen existieren werden, die unsere Arbeit verrichten und uns als Arbeitskräfte überflüssig machen werden.[86] Die einfache Kopierbarkeit von Algorithmen macht es tatsächlich schwierig, sie im Rechtsverkehr zu identifizieren. Es kann praktisch schwierig sein zu bestimmen, welche der zahlreichen Kopien eines auf Servern in einem Rechenzentrum laufenden Algorithmus eine Dienstleistung erbracht hat. Ist dieser Umstand ein unüberwindbares Hindernis für das Identitätskriterium?

Statt dafür zu argumentieren, dass die Möglichkeit von Kopien die Identifikation im Rechtsverkehr nicht erheblich erschwert, möchten wir zeigen, dass das Problem der Kopierbarkeit auch für biologische Organismen existiert – wenn auch erst theoretisch. In der philosophischen Literatur werden im Rahmen der Debatte um die Personale Identität viele Ge-

---

84 Eine wichtige Ausnahme stellen Schimpansen dar, denen im April 2015 implizit von einer Richterin am New York Supreme Court Rechtspersönlichkeit zuerkannt wurde, und die Grundrechte auf Unversehrtheit und Bewegungsfreiheit besitzen können, *Fasel/Mannino/Baumann/Blattner*, Grundrechte (Fn. 4), S. 6 f.

85 Wir vertreten damit eine in der Philosophie des Geistes „Physikalismus" genannte Position, wonach das Bewusstsein physikalisch ist. Die von uns vertretene Position ist aber relativ breit und umfasst auch *Russellian Monism*, eine Position die Bewusstsein als intrinsische Natur physikalischer Eigenschaften qualifiziert. Ausgeschlossen wird also bloss Eigenschafts- und Substanzdualismus. Für die vermutlich beste Übersicht über alle Positionen in der Philosophie des Geistes siehe *D. Chalmers*, The Character of Consciousness, Oxford 2010, S. 111 ff.

86 *R. Hanson*, If Uploads Come First, Extropy 6, 2 (1994), S. 10 ff., abrufbar unter: http://hanson.gmu.edu/uploads.html (zuletzt abgerufen am 12.09.2016).

dankenexperimente diskutiert, in denen Menschen „kopiert" werden und deren Identität als Folge davon nur schwierig feststellbar ist. Zu den prominentesten Szenarien gehören Gehirntransplantation und Mind-Uploading: Einer Person wird eine der beiden Hirnhälften entfernt. Sie wird dann in den leeren Schädel einer kurz zuvor verstorbenen Person transplantiert. Nach der Operation existieren zwei unabhängige, lebensfähige Personen, die je eine Hirnhälfte der ursprünglichen Person besitzen.[87] Das ist der eindeutigste Fall einer menschlichen Kopie. Mind-Uploading hingegen ist ein populärer Begriff für das Übertragen aller Eigenschaften eines menschlichen Bewusstseins (Erinnerungen, Persönlichkeit und andere Eigenschaften) auf einen Computer.[88] Eine bestimmte Form des Mind-Uploadings ist die Gehirnsimulation, die auf einem gescannten, menschlichen Gehirn basiert. Bei diesem Prozess wird ebenfalls eine Art digitale Kopie eines Menschen kreiert.

In beiden Szenarien wird ein Mensch, bzw. der Teil des Menschen, der für seine Persönlichkeit verantwortlich ist, kopiert. Im zweiten Fall könnte dieser Prozess beliebig wiederholt werden, um zahlreiche Kopien zu erstellen. Trotz dieser beiden (vorerst noch theoretischen, aber möglicherweise bald realen)[89] Möglichkeiten, Menschen zu kopieren, stellen wir ihren Status als Rechtspersonen nicht in Frage. Zwar stellen uns solche Fälle vor schwierige epistemische und juristische Probleme, aber diese sind überwindbar[90] und sollten nicht dazu führen, dass wir dieses Kriterium bei Menschen als nicht erfüllt betrachten. Dementsprechend sollten wir auch künstlichen Intelligenzen den Status als Rechtspersonen aufgrund ihrer einfachen Kopierbarkeit nicht voreilig verweigern.

Reuters Kriterien können als notwendige, nicht aber als hinreichende Kriterien verstanden werden, denn die meisten Säugetiere erfüllen diese

87 Es ist tatsächlich möglich, einer Person im Rahmen einer Hemisphärektomie eine ganze Hirnhälfte zu entfernen, ohne dass die Person verstirbt, vgl. dazu *J. Erhardt*, Strafrechtliche Verantwortung und personale Identität, 2014, S. 40 ff., abrufbar unter http://www.zb.unibe.ch/download/eldiss/13erhardt_j.pdf (zuletzt abgerufen am 12.09.2016) m.w.N.

88 http://www.minduploading.org/ (zuletzt abgerufen am 12.09.2016).

89 Der Chirurg Sergio Canavero ging Anfangs 2015 davon aus, dass die erste menschliche Gehirntransplantation innerhalb der nächsten zwei Jahre gelingen werde, vgl. http://www.theguardian.com/society/2015/feb/25/first-full-body-transplant-two-years-away-surgeon-claim (zuletzt abgerufen am 12.09.2016).

90 Für eine eingehende Auseinandersetzung mit den juristischen Problemen der personalen Identität siehe *Erhardt*, Verantwortung (Fn. 87).

ebenfalls. Sie werden deshalb aber nicht als Rechtspersonen qualifiziert. Vielmehr scheint es angebracht, sie als Ergänzungen zum alltagssprachlichen Personenbegriff zu verstehen. So wird in Art. 11 ZGB Rechtsfähigkeit „jedermann" zugeschrieben, was wohl als „jeder Person" in einem alltagssprachlichen Sinn verstanden werden muss.[91] Zwar wurde dieser Begriff durch die juristische Praxis und die Rechtswissenschaft an den Randbereichen des Lebensbeginns und -endes verfeinert, die allgemeinen Anwendungsbedingungen des Konzepts scheinen aber immer noch durch den Personenbegriff des Alltags bestimmt. Wie wir gesehen haben, werden einige Kriterien für Persönlichkeit von existierenden künstlichen Intelligenzen schon teilweise oder ganz erfüllt, und einer vollständigen Erfüllung durch zukünftige künstliche Intelligenzen steht nichts im Wege. Deshalb können wir abschliessend sagen, dass zwar noch keine existierenden künstlichen Intelligenzen als Rechtspersonen gelten,[92] sich dies aber in den nächsten Jahrzehnten ändern dürfte.

III. Folgebetrachtungen

Existierende künstliche Intelligenzen sind noch keine Rechtspersonen und haben auch noch keinen Personenstatus. Wir denken, dass es dafür gute Gründe gibt. Es soll aber darauf hingewiesen werden, dass bei der Ausdehnung der Rechtssphäre auf andere biologische oder nicht-biologische Wesen starke kognitive Verzerrungen, sogenannte *biases*,[93] am Werk sind. Wir sollten uns davor hüten, dem *status quo bias* zu verfallen und die Messlatte für Personenstatus unvernünftig hoch anzusetzen, z.B. indem wir Alltagsbegriffe wie „Freiheit" mit problematischen metaphysischen Annahmen anreichern. Im Zweifel scheint es vernünftig, anderen Wesen eher grosszügig Personenstatus zuzuschreiben. Wenn wir dies fälschlicherweise tun, verlieren wir wenig, wenn wir aber Personen fälschlicherweise als Nicht-Personen qualifizieren, kann dies zu ethisch höchst problematischen Konsequenzen führen.

---

91 Die Deutung „jeder Mensch" können wir ausschliessen, weil Föten keine Rechtspersönlichkeit zukommt, obwohl sie Menschen sind. Sie werden beispielsweise nicht wie Rechtspersonen durch die Tötungsdelikte in Art. 111 ff. StGB geschützt.

92 Zu diesem Resultat kommt auch *Solum*, Personhood (Fn. 69), S. 1231.

93 *D. Kahneman*, Thinking, Fast and Slow, New York 2011.

Existierende und in naher Zukunft entstehende künstliche Intelligenzen sind noch keine Rechtspersonen und können daher auch noch keine strafrechtliche Verantwortung tragen. Wir sollten daraus aber nicht folgern, dass sie keinerlei Schutz verdienen. Die Möglichkeit leidender künstlicher Intelligenzen ist eine erschreckende Perspektive.[94] Wir sollten deshalb die Vorstellung ernst nehmen und für künstliche Intelligenzen ähnliche Schutzmechanismen bedenken, wie sie bei Föten und Tieren schon existieren. Wir wissen nicht, ob existierende künstliche Intelligenzen schon leiden können und daher schutzwürdig sind. Wir zweifeln aber nicht daran, dass jetzt der richtige Zeitpunkt ist, sich diese Frage zu stellen.

Wenn künstliche Intelligenzen grundsätzlich zukünftig den Status der Rechtsperson erreichen können, dann stellt sich eine Reihe juristischer Folgefragen. Forschung im Bereich der strafrechtlichen Verantwortung künstlicher Intelligenzen muss beispielsweise die Frage beantworten, inwiefern das Verhältnis zwischen Programmierer und künstlicher Intelligenz bei deliktischem Verhalten im Rahmen der Regeln der Täterschaft und Teilnahme zu beurteilen ist. Ebenfalls von grossem Interesse ist, wie sich die subjektive Seite des Tatbestandes bei künstlichen Intelligenzen ausgestaltet. Wie ist beispielsweise die Unterscheidung zwischen bewusster Fahrlässigkeit und Eventualvorsatz bei künstlichen Intelligenzen zu bestimmen? Diese und weitere Fragen verdienen eine ausführliche Untersuchung und können nicht im Rahmen dieses Aufsatzes behandelt werden. Einige Hinweise müssen an dieser Stelle genügen.

Für strafrechtliche Verantwortlichkeit ist Rechtspersönlichkeit allein nicht genug. Das Strafrecht stellt spezifische Anforderungen an Rechtssubjekte, wenn es um die Zuschreibung strafrechtlicher Verantwortlichkeit geht, einerseits im Hinblick auf die Handlungsfähigkeit und andererseits im Hinblick auf die Schuldfähigkeit.

Gewiss sind die von heutigen künstlichen Intelligenzen einbezogenen Handlungsfolgen noch stark beschränkt. Sie werden kaum das Wohlbefinden von Akteuren in ihrer Umgebung modellieren und bei ihrer Entscheidung berücksichtigen. Die Komplexität dieser Aufgabe ist überwältigend und vorerst noch nicht in Reichweite künstlicher Intelligenzen. Deshalb ist die Finalität ihrer Handlungen wohl noch zu rudimentär, um die Anforderungen des strafrechtlichen Handlungsbegriffs vollständig zu erfüllen. Sie

94 *R. Hanson*, The Age of Em: Work, Love and Life when Robots Rule the Earth, Oxford 2016.

müssten dazu die sozialen Folgen ihrer Handlungen mindestens in ihren Grundzügen beurteilen können. Die schon erwähnten Versuche, durch neuronale Netze Emotionen zu erkennen, sind aber erste, wichtige Schritte in diese Richtung.[95] Wenn Algorithmen die Emotionen der Akteure in ihrer Umgebung erkennen und in einer Skala, die das Wohlbefinden erfasst, einordnen können, so hätten sie einen wesentlichen Schritt hin zum Erfassen der sozialen Folgen ihrer Handlungen gemacht. Ein genügend ausgeprägtes Verständnis der sozialen Folgen ihres Handelns hätte die künstliche Intelligenz wohl bereits dann, wenn sie das Wohlbefinden ihrer Umgebung *überhaupt* erfasst. Dass sie ihm auch einen intrinsischen Wert beimisst, kann hingegen – wie beim Menschen – nicht gefordert werden.

Am interessantesten dürfte sodann die Frage sein, ob zukünftige künstliche Intelligenzen grundsätzlich überhaupt schuldfähig sein können. Bei Menschen gilt Schuldfähigkeit als Norm, die in Ausnahmefällen aufgehoben werden kann. Bei künstlichen Intelligenzen müsste abgeklärt werden, ob sie überhaupt je schuldfähig sein könnten. Hierzu nur so viel: Damit künstliche Intelligenzen schuldfähig sind, müssen sie zunächst über die Möglichkeit verfügen, zu erkennen, was in einer Situation rechtlich geboten ist. Ähnlich wie ein Schachcomputer die Schachregeln kennt, sollte eine künstliche Intelligenz mindestens wie ein juristischer Laie die gebotenen Handlungen erkennen können. In der Form selbstfahrender Autos haben wir heute schon künstliche Intelligenzen, die rechtliche Regeln, nämlich die Verkehrsregeln, erkennen und befolgen können. Die Kenntnis des rechtlich Gebotenen mag sich bei künstlichen Intelligenzen zwar subjektiv anders anfühlen als bei Menschen – ein Algorithmus wird bei der Verletzung der Rechtsordnung wohl nicht evolutionär entstandene Emotionen wie ein „schlechtes Gewissen" fühlen.[96] Trotzdem dürfte es angemessen sein, Kenntnis des rechtlich Gebotenen als Einsicht zu bezeichnen und als Grundlage für die Schuldfähigkeit zu akzeptieren. Denn sie erfüllt eine ähnliche Rolle, wie die sich in einem schlechten Gewissen manifestierende Kenntnis des Gebotenen beim Menschen: Sie bietet einen epistemischen Zugang zu normativen Tatsachen. Ob dieser Zugang in Form von evolutionär entstandenen Emotionen bei Menschen oder in Form eines einfachen Zugriffs auf eine Normdatenbank bei künstlichen Intelligenzen geschieht, dürfte wohl keine Rolle spielen. Beide Formen sind notwendige

95 *Dai/Leike/MacAuslan*, Emotion (Fn. 72).

96 *Seelmann*, Personalität (Fn. 3), S. 575; *G. Stratenwerth*, Schweizerisches Strafrecht – Allgemeiner Teil I: Die Straftat, 4. Aufl., Bern 2011, S. 305.

Bedingungen dafür, dass ein Akteur die Möglichkeit hat, das rechtlich Gebotene zu tun.

Der zweite Teil der Schuldfähigkeit ist die Möglichkeit, das Gebotene zu tun und das Verbotene zu unterlassen. Diese Komponente wird oft als Freiheit, nach der Einsicht des Gebotenen zu handeln, bezeichnet. Wie wir im terminologischen Teil festgestellt haben, können künstliche Intelligenzen in einem kompatibilistischen Sinne frei sein. Zwar sind sie – wie vermutlich auch Menschen – kausal durch die Naturgesetze und die Anfangsbedingungen des Universums determiniert, aber sie sprechen rational auf Gründe an und können gemäss ihrem Entscheidalgorithmus urteilen und Handlungsentscheide treffen. In diesem Sinne sind künstliche Intelligenzen im relevanten Sinne frei, das rechtlich Gebotene zu tun. Bemerkenswert an dieser kompatibilistischen Argumentation bezüglich Willensfreiheit ist, dass sie auch den Menschen vor der Negierung seiner Freiheit schützt, die sich auf physikalischen, biologischen oder neurologischen Determinismus stützt. Wenn etwa gesagt wird, dass die neuen Kenntnisse der Hirnforschung dazu führen sollten, dass Handelnde nicht für ihr Handeln strafrechtlich verantwortlich gemacht werden dürfen,[97] geht das aus der Sicht eines Kompatibilisten an der Sache vorbei. Willensfreiheit ist nicht eine Frage der Freiheit von (neuronaler) Determination, sondern eine Frage der Fähigkeit, angemessen auf Gründe anzusprechen und eigene Entscheidungen treffen zu können. Es gibt also aus kompatibilistischer Sicht keine Gründe, Menschen und künstlichen Intelligenzen die zweite Komponente der Schuldfähigkeit aufgrund des Determinismus abzusprechen.

Abschliessend können wir feststellen, dass die Einsichtsfähigkeit vorerst noch ein Hindernis für die Schuldfähigkeit künstlicher Intelligenzen darstellt. Existierende Algorithmen verfügen über keine genügend breite Verbotskenntnis, um rechtlich gebotene Handlungen identifizieren zu können. Dies dürfte sich aber ändern, sobald künstliche Intelligenzen vermehrt mit Menschen und ihrer Sozialsphäre interagieren. Dann müssen die Entwickler sicherstellen, dass sie die entsprechenden rechtlichen Normen, mindestens in denselben groben Zügen erkennen (und befolgen) können, wie dies von Menschen gefordert wird. Die zweite Komponente der Schuldfähigkeit, die Freiheit das Gebotene zu tun, bereitet keine besonde-

97 *G. Roth/M. Lück/D. Strüber*, „Freier Wille“ und Schuld von Gewaltstraftätern aus Sicht der Hirnforschung und Neuropsychologie, Neue Kriminalpolitik 18, 2 (2006), S. 55 ff.

ren Schwierigkeiten, sofern man eine kompatibilistische Konzeption der Willensfreiheit vertritt.

C. Konklusion

Wenn es um die schwierigen juristischen Probleme der künstlichen Intelligenz geht, zeigen sich viele Rechtswissenschaftler pessimistisch, was deren Auflösung mit den existierenden Mitteln des Rechts angeht. So schreiben etwa Gless und Weigend: „Das Vordringen mehr oder weniger segensreicher intelligenter Agenten in unser Leben ist eines jener Phänomene, das die Rechtsordnung und speziell das Strafrecht vor praktische, aber auch vor grundsätzliche Fragen stellt, die sich mit den hergebrachten Lehren nicht wirklich beantworten lassen."[98] Auch Andreas Matthias geht davon aus, dass die bisherigen Werkzeuge der Verantwortungszuschreibung nicht ausreichen, um Situationen, in denen fortgeschrittene künstliche Intelligenzen eine Rolle spielen, befriedigend rechtlich zu beurteilen: „[S]ociety must decide between not using this kind of machine any more (which is not a realistic option), or facing a responsibility gap, which cannot be bridged by traditional concepts of responsibility ascription."[99] In seinem wichtigen Aufsatz über die Rechtspersönlichkeit künstlicher Intelligenzen zieht Lawrence Solum ein ähnliches Fazit: „Our theories of personhood cannot provide an a priori chart for the deep waters at the borderline of status. An answer to the question whether artificial intelligences should be granted some form of legal personhood cannot be answered until our form of life gives the question urgency."[100]

Wir sind weniger pessimistisch. Im vorliegenden Aufsatz haben wir versucht, einige schwierige Fragen der künstlichen Intelligenz zu beantworten. Wir haben versucht zu zeigen, dass Schwierigkeiten bei der Anwendung der relevanten philosophischen und juristischen Konzepte oft nicht primär durch deren Vagheit oder Unklarheit entstehen, sondern häufig durch metaphysisch fragwürdige Interpretationen alltäglicher Begriffe wie „Überzeugung", „Freiheit", „Rationalität" oder „Selbstbewusstsein". Wählt man eher deflationäre und metaphysisch weniger anspruchsvolle

98 *Gless/Weigend*, Intelligente Agenten (Fn. 46), S. 588.
99 *A. Matthias*, The Responsibility Gap. Ascribing Responsibility for the Actions of Learning Automata, Ethics and Information Technology 6 (2004), S. 175 ff.
100 *Solum*, Personhood (Fn. 69), S. 1287.

Interpretationen dieser Begriffe, so wird ersichtlich, wie künstliche Intelligenzen unter die entsprechenden Prädikate fallen können. Damit ist der Bogen zu den wichtigen juristischen Konzepten geschlagen. Dann kann man damit beginnen, zu untersuchen, inwiefern zukünftige künstliche Intelligenzen Personen oder Rechtspersonen sein können, Handlungen im strafrechtlichen Sinne begehen und schuldhaft handeln können. Wir sind zum Urteil gelangt, dass existierende künstliche Intelligenzen noch keine Personen bzw. Rechtspersonen sind und deshalb auch noch nicht im Rahmen des Strafrechts verantwortlich gemacht werden können. Mit einer Reihe von Gedankenexperimenten und hypothetischen technischen Erweiterungen, die von existierenden künstlichen Intelligenzen ausgehen, haben wir aber zu zeigen versucht, wie sich künstliche Intelligenzen dank des technischen Fortschritts zunehmend der Grenze der Rechtspersönlichkeit nähern könnten. Schon heute erfüllen Algorithmen einige der relevanten Kriterien und es ist absehbar, dass die Anzahl der erfüllten Kriterien kontinuierlich wachsen wird. Mit anderen Worten: Wir denken nicht, dass künstliche Intelligenz, wie in Filmen gerne porträtiert, mit einem grossen Sprung plötzlich „erwacht" und zur Rechtsperson wird. Stattdessen erwarten wir eine graduelle Entwicklung, in der ständig mehr relevante Prädikate auf Algorithmen anwendbar werden.

Selbst wenn man aber zeigen kann, dass Verantwortungskonzepte des Strafrechts auf künstliche Agenten *anwendbar* sein könnten, und diese grundsätzlich im Rahmen ihrer kognitiven Fähigkeiten verantwortlich gemacht werden können, ist damit noch nicht gesagt, dass das Strafrecht auch ideal *geeignet* ist, um mit künstlichen Intelligenzen umzugehen. Tatsächlich ist höchst fraglich, ob beispielsweise die Strafen und Sanktionen des Strafrechts bei künstlichen Intelligenzen sinnvoll sind.[101] Bei künstlichen Intelligenzen rücken aber durchaus kriminalpolitische Überlegungen in den Mittelpunkt, die bereits heute aktuell sind. Bei einer künstlichen Intelligenz, die der klassischen Entscheidungstheorie folgt, können die Kosten einer widerrechtlichen Handlung durch Erhöhung der Strafandrohung beliebig erhöht werden. Anders als biologische Menschen, die keiner bestimmten Entscheidungstheorie folgen, wird eine solche künstliche Intelligenz für den Erwartungswert einer Handlung stets die Wahrscheinlichkeit, von den Behörden erwischt zu werden, mit der Höhe der Strafandrohung

101 Vgl. dazu *Gless/Weigend*, Intelligente Agenten (Fn. 46), S. 577 f.

(in Nutzenpunkten ausgedrückt) multiplizieren.[102] Durch das beliebige Erhöhen der Strafandrohung sollten so – anders als beim Menschen – prinzipiell fast alle Delikte verhindert werden können, weil der Erwartungswert der Handlung für die künstliche Intelligenz enorm tief wird. Keine künstliche Intelligenz würde eine Handlung mit so tiefem Erwartungswert wählen. Soll der Gesetzgeber deshalb für künstliche Intelligenzen andere Strafandrohungen bestimmen? Das ist nur eine der vielen Fragen, die wir uns in den nächsten Jahren stellen werden müssen.

102 *J.L. Bermúdez*, Decision Theory and Rationality, Oxford 2009, S. 20 ff.