Malte Rehbein

Die „Asilomar AI Principles“ zu Künstlicher Intelligenz

Die im vergangenen Jahr veröffentlichten und von zahlreichen Wissenschaftlerinnen und Wissenschaftlern sowie weiteren Shareholdern aus Wirtschaft und Technik unterzeichneten Asilomar AI Principles[1] *sind ein Vorschlag für ein Regelwerk zum Umgang mit Künstlicher Intelligenz (KI). Sie umfassen 23 Imperative „ranging from research strategies to data rights to future issues including potential super-intelligence“[2] und können als eine Reaktion auf die beschleunigte technologische Entwicklung in diesem Bereich verstanden werden. Die Folgen dieser Entwicklung bezeichnet das organisierende* Future-of-Life-Institute *zu Recht als „major change […] across every segment of society“[3]. Jahrzehntelange Verheißungen der KI scheinen kurz vor dem Durchbruch zu stehen.*

Eine ethische Bewertung, rechtliche Regelung und konsequente politische Regulierung, die sowohl den Einsatz als auch bereits die Entwicklung von KI umfassen, sind dringend geboten. Diese Notwendigkeit ist in Deutschland inzwischen auch in der breiten Öffentlichkeit angekommen, wie etwa die Gründung des *Weizenbaum-Instituts* in Berlin und die Aufnahme der KI als *Schlüsseltechnologie* im Koalitionsvertrag von Union und SPD zeigen. Es ist zu begrüßen, dass sich mit Asilomar auch die wissenschaftlich-wirtschaftliche Prominenz diesem Thema widmet. Die vorgeschlagenen Prinzipien sind jedoch von ihrer grundsätzlichen Anlage und im Detail zu kritisieren.

Die 23 Imperative entstanden aus einer im Januar 2017 im kalifornischen Asilomar veranstalteten interdisziplinären Tagung *„Beneficial AI“* heraus. Nützlichkeit als Tagungsmotto gab dabei Tenor und Stoßrichtung der Prinzipien bereits vor:

> *„Artificial intelligence has already provided beneficial tools that are used every day by people around the world. Its continued development, guided by the following principles, will offer amazing opportunities to help and empower people in the decades and centuries ahead.“*[4]

Mit dieser technikdeterministischen und rein utilitaristischen Sichtweise auf den Einsatz von KI-Technologien bleiben viele Fragen offen. Denn den „amazing opportunities“ stehen auch zahlreiche Warnungen vor den Folgen eines (unregulierten) Ausbaus von KI-Technologien gegenüber. Führende Wissenschaftler wie *Stephen Hawking* halten ein exponentielles Wachstum von KI für möglich, bis hin zu einer verselbstständigten „Super-Intelligenz“, die den Menschen letztlich überflüssig mache. Bedrohungen werden auch aus der Zivilgesellschaft heraus aufgezeigt: „Slaughterbots“ am Beispiel autonomer Waffensysteme,[5] die britische Serie „Black Mirror“ und zahlreiche kulturkritische Literatur illustrieren in ihren Szenarien künftige Welten, die in ihren Dystopien erschrecken, zugleich aber äußerst realistisch wirken.

Zwischen profitabel und brandgefährlich

Der Einsatz von KI bietet ein riesiges Geschäftspotenzial. Es ist eine Technologie, die hohe finanzielle Profite ermöglicht; es ist ebenfalls eine Technologie, mit der Macht akkumuliert werden kann. Für wirtschaftliche wie politische Prozesse sind Daten und Informationen die Leitwährung,[6] für KI-Systeme sind sie ihr Futter, ohne die Maschinen nicht lernen und KI-Systeme nicht existieren könnten. Wer Daten *und* KI-Systeme kontrolliert, wird Macht ausüben können. Die Entwicklungen in China in Richtung einer *Big-Data-Diktatur* sind da sicherlich nur der Anfang.[7]

Bezeichnend für die Pole zwischen profitabel und brandgefährlich stehen die Aussagen von *Tesla*-Chef *Elon Musk,* der KI, vor allem in ihrer *starken* Form als *Artificial General Intelligence (AGI)* charakterisiert als „the most serious threat to the survival of the human race“.[8] Tritt eine solche „Fernwirkung“ (Jonas, s. u.) der Technologie ein, wäre es in der Tat ein „profound change in the history of life on Earth“ (Principle No. 20) – die radikalste Form der von der Gesellschaft geduldeten technologischen Folgen, nämlich das Ende menschlichen Lebens. Wer hier argumentiert, das Aufzeigen von Drohkulissen sei als Kulturkritik Teil der menschlichen Geschichte, habe sich aber nie bewahrheitet, möge sich vergegenwärtigen, dass in der Geschichte der Menschheit sehr wohl ganze Gesellschaften untergegangen sind, auch selbstverschuldet![9]

Musk, einer der „Endorser“ der Prinzipien, steht damit für das Dilemma, das typisch für komplexe ethische Fragen zu sein scheint. Offensichtlich verdient er durch sein Unternehmen Tesla, das für sehr fortschrittliche Technologien eben unter Nutzung von KI steht, nicht gerade wenig Geld. Es nimmt also kaum Wunder, wenn die Unterzeichner der Prinzipien Entwicklung und Einsatz von KI grundsätzlich bejahen und nicht grundsätzlich in Frage stellen. Dass erhebliche wirtschaftliche Interessen im Spiel sind, mag dafür ein Grund sein und ist bei einer Weiterentwicklung ethischer Grundsätze unbedingt mitzudenken.

Spricht Asilomar nun von „beneficial intelligence“ (Principle No. 1), so schließt sich unmittelbar die Frage an, wie *nützlich* zu definieren ist und wer bestimmt, was für wen als nützlich gilt – falls das überhaupt bestimmbar ist. Die grundlegende Prämisse des Utilitarismus ist zu diskutieren und kritisch zu hinterfragen, denn der Zweck heiligt nicht immer die Mittel. Schon gar nicht, wenn über den Zweck kein Konsens besteht, sondern die Zweckbestimmung von wenigen beherrscht wird. So ist aus Sicht der Rüstungsindustrie ein autonomes, KI-basiertes Waffensystem sicherlich sehr nützlich, wenn sich damit Geld verdienen lässt. Ebenso ist es einem totalitären Regime nützlich, wenn sich damit gezielt Dissidenten töten lassen. Potenzielle Opfer und demokratisch Gesinnte werden die Nützlichkeit anders bewerten.

Meine Überlegungen sind angeregt von der Verantwortungsethik, die Hans Jonas vor fast 40 Jahren formuliert hat. Nach Jonas' Ansatz der „Heuristik der Furcht“ ist bei menschlichen Entscheidungen zunächst von den potenziellen Folgen für die Zukunft auszugehen, die diese Entscheidung nach sich ziehen könnte. Jonas' Motiv, „die Unversehrtheit seiner [des Menschen] Welt und seines Wesens gegen die Übergriffe seiner Macht zu bewahren“,[10] und sein Imperativ „Handle so, daß die Wirkungen deiner Handlung verträglich sind mit der Per-

manenz echten menschlichen Lebens auf Erden“[11] können hilfreiche Leitbilder für den Umgang mit KI bilden. Es sollten aber nicht die einzigen bleiben, denn mit Jonas lässt sich zwar die Frage, wie eine zukünftige Welt *nicht aussehen darf,* recht gut beantworten, nicht aber die für mich wichtige Frage, wie sie *gestaltet sein soll.*

Um ethische und rechtliche Prinzipien für KI herzuleiten, sind also zwei Sichtweisen zu kombinieren: die dystopische (Jonas Ansatz der Heuristik der Furcht) und die utopische Zukunftsvorstellung. Bei letzterer stellt sich allerdings die Frage, wer die *schöne neue Welt* definiert. Wer sagt, was *gut* und *beneficial* ist und erklärt, warum das für alle gilt? Wie sieht also die Welt aus, in der wir zukünftig leben wollen? Diese Frage beantworten die Asilomar-Prinzipien nicht, und das ist problematisch. Sie unterstellen die Existenz einer allgemeingültigen und breit akzeptierten Zustimmung zu einer Zukunftskonzeption, die einem Technologiedeterminismus unterliegt, zugleich aber unbestimmt bleibt. Nehmen wir dies unkritisch hin, so laufen wir Gefahr, dass nur eine Silicon-Valley-Avantgarde bestimmt, wie wir künftig leben werden. Daraus wiederum erwächst unter anderem die Gefahr eines Totalitarismus durch eben diese Avantgarde.

Wer bestimmt also, was *gut* ist, wenn gleichzeitig die Technologie allumfassend ist, *alle* betrifft, nicht nur die, die ein bestimmtes KI-basiertes Produkt kaufen oder nutzen? Gibt es darüber Konsens? Wie findet man einen solchen Konsens auf einer globalen Ebene? Und wird es auch in der Zukunft Konsens sein, wenn Dinge nicht mehr rückgängig zu machen sind, die in der heutigen Gegenwart Konsens waren?

Wie bindend können Regeln sein?

Selbst wenn man sich jetzt auf ein Regelwerk festlegt, welches KI beispielsweise durch *Einbau* ethischer Prinzipien kontrollieren soll: Wer sagt, dass die KI selbst sich nicht über dieses Regelwerk hinwegsetzen und beginnen wird, eigenen Maßstäben, auch ethischen, zu folgen? Als Analogie mag Nordkorea dienen, ein Staat, der sich selbst aus der Weltgemeinschaft ausschließt und damit bestehenden Regelwerken entzieht und trotz der Regelwerke, Sanktionen und internationaler Ächtung Massenvernichtungswaffen entwickelt. Man könnte auch anders fragen: Wenn mit einem KI-System einem *Wesen* Intelligenz zugeschrieben und letztlich eine eigene Urteilsfähigkeit zugesprochen wird, warum sollte sich dieses Wesen dann den Regeln von Menschen unterwerfen, die ihm unterlegen erscheinen müssen? So sind auch militärisch höchst potente Staaten wie die USA und Russland von der Völkergemeinschaft nur schwer zu kontrollieren und dies auf Grund ihrer militärischen Überlegenheit. Bei KI ist zusätzlich zu bedenken, dass sie sich einer Kontrolle mit einer Aktionsgeschwindigkeit entziehen könnte, die eine menschliche (Gegen-) Reaktion unmöglich macht.

Ich habe auch keine Antwort auf diese Fragen, plädiere aber dafür, dass die Diskussion darüber geführt werden muss, und zwar von allen Betroffenen, nicht nur von profitierenden Technologen und Unternehmern. Dringend! Ethische Prinzipien für KI, bzw. Technik und Gesellschaftsentwicklung im Allgemeinen, brauchen dabei zwei klar illustrierte Szenarien als Orientierung: „die vorausgedachte Gefahr“[12] und eine Zukunftsvorstellung. Diese Zukunftsvorstellung müsste ausgehandelt werden.[13] Weniger ist hierbei an eine Utopie zu denken, sondern eher ist eine Charta der künftigen *conditio humana* zu zeichnen, die selbst nicht statisch, sondern wandelbar zu sehen ist: Was bedeutet (uns) das Menschsein, und was von dem technisch Machbaren kann noch als menschlich zugelassen werden? Denn es geht mit den Worten von Hans Jonas „nicht nur um physisches Überleben, sondern auch um Unversehrtheit des [menschlichen] Wesens.“[14]

Versuche ethischer und rechtlicher Regelungen für KI laufen ins Leere

Noch ist die dystopische Sicht nicht jedem klar, die nach Jonas' Konzeption der „Fernwirkung der Technik“ Kollateralschäden sowie Auswirkungen auf die Zukunft berücksichtigen muss. Auch der utopische Zukunftsentwurf, die Charta, ist nicht verhandelt. Das zeigt der in Asilomar angeregte Grundsatz deutlich:

> *„Teams developing AI systems should actively cooperate to avoid corner-cutting on safety standards“ (Principle No. 5).*

Die „safety standards“ oder allgemeiner, gesetzliche Regelungen und Limitierungen, gibt es noch nicht, und es wird schwer sein sie zu finden, insbesondere solange die technologische Entwicklung so viel schneller voranschreitet als eine Gesetzgebung reagieren und Rahmen setzen kann. Dieses von Soziologen in Bezug auf gesellschaftliche Anpassungs- und Veränderungsprozesse als „cultural lag“ bezeichnete Phänomen müsste eigentlich zu einem *Moratorium* führen, d. h. Entwicklungen so lange zu stoppen, bis die dringenden Fragen geklärt und auf breiter und globaler gesellschaftlicher Basis ausgehandelt sind. Das ist aber auf Grund der Dynamik technologischer Entwicklung und

Malte Rehbein

Prof. Dr. **Malte Rehbein** erforscht und lehrt formale und computergestützte Methoden und ihre Anwendungsmöglichkeiten für geistes- und kulturwissenschaftliche Fragestellungen (Digital Humanities) an der Universität Passau. Seine wissenschafts- und gesellschaftskritischen Äußerungen entstammen einer gewachsenen Sorge um die Gestaltung unserer Zukunft.
Website: *http://www.phil.uni-passau.de/dh/lehrstuhlteam/prof-dr-malte-rehbein/*

ihrer engen Verzahnung mit ökonomischen und machtpolitischen Interessen kaum vorstellbar.

Daher wirft Asilomar mehr Fragen auf als beantwortet werden. Ein Beispiel: Wir wollen keine autonom handelnden Waffensysteme. Ich zumindest nicht. Die Unterzeichner der Asilomar-Prinzipien hingegen haben offenbar nichts gegen solche Waffen. Das Einzige, das sie problematisieren, ist ein Wettrüsten mit diesen Systemen (Principle No. 18). An diesem Beispiel lässt sich gut illustrieren, wie schwierig es ist und weiterhin sein wird, so etwas wie einen gesellschaftlichen Konsens zu erwirken, wenn die Diskussion von ökonomischen und machtpolitischen Interessen dominiert wird.

Die Utopie-Prämisse

Asilomar versuchte, dies für sich im Kleinen so zu lösen:

> *„This consensus allowed us to set a high bar for inclusion in the final list [of principles]: we only retained principles if at least 90 % of the attendees [of the conference] agreed on them".*

Diese „high bar" der Inklusion gilt damit aber auch für kritische Stimmen und Minderheiten, die folglich ausgeschlossen werden! Da ist es kaum verwunderlich, dass „An arms race in lethal autonomous weapons should be avoided" (Principle No. 18) aufgenommen, aber kein grundsätzlicher Bann von autonomen Waffen ausgesprochen wurde. Hier zeigt sich das Problem der exklusiven Utopie-Prämisse, die Asilomar angewandt hat: Es erfordert 90 % Zustimmung, um einen Technologieeinsatz zu verbieten. Eine Prämisse, die dem europäischen, leider zunehmend ausgehöhlten Vorsorgeprinzip entspräche, würde hingegen 90 % Zustimmung erfordern, um einen solchen Einsatz zu erlauben. Die Frage müsste nach meiner Auffassung also lauten: Sind 90 % der Menschen für autonome Waffensysteme? Dann könnte man sie erlauben. Aber sie sollte eben nicht heißen: Findet sich eine Minderheit von 10 %, die einen Bann verhindert?

Die Schärfe der Problematik steckt weiterhin oft im Detail. So ist die Formulierung von „The application of AI to personal data must not unreasonably curtail people's real or perceived liberty" (Principle No. 13) äußerst subtil: Der Einsatz von KI darf gemäß den Unterzeichnern von Asilomar also durchaus menschliche Rechte beschneiden. Er solle es nur nicht in einer Weise tun, die als „unreasonably" (wahlweise zu übersetzen mit unvernünftig, unangemessen oder übertrieben)[15] charakterisiert wird. Nach meinem Verständnis einer demokratischen Gesellschaft sind Freiheitsrechte ein hohes Gut und gesetzlich geregelt. Eine Einschränkung ist den Gerichten vorbehalten – unter hohen Auflagen. Eine Abweichung davon ist Notstand und dessen Dauerzustand Totalitarismus.

Der Einsatz von KI, auch in ihrer schwachen Form, kann globale Auswirkungen haben, insbesondere wenn man die KI mit exekutiven Mitteln wie Waffen ausstattet oder ihr Zugriff auf kritische Infrastruktur erlaubt. Jonas spricht davon, dass die Reichweite menschlichen Handelns und daher menschlicher Verantwortung nicht mehr eng umschrieben werden kann.[16] Daher bedarf es globaler Regulierung und Mitspracherecht auch derer, die nicht unmittelbar zu den Profiteuren zählen. Der bereits herausgebildete *digital divide* innerhalb von Gesellschaften und zwischen Gesellschaften darf nicht weiter dazu führen, dass eine Gruppe über die Zukunft einer anderen bestimmt. Zudem muss die Bereitschaft da sein, nicht nur den Einsatz von KI in bestimmten Szenarien zu ächten, sondern gegebenenfalls schon die Entwicklung solcher Technologieformen zu unterbinden, wenn das Technologierisiko zu groß ist, was vor allem bei starker KI der Fall zu sein scheint. Eine zentrale Frage wird weiterhin sein, wie unsere gegenwärtigen Wirtschafts- und Gesellschaftssysteme umzugestalten sind, damit sie eine globale und auch langfristige, nachhaltige Perspektive begünstigen und nicht von den kurzfristigen technologischen und ökonomischen Interessen weniger dominiert werden.

Anmerkungen und Referenzen

1 *https://futureoflife.org/ai-principles/*
2 *https://futureoflife.org/bai-2017/*
3 *https://futureoflife.org/2017/01/17/principled-ai-discussion-asilomar/*
4 *https://futureoflife.org/ai-principles/*
5 *Stop Autonomous Weapons (2017) Slaughterbots. https://www.youtube.com/watch?v=9CO6M2HsoIA&t=.*
6 *Neben den Daten, die Menschen von sich direkt oder indirekt preisgeben (etwa durch soziale Medien, Nutzung von Suchmaschinen, Internet of Things) ist die Bedeutung einer Sammlung von Daten durch Sensoren (angefangen mit Überwachungskameras) nicht zu unterschätzen. Auch aktuelle Entwicklungen in Richtung „Neuro-Daten" sind zu beobachten: Schnabel U (2017) Attacke aufs Gedankenstübchen; Forscher fordern, Neurotechniken besser zu regulieren. Die Zeit 30.11.2017, S. 37.*
7 *Assheuer T (30.11.2017) Die Big-Data-Diktatur. Die Zeit 30.11.2017, S. 47.*
8 *Gibbs S (2014) Elon Musk: artificial intelligence is our biggest existential threat. The Guardian 27.10.2014, https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat.*
9 *Diamond J (2010) Kollaps; Warum Gesellschaften überleben oder untergehen. 4. Aufl. Fischer-Taschenbuch-Verlag, Frankfurt am Main.*
10 *Jonas H (1979) Das Prinzip Verantwortung; Versuch einer Ethik für die technologische Zivilisation. Insel-Verlag, Frankfurt am Main, S. 9.*
11 *Ebd., S. 35.*
12 *Ebd., S. 7.*
13 *Vgl. etwa die Leitfrage „Welche digitale Gesellschaft wollen wir werden?" der Bonner Gespräche zur politischen Bildung im März 2018 „Künstliche Intelligenz, Big Data und digitale Gesellschaft – Herausforderungen für die politische Bildung" (http://www.bpb.de/veranstaltungen/format/kongress-tagung/242756/kuenstliche-intelligenz-big-data-und-digitale-gesellschaft-herausforderungen-fuer-die-politische-bildung).*
14 *Jonas, Prinzip Verantwortung, S. 8.*
15 *In der nachträglich veröffentlichten Übersetzung der Prinzipien ins Deutsche wurde „unangemessen" gewählt.*
16 *Jonas, Prinzip Verantwortung, S. 15.*