STIFTUNG FÜR
EFFEKTIVEN
ALTRUISMUS

Künstliche Intelligenz: Chancen und Risiken

Diskussionspapier

Künstliche Intelligenz (KI) und immer komplexer werdende Algorithmen beeinflussen unser Leben und unsere Zivilisation stärker denn je. Die KI-Anwendungsbereiche sind vielfältig und die Möglichkeiten weitreichend: Insbesondere aufgrund von Verbesserungen in der Computerhardware übertreffen gewisse KI-Algorithmen bereits heute die Leistungen menschlicher Experten/innen. Ihr Anwendungsgebiet wird künftig weiter wachsen und die KI-Leistungen werden sich verbessern. Konkret ist zu erwarten, dass sich die entsprechenden Algorithmen in immer stärkerem Ausmaß selbst optimieren — auf übermenschliches Niveau. Dieser technologische Fortschritt stellt uns wahrscheinlich vor historisch beispiellose ethische Herausforderungen. Nicht wenige Experten/innen sind der Meinung, dass von der KI neben globalen Chancen auch globale Risiken ausgehen, welche diejenigen etwa der Nukleartechnologie — die historisch ebenfalls lange unterschätzt wurde — übertreffen werden. Eine wissenschaftliche Risikoanalyse legt zudem nahe, dass hohe potenzielle Schadensausmaße auch dann sehr ernst zu nehmen sind, wenn die Eintretenswahrscheinlichkeiten tief wären.

12. Dezember 2015

Diskussionspapier der Stiftung für Effektiven Altruismus.

Bevorzugte Zitation: Mannino, A., Althaus, D., Erhardt, J., Gloor, L., Hutter, A. und Metzinger, T. (2015). Künstliche Intelligenz: Chancen und Risiken. Diskussionspapiere der Stiftung für Effektiven Altruismus (2): 1-17.

Erstveröffentlichung, 12. Dezember 2015.

www.ea-stiftung.org

Inhaltsverzeichnis

Adriano Mannino, Philosoph & Co-Präsident, Stiftung für Effektiven Altruismus

David Althaus, Wissenschaftlicher Mitarbeiter, Stiftung für Effektiven Altruismus

Dr. Jonathan Erhardt, Wissenschaftlicher Mitarbeiter, Stiftung für Effektiven Altruismus

Lukas Gloor, Wissenschaftlicher Mitarbeiter, Stiftung für Effektiven Altruismus

Dr. Adrian Hutter, Departement Physik, Universität Basel

Prof. Thomas Metzinger, Professor für Philosophie, Universität Mainz

Künstliche Intelligenz: Chancen und Risiken

Executive Summary

Künstliche Intelligenz (KI) und immer komplexer werdende Algorithmen beeinflussen unser Leben und unsere Zivilisation stärker denn je. Die KI-Anwendungsbereiche sind vielfältig und die Möglichkeiten weitreichend: Insbesondere aufgrund von Verbesserungen in der Computerhardware übertreffen gewisse KI-Algorithmen bereits heute die Leistungen menschlicher Experten/innen. Ihr Anwendungsgebiet wird künftig weiter wachsen und die KI-Leistungen werden sich verbessern. Konkret ist zu erwarten, dass sich die entsprechenden Algorithmen in immer stärkerem Ausmaß selbst optimieren — auf übermenschliches Niveau. Dieser technologische Fortschritt stellt uns wahrscheinlich vor historisch beispiellose ethische Herausforderungen. Nicht wenige Experten/innen sind der Meinung, dass von der KI neben globalen Chancen auch globale Risiken ausgehen, welche diejenigen etwa der Nukleartechnologie — die historisch ebenfalls lange unterschätzt wurde — übertreffen werden. Eine wissenschaftliche Risikoanalyse legt zudem nahe, dass hohe potenzielle Schadensausmaße auch dann sehr ernst zu nehmen sind, wenn die Eintretenswahrscheinlichkeiten tief wären.

Aktuell

In engen, gut erprobten Anwendungsbereichen (z.B. bei selbstfahrenden Autos und in Teilbereichen der medizinischen Diagnostik) konnte die Überlegenheit von KIs gegenüber Menschen bereits nachgewiesen werden. Ein vermehrter Einsatz dieser Technologien birgt großes Potenzial (z.B. bedeutend weniger Unfälle im Straßenverkehr und weniger Fehler bei der medizinischen Behandlung von Patienten/innen bzw. Erfindung vieler neuartiger Therapien). In komplexeren Systemen, wo mehrere Algorithmen mit hoher Geschwindigkeit interagieren (z.B. im Finanzmarkt oder bei absehbaren militärischen Anwendungen) besteht ein erhöhtes Risiko, dass die neuen KI-Technologien unerwartet systemisch fehlschlagen oder missbraucht werden. Es droht ein KI-Wettrüsten, das die Sicherheit der Technologieentwicklung ihrem Tempo opfert. In jedem Fall relevant ist die Frage, welche Ziele bzw. ethischen Werte einem KI-Algorithmus einprogrammiert werden sollen und wie technisch garantiert werden kann, dass die Ziele stabil bleiben und nicht manipuliert werden können. Bei selbstfahrenden Autos stellt sich etwa die Frage, wie der Algorithmus entscheiden soll, falls ein Zusammenstoß mit mehreren Fußgängern nur so vermieden werden kann, dass die eine Autoinsassin gefährdet wird — und wie sichergestellt werden kann, dass die Algorithmen der selbstfahrenden Autos nicht systemisch versagen.

Maßnahme 1 Die Förderung eines sachlich-rationalen Diskurses zum KI-Thema ist vonnöten, damit Vorurteile abgebaut werden können und der Fokus auf die wichtigsten und dringendsten Sicherheitsfragen gelegt werden kann.

Maßnahme 2 Die gesetzlichen Rahmenbedingungen sollen den neuen Technologien entsprechend angepasst werden. KI-Hersteller sind zu verpflichten, mehr in die Sicherheit und Verlässlichkeit der Technologien zu investieren und Prinzipien wie Vorhersagbarkeit, Transparenz und Nicht-Manipulierbarkeit zu beachten, damit das Risiko unerwarteter Katastrophenfälle minimiert werden kann.

Mittelfristig

Die Fortschritte in der KI-Forschung ermöglichen es, mehr und mehr menschliche Arbeit von Maschinen erledigen zu lassen. Viele Ökonomen/innen gehen davon aus, dass die zunehmende Automatisierung bereits innerhalb der nächsten 10-20 Jahre zu einer massiven Erhöhung der Arbeitslosigkeit führen könnte. (Sie tun dies durchaus im Bewusstsein, dass sich ähnliche Prognosen in der Vergangenheit nicht bewahrheitet haben, denn die aktuellen Entwicklungen sind von neuartiger Qualität und es wäre unverantwortlich, die Augen vor der Möglichkeit zu verschließen, dass die Prognosen irgendwann

zutreffen: Selbst tiefe Wahrscheinlichkeiten auf ein sehr hohes Schadensausmaß sind im Rahmen einer wissenschaftlichen Risikoanalyse zu berücksichtigen und für unser Handeln hochrelevant.) Durch die fortschreitende Automatisierung wird der Lebensstandard im statistischen Durchschnitt steigen. Es ist jedoch nicht garantiert, dass alle Menschen — oder auch nur eine Mehrheit der Menschen — davon profitieren werden.

Maßnahme 3 Können wir gesellschaftlich sinnvoll mit den Folgen der KI-Automatisierung umgehen? Sind die aktuellen Sozialsysteme dafür geeignet? Diese Fragen sind ausführlich zu klären. Gegebenenfalls sind neuartige Maßnahmen zu ergreifen, um die negativen Entwicklungen abzufedern bzw. positiv zu wenden. Modelle eines bedingungslosen Grundeinkommens oder einer negativen Einkommenssteuer etwa sind zur gerechteren Verteilung der Produktivitätsgewinne prüfenswert.

Langfristig

Viele KI-Experten/innen halten es für plausibel, dass noch in diesem Jahrhundert KIs erschaffen werden, deren Intelligenz der menschlichen in allen Bereichen weit überlegen ist. Die Ziele solcher KIs, welche prinzipiell alles Mögliche zum Gegenstand haben können (menschliche, ethisch geprägte Ziele stellen eine winzige Teilmenge aller möglichen Ziele dar), würden die Zukunft unseres Planeten maßgeblich beeinflussen — was für die Menschheit ein existenzielles Risiko darstellen könnte. Unsere Spezies hat deshalb eine dominante Stellung inne, weil sie (aktuell) über die am höchsten entwickelte Intelligenz verfügt. Es ist aber wahrscheinlich, dass bis zum Ende des Jahrhunderts KIs entwickelt werden, deren Intelligenz sich zu der unseren so verhält wie die unsere zu derjenigen etwa der Schimpansen. Zudem ist die Möglichkeit nicht auszuschließen, dass KIs in Zukunft auch phänomenale Zustände entwickeln, d.h. (Selbst-)Bewusstsein und besonders auch subjektive Präferenzen und Leidensfähigkeit, was uns mit neuartigen ethischen Herausforderungen konfrontiert. Angesichts der unmittelbaren Relevanz der Thematik und dessen, was längerfristig auf dem Spiel steht, sind Überlegungen zur KI-Sicherheit sowohl in der Politik als auch in der Forschung aktuell stark unterrepräsentiert.

Maßnahme 4 Es gilt, institutionell sicherheitsfördernde Maßnahmen auszuarbeiten, beispielsweise die Vergabe von Forschungsgeldern für Projekte, die sich auf die Analyse und Prävention von Risiken der KI-Entwicklungen konzentrieren. Die Politik muss insgesamt mehr Ressourcen für die kritische, wissenschaftlich-ethische Begleitung folgenschwerer Technologieentwicklungen bereitstellen.

Maßnahme 5 Bestrebungen zur internationalen Forschungskollaboration (analog etwa zum CERN in der Teilchenphysik) sind voranzutreiben. Internationale Koordination ist im KI-Bereich besonders deshalb essenziell, weil sie das Risiko eines technologischen Wettrüstens minimiert. Ein Verbot jeder risikobehafteten KI-Forschung wäre nicht praktikabel und würde zu einer schnellen und gefährlichen Verlagerung der Forschung in Länder mit tieferen Sicherheitsstandards führen.

Maßnahme 6 Forschungsprojekte, die selbstoptimierende neuromorphe, d.h. gehirnanaloge KI-Architekturen entwickeln oder testen, die mit hoher Wahrscheinlichkeit über Leidensfähigkeit verfügen werden, sollten unter die Aufsicht von Ethikkommissionen gestellt werden (in Analogie zu den Tierversuchskommissionen).

Einleitung

Das Streben nach Wissen zieht sich als roter Faden durch die Menschheitsgeschichte. Wenn sich Gesellschaften in ihrer Struktur und ihrer Dynamik stark änderten, beruhte dies in den meisten Fällen auch auf neuen technologischen Erfindungen. Zwischen der ersten Verwendung von Steinwerkzeugen bis zum entwicklungsgeschichtlichen "Großen Schritt nach Vorne", als Homo sapiens die Kunst erfand und anfing, Höhlenwände zu bemalen, lagen rund zwei Millionen Jahre. Bis zum Ackerbau und zur Sesshaftigkeit dauerte es einige zehntausend Jahre. Die ersten Symbole erschienen wenige tausend Jahre danach, später entwickelten sich die ersten Schriften. Im 17. Jahrhundert wurde das Mikroskop erfunden. Die Industrialisierung im 19. Jahrhundert ermöglichte die ersten Millionenstädte. Nur ein Jahrhundert später wurde das Atom gespalten und Menschen flogen zum Mond. Der Computer wurde erfunden, und seither verdoppelten sich Maßzahlen zur Rechenleistung und Energieeffizienz von Computern in regelmäßigen Zeitabständen [1]. Der technologische Fortschritt entwickelt sich oft exponentiell. Für die geistigen Fähigkeiten des Menschen gilt dies nicht.

Im Verlauf des letzten Jahres betonten zahlreiche renommierte Wissenschaftler/innen und Unternehmer/innen die dringliche Bedeutung der KI-Thematik, und wie wichtig es sei, dass sich Entscheidungsträger mit den Prognosen der KI-Forschung auseinandersetzen [2]. Zu den Exponenten dieser Bewegung zur KI-Sicherheit gehören beispielsweise Stuart Russell [3], Nick Bostrom [4], Stephen Hawking [5], Sam Harris [6], Max Tegmark [7], Elon Musk [8], Jann Tallinn [9] und Bill Gates [10].

In spezifischen Gegenstandsbereichen (d.h. *domänenspezifisch*) haben künstliche Intelligenzen schon wiederholt das menschliche Niveau erreicht oder gar übertroffen. 1997 besiegte der Computer *Deep Blue* den amtierenden Weltmeister Garry Kasparov im Schach [11], 2011 besiegte *Watson* die zwei besten menschlichen Spieler/innen der auf Sprachverständnis beruhenden Spielshow Jeopardy! [12], und 2015 wurde mit *Cepheus* die erste Pokervariante — *Fixed Limit Holdem heads-up* — spieltheoretisch komplett gelöst [13]. Künstliche neuronale Netzwerke können inzwischen mit menschlichen Experten/innen bei der Diagnose von Krebszellen konkurrieren [14] und nähern sich etwa auch beim Erkennen handgeschriebener chinesischer Schriftzeichen dem menschlichen Niveau an [15]. Schon 1994 erreichte ein selbstlernendes Backgammon-Programm die Spielstärke der weltbesten Spieler/innen, indem es Strategien fand, die von Menschen zuvor noch nie angewandt wurden [16]. Mittlerweile existieren sogar Algorithmen, die verschiedenste Computerspiele von Grund auf selbstständig erlernen können und dabei (über)menschliches Niveau erreichen [17, 18]. Damit kommen wir langsam einer *generellen Intelligenz* näher, die zumindest prinzipiell Probleme jeglicher Art selbständig lösen kann.

Größere Macht verleiht größere Verantwortung. Technologie ist bloß ein Mittel; entscheidend ist, wie wir sie verwenden. Schon die Anwendung existierender KIs stellt uns vor erhebliche ethische Herausforderungen, die im nächsten Teil dieses Diskussionspapiers erläutert werden. Im Kapitel danach werden Entwicklungen besprochen, die erwarten lassen, dass Fortschritte in der KI-Forschung mittelfristig die wirtschaftliche Automatisierung so weit vorantreiben werden, dass es auf dem Arbeitsmarkt zu großen Umstrukturierungen kommen wird. In den beiden letzten Kapiteln geht es dann um langfristige und existenzielle Risiken der KI-Forschung im Zusammenhang mit der möglichen Erschaffung (über)menschlicher Intelligenz und künstlichen Bewusstseins.

Vorteile und Risiken gängiger KIs

Unser Leben und unsere Zivilisation werden in immer größerem Maße von Algorithmen und domänenspezifischen künstlichen Intelligenzen (KIs) beeinflusst und beherrscht [19]: Man denke nur an Smartphones, den Flugverkehr [20] oder Internetsuchmaschinen [21]. Auch die Finanzmärkte sind auf immer komplexer werdende Algorithmen angewiesen, die wir immer weniger verstehen [22, 23]. Meist verläuft der Einsatz solcher Algorithmen ohne Zwischenfälle, doch es besteht immer die Möglichkeit, dass ein unwahrscheinliches *Black-Swan-Ereignis* [24] eintritt, welches das ganze System ins Chaos zu stürzen droht. So kam es beispielsweise 2010 in den USA zu einem für die Finanzwelt schockierenden Börsencrash [25], weil Computer-Algorithmen auf unvorhergesehene Art und Weise mit dem Finanzmarkt interagierten [26]. Innerhalb von Minuten verloren bedeutsame Aktien mehr als 90% ihres Wertes und schnellten dann wieder auf den Ausgangswert hoch. Bei militärischen Anwendungen wäre die Wahrscheinlich-

keit höher, dass eine solche "Rückkehr zur Ausgangssituation" ausbleibt[27]. Um verheerende Fehlfunktionen dieser Art zu verhindern, scheint es generell ratsam, wesentlich mehr in die Sicherheit und Verlässlichkeit von KIs zu investieren. Leider bestehen zurzeit wirtschaftliche Anreize, KI-Leistungssteigerungen gegenüber der KI-Sicherheit zu priorisieren.

Vier Kriterien zur Konstruktion von KIs

Sicherheit ist bei jeder Art von Maschine essenziell, doch die Konstruktion domänenspezifischer KIs geht mit neuartigen ethischen Herausforderungen einher, sobald diese ehemals von Menschen ausgeführte kognitive Arbeit mit sozialer Dimension übernehmen. Man denke beispielsweise an einen Algorithmus, der die Kreditwürdigkeit von Bankkunden beurteilt und dabei (ohne dass dies explizit einprogrammiert war) gewissen Bevölkerungsgruppen gegenüber diskriminierende Entscheidungen fällt. Sogar Technologien, die im Grunde genommen nur bestehende Tätigkeiten ersetzen, können die Maschinenethik [28] vor interessante Herausforderungen stellen: Selbstgesteuerte Fahrzeuge beispielsweise werfen die Frage auf, nach welchen Kriterien bei einem drohenden Unfall entschieden werden soll. Sollten die Fahrzeuge beispielsweise das Überleben der Insassen/innen am höchsten priorisieren oder sollte es bei einem unausweichlichen Unfall darum gehen, die Opferzahl insgesamt möglichst gering zu halten [29]?

Deshalb haben der KI-Theoretiker Eliezer Yudkowsky und der Philosoph Nick Bostrom vier Prinzipien vorgeschlagen, welche die Konstruktion neuer KIs leiten sollten [30]: Die Funktionsweise einer KI sollte 1) *nachvollziehbar* und 2) ihre Handlungen *prinzipiell vorhersagbar* sein; beides in einem Zeitfenster, das den verantwortlichen Experten/innen im Falle einer möglichen Fehlfunktion genügend Raum zur Reaktion und Veto-Kontrolle bietet. Zudem sollten KIs 3) sich nicht einfach *manipulieren* lassen, und falls doch ein Unfall geschieht, sollte 4) die *Verantwortlichkeit* klar bestimmt sein.

Vorteile (domänenspezifischer) künstlicher Intelligenz

Algorithmen und domänenspezifische KIs bringen grundsätzlich sehr viele Vorteile mit sich. Sie haben unser Leben zum Positiven beeinflusst und werden dies, sofern die nötigen Vorkehrungen getroffen werden, in Zukunft auch weiterhin tun. Im Folgenden werden zwei instruktive Beispiele diskutiert.

Selbstfahrende Autos sind schon lange keine Science-Fiction mehr [31, 32] und werden in absehbarer Zeit auch kommerziell erhältlich sein: Das von Google entwickelte Auto *Google Driverless Car*, das von KI-Algorithmen vollständig autonom gesteuert wird, unternahm die ersten Testfahrten in den USA schon 2011 [33, 34]. Neben der für Arbeit oder Entspannung gewonnenen Zeit besteht ein zweiter Vorteil selbstfahrender Autos in ihrer erhöhten Sicherheit. 2010 beispielsweise starben weltweit 1,24 Millionen Menschen in Verkehrsunfällen, beinahe ausschliesslich aufgrund menschlichen Versagens [35]. Zahlreiche Menschenleben könnten also jedes Jahr gerettet werden, denn selbstfahrende Autos sind bereits jetzt nachweislich sicherer als von Menschen gesteuerte Fahrzeuge [36, 37].

Allerdings stehen übermäßig viele Menschen selbstgesteuerten Autos immer noch skeptisch gegenüber, wohl weil sie deren Risiken sowie die eigenen Fahrfähigkeiten überschätzen. Eine Studie kam beispielsweise zum Schluss, dass 93% aller amerikanischen Autofahrer/innen glauben, dass sie generell bessere Fahrfähigkeiten besitzen als der Median [38] — was statistisch unmöglich ist. Unrealistischer Optimismus [39] und die Kontrollillusion [40] veranlassen Menschen vermutlich auch dazu, die Risiken zu unterschätzen, wenn sie selbst am Steuer sitzen [41, 42].

Auch Ärzte überschätzen ihre Fähigkeiten [43], was zu tödlichen Irrtümern führen kann. Allein in den USA sterben jährlich schätzungsweise zwischen 44'000 und 98'000 Menschen in Krankenhäusern aufgrund von Behandlungsfehlern [44]. In diesem Zusammenhang ist die von IBM entwickelte KI Watson [45] zu begrüßen, die 2011 die besten menschlichen Spieler/innen in der Quiz-Show *Jeopardy!* besiegte und dadurch Berühmtheit erlangte [12]. Watson ist Menschen nicht nur in Quiz-Shows überlegen: Seit 2013 können Krankenhäuser Watson mieten, um beispielsweise Krebsdiagnosen zu tätigen. Da "Doktor Watson" innert kürzester Zeit enorme Mengen an Information aufnehmen und kombinieren kann, ist er menschlichen Kollegen diagnostisch teilweise überlegen [46, 47].

Dass eine aktuelle KI akkuratere Krankheitsdiagnosen als menschliche Ärzte tätigen kann, mag erstaunen. Doch seit Langem ist bekannt, dass *statistisches Schlussfolgern* klinischem Schlussfolgern, d.h. den Urteilen menschlicher Experten/innen, meist überlegen ist [48, 49]. Und natürlich sind KIs wie Watson geradezu gemacht für statistisches Schlussfolgern. Computer bei Diagnosen (nicht) zu Rate zu ziehen, kann folglich über Menschenleben entscheiden.

Kognitive Verzerrungen — Irren ist menschlich

Ein Grund, weshalb menschliche Experten/innen im statistischen Urteilen weniger kompetent sind als KIs, besteht in der oben erwähnten, allzu menschlichen Tendenz, die eigenen Fähigkeiten zu überschätzen. Diese Tendenz wird als *overconfidence bias* bezeichnet [50]. Der *overconfidence bias* ist nur einer von etlichen kognitiven Verzerrungen [51, 52], die das menschliche Denken systematisch in die Irre führen können. KIs hingegen können so konstruiert werden, dass sie keine kognitiven Verzerrungen aufweisen. Prinzipiell könnte gesteigertes Vertrauen in die Prognosen von KIs, sofern diese sicher und nach nachvollziehbaren Kriterien konstruiert sind, auch zu einer deutlichen Rationalitätssteigerung bei vielen gesellschaftlichen und politischen Herausforderungen führen. Das Problem bestünde hier darin, die Stärken der KI zu nutzen, ohne menschliche Handlungsautonomie an die entsprechenden Systeme abzugeben.

Zusammenfassung und Ausblick

Irrationale Ängste vor neuartigen, im Grunde vorteilhaften Technologien sind nach wie vor weit verbreitet [53]. Derartige Technophobie mag auch ein Grund dafür sein, dass Watson oder selbstfahrende Autos skeptisch betrachtet werden. Bedenken hinsichtlich neuartiger Technologien sind aber nicht immer irrational. Die meisten Technologien lassen sich zum Wohle der Menschheit einsetzen, können jedoch auch zur Gefahr werden, wenn sie in die falschen Hände gelangen oder wenn nicht genügend Rücksicht auf Sicherheit und unbeabsichtigte Nebeneffekte genommen wird.

Ähnlich verhält es sich auch mit künstlicher Intelligenz: Selbstgesteuerte Autos könnten unser Leben erleichtern und Menschenleben retten, aber komplexe Computeralgorithmen können auch die Börse abstürzen lassen. Obwohl die meisten domänenspezifischen KIs der nahen Zukunft relativ einfach sicher gestaltet werden können, gilt es langfristige Entwicklungen zu beachten: In nicht allzu ferner Zukunft könnte die künstliche Intelligenz prinzipiell sogar, ähnlich wie die Biotechnologie (etwa durch die mögliche Synthetisierung neuartiger Viren), eine existenzielle Bedrohung darstellen [54, 55, 4].

> **Empfehlung 1 — Verantwortungsvoller Umgang:** Wie bei allen anderen Technologien sollte auch bei der Erforschung der KI genau darauf geachtet werden, dass die (potenziellen) Vorteile die (potenziellen) Nachteile klar überwiegen. Die Förderung eines sachlich-rationalen Diskurses ist vonnöten, damit irrationale Vorurteile und Ängste abgebaut und veraltete gesetzliche Rahmenwerke den neuen Technologien entsprechend reformiert werden können. Bei jeder großflächigen Anwendung von KIs sollten die oben erläuterten vier Prinzipien eingehalten werden [30]. ■

Automatisierung und Arbeitslosigkeit

Angesichts der Erfolge im Bereich des maschinellen Lernens und der Robotik in den letzten Jahren scheint es bloß eine Frage der Zeit zu sein, bis auch komplexe Arbeiten, die hohe Intelligenz erfordern, umfassend von Maschinen übernommen werden können [56].

Wenn Maschinen in vielen Aufgabenbereichen Arbeiten schneller, zuverlässiger und billiger erledigen werden als menschliche Arbeiter, dann hätte dies weitreichende Auswirkungen auf den Arbeitsmarkt. Ökonomen/innen wie Cowen [57], McAfee und Brynjolfsson [58] sagen vorher, dass der technologische Fortschritt die Einkommensschere noch viel stärker öffnen wird und dass es zu großflächigen Lohnsenkungen sowie massiv erhöhter Arbeitslosigkeit kommen könnte.

Eine 2013 erschienene Analyse kommt zum Schluss, dass 47% aller Jobs in den USA in 10-20 Jahren mit hoher Wahrscheinlichkeit automatisierbar sein werden [59]. Am schwierigsten zu automatisieren sind Tätigkeiten, die hohe soziale Intelligenz (z.B. PR-Beratung), Kreativität (z.B. Mode-Design) oder Feingefühl und Flexibilität bei den Bewegungen (z.B. Chirurgie) erfordern. In diesen Bereichen ist der Stand der KI-Forschung noch weit vom Niveau menschlicher Experten/innen entfernt.

Vor- und Nachteile der Automatisierung durch Computer

Insbesondere diejenigen Menschen und Länder werden vom technologischen Fortschritt profitieren, die es verstehen, von den neuen technologischen Möglichkeiten und der damit verbundenen Datenflut (*Big Data*) Gebrauch zu machen [60]. Dies sind insbesondere Länder mit gut ausgebildeten Computerspezialisten. Ausserdem wird es in Zukunft immer wichtiger werden, dass Menschen ein treffendes Bild der Vor- und Nachteile verschiedener Computeralgorithmen im Vergleich mit rein menschlicher Ent-

scheidungsfindung und Arbeitsleistung haben, wofür gute Bildung zentral ist [61].

Auch in der Unterhaltungsindustrie wird es zu weitreichenden Neuerungen kommen: Mit verbesserter Graphik, neuen Unterhaltungstechnologien und neuen Funktionen für mobile Geräte, die alle zunehmend billiger werden, erhöht sich auch der Suchtfaktor von Videospielen und von Internetzugang [62]. Die sozialen und psychologischen Auswirkungen dieser Entwicklung sind noch wenig erforscht, aber es deutet einiges darauf hin, dass diese Trends unser Sozialverhalten [63], unsere Aufmerksamkeitsspannen und die Art, wie Kinder aufwachsen, nachhaltig verändern [64]. In absehbarer Zukunft, wenn ausgeklügelte virtuelle Realitäten auch für Nicht-Wissenschaftler/innen erlebbar sein werden und immer tiefer in unsere Lebenswelt eindringen werden, könnte dieser Effekt noch viel stärker zum Tragen kommen. Die Auswirkungen häufiger Immersionen in virtuelle Realitäten, oder von Verfahren wie Ganzkörper-Illusionen, bei denen das subjektive Selbstgefühl zeitweise auf einen virtuellen Avatar projiziert wird [65], dürften erheblich sein.

Für den Bereich der Bildung schliesslich bietet die Unterhaltungsindustrie über die Gamifizierung von Lerninhalten große Chancen [66]; gleichzeitig besteht das Risiko, dass sich der Anteil der Jugendlichen erhöht, die wegen pathologischen Videospiel- oder Internetkonsums [67] Mühe beim Abschliessen einer Ausbildung haben.

Utopien und Dystopien

Der technologische Fortschritt steigert die Produktivität einer Gesellschaft [68], was den durchschnittlichen Lebensstandard erhöht [69]. Wenn mehr Arbeit von Maschinen erledigt wird, schafft dies Raum für Freizeit und Selbstverwirklichung der Menschen — zumindest für diejenigen Menschen, welche in der Lage sind, davon zu profitieren. Eine Schattenseite der zunehmenden Automatisierung könnte jedoch darin bestehen, dass der gewonnene Produktivitätszuwachs mit zunehmender sozialer Ungleichheit einher geht, so dass ein Anstieg des *durchschnittlichen* Lebensstandards nicht mit einem Anstieg der Lebensqualität des *Medians* zusammenfällt. Experten/innen wie der MIT-Wirtschaftsprofessor Erik Brynjolfsson befürchten aus diversen Gründen [70] gar, dass der technologische Fortschritt die Situation für eine Mehrheit der Menschen zu verschlechtern droht.

In einer kompetitiven Weltwirtschaft, in der die KI-Technologie so weit fortgeschritten ist, dass viele Tätigkeiten von Maschinen ausgeführt werden können, wird der Lohn für automatisierbare menschliche Arbeit zunehmend sinken [58]. Ohne Regulierung könnte das Lohnniveau für viele Menschen unter das Existenzminimum sinken. Die soziale Ungleichheit könnte stark zunehmen, wenn der wirtschaftliche Output sich zwar erhöht, es aber ohne Lohnzahlungen keine Umverteilung mehr gäbe. Um dieser Entwicklung entgegenzuwirken, schlagen McAfee und Brynjolfsson vor, dass bestimmte von Menschen ausgeführte Tätigkeiten subventioniert werden könnten. Weitere Möglichkeiten, die Vorteile des technologischen Fortschritts auf die Gesamtbevölkerung zu verteilen, sind das bedingungslose Grundeinkommen und die negative Einkommenssteuer [71, 72].

Einige Experten/innen warnen auch vor Zukunftsszenarien, in denen die Veränderungen noch gravierender sind. Der Ökonom Robin Hanson hält es beispielsweise für plausibel, dass es noch in diesem Jahrhundert möglich sein wird, menschliche Gehirnsimulationen, sogenannte *whole brain emulations (WBEs)* [73], digital in virtueller Realität laufen zu lassen. WBEs wären duplizierbar und könnten, sofern genügend Hardware vorhanden ist, um ein Vielfaches schneller laufen als ein biologisches Gehirn — was einen enormen Effizienzgewinn beim Arbeiten zur Folge hätte [74]. Hanson prognostiziert, dass es in einem solchen Fall eine “Bevölkerungsexplosion” unter WBEs geben würde, weil diese in vielen Bereichen als enorm kosteneffektive Arbeiter eingesetzt werden könnten [75]. Hansons Spekulationen sind umstritten [61], und es sollte nicht davon ausgegangen werden, dass sie die *wahrscheinlichste* Zukunft skizzieren. Aktuell ist die Forschung — beispielsweise das *Blue Brain Project* an der ETH Lausanne — noch weit entfernt von den ersten Gehirnsimulationen, geschweige denn davon, diese auch in Echtzeit (oder gar beschleunigt) mit Inputs einer virtuellen Realität zu versorgen. Es ist dennoch von Bedeutung, die Hardware-Entwicklung in Bezug auf die Möglichkeit von WBEs im Auge zu behalten. Falls das von Hanson skizzierte Szenario eintrifft, wäre dies nämlich von hoher ethischer Relevanz: Zum einen könnten viele durch komplexe Simulationen ersetzte Menschen arbeitslos werden. Zum anderen stellt sich die Frage, unter welchen Bedingungen die eingesetzten WBEs phänomenales Bewusstsein und subjektive Präferenzen hätten, d.h. ob sie bei ihrer (möglicherweise forcierten) Arbeitstätigkeit auch Leid empfinden würden.

Empfehlung 2 — Vorausschauend handeln: Wie etwa auch bei der Problematik des Klimawandels sollten für Forscher/innen und Entscheidungsträger/innen Anreize geschaffen werden, sich mit KI-Zukunftsszenarien auseinanderzusetzen. Dadurch können die Grundlagen für vorsorgliche Maßnahmen geschaffen werden. Insbesondere sollten im Bereich der KI-Folgenabschätzung und -Sicherheit entsprechende Fachtagungen durchgeführt, Expertenkommissionen gebildet und Forschungsprojekte finanziert werden. ■

Empfehlung 3 — Bildung: Gezielte Anpassungen der Bildungsinhalte könnten helfen, die Menschen besser auf die neuartigen Herausforderungen vorzubereiten. EDV- und Programmierkenntnisse beispielsweise gewinnen stark an Relevanz, während auswendig gelerntes Wissen an Wert verliert. Die Gamifizierung von Lerninhalten bietet ein großes Potenzial, das zu fördern ist. Die sozialen und psychologischen Auswirkungen des Internets sollten weiter untersucht werden und dem pathologischen Konsum von Videospielen und Online-Medien ist vorzubeugen. ■

Empfehlung 4 — Offenheit gegenüber neuen Maßnahmen: Die Subventionierung menschlicher Arbeit, ein bedingungsloses Grundeinkommen sowie eine negative Einkommenssteuer wurden als mögliche Maßnahmen vorgeschlagen, um die negativen Auswirkungen der zunehmenden Automatisierung sozial abzufedern. Es gilt zu klären, welche weiteren Optionen existieren und welches Maßnahmenpaket maximal zielführend ist. Dazu müssen Vor- und Nachteile systematisch analysiert und auf politischer Ebene diskutiert werden. Fördergelder sollten investiert werden, um die dabei aufgeworfenen empirischen Fragen zu beantworten. ■

Generelle Intelligenz und Superintelligenz

Die "generelle Intelligenz" misst die Fähigkeiten eines Akteurs, seine Ziele in einer umfassenden Menge an unbekannten Umgebungen zu erreichen [76, 77]. Diese Art von Intelligenz kann ein (Katastrophen-)Risiko darstellen, wenn die Ziele des Akteurs nicht mit den unseren übereinstimmen. Wenn eine generelle Intelligenz ein übermenschliches Niveau erreicht, dann ist von *Superintelligenz* die Rede: Eine Superintelligenz ist der menschlichen Intelligenz in jeder Hinsicht überlegen, einschliesslich wissenschaftlicher Kreativität, gesundem "Menschenverstand" und Sozialkompetenz. Diese Definition für Superintelligenz lässt offen, ob eine Superintelligenz Bewusstsein hätte oder nicht [78, 79].

Komparative Vorteile genereller künstlicher Intelligenz gegenüber dem Menschen

Menschen sind intelligente zweibeinige "Bio-Roboter", die ein bewusstes Selbstmodell besitzen und von der Evolution über Jahrmilliarden hervorgebracht wurden. Diese Tatsache wurde als Argument dafür ins Feld geführt [80, 81, 82], dass die Erschaffung künstlicher Intelligenz nicht allzu schwer sein dürfte, da die KI-Forschung im Gegensatz zur Evolution, die nur in langsamen und ungezielt-verschwenderischen Generationenschritten fortschreitet, viel schneller und zielgerichteter verlaufen kann. Neben der Tatsache, dass die Evolution einen *KI-Machbarkeitsnachweis* liefert, ermöglicht sie der gezielten menschlichen Forschung natürlich auch, bei biologischem *Design* Anleihen zu machen und entsprechend schneller voranzuschreiten.

Im Vergleich zum biologischen Gehirn der Menschen bietet die Computerhardware nämlich mehrere Vorteile [4, S. 60]: Die Grundelemente (moderne Mikroprozessoren) "feuern" millionenfach schneller als Neuronen; die Signale werden millionenfach schneller übertragen; und ein Computer kann insgesamt über bedeutend mehr Grundelemente verfügen — Supercomputer können die Größe einer Fabrikhalle annehmen. Auch bezüglich der Softwarekomponenten hätte eine digitale Intelligenz der Zukunft einem biologischen Hirn gegenüber große Vorteile [4, S. 60–61]: Software lässt sich beispielsweise leicht editieren oder vervielfachen, damit die Vorzüge eines Designs gleich in mehrfacher Weise genutzt werden können. Eine künstliche Intelligenz kann mit großen Datenbanken versorgt werden, so dass potenziell relevante Information jederzeit abgerufen werden kann. In wenigen wichtigen Bereichen, beispielsweise bei der Energieffizienz, der Resilienz gegenüber rein physikalischen Beschädigungen und der *graceful degradation* [83], bleibt die künstliche Hardware noch hinter dem menschlichen Gehirn zurück. Insbesondere gibt es noch keinen direkten Zusammenhang zwischen thermodynamischer Effizienz und Komplexitätsreduktion auf der Ebene der Informationsverarbeitung [84, 85]. In den kommenden Jahrzehnten wird die Computerhardware jedoch kontinuierlich weiterentwickelt werden.

Angesichts der genannten komparativen Vorteile und

der prognostizierten rasanten Verbesserung von Hardware [86] und Software scheint es wahrscheinlich, dass die menschliche Intelligenz dereinst von Maschinen überflügelt wird. Es gilt, herauszufinden beziehungsweise genauer abzuschätzen, wie und wann das der Fall sein könnte und worin die Implikationen eines solchen Szenarios bestehen.

Zeithorizonte

Verschiedene Experten/innen auf dem Gebiet der künstlichen Intelligenz haben sich der Frage gewidmet, wann die erste Maschine das menschliche Intelligenzniveau erreichen wird. Eine Umfrage unter den hundert erfolgreichsten KI-Experten/innen, gemessen anhand eines Zitationsindex, ergab, dass eine Mehrheit dieser Experten/innen es für wahrscheinlich hält, dass dies bereits in der ersten Hälfte dieses Jahrhunderts der Fall sein wird [4, S. 19]. Eine Mehrheit der Experten/innen geht weiterhin davon aus, dass Menschen dereinst eine Superintelligenz erschaffen werden, falls der technologische Fortschritt (infolge globaler Katastrophen) keine schweren Rückschläge erfahren wird [4, S. 20]. Die Varianz der zeitlichen Abschätzungen ist hoch: Manche Experten/innen sind sich sehr sicher, dass es spätestens 2040 Maschinen mit mindestens menschlichem Intelligenzniveau geben wird, während (wenige) andere denken, dass dieses Niveau gar nie erreicht werden wird. Selbst wenn man etwas konservativere Annahmen trifft, weil man einbeziehen möchte, dass menschliche Experten/innen die Tendenz haben, sich bei ihren Schätzungen *zu* sicher zu sein [87, 88], wäre es immer noch völlig verfehlt, die Superintelligenz-Thematik als "Science-Fiction" einzustufen: Denn auch konservative Annahmen implizieren, dass die Wahrscheinlichkeit nicht vernachlässigbar ist, dass eine KI menschlichen Intelligenzniveaus noch in diesem Jahrhundert entwickelt wird.

Ziele einer generellen KI

Als rationaler Akteur strebt eine künstliche Intelligenz genau das an, was ihre Ziele/ihre Zielfunktion besagen [89]. Ob eine künstliche Intelligenz *ethisch* vorgehen wird, d.h. ob sie Ziele haben wird, die nicht im Konflikt mit den Interessen von Menschen und anderen leidensfähigen Wesen stehen, ist völlig offen: Eine künstliche Intelligenz kann alle möglichen Ziele verfolgen [90]. Es wäre ein fehlerhafter Anthropomorphismus, davon auszugehen, dass sich jede Art Superintelligenz wie (typische) Menschen für ethische Fragen interessieren würde. Wenn wir eine künstliche Intelligenz bauen, legen wir explizit oder implizit auch ihr Ziel fest.

Manchmal werden diese Forderungen dahingehend kritisiert, dass jegliche Versuche, das Ziel einer künstlichen Intelligenz nach menschlichen Wertmaßstäben zu richten, einer "Versklavung" gleichkommen, weil der KI unsere menschlichen Werte *aufgezwungen* würden [91]. Diese Kritik beruht allerdings auf Missverständnissen. Der Ausdruck "aufzwingen" suggeriert, dass schon ein bestimmtes, "wahres" Ziel existiert, das eine KI vor ihrer Erschaffung hätte. Diese Vorstellung ist jedoch unsinnig: Es gibt keinen "Geist in der Maschine", kein Ziel, das von den Prozessen unabhängig ist, die einen Akteur hervorgebracht haben. Der Prozess, der eine Intelligenz hervorbringt, bestimmt unweigerlich die Funktionsweise und die Ziele dieser Intelligenz. *Falls* wir eine Superintelligenz zu bauen beabsichtigen, sind wir, und nichts/niemand sonst, für deren (Haupt-)Ziele verantwortlich. Weiterhin ist es auch nicht der Fall, dass eine KI durch die Ziele, die wir ihr unweigerlich mitgeben, in irgendeiner Weise eine Schädigung erfahren muss. (Die Möglichkeit, in einem ethisch relevanten Sinn geschädigt zu werden, setzt zudem voraus, dass Bewusstsein vorliegt — eine Voraussetzung, die bei einer Superintelligenz auch nicht erfüllt sein muss.) Ganz analog formen wir *nolens volens* die Werte beziehungsweise Ziele biologischer Kinder — d.h. biologischer Intelligenzen —, die wir hervorbringen. Selbstverständlich impliziert dies nicht, dass Kinder dadurch in unethischer Weise "versklavt" würden. Ganz im Gegenteil: Wir haben die starke ethische Pflicht, unseren biologischen Kindern grundlegende ethische Werthaltungen mitzugeben. Dasselbe gilt für alle künstlichen Intelligenzen, die wir hervorbringen.

Der Informatikprofessor Stuart Russell betont [3], dass die Einprogrammierung ethischer Ziele eine große Herausforderung darstellt, sowohl auf technischer Ebene (Wie werden komplexe Ziele in einer Programmiersprache so erfasst, dass keine unbeabsichtigten Ergebnisse resultieren?) als auch auf ethischer, moralphilosophischer Ebene (Welche Ziele eigentlich?). Das erste von Russell erwähnte Problem wird in der Fachliteratur auch als *Value-Loading-Problem* bezeichnet [92].

Obwohl der Raum möglicher Ziele einer Superintelligenz riesig ist, können wir einige verlässliche Aussagen über ihre Handlungen treffen. Es existiert nämlich eine Reihe instrumentell rationaler Zwischenziele, die für Akteure mit unterschiedlichsten Endzielen nützlich sind. Dazu gehören Ziel- und Selbsterhaltung, Intelligenzerhöhung, Erkenntnisfortschritt und physische Ressourcenakkumulation [93]. Wenn das Ziel einer KI verändert wird, ist das für die Erreichung ihres ursprünglichen Ziels unter Umständen gleich negativ (oder negativer), wie wenn sie zer-

stört würde. Intelligenzerhöhung ist wichtig, weil sie nichts anderes bedeutet als die Erhöhung der Fähigkeit, Ziele in variierenden Umgebungen zu erreichen — deshalb besteht die Möglichkeit einer sogenannten *Intelligenzexplosion*, bei der eine KI in kurzer Zeit durch rekursive Selbstverbesserung stark an Intelligenz gewinnt [94, 95]. (Die Grundidee der rekursiven Selbstverbesserung wurde erstmals von I. J. Good konzeptualisiert [96]; mittlerweile existieren dazu konkrete Algorithmen [97].) Ressourcenakkumulation und die Erfindung neuer Technologien verleihen der KI mehr Macht, was auch der besseren Zielerreichung dient. Falls die Zielfunktion einer neu entstandenen Superintelligenz dem Wohl leidensfähiger Wesen keinen Wert zuschreibt, würde sie, wo immer es für ihre (Zwischen-)Zielerreichung nützlich wäre, rücksichtslos Tod und Leid verursachen.

Man könnte zur Annahme geneigt sein, dass eine Superintelligenz keine Gefahr darstellt, weil es sich nur um einen Computer handelt, dem man wortwörtlich den Stecker ziehen könnte. *Per definitionem* wäre eine Superintelligenz jedoch nicht dumm: Wenn die Gefahr besteht, dass ihr der Stecker gezogen wird, dann würde sie sich vorerst einmal so verhalten, wie dies von den Machern gewünscht wird, bis sie herausgefunden hat, wie sie das Risiko einer unfreiwilligen Deaktivierung minimieren kann [4, S. 117]. Einer Superintelligenz könnte es zudem möglich sein, die Sicherheitssysteme von Großbanken und nuklearen Waffenarsenalen mittels bisher unbekannter Sicherheitslücken (sogenannten *zero day exploits*) zu umgehen und die Weltbevölkerung auf diese Weise zu erpressen und zur Kooperation zu zwingen. Wie bereits zu Beginn erwähnt, könnte auch hier eine “Rückkehr zur Ausgangssituation” nicht mehr möglich sein.

Was auf dem Spiel steht

Im besten Fall könnte eine Superintelligenz zahlreiche Probleme der Menschheit lösen, d.h. uns helfen, die großen wissenschaftlichen, ethischen, ökologischen und ökonomischen Herausforderungen der Zukunft zu bewältigen. Wenn sich die Ziele einer Superintelligenz allerdings nicht mit unseren Präferenzen beziehungsweise den Präferenzen aller empfindungsfähigen Wesen decken, dann wird sie zu einer existenziellen Bedrohung und kann möglicherweise mehr Leid verursachen, als es ohne sie je gegeben hätte [98].

Rationales Risikomanagement

In Entscheidsituationen, in denen potenziell sehr viel auf dem Spiel steht, sind die folgenden Prinzipien wichtig:

1. Teure Vorkehrungen zu treffen lohnt sich selbst bei geringen Risikowahrscheinlichkeiten, wenn es hinreichend viel zu gewinnen/verlieren gibt [89].
2. Wenn in einem Gebiet unter Experten/innen wenig Konsens besteht, ist epistemische Bescheidenheit ratsam, d.h. man sollte kein allzu großes Vertrauen in die Zuverlässigkeit der eigenen Meinung haben.

Die Risiken der KI-Forschung sind globaler Natur. Misslingt den KI-Forschenden der erste Versuch, einer Superintelligenz ethische Ziele zu verleihen, so gibt es womöglich keine zweite Chance mehr. Es ist durchaus vertretbar, die längerfristigen Risiken der KI-Forschung als noch größer einzuschätzen als diejenigen der Klimaerwärmung. Im Vergleich dazu erhielt die Thematik jedoch noch kaum Aufmerksamkeit. Wir weisen mit diesem Diskussionspapier darauf hin, dass es sich deshalb umso mehr lohnt, erhebliche Ressourcen in die Sicherheit der KI-Forschung zu investieren.

Wenn die hier erörterten Szenarien (vielleicht geringe, aber) mehr als bloss infinitesimale Eintrittswahrscheinlichkeit haben, dann sollte künstliche Intelligenz und die damit assoziierten Chancen und Risiken zu den globalen Prioritäten gehören. Die Wahrscheinlichkeit eines guten Ausgangs der KI-Forschung kann u.a. durch folgende Maßnahmen maximiert werden:

Empfehlung 5 — Information: Eine wirksame Verbesserung der Sicherheit künstlicher Intelligenz beginnt mit der Aufklärung seitens der sich mit KI beschäftigenden Experten/innen, Investoren und Entscheidungsträger. Informationen zu den mit KI-Fortschritten assoziierten Risiken müssen einfach zugänglich gemacht werden. Organisationen, welche dieses Anliegen unterstützen, sind das Future of Humanity Institute (FHI) der Universität Oxford, das Machine Intelligence Research Institute (MIRI) in Berkeley, das Future of Life Institute (FLI) in Boston, sowie im deutschsprachigen Raum die Stiftung für Effektiven Altruismus (EAS). ■

Empfehlung 6 — KI-Sicherheit: In den vergangenen Jahren war ein eindrücklicher Anstieg der Investitionen in die KI-Forschung zu beobachten [86]. Die Erforschung der KI-Sicherheit hingegen ist vergleichsweise weit zurückgeblieben. Die weltweit einzige Organisation, die der Erforschung der theoretischen und technischen Probleme der KI-Sicherheit höchste Priorität beimisst, ist das Machine Intelligence Research Institute (MIRI). Bei der Vergabe von Forschungsgeldern im KI-Bereich sollte gefordert werden, dass sicherheitsrelevante Aspekte der Forschungsprojekte ausgewiesen und entsprechende Vorkehrungen getroffen werden. Ein Verbot jeder risikoreichen KI-Forschung wäre nicht praktikabel und würde zu einer schnellen und gefährlichen Verlagerung der Forschung in Länder mit tieferen Sicherheitsstandards führen. ■

Empfehlung 7 — Globale Kooperation und Koordination: Ökonomische und militärische Anreize schaffen ein kompetitives Klima, in dem es mit an Sicherheit grenzender Wahrscheinlichkeit zu einem gefährlichen Wettrüsten kommen wird. Dabei würde die Sicherheit der KI-Forschung zugunsten von schnelleren Fortschritten und Kostensenkungen reduziert. Verstärkte internationale Kooperation kann dieser Dynamik entgegenwirken. Gelingt die internationale Koordination, lässt sich auch ein "Race to the Bottom" der Sicherheitsstandards (durch Abwanderung der wissenschaftlichen und industriellen KI-Forschung oder Androhung derselben) eher vermeiden. ■

Künstliches Bewusstsein

Menschen und viele nichtmenschliche Tiere haben phänomenales Bewusstsein — es fühlt sich subjektiv-innerlich in bestimmter Weise an, ein Mensch oder ein nichtmenschliches Tier zu sein [99]. Sie haben Sinneseindrücke, ein (rudimentäres oder ausgeprägtes) Ich-Gefühl, empfinden Schmerzen bei körperlicher Schädigung, und können psychisches Leid oder Freude verspüren (vgl. etwa die Depressionsstudien bei Mäusen [100]). Kurzum: Sie sind *empfindungsfähige* Wesen. Dies hat zur Folge, dass sie in einem für sie selbst relevanten Sinn *geschädigt* werden können. Im KI-Kontext stellt sich dazu die Frage: Kann es auch Maschinen geben, deren materiell-funktionale Struktur ein leidvolles "Innenleben" realisieren kann? Für den Leidbegriff liefert der Philosoph und Kognitionswissenschaftler Thomas Metzinger vier Kriterien, die bei Maschinen entsprechend auch erfüllt sein müssten:

1. Bewusstsein.
2. ein phänomenales Selbstmodell.
3. die Fähigkeit zur Darstellung negativer Valenzen (d.h. verletzter subjektiver Präferenzen) innerhalb des Selbstmodells.
4. Transparenz (d.h. Wahrgenommenes fühlt sich unwiderruflich "real" an — das System ist also gezwungen, sich mit dem Inhalt seines bewussten Selbstmodells zu identifizieren) [101, 102].

Etwas präziser ist zwischen zwei verwandten Fragen zu unterscheiden: Erstens, ob Maschinen überhaupt je Bewusstsein und Leidensfähigkeit entwickeln könnten; und zweitens, falls die erste Frage zu bejahen ist, welche *Typen* von Maschinen Bewusstsein haben (werden).

Diese beiden Fragen werden von Philosophen/innen und KI-Experten/innen gleichermaßen untersucht. Ein Blick auf den Stand der Forschung zeigt, dass die erste Frage einfacher zu beantworten ist als die zweite. Es existiert unter Experten/innen ein relativ solider Konsens darüber, dass Maschinen prinzipiell über Bewusstsein verfügen können und dass Maschinenbewusstsein zumindest in *neuromorphen* Computern möglich ist [103, 104, 105, 106, 107, 108, 109]. Solche Computer haben Hardware derselben funktionalen Organisation wie ein biologisches Gehirn [110]. Schwieriger ist die zweite Frage zu beantworten: Welche Typen von Maschinen, neben neuromorphen Computern, können Bewusstsein haben? In diesem Bereich ist der wissenschaftliche Konsens weniger ausgeprägt [111]. Es ist beispielsweise umstritten, ob reine Simulationen — etwa das simulierte Gehirn des *Blue Brain Project* — Bewusstsein haben können. Die Frage wird zwar von verschiedenen Experten/innen positiv beantwortet [109, 105], von einigen aber auch verneint [111, 112].

Angesichts der Unsicherheit unter Experten/innen scheint es angebracht, eine *vorsichtige* Position zu vertreten: Nach dem heutigen Wissensstand ist es zumindest denkbar, dass viele hinreichend komplexe Computer, darunter auch nicht-neuromorphe, leidensfähig sein werden.

Diese Überlegungen haben weitreichende ethische Konsequenzen. Wenn Maschinen kein Bewusstsein haben könnten, so wäre es ethisch unbedenklich, sie als Arbeitskräfte auszubeuten und ihnen riskante Tätigkeiten wie die Entschärfung von Minen oder die Handhabung gefährlicher Stoffen aufzutragen [4, S. 167]. Wenn hinrei-

chend komplexe künstliche Intelligenzen aber mit einiger Wahrscheinlichkeit Bewusstsein und subjektive Präferenzen haben werden, so sind ähnliche ethisch-rechtliche Sicherheitsvorkehrungen zu treffen wie bei Menschen und vielen nichtmenschlichen Tieren [113]. Wenn etwa das virtuelle Gehirn des *Blue Brain Project* Bewusstsein haben wird, dann wäre es ethisch beispielsweise hochproblematisch, es (und mit ihm zahlreiche Kopien beziehungsweise "Klone") in depressive Zustände zu versetzen, um Depression systematisch zu erforschen. Metzinger warnt davor, dass bewusste Maschinen für Forschungszwecke missbraucht werden könnten und als "Bürger zweiter Klasse" nicht nur keine Rechte haben [1] und als austauschbare experimentelle Werkzeuge benutzt werden könnten, sondern dass sich diese Tatsache auch negativ auf der Ebene ihres inneren Erlebens widerspiegeln könnte [106]. Diese Aussicht ist deshalb besonders besorgniserregend, weil es denkbar ist, dass künstliche Intelligenzen dereinst in riesiger Anzahl erschaffen werden [4, 75]. So könnte in einem Worst-Case-Szenario eine astronomische, historisch beispiellose Opferzahl und Leidmenge resultieren.

Diese dystopischen Szenarien deuten auf eine wichtige Implikation technologischer Fortschritte hin: Selbst wenn uns nur "geringe" ethische Fehler unterlaufen, etwa indem wir gewisse Computer fälschlicherweise als unbewusst oder moralisch unbedeutend klassifizieren, kann dies aufgrund historisch beispielloser technologischer Macht zu historisch beispiellosen Katastrophen führen. Wenn sich die Gesamtzahl empfindungsfähiger Wesen stark erhöht, dann reicht eine marginale Verbesserung unserer ethischen Werte und empirischen Einschätzungen nicht aus — beide müssten sich *massiv* verbessern, um der stark erhöhten Verantwortung gerecht werden zu können. Daher sollten wir angesichts unserer Unsicherheit bezüglich des Maschinenbewusstseins im KI-Bereich besonders große Vorsicht walten lassen. Nur so bleibt die Möglichkeit intakt, potenzielle Katastrophenszenarien der beschriebenen Art zu vermeiden.

Empfehlung 8 — Forschung: Um ethische Entscheidungen treffen zu können, ist es unabdingbar, zu wissen, welche natürlichen und künstlichen Systeme über Bewusstsein und insbesondere Leidensfähigkeit verfügen. Gerade im Bereich des Maschinenbewusstseins besteht aber noch große Unsicherheit. Es scheint deshalb sinnvoll, entsprechende interdisziplinäre Forschung zu fördern (Philosophie, Neurowissenschaft, Computerwissenschaft). ■

Empfehlung 9 — Regulierung: Es ist mittlerweile gängige Praxis, Experimente an lebenden Testsubjekten durch Ethikkommissionen prüfen zu lassen [114, 115]. Aufgrund der Möglichkeit, dass neuromorphe Computer und simulierte Lebewesen auch Bewusstsein beziehungsweise eine subjektive Innenperspektive entwickeln, sollte Forschung an ihnen ebenfalls unter der strengen Aufsicht von Ethikkommissionen erfolgen. Die (unerwartete) Erschaffung leidensfähiger künstlicher Wesen sollte vermieden oder hinausgezögert werden, insbesondere weil diese in sehr großer Zahl auftreten könnten und zunächst — in Ermangelung einer gesellschaftlich-politischen Interessenvertretung — wohl rechtlos dastünden. ■

Zusammenfassung

Bereits heute existieren Erstversionen neuer KI-Technologien mit überraschendem Potenzial, seien es die selbstgesteuerten Fahrzeuge, Watson als Hilfe bei der medizinischen Diagnostik, oder die neusten vom US-Militär getesteten Drohnen. In absehbarer Zeit werden diese Anwendungen marktreif für den großflächigen Einsatz sein. Spätestens dann braucht es gut durchdachte gesetzliche Rahmenbedingungen, um das Potenzial dieser technologischen Möglichkeiten so zu verwirklichen, dass die Risiken einer negativen Gesamtentwicklung möglichst gering bleiben.

Je größer der Fortschritt in zentralen Bereichen der KI-Technologie, desto wichtiger und dringender wird das rational-vorausschauende Angehen der dabei entstehenden Herausforderungen. Auch die Forscher/innen und die Entwickler/innen neuer Technologien tragen Verantwortung dafür, wie ihre Beiträge die Welt verändern werden. Im Gegensatz zu Politik und Gesetzgebung, die den neuesten Entwicklungen i.d.R. nachhinken, sind die KI-Forscher/innen und KI-Entwickler/innen direkt am Geschehen beteiligt; sie sind diejenigen, die sich am besten mit der Materie auskennen.

Leider bestehen starke wirtschaftliche Anreize, die Entwicklung neuer Technologien möglichst schnell voranzu-

[1] Vereinigungen wie "People for the ethical treatment of reinforcement learners" (PETRL) sprechen sich dafür aus, dass künstliche Intelligenzen, sofern sie empfindungsfähig sind, die gleiche moralische Berücksichtigung erhalten sollten wie "biologische Intelligenzen": http://petrl.org/.

treiben, ohne dass Zeit für teure Risikoanalysen "verloren" geht. Diese ungünstigen Rahmenbedingungen erhöhen das Risiko, dass uns die Kontrolle über KI-Technologien und deren Verwendung mehr und mehr entgleiten wird. Dem ist auf möglichst vielen Ebenen entgegenzuwirken: Politisch; in der Forschung selbst; und allgemein bei allen Individuen, die sich auf relevante Weise mit dem Thema beschäftigen können. Eine Kernvoraussetzung dafür, dass die KI-Entwicklung in möglichst vorteilhafte Bahnen gelenkt wird, wird sein, dass die Thematik nicht nur unter wenigen Experten/innen, sondern auch im breiten öffentlichen Diskurs als große (möglicherweise größte) bevorstehende Herausforderung erkannt wird.

Neben den genannten konkreteren Forderungen möchten wir mit diesem Diskussionspapier deshalb auch einen wesentlichen Anstoß und ein Plädoyer dafür liefern, dass das Thema "Risiken und Chancen der KI", wie der Klimawandel oder die Verhinderung kriegerischer Konflikte, möglichst bald als globale Priorität erkannt wird.

Danksagung

Wir bedanken uns bei all jenen, die uns bei der Recherche oder beim Verfassen des Diskussionspapiers behilflich waren. Besonders hervorzuheben sind hierbei Kaspar Etter und Massimo Mannino für ihre Ratschläge zum Aufbau des Papiers; Prof. Oliver Bendel für Anstöße zum Kapitel "Vorteile und Risiken gängiger KIs"; und Prof. Jürgen Schmidhuber für Inputs zu den Kapiteln "Generelle Intelligenz und Superintelligenz" und "Künstliches Bewusstsein", sowie für seine Inputs zum aktuellen Forschungsstand verschiedener KI-Bereiche.

Unterstützer/innen

Die Kernpunkte des Diskussionspapiers werden getragen von:

- **Prof. Dr. Fred Hamker**, Professor für Künstliche Intelligenz, Technische Universität Chemnitz
- **Prof. Dr. Dirk Helbing**, Professor für Computational Social Science, ETH Zürich
- **Prof. Dr. Malte Helmert**, Professor für Künstliche Intelligenz, Universität Basel
- **Prof. Dr. Manfred Hild**, Professor für Digitale Systeme, Beuth Hochschule für Technik (Berlin)
- **Prof. Dr. Dr. Eric Hilgendorf**, Leiter Forschungsstelle RobotRecht, Universität Würzburg
- **Prof. Dr. Marius Kloft**, Professor für Maschinelles Lernen, Humboldt Universität Berlin
- **Prof. Dr. Jana Koehler**, Professorin für Informatik, Hochschule Luzern
- **Prof. Dr. Stefan Kopp**, Professor für Social Cognitive Systems, Universität Bielefeld
- **Prof. Dr. Dr. Franz Josef Radermacher**, Professor für Datenbanken und Künstliche Intelligenz, Universität Ulm

Literatur

[1] Koomey, J. G., Berard, S., Sanchez, M. & Wong, H. (2011). Implications of Historical Trends in the Electrical Efficiency of Computing. *IEEE Annals of the History of Computing*, *33*(3), 46–54.

[2] Brockman, J. (2015). *What to Think About Machines That Think: Today's Leading Thinkers on the Age of Machine Intelligence*. Harper Perennial.

[3] Russell, S. (2015). Will They Make Us Better People? (http://edge.org/response-detail/26157)

[4] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press.

[5] BBC. (2015a). Stephen Hawking Warns Artificial Intelligence Could End Mankind. (http://www.bbc.com/news/technology-30290540)

[6] Harris, S. (2015). Can We Avoid a Digital Apocalypse? (https://edge.org/response-detail/26177)

[7] The Independent. (2014). Stephen Hawking: 'Transcendence Looks at the Implications of Artificial Intelligence — But Are We Taking AI Seriously Enough?' (http://www.independent.co.uk/news/science/stephen-hawking-transcendence-looks-at-the-implications-of-artificial-intelligence--but-are-we-taking-ai-seriously-enough-9313474.html)

[8] The Guardian. (2014). Elon Musk Donates $10m to Keep Artificial Intelligence Good for Humanity. (http://www.theguardian.com/technology/2015/jan/16/elon-musk-donates-10m-to-artificial-intelligence-research)

[9] SBS. (2013). Artificial Irrelevance: The Robots Are Coming. (http://www.sbs.com.au/news/article/2012/07/18/artificial-irrelevance-robots-are-coming)

[10] BBC. (2015b). Microsoft's Bill Gates Insists AI Is a Threat. (http://www.bbc.com/news/31047780)

[11] Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*. Penguin.

[12] PCWorld. (2011). IBM Watson Vanquishes Human Jeopardy Foes. (http://www.pcworld.com/article/219893/ibm_watson_vanquishes_human_jeopardy_foes.html)

[13] Bowling, M., Burch, N., Johanson, M. & Tammelin, O. (2015). Heads-up Limit Hold'em Poker Is Solved. *Science*, *347*(6218), 145–149.

[14] Ciresan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. (2013). Mitosis Detection in Breast Cancer Histology Images Using Deep Neural Networks. MICCAI 2013. (http://people.idsia.ch/~juergen/deeplearningwinsMICCAIgrandchallenge.html)

[15] Ciresan, D., Meier, U. & Schmidhuber, J. (2012). Multi-Column Deep Neural Networks for Image Classification. *Computer Vision and Pattern Recognition 2012*, 3642–3649.

[16] Tesauro, G. (1994). TD-Gammon, a Self-Teaching Backgammon Program, Achieves Master-Level Play. *Neural Computation*, *6*(2), 215–219.

[17] Koutník, J., Cuccu, G., Schmidhuber, J. & Gomez, F. (2013). Evolving Large-Scale Neural Networks for Vision-Based Reinforcement Learning. In *Proceedings of the 15th Annual Conference on Genetic and Evolutionary Computation* (S. 1061–1068). ACM.

[18] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Ostrovski, G. u. a. (2015). Human-Level Control Through Deep Reinforcement Learning. *Nature*, *518*(7540), 529–533.

[19] Slavin, K. (2012). How Algorithms Shape Our World. (http://ed.ted.com/lessons/kevin-slavin-how-algorithms-shape-our-world)

[20] Tagesanzeiger. (2008). Computer-Panne legt US-Flugverkehr lahm. (http://www.tagesanzeiger.ch/ausland/amerika/ComputerPanne-legt-USFlugverkehr-lahm/story/13800972)

[21] Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. (http://ilpubs.stanford.edu:8090/422/)

[22] Wired. (2010). Algorithms Take Control of Wall Street. (http://www.wired.com/2010/12/ff_ai_flashtrading/all/)

[23] Lin, T. C. (2012). The New Investor. *UCLA L. Rev. 60*, 678–735.

[24] Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable Fragility*. Random House.

[25] Lauricella, T. & McKay, P. (2010). Dow Takes a Harrowing 1,010.14-point Trip. *Wall Street Journal (May 7, 2010)*.

[26] Securities, U., Commission, E. & the Commodity Futures Trading Commission. (2010). Findings Regarding the Market Events of May 6, 2010. *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*.

[27] Spiegel. (2015). Denkende Waffen: Künstliche-Intelligenz-Forscher Warnen vor Künstlicher Intelligenz. (http://www.spiegel.de/netzwelt/netzpolitik/elon-musk-und-stephen-hawking-warnen-vor-autonomen-waffen-a-1045615.html)

[28] Bendel, O. (2013). Towards Machine Ethics. In *Technology Assessment and Policy Areas of Great Transitions* (S. 343–347). Proceedings from the PACITA 2013 Conference in Prague.

[29] Goodall, N. J. (2014). Machine Ethics and Automated Vehicles. In *Road Vehicle Automation: Lecture Notes in Mobility* (S. 93–102). Springer International Publishing.

[30] Bostrom, N. & Yudkowsky, E. (2013). The Ethics of Artificial Intelligence. In *Cambridge Handbook of Artificial Intelligence*. Cambridge University Press.

[31] Dickmanns, E. D., Behringer, R., Dickmanns, D., Hildebrandt, T., Maurer, M., Thomanek, F. & Schiehlen, J. (1994). The Seeing Passenger Car 'VaMoRs-P'. In *International Symposium on Intelligent Vehicles 94* (S. 68–73).

[32] Dickmanns, E. (2011). Evening Keynote: Dynamic Vision as Key Element for AGI. 4th Conference on Artificial General Intelligence, Mountain View, CA. (https://www.youtube.com/watch?v=YZ6nPhUG2i0)

[33] Thrun, S. (2011). Google's Driverless Car. (http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car)

[34] Forbes. (2012). Nevada Passes Regulations for Driverless Cars. (http://www.forbes.com/sites/alexknapp/2012/02/17/nevada-passes-regulations-for-driverless-cars/)

[35] Organization, W. H. u. a. (2013). *WHO Global Status Report on Road Safety 2013: Supporting a Decade of Action*. World Health Organization.

[36] Simonite, T. (2013). Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *MIT Technology Review, Oct*, *25*.

[37] CNBC. (2014). Self-Driving Cars Safer Than Those Driven by Humans: Bob Lutz. (http://www.cnbc.com/id/101981455)

[38] Svenson, O. (1981). Are We All Less Risky and More Skillful Than Our Fellow Drivers? *Acta Psychologica*, *9*(6), 143–148.

[39] Weinstein, N. D. (1980). Unrealistic Optimism about Future Life Events. *Journal of Personality and Social Psychology*, *39*(5), 806.

[40] Langer, E. J. (1975). The Illusion of Control. *Journal of Personality and Social Psychology*, *32*(2), 311.

[41] Von Hippel, W. & Trivers, R. (2011). The Evolution and Psychology of Self-Deception. *Behavioral and Brain Sciences*, *34*(1), 1–56.

[42] Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books.

[43] Berner, E. S. & Graber, M. L. (2008). Overconfidence as a Cause of Diagnostic Error in Medicine. *The American Journal of Medicine*, *121*(5), S2–S23.

[44] Kohn, L. T., Corrigan, J. M., Donaldson, M. S. u. a. (2000). *To Err Is Human: Building a Safer Health System*. National Academies Press.

[45] The New York Times. (2010). What Is IBM's Watson? (http://www.nytimes.com/2010/06/20/magazine/20Computer-t.html)

[46] Wired. (2013). IBM's Watson Is Better at Diagnosing Cancer Than Human Doctors. (http://www.wired.co.uk/news/archive/2013-02/11/ibm-watson-medical-doctor)

[47] Forbes. (2013). IBM's Watson Gets Its First Piece Of Business In Healthcare. (http : / / www . forbes . com / sites / bruceupbin/2013/02/08/ibms-watson-gets-its-first-piece-of-business-in-healthcare/)

[48] Dawes, R. M., Faust, D. & Meehl, P. E. (1989). Clinical Versus Actuarial Judgment. *Science*, *243*(4899), 1668–1674.

[49] Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E. & Nelson, C. (2000). Clinical Versus Mechanical Prediction: A Meta-Analysis. *Psychological Assessment*, *12*(1), 19.

[50] West, R. F. & Stanovich, K. E. (1997). The Domain Specificity and Generality of Overconfidence: Individual Differences in Performance Estimation Bias. *Psychonomic Bulletin & Review*, *4*(3), 387–392.

[51] Tversky, A. & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131.

[52] Pohl, R. (Hrsg.). (2004). *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory*. Psychology Press.

[53] Brosnan, M. J. (2002). *Technophobia: The Psychological Impact of Information Technology*. Routledge.

[54] Yudkowsky, E. (2008). Artificial Intelligence as a Positive and Negative Factor in Global Risk. *Global Catastrophic Risks*, *1*, 303.

[55] Bostrom, N. (2002). Existential Risks. *Journal of Evolution and Technology*, *9*(1).

[56] Smith, A. & Anderson, J. (2014). AI, Robotics, and the Future of Jobs. Pew Research Center.

[57] Cowen, T. (2013a). *Average Is Over: Powering America Beyond the Age of the Great Stagnation*. Penguin.

[58] Brynjolfsson, E. & McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. WW Norton & Company.

[59] Frey, C. B. & Osborne, M. A. (2013). The Future of Employment: How Susceptible Are Jobs to Computerisation? *Oxford Martin Programme on Technology and Employment*. (https://web.archive.org/web/20150109185039/http://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf)

[60] Helbing, D. (2015). *Thinking Ahead — Essays on Big Data, Digital Revolution, and Participatory Market Society*. Springer.

[61] Cowen, T. (2013b). EconTalk Episode with Tyler Cowen: Tyler Cowen on Inequality, the Future, and Average is Over. (http://www.econtalk.org/archives/2013/09/tyler_cowen_on.html)

[62] Griffiths, M., Kuss, D. & King, D. (2012). Video Game Addiction: Past, Present and Future. *Current Psychiatry Reviews*, *8*(4), 308–318.

[63] Srivastava, L. (2010). Mobile Phones and the Evolution of Social Behaviour. *Behavior & Information Technology*, *24*(2), 111–129.

[64] Prensky, M. (2001). Do They Really Think Differently? *On the Horizon*, *47*(2).

[65] Metzinger, T. (2015a). Virtuelle Verkörperung in Robotern. *SPEKTRUM*, *2*, 48–55.

[66] Kapp, K. M. (2012). *The Gamification of Learning and Instruction: Game-Based Methods and Strategies for Training and Education*. Pfeiffer.

[67] Bavelier, D., Green, S., Hyun Han, D., Renshaw, P., Merzenich, M. & Gentile, D. (2011). Viewpoint: Brains on Video Games. *Nature Reviews Neuroscience*, *12*, 763–768.

[68] Fagerberg, J. (2000). Technological Progress, Structural Change and Productivity Growth: A Comparative Study. *Structural Change and Economic Dynamics*, *11*(4), 393–411.

[69] Galor, O. & Weil, D. N. (1999). From Malthusian Stagnation to Modern Growth. *American Economic Review*, 150–154.

[70] Brynjolfsson, E. (2014). EconTalk Episode with Erik Brynjolfsson: Brynjolfsson on the Second Machine Age. (http://www.econtalk.org/archives/2014/02/brynjolfsson_on.html)

[71] Hughes, J. J. (2014). Are Technological Unemployment and a Basic Income Guarantee Inevitable or Desirable? *Journal of Evolution and Technology*, *24*(1), 1–4.

[72] Krugman, P. (2013). Sympathy for the Luddites. *New York Times*, *13*. (http://www.nytimes.com/2013/06/14/opinion/krugman-sympathy-for-the-luddites.html)

[73] Bostrom, N. & Sandberg, A. (2008). Whole Brain Emulation: A Roadmap. Oxford: Future of Humanity Institute.

[74] Hanson, R. (2012). Extraordinary Society of Emulated Minds. (http://library.fora.tv/2012/10/14/Robin_Hanson_Extraordinary_Society_of_Emulated_Minds)

[75] Hanson, R. (1994). If Uploads Come First. *Extropy*, *6*(2), 10–15.

[76] Legg, S. & Hutter, M. (2005). A Universal Measure of Intelligence for Artificial Agents. In *International Joint Conference on Artificial Intelligence* (Bd. 19, S. 1509). Lawrence Erlbaum Associates ltd.

[77] Hutter, M. (2007). Universal Algorithmic Intelligence: A Mathematical Top-Down Approach. In *Artificial General Intelligence* (Bd. 6, *2*, S. 227–290). Springer.

[78] Bostrom, N. (1998). How Long Before Superintelligence? *International Journal of Future Studies*, *2*.

[79] Schmidhuber, J. (2012). Philosophers & Futurists, Catch Up! Response to The Singularity. *Journal of Consciousness Studies*, *19*(1-2), 173–182.

[80] Moravec, H. (1998). When Will Computer Hardware Match the Human Brain. *Journal of Evolution and Technology*, *1*(1), 10.

[81] Moravec, H. (2000). *Robot: Mere Machine to Transcendent Mind*. Oxford University Press.

[82] Shulman, C. & Bostrom, N. (2012). How Hard Is Artificial Intelligence? Evolutionary Arguments and Selection Effects. *Journal of Consciousness Studies*, *19*(7-8), 103–130.

[83] Sengupta, B. & Stemmler, M. (2014). Power Consumption During Neuronal Computation. *Proceedings of the IEEE*, *102*(5), 738–750.

[84] Friston, K. (2010). The Free-Energy Principle: A Unified Brain Theory? *Nature Reviews Neuroscience*, *11*, 127–138.

[85] Sengupta, B., Stemmler, M. & Friston, K. (2013). Information and Efficiency in the Nervous System — A Synthesis. *PLoS Comput Biol*, *9*(7).

[86] Eliasmith, C. (2015). On the Eve of Artificial Minds. In T. Metzinger & J. M. Windt (Hrsg.), *Open mind*. MIND Group. (http://open-mind.net/papers/@@chapters?nr=12)

[87] Armstrong, S., Sotala, K. & ÓhÉigeartaigh, S. S. (2014). The Errors, Insights and Lessons of Famous AI Predictions — And What They Mean for the Future. *Journal of Experimental & Theoretical Artificial Intelligence*, *26*(3), 317–342.

[88] Brenner, L. A., Koehler, D. J., Liberman, V. & Tversky, A. (1996). Overconfidence in Probability and Frequency Judgments: A Critical Examination. *Organizational Behavior and Human Decision Processes*, *65*(3), 212–219.

[89] Peterson, M. (2009). *An Introduction to Decision Theory*. Cambridge University Press.

[90] Armstrong, S. (2013). General Purpose Intelligence: Arguing the Orthogonality Thesis. *Analysis and Metaphysics*, (12), 68–84.

[91] Noë, A. (2015). The Ethics Of The 'Singularity'. (http://www.npr.org/sections/13.7/2015/01/23/379322864/the-ethics-of-the-singularity)

[92] Bostrom, N. (2012). The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines*, *22*(2), 71–85.

[93] Omohundro, S. M. (2008). The Basic AI Drives. In *Proceedings of the First AGI Conference, 171, Frontiers in Artificial Intelligence and Applications* (Bd. 171, S. 483–492).

[94] Solomonoff, R. (1985). The Time Scale of Artificial Intelligence: Reflections on Social Effects. *Human Systems Management*, *5*, 149–153.

[95] Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies*, *17*(9-10), 7–65.

[96] Good, I. J. (1965). Speculations Concerning the First Ultraintelligent Machine. In *Advances in Computers* (S. 31–88). Academic Press.

[97] Schmidhuber, J. (2006). Gödel Machines: Fully Self-Referential Optimal Universal Self-Improvers. In *Artificial General Intelligence* (S. 119–226).

[98] Tomasik, B. (2011). Risks of Astronomical Future Suffering. Foundational Research Institute. (http://foundational-research.org/publications/risks-of-astronomical-future-suffering/)

[99] Nagel, T. (1974). What Is it Like to Be a Bat? *The Philosophical Review*, 435–450.

[100] Durgam, R. (2001). Rodent Models of Depression: Learned Helplessness Using a Triadic Design in Tats. *Curr Protoc Neurosci*, (8).

[101] Metzinger, T. (2012). Two Principles for Robot Ethics. In H. E & G. J-P (Hrsg.), *Robotik und Gesetzgebung* (S. 263–302). NOMOS. (http://www.blogs.uni-mainz.de/fb05philosophie/files/2013/04/Metzinger_RG_2013_penultimate.pdf)

[102] Metzinger, T. (2015b). *Empirische Perspektiven aus Sicht der Selbstmodell-Theorie der Subjektivität: Eine Kurzdarstellung mit Beispielen*. Selbstverlag. (http://www.amazon.de/Empirische-Perspektiven-Sicht-Selbstmodell-Theorie-Subjektivitat-ebook/dp/B01674W53W)

[103] Moravec, H. P. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press.

[104] Chalmers, D. J. (1995). Absent Qualia, Fading Qualia, Dancing Qualia. *Conscious Experience*, 309–328.

[105] Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.

[106] Metzinger, T. (2014). *Der Ego Tunnel. Eine neue Philosophie des Selbst: Von der Hirnforschung zur Bewusstseinsethik*. Piper.

[107] Metzinger, T. (2015c). What If They Need to Suffer? (https://edge.org/response-detail/26091)

[108] Dennett, D. C. (1993). *Consciousness Explained*. Penguin UK.

[109] Bostrom, N. (2003). Are We Living in a Computer Simulation? *The Philosophical Quarterly*, *53*(211), 243–255.

[110] Hasler, J. & Marr, B. (2013). Finding a Roadmap to Achieve Large Neuromorphic Hardware Systems. *Frontiers in Neuroscience*, *7*(118).

[111] Koch, C. (2014). What it Will Take for Computers to Be Conscious, MIT Technology Review. (http://www.technologyreview.com/news/531146/what-it-will-take-for-computers-to-be-conscious/)

[112] Tononi, G. (2015). Integrated Information Theory. *Scholarpedia*, *10*(1), 4164. (http://www.scholarpedia.org/article/Integrated_Information_Theory)

[113] Singer, P. (1988). Comment on Frey's 'Moral Standing, the Value of Lives, and Speciesism'. *Between the Species: A Journal of Ethics*, *4*, 202–203.

[114] Swissethics, Verein anerkannter Ethikkommissionen der Schweiz. (o.d.). (http://www.swissethics.ch/)

[115] Senatskommission für Tierexperimentelle Forschung. (2004). Tierversuche in der Forschung. (http://www.dfg.de/download/pdf/dfg_im_profil/geschaeftsstelle/publikationen/dfg_tierversuche_0300304.pdf, publisher=Deutsche Forschungsgemeinschaft)

Die Stiftung für Effektiven Altruismus (EAS) ist eine unabhängige Denkfabrik und Projektschmiede im Schnittbereich von Ethik und Wissenschaft. Die Resultate ihrer Arbeit macht sie im Rahmen von Diskussionspapieren der Gesellschaft und Politik zugänglich. Sie bietet zudem Spenden- und Karriereberatung an. Der Effektive Altruismus (EA) stellt das Leitkonzept der Stiftung dar: Unsere Ressourcen — Zeit und Geld — sind limitiert. Wie können wir sie so einsetzen, dass das meiste Leid verhindert und die meisten Leben gerettet werden? Und welche rationalen Gründe sprechen überhaupt dafür, Ressourcen in eine nachhaltig-effektive Leidminderung zu investieren? Diesen Fragen gehen wir aus philosophischer, ökonomischer sowie kognitions- und sozialpsychologischer Sicht nach.

www.ea-stiftung.org