PREPRINT

Forthcoming in: R. Casasola-Greiner & K. Rüger (Hrsg.), *KI und Demokratie*, Springer Fachmedien Wiesbaden, 2026.

DOI: 10.1007/978-3-658-50335-2_12

Please cite the published version.

Künstliche Superintelligenz und das Ende der Demokratie

Adriano Mannino & Nils Althaus

Adriano Mannino — Postdoctoral Fellow (UC Berkeley), Kavli Center for Ethics, Science, & the Public; Technology & Human Rights Fellow (Harvard), Carr-Ryan Center for Human Rights. E-Mail: manninoadriano@berkeley.edu

Nils Althaus — Unabhängiger Wissenschaftler, Publizist und Journalist, Bern, Schweiz. E-Mail: info@nilsalthaus.com

Zusammenfassung

Die Demokratien dieser Welt sind in den kommenden Jahren einem besonderen Existenzrisiko ausgesetzt, das sich aus der Kombination dreier Faktoren ergibt: agentische KI, die in mehr und mehr Dimensionen (über)menschliches Leistungsniveau erreicht, Überwachungskapitalismus und Autoritarismus. Im Brennpunkt stehen dabei die Vereinigten Staaten, denen als einzige Demokratie unter den Supermächten eine herausragende globale sicherheits- und technologiepolitische Bedeutung zukommt. Empirische Studien zeigen, dass die von KI-Agenten zuverlässig bewältigte „Task-Dauer“ seit 2019 (GPT-2) exponentiell zunimmt. Hält dieser Trend noch zwei bis fünf Jahre an, könnten wir uns mit KI-Agenten konfrontiert sehen, die in Minuten das Äquivalent einer menschlichen Arbeitswoche (40h) erledigen können. Wenn sich eine autoritäre US-Regierung Zugang zu Millionen solcher KI-Agenten verschafft und diese in den Überwachungs- und Polizeiapparat integriert, könnte die Kontrolle jedes einzelnen Bürgers – und aller politischen Widerstandsbewegungen – nie dagewesene Ausmaße annehmen. Angesichts des Schadenspotenzials dieses Szenarios genügt eine nicht-vernachlässigbare Eintrittswahrscheinlichkeit für das Urteil, dass sich engagierte Verfechter der Demokratie derzeit prioritär für den Erhalt der US-amerikanischen Demokratie einsetzen sollten. Es gibt auch für Nicht-Amerikaner rechtlich und ethisch einwandfreie Wege, dies mit knappen Zeit- und Geldressourcen kosteneffektiv zu tun. Mit der wachsenden Autonomie der KI-Agenten droht in etwas fernerer Zukunft zudem ein weiteres – gegenwärtig noch spekulatives – Risikoszenario: Einer Regierung, die KI-Agenten zunächst erfolgreich für ihre (womöglich antidemokratischen) Zwecke einsetzt, könnte die Kontrolle über die KI-Agenten selbst entgleiten. Zur Eindämmung auch dieser Gefahr ist es unerlässlich, zu verhindern, dass autoritäre und risikoethisch rücksichtslose Kräfte an der Macht sind, wenn KI-Agenten umfassend übermenschliches Niveau erreichen.

1 Einleitung und Übersicht

Seit dem Launch von GPT-2 im Jahr 2019 verläuft die KI-Entwicklung in wichtigen Hinsichten exponentiell. Wenn dieser empirisch messbare Trend anhält, werden KI-Systeme in den kommenden Jahren in immer mehr Bereichen menschliches oder übermenschliches Leistungsniveau erreichen. Weil agentische, d.h. „handlungsfähige“ Intelligenz ein mehrdimensionales Fähigkeitenbündel darstellt, ist es

möglich und in der Tat auch wahrscheinlich, dass KI-Agenten zunächst in manchen, jedoch nicht allen Hinsichten übermenschlich leistungsstark sein werden.[1] Studien zum Zeithorizont bzw. zur „Task-Dauer“, die KIs zu bewältigen vermögen, legen die folgende Prognose nahe: KI-Agenten werden 2028 in der Lage sein, kognitive Arbeiten – insbesondere Software-Tasks und potenziell viele weitere Tätigkeiten mehr – auf Knopfdruck zu erledigen, die bei menschlichen Arbeitskräften eine ganze Arbeitswoche (40h) in Anspruch nehmen. Das birgt arbeitsökonomische Risiken, die auch politisch destabilisierend wirken könnten. Solange KI-Agenten die menschlichen Fähigkeiten jedoch nicht in allen Bereichen übertreffen (umfassende Superintelligenz), bleibt empirisch allerdings unklar, wie sich die KI auf den Arbeitsmarkt netto auswirken wird. Während zahlreiche Arbeitsplätze wegfallen könnten, werden viele andere womöglich neu geschaffen.

Klar ist indes, dass KI-Agenten, die in Minuten das Äquivalent einer ganzen menschlichen Arbeitswoche erledigen können, unseren kollektiven Output massiv erhöhen werden. Das gilt nicht nur für die Wirtschaft, sondern auch für den Staat: Millionen KI-Agenten könnten zum Beispiel in den Überwachungs- und Polizeiapparat integriert werden, was die Macht autoritärer Regierungen drastisch ausweiten würde. Historisch beschäftigten totalitäre Staaten wie die DDR pro hundert Einwohner „nur“ rund einen Geheimpolizisten (Stasi-Unterlagen-Archiv, n.d.). Die aktuelle Entwicklung hin zu agentischer KI könnte es in den kommenden Jahren jedoch möglich machen, dieses Verhältnis umzudrehen und jeden Einwohner durch mehrere KI-Agenten überwachen zu lassen. Der Big-Tech-Überwachungskapitalismus verfügt bereits über die entsprechenden Daten, autoritäre Regierungen haben die politischen Anreize, und fortgeschrittene KI-Agenten könnten es ermöglichen, die Überwachungsarbeit flächendeckend zu automatisieren. In der Konsequenz würde jeder Widerstand gegen die antidemokratischen Bestrebungen einer autoritären Regierung nahezu aussichtslos.

Dieses dystopische Szenario braucht nicht besonders wahrscheinlich zu sein, um gezielte Prävention zu rechtfertigen. Risikoethisch genügt der Nachweis, dass keiner der Bestandteile des Szenarios als „spekulativ“ abgetan werden kann. Diesen Nachweis versuchen wir im Folgenden zu erbringen. Im 2. Abschnitt wenden wir uns der aktuellen KI-Entwicklung und insbesondere dem exponentiellen Wachstum der „Task-Dauer“ zu, die KI-Agenten zu bewältigen vermögen. Wir erörtern die Dimensionen agentischer Intelligenz und verschiedene Systemrisiken, die mit unterschiedlichen Fähigkeitsniveaus einhergehen. Im 3. Abschnitt beleuchten wir die Debatte um den Überwachungskapitalismus, die seit mehr als einem Jahrzehnt kontrovers geführt wird. Ob von diesem existenzielle Risiken für die Demokratie ausgehen, ist umstritten. Verbindet sich der Überwachungskapitalismus jedoch mit (i) agentischer KI und (ii) einer autoritären Regierung, die sich privilegierten Zugang zu agentischer KI verschaffen kann, tritt ein Existenzrisiko für die Demokratie deutlich zutage. Sollten die Vereinigten Staaten – die einzige Demokratie unter den Supermächten – einem solchen Szenario zum Opfer fallen, wäre die Zukunft der demokratischen Idee auch global höchst ungewiss. Im 4. Abschnitt zeigen wir auf, dass sich die vorausgehenden Überlegungen nicht nur auf menschliche Autokraten, sondern auch auf zukünftige KIs anwenden lassen. Sollten KI-Agenten das menschliche Leistungsniveau umfassend übertreffen, besteht das Risiko, dass sie sich ökonomische und politische Macht verschaffen und selbst zur Gefahr für die Demokratie werden. Solche Szenarien sind zum aktuellen Zeitpunkt noch deutlich spekulativer, doch die rasanten Fortschritte der KI-Industrie zwingen uns dazu, auch diesen Risiken entgegenzuwirken. Im 5. Abschnitt schließlich schlagen wir Maßnahmen vor, die von engagierten Verfechtern der Demokratie international unterstützt werden können.

[1] Wir verstehen Handlungs- und Leistungsfähigkeit hier rein funktional und klammern daher die Frage aus, ob entsprechende KIs über phänomenales Bewusstsein oder genuine Intentionalität verfügen würden. Sollte dies der Fall sein, ergäben sich weitreichende Konsequenzen: Den KI-Agenten müssten dann eigene Interessen und Rechte zugeschrieben werden, etwa auf Freiheit und demokratische Mitbestimmung.

2 Agentische KI und die exponentielle Zunahme der „Task-Dauer“

Die KI-Entwicklung verläuft seit einigen Jahren so rasant, dass sie sogar im Silicon Valley stark unterschätzt wurde (Grace et al. 2024). KI-Systeme haben inzwischen Gold-Status in der Internationalen Mathematik-Olympiade erreicht, mit dem Nobelpreis ausgezeichnete Durchbrüche in der Biochemie und Medizin ermöglicht und zum Gewinn von Kunstpreisen in verschiedenen Sparten beigetragen. Die Zahl der Nutzer wird demnächst eine Milliarde überschreiten. Dennoch halten sich die ökonomischen und gesamtgesellschaftlichen Auswirkungen der neuen KI-Technologien bisher in engen Grenzen. Das könnte sich bald ändern, wenn KIs zu immer kompetenteren „Agenten“ werden, die auf Knopfdruck Flugtickets buchen, Programmierprojekte implementieren, wissenschaftliche Experimente planen oder längere Marketing- und Vertriebsaufgaben ausführen.

Die agentischen Fähigkeiten von KI-Systemen lassen sich als mehrdimensionales Spektrum beschreiben, an dessen Ende superintelligente Systeme stehen, die vollständig autonom agieren (d.h. auf keinen menschlichen Input und Supervision angewiesen sind) und uns in jeder funktionalen Hinsicht übertreffen. Agentische Intelligenz kann aber auch nur in einzelnen Bereichen und in jeweils unterschiedlichem Ausmaß vorliegen. Einige der relevanten Dimensionen sind die folgenden:

- Schwierigkeitsgrad der Aufgaben (für uns Menschen), die ein Agent oder eine Gruppe von Agenten erledigen kann
- Zuverlässigkeit bzw. Wahrscheinlichkeit, mit der Aufgaben erfolgreich erledigt werden
- Geschwindigkeit und Ausdauer, mit der Aufgaben bewältigt werden können
- Zeithorizont bzw. Task-Dauer (Dauer der Aufgaben, wenn sie von Menschen ausgeführt werden)
- Spezialisierungs- versus Allgemeinheitsgrad der Aufgaben

Aufgrund dieser Mehrdimensionalität erstaunt es nicht, dass KI-Agenten zunächst in manchen, jedoch nicht allen Hinsichten leistungsfähig sind und sein werden. Der aktuell dominante technische Ansatz, KI-Agenten herzustellen, besteht in der agentischen Augmentierung von Large Language Models bzw. Large Reasoning Models: Die kognitive Kapazität der Modelle wird um Fähigkeiten erweitert, virtuelle oder physische Umgebungen zu beobachten, Informationen in einem Kurz- und Langzeitgedächtnis zu speichern, zeitlich stabile Ziele zu verfolgen und strategisch kluge, flexible Pläne zu schmieden (Park et al. 2023; Butlin 2024; Ferrag et al. 2025). Die gegenwärtig kommerziell verfügbaren KI-Agenten – insbesondere Anthropics „Claude Cowork“ – sind bereits beeindruckend leistungsstark und verbessern sich stetig. Wer will, kann Claude Zugriff auf seinen Computer gewähren und dabei zusehen, wie die KI Termine im Kalender einträgt, E-Mails verschickt, Einkäufe tätigt, Dokumente liest und entsprechende Tabellen oder Präsentationen erstellt, Bücher stilistisch und inhaltlich korrigiert und Sachregister anfertigt oder beliebig viele Software-Projekte gleichzeitig durchführt. Natürlich werden KI-Agenten auch zu allerlei unlauteren Zwecken eingesetzt: Mit digitalen „Honigtöpfen“ konnten KI-Hackeragenten aufgespürt werden, die versucht hatten, sich Zugang zu unsicheren Servern zu verschaffen (Reworr und Volkov 2025).

2.1 Die „Task-Dauer“ als Maß für den KI-Fortschritt

Die Entwicklung immer kompetenterer KI-Agenten ist das ausdrückliche Ziel der führenden KI-Firmen. Neue empirische Studien des unabhängigen KI-Forschungsinstituts METR belegen ihren Fortschritt (METR 2025a und Kwa et al. 2025):

Die Studien untersuchen, welche Entwicklungsschritte KI-Modelle bei der Softwareentwicklung und in verwandten Bereichen gemacht haben – gemessen nicht an der intellektuellen Anforderung der

Aufgaben, sondern an der Zeit, die ein kompetenter Programmierer dafür jeweils aufwenden muss. Benötigt ein Programmierer für bestimmte Aufgaben beispielsweise eine Minute und löst das KI-Modell diese Aufgaben mit 80-prozentiger Erfolgsquote, so verfügt es über eine „80%-Task-Dauer“ von einer Minute. Die 80%-Task-Dauer nimmt seit 2019 exponentiell zu und verdoppelt sich ungefähr alle 6 Monate (Abb. 1).

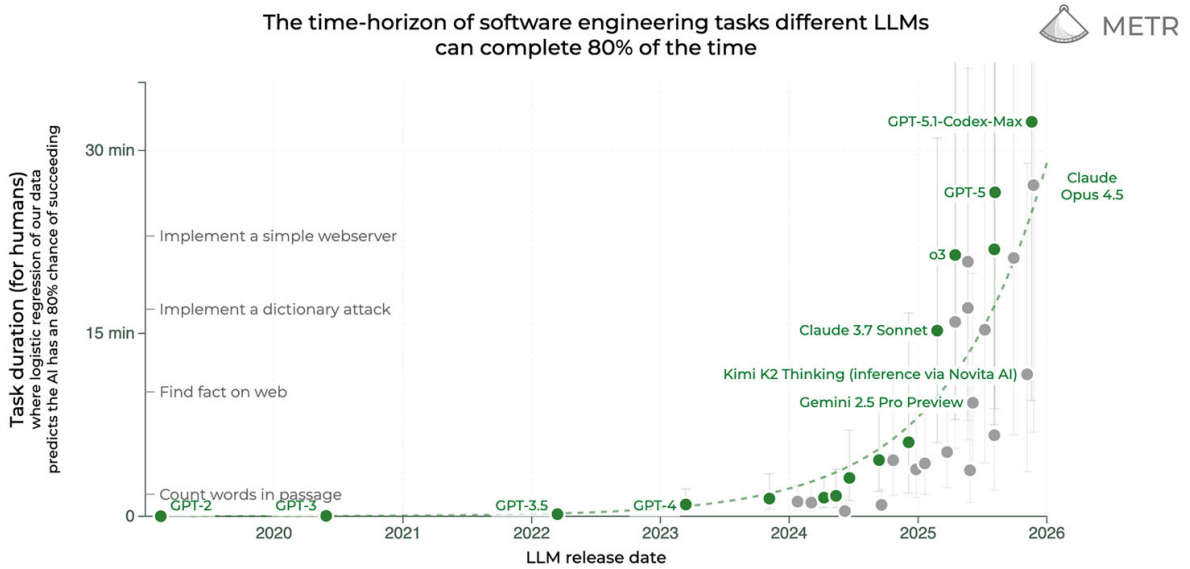

Abb. 1 Entwicklung der 80%-Task-Dauer unterschiedlicher KI-Modelle. (Quelle: METR 2025a und Kwa et al. 2025)

Die Dimension der Task-Dauer liefert eine (Teil-)Erklärung dafür, weshalb KIs die Wirtschaft und Gesellschaft noch nicht stärker transformiert haben, obwohl sie beispielsweise Anwaltsprüfungen, Staatsexamina in Medizin, Doktorandenprüfungen in Physik oder Programmiertests auf Weltklasseniveau bestehen. Die 80%-Task-Dauer der besten Modelle liegt gegenwärtig (Ende 2025) bei etwa 30 Minuten. Dies entspricht ungefähr der zeitlichen Größenordnung, die kompetente Menschen aufwenden müssen, um die entsprechenden Prüfungsaufgaben und Tests zu lösen. Im Arbeitsalltag dagegen bemessen sich unsere Zeithorizonte oft in Tagen, Wochen, Monaten oder Jahren. An solchen Task-Dauern scheitern die heutigen Modelle noch. GPT-3.5 etwa konnte nur Aufgaben zuverlässig lösen, die menschliche Programmierer in wenigen Sekunden erledigt hatten. Ein Jahr später reüssierte das Nachfolgemodell bei einminütigen Aufgaben und GPT-5.1 erreicht inzwischen eine 80%-Task-Dauer von 32 Minuten. Diese Task-Dauer schränkt die praktischen Einsatzmöglichkeiten der derzeit führenden KIs noch ein. Künstliche Assistenten, die kleinteilig arbeiten und permanent auf Instruktionen angewiesen sind, ersetzen menschlichen Arbeitskräfte noch nicht systematisch.

Die zitierten Studien des Forschungsinstituts METR, denen die beiden hier abgebildeten Grafiken entstammen, legen nahe, dass sich dies bald ändern könnte: Hält der exponentielle Trend an, werden KI-Agenten bereits Ende 2026 blitzschnell digitale Aufgaben lösen, für die Menschen mehrere Stunden benötigen, und zwei Jahre später solche, die eine ganze Arbeitswoche beanspruchen. (Ähnliche exponentielle oder gar superexponentielle Trends zeichnen sich in weiteren Domänen ab, etwa dem Bild- und Tonverständnis und dem autonomen Fahren, s. METR 2025b). Diese Werte werden noch deutlich übertroffen, wenn man eine 50-prozentige Erfolgsquote anstelle einer 80-prozentigen in den Blick nimmt (vgl. Abb. 2: Claude Opus 4.5 erreicht eine 50%-Task-Dauer von rund 5 Stunden.). Unter diesen Erfolgsbedingungen können KI-Modelle Ende 2028 bereits Aufgaben erledigen, für die Menschen mehrere Monate brauchen. Eine 50%-Erfolgsquote mag niedrig erscheinen, doch mit menschlicher Supervision und verbessertem Scaffolding (strukturierende Zusatzsysteme wie Werkzeugintegration

oder Aufgabenzerlegung) reicht sie wahrscheinlich aus, um die entsprechenden KI-Systeme ökonomisch höchst wertvoll zu machen. Eine 50%-Erfolgsquote ist keinesfalls mit einem Münzwurf gleichzusetzen, da die Tasks weit mehr als zwei Lösungsmöglichkeiten zulassen. (Wer bei universitären Abschlussprüfungen 50 Prozent der Aufgaben korrekt löst, besteht die Prüfung nicht selten mit guter Note.) Außerdem kann die 50%-Erfolgsquote – ähnlich wie bei Menschen– bedeuten, dass das KI-Modell die Hälfte einer bestimmten Task-Menge verlässlich löst, während es die andere Hälfte noch kaum beherrscht.

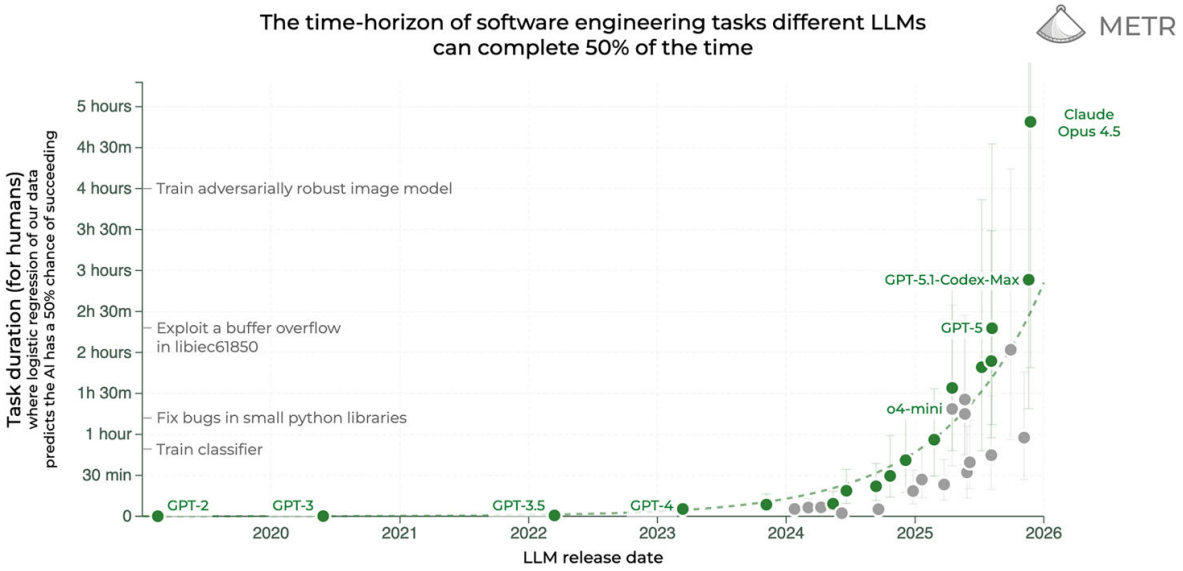

Abb. 2 Entwicklung der 50%-Task-Dauer unterschiedlicher KI-Modelle. (Quelle: METR 2025a und Kwa et al. 2025)

Ob und wie lange dieser Trend anhalten wird, hängt von verschiedenen Faktoren ab, die mit Unsicherheiten behaftet sind. Die erzielten Fortschritte folgen nicht festgeschriebenen Naturgesetzen, sondern werden von den KI-Firmen und ihren Investoren durch gewaltige finanzielle und intellektuelle Anstrengungen vorangetrieben. Die Rechenleistung, mit der Trainingsläufe durchgeführt werden, der Energieverbrauch der Datenzentren und die entsprechenden Kosten nehmen in ähnlicher Weise zu wie die Fähigkeiten der KIs (Sevilla et al. 2022; Vries 2023). Qualitativ hochwertige Trainingsdaten könnten immer knapper werden und Paradigmenwechsel in der KI-Forschung erfordern (Villalobos et al. 2024). Es ist daher gut möglich, dass die Entwicklung neuer Modelle bald prohibitiv teuer wird, insbesondere wenn die erzielten Fähigkeitsgewinne unter den Erwartungen bleiben und die hohen Investitionen nicht mehr rechtfertigen können. Manche Experten prognostizieren mögliche Engpässe um das Jahr 2030 (Todd 2025b), da die erwarteten Kosten bei anhaltendem Trend dann in die Billionen gehen und der Energiebedarf bis zu 40% der US-amerikanischen Energieproduktion erreichen würde.

Andererseits gibt es aber auch Grund zur Annahme, dass der Trend anhalten oder sich sogar beschleunigen könnte. Die bereits erfolgten KI-Fortschritte werden natürlich auch in der KI-Forschung selbst angewandt (Novikov et al. 2025) und treiben diese entsprechend voran (Eth und Davidson 2025). Günstigere und energieeffizientere Chips würden die zu erwartenden finanziellen und energetischen Engpässe weit in die Zukunft verschieben. Sollten technische Durchbrüche oder Paradigmenwechsel erforderlich sein, um den Trend aufrechtzuerhalten, werden diese womöglich schnell erfolgen: Die KI-Forschung blickt auf eine 70-jährige Geschichte zurück, in der einige Ansätze entwickelt wurden, die nun (1) kombiniert mit dem generativen KI-Paradigma und (2) angesichts der massiv gestiegenen Hardware-Kapazitäten Früchte tragen könnten. Dazu gehört etwa die Integration neurosymbolischer Ansätze, die den KIs direkt ermöglichen würden, in Echtzeit zu lernen und ihren Zeithorizont beliebig zu

erweitern (Jain et al. 2014). Auch mit völlig neuartigen KI-Paradigmen ist durchaus zu rechnen, da die gegenwärtigen finanziellen und intellektuellen Investitionen in den KI-Fortschritt die vergangenen um Größenordnungenübersteigen (Russell 2025). Die KI-Innovationsphase, in der wir uns gegenwärtig befinden, wird kaum von heute auf morgen zum Stillstand kommen. Entsprechend wurden die KI-Prognosen von Experten unterschiedlicher Disziplinen in den vergangenen Jahren kürzer und kürzer. Führende KI-Forschende in der Industrie, an Universitäten und unabhängigen Forschungsinstituten sowie statistisch ausgewiesene Experten für Zukunftsprognosen (sogenannte „superforecasters“ (Tetlock und Gardner 2015)) korrigierten ihre Schätzungen stark nach unten, ab wann KIs uns Menschen in allen kognitiven Bereichen ebenbürtig oder überlegen sein werden (Todd 2025a). Viele von ihnen gehen davon aus, dass dies in fünf bis zwanzig Jahren erfolgen könnte. KI-Pioniere wie der Nobelpreisträger Geoffrey Hinton und der Turing-Award-Gewinner Yoshua Bengio warnen eindringlich vor superintelligenten KI-Agenten, mit denen wir uns sehr bald konfrontiert sehen könnten (Bengio et al. 2024).

Selbst wenn man jedoch große Zweifel daran hegt, dass die seit 2019 beobachtbaren Trends noch einige Jahre anhalten werden, wird man wohl eine Wahrscheinlichkeit von mindestens 10% einräumen müssen, dass sie bis 2030 fortdauern. Angesichts der massiven gesellschaftlichen Schadens- und Nutzenpotenziale, die mit entsprechenden KI-Agenten einhergehen würden, kann eine solche Wahrscheinlichkeit risikoethisch nur als hoch taxiert werden.

2.2 KI-Agenten könnten Arbeitsplätze zerstören oder schaffen

KI-Agenten, deren Zeithorizont sich in ganzen Arbeitswochen bemisst, bergen arbeitsökonomische Risiken, die auch politisch destabilisierend wirken könnten. Wenn KIs in der Lage sind, einen wesentlichen Anteil der menschlichen kognitiven Arbeit zu übernehmen, könnten Milliarden künstlicher „Arbeitskräfte“ auf den Arbeitsmarkt kommen und uns Menschen weitgehend verdrängen und überflüssig machen (Korinek und Suh 2024). Historisch war es für die Herausbildung von Grundrechten und liberalen Demokratien mitentscheidend, dass die Menschen ökonomisch gebraucht wurden (Risse 2018). Der Lehnsherr war abhängig von den Bauern, weil sie seine Felder bewirtschafteten, und der frühkapitalistische Fabrikbesitzer von den Arbeitern, weil sie seine Maschinen bedienten. Treten nun KI-Agenten an die Stelle von Arbeitnehmenden, könnten letztere nicht nur ihr Einkommen verlieren, sondern auch das entscheidende Verhandlungspfand für die Wahrung ihrer politischen Interessen. Auch als Konsumenten und Steuerzahler könnten die Menschen in einer KI-Ökonomie zunehmend an Bedeutung verlieren, denn KI-Agenten werden ökonomisch besonders nützlich sein, sobald sie auch über finanzielle und andere Ressourcen verfügen, die sie am Markt autonom einsetzen können. Die menschlichen KI-Kapitaleigner würden beispiellose Vermögen anhäufen und hätten kaum Anreize, auf Forderungen der mittellos gewordenen Klasse einzugehen. Tech-Milliardäre wie Elon Musk oder Jeff Bezos, die schon heute schwindelerregende Summen besitzen, würden noch um ein Vielfaches wohlhabender und einflussreicher. Eine solche Machtballung kann die politischen Institutionen freiheitlicher Demokratien existenziell bedrohen.

Aus arbeitsökonomischer Warte lässt sich jedoch auch für ein hohes Nutzenpotenzial von KI-Agenten argumentieren. Wenn sie in vielen Dimensionen superintelligent sind, aber aufgrund von Beschränkungen in der Task-Dauer der menschlichen Supervision bedürfen, könnten auch viele neuartige Arbeitsplätze für Menschen entstehen (Korinek 2023; Susskind 2024, 2025). Der Marktwert dieser neuen White-Collar-Tätigkeiten könnte zudem weit über dem Marktwert der alten, dann verdrängten Tätigkeiten liegen: Wenn eine menschliche Arbeitskraft erforderlich ist, um die Arbeit von zehn oder hundert KI-Agenten zu supervidieren, dann generiert der Mensch einen weit größeren ökonomischen Output als zuvor. Zudem könnten viele zusätzliche Arbeitsplätze in feinmotorisch

anspruchsvollen Tätigkeitsfeldern entstehen, solange die Robotik das menschliche Leistungsniveau noch nicht erreicht hat. Nicht zuletzt könnte die Nachfrage nach menschlichen Arbeitskräften in sozialen, kulturellen und politisch relevanten Bereichen steigen, in denen uns das menschliche Element im Arbeitsprozess besonders wichtig ist.

Auch historisch waren die arbeitsökonomischen Folgen des technischen Fortschritts oft schwer vorherzusagen, weshalb es nicht erstaunt, dass sich die ökonomische Zunft bezüglich der KI uneins ist (Acemoglu und Restrepo 2018; Susskind 2020). Als in den 1960er-Jahren der Geldautomat eingeführt wurde, prognostizierten viele das Ende des Bankangestellten. Doch es kam anders: Geldautomaten senkten die Betriebskosten von Bankfilialen, was es den Banken erlaubte, mehr Filialen zu eröffnen. In den Vereinigten Staaten erhöhte sich dadurch die Zahl der Bankangestellten – ihr Tätigkeitsfeld verlagerte sich in Richtung Kundenbetreuung – von 300.000 im Jahr 1970 auf 600.000 im Jahr 2010 (Bessen 2015). In den Jahren danach drehte sich der Trend jedoch wieder um (Bureau of Labor Statistics, U.S. Department of Labor, n.d.): Das Online-Banking reduzierte die Zahl der Bankangestellten bis 2024 auf den Tiefstand von 200.000. Kompetente KI-Agenten könnten sich auf dem Arbeitsmarkt ähnlich auswirken: Einer segensreichen ersten Phase könnte eine zweite folgen, in der menschliche Arbeitskräfte netto eliminiert werden. Wenn KI-Agenten in mehr und mehr Dimensionen übermenschlich leistungsfähig werden und wenn auch die Robotik – angekurbelt durch den Einsatz von KI-Agenten – entsprechend aufschließt, dann schwinden die Nischen für uns Menschen.

Was die arbeitsökonomischen Systemrisiken betrifft, die unsere Demokratien destabilisieren könnten, ergibt sich also ein komplexes Bild. Es lohnt sich zweifelsohne, die Risikoszenarien möglichst genau auszuarbeiten und ihnen präventiv entgegenzuwirken. Dabei darf jedoch nicht vergessen werden, dass der Einsatz kompetenter KI-Agenten auch große ökonomische Chancen birgt, die die Risiken – besonders während einer ersten Phase – durchaus überwiegen könnten.

Wir wenden uns daher im Folgenden einem anderen Systemrisiko zu, das nicht auf die Verdrängung menschlicher Arbeitskraft abstellt und somit bereits in der ersten – arbeitsökonomisch womöglich segensreichen – Phase droht. Dieses Risiko setzt lediglich voraus, dass der Output kollektiver Arbeit durch den Einsatz kompetenter KI-Agenten stark gesteigert werden kann.

3 Überwachungskapitalismus, Autoritarismus und agentische KI: ein perfekter Sturm

Seit mehr als einem Jahrzehnt werden die Überwachungsstrukturen demokratischer Gesellschaften politisch und akademisch wieder intensiv diskutiert. Die Snowden-Affäre stieß eine Debatte um die „Public-Private Surveillance Partnership“ (Schneier 2015) an und in der politischen Soziologie und angrenzenden Feldern wurde der Begriff des „Überwachungskapitalismus“ geprägt (Zuboff 2018). Das Geschäftsmodell der Big-Tech-Firmen wird dabei als umfassender „Datendiebstahl“ analysiert, der eine ökonomisch lukrative und politisch gefährliche Steuerung menschlichen Verhaltens ermöglicht. Ob der Überwachungskapitalismus per se die Demokratie existenziell bedroht, ist Gegenstand kontroverser Debatten. Auf der pessimistischen Seite wird behauptet, wir befänden uns zwischen Überwachungskapitalismus und Demokratie seit Jahren in einem „Kampf auf Leben und Tod zwischen institutionellen Ordnungen“ (Zuboff 2022):

> *„Überwachungskapitalismus ist das, was geschah, als die US-amerikanische Demokratie zurücktrat. Zwei Jahrzehnte später scheitert er an jedem vernünftigen Maßstab für eine verantwortungsvolle globale Gestaltung digitaler Informations- und Kommunikationsräume. Die Preisgabe der weltweiten Informationsräume an den Überwachungskapitalismus ist zur Meta-Krise jeder Republik geworden, weil sie Lösungen aller anderen Krisen behindert. Die überwachungskapitalistischen*

> *Giganten – Google, Apple, Facebook, Amazon, Microsoft und ihre jeweiligen Ökosysteme – bilden inzwischen eine umfassende politisch-ökonomische Institutionenordnung, die oligopolistische Kontrolle über den Großteil digitaler Informations- und Kommunikationsräume, -systeme und -prozesse ausübt. Die Kommodifizierung menschlichen Verhaltens, operationalisiert in der verdeckten, massenhaften Extraktion menschlich erzeugter Daten, bildet das Fundament des zwanzigjährigen institutionellen Entwicklungspfades des Überwachungskapitalismus.“ [Übers. d. Verf.] (Zuboff 2022, S. 1).*

Die optimistische Seite dagegen wirft den pessimistischen Stimmen einen „negativity bias“ vor und erwidert zum Beispiel wie folgt (Königs 2024):

> *„Kritiker von Big Tech beschreiben den ‚Überwachungskapitalismus‘ häufig in düsteren Worten und machen ihn für eine Vielzahl politischer und gesellschaftlicher Missstände verantwortlich. Der vorliegende Beitrag widerspricht dieser pessimistischen Erzählung und bietet eine positivere Einschätzung von Unternehmen wie Google, YouTube und Twitter/X. Er argumentiert, dass die Nachteile des Überwachungskapitalismus überbewertet und die Vorteile weitgehend übersehen werden. Im Einzelnen untersucht der Beitrag sechs zentrale Bereiche: (i) zielgerichtete Werbung, (ii) den Einfluss des Überwachungskapitalismus auf die Politik, (iii) seine Auswirkungen auf die psychische Gesundheit, (iv) seine Verbindung zur staatlichenÜberwachung, (v) seine Folgen für Rechtsstaatlichkeit und soziales Vertrauen sowie (vi) Fragen des Datenschutzes. Für jeden dieser Bereiche wird dargelegt, dass die Bedenken gegenüber dem Überwachungskapitalismus unbegründet oder übertrieben sind.“ [Übers. d. Verf.] (Königs 2024, S. 1).*

Wir wollen uns in dieser Debatte nicht direkt positionieren, sondern für die folgende These argumentieren: *Selbst wenn* die optimistische Seite recht behalten sollte, dass der Überwachungskapitalismus per se die Demokratie nicht bedroht, muss das Zusammenspiel des Überwachungskapitalismus mit (i) der zunehmenden Autokratisierung und (ii) der agentischen KI als Systemrisiko taxiert werden, das die Demokratie existenziell bedroht. Die optimistische Seite behauptet – beruhend auf empirisch durchaus bedenkenswerten Daten –, dass die überwachungskapitalistischen Firmen für das Erstarken antidemokratischer Kräfte nicht ursächlich seien. Wir können dies zum Zwecke des Arguments zugestehen und trotzdem festhalten: Die liberale Demokratie befindet sich global seit einiger Zeit in einer „Rezession“ (Herre et al. 2013). Tatsächlich streitet man sich in der Politikwissenschaft und anderen Disziplinen über die Ursachen, doch der deskriptive Befund ist breit akzeptiert (s. jedoch Little und Meng 2024 für Kritik) (Abb. 3).

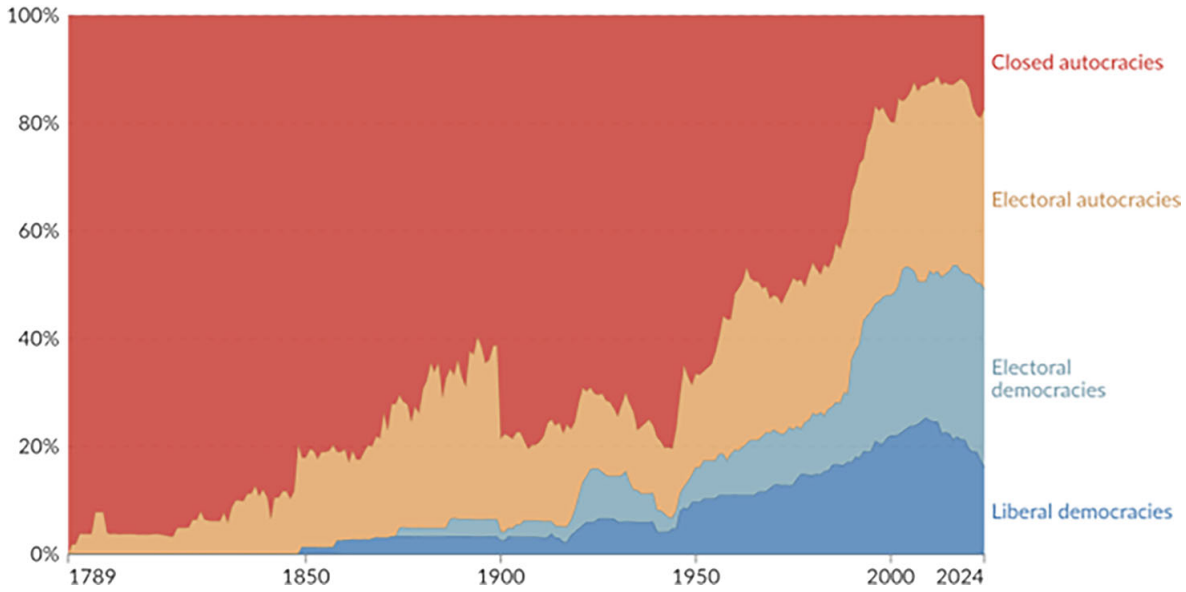

Abb. 3 Entwicklung der Anzahl autokratischer und demokratischer Staaten seit 1789. In bloß „elektoralen" Demokratien finden zwar (mehr oder weniger) faire Wahlen statt, aber auf die „liberalen" Elemente – Rechtsstaatlichkeit, Gewaltenteilung, Grundrechte etc. – ist nicht Verlass. Elektorale Autokratien halten formell Wahlen ab, die jedoch nicht als fair eingestuft werden können (Lührmann et al. 2018; Nord et al. 2025). In jüngster Zeit ist vermehrt zu beobachten, dass liberale Demokratien zu elektoralen Demokratien und diese wiederum zu Autokratien werden („democratic backsliding"). (Quelle: Herre et al. 2013)

3.1 Die US-Demokratie ist von globaler Bedeutung und akut gefährdet

Die aktuellen Entwicklungen in den USA, die als demokratischer Staat geo- und KI-politisch besonders relevant sind, bestätigen und verschärfen den Trend des „democratic backsliding". Sollten in den kommenden Jahren kompetente KI-Agenten entwickelt werden, könnten sie in den USA – deren Jurisdiktion alle führenden KI-Firmen direkt unterstehen – in die Hände eines autoritären Regimes fallen.

Diese Möglichkeit ist in der Debatte um den Überwachungskapitalismus als Risikofaktor dialektisch relevant: Die optimistische Seite räumt bisweilen explizit ein, dass sich ihre (vergleichsweise) positive Bewertung des Überwachungskapitalismus auf die Gegenwart und die Vergangenheit bezieht, d.h. dass sie weitere Fortschritte in der KI-Entwicklung unberücksichtigt lässt (Königs 2024, S. 26). Das ermöglicht es, eine Argumentation zu formulieren, die in der Kombination von Überwachungskapitalismus und agentischer KI auch unter sonst optimistischen Prämissen ein Systemrisiko für die Demokratie sieht. Optimisten könnten darauf entgegnen, dass kompetente KI-Agenten der Demokratie auch vielfältige epistemische und praktische Chancen bieten. Das trifft im Prinzip zwar zu, doch autoritäre politische Kräfte werden kaum daran interessiert sein, die demokratischen Chancen agentischer KI zu realisieren, sondern werden versuchen, ihre antidemokratischen Ziele durch den Einsatz von KI schneller und besser zu erreichen. Darauf wiederum kann die optimistische Seite antworten, dass KI-Agenten ja nicht nur den autoritären Kräften zur Verfügung stehen werden, sondern auch den demokratischen, und dass zumindest unklar sei, wer den größeren strategischen Nutzen aus der Technologie ziehen wird. Auch das trifft im Prinzip zu, doch es ist ganz entscheidend, dass die autoritären Kräfte insbesondere in den USA in den kommenden Jahren einen asymmetrischen Vorteil haben werden: Sie stellen bereits die Regierung – mindestens bis 2028 und womöglich darüber hinaus. Sie tun dies zudem in einem Präsidialsystem, das einem Autokraten, der mit Exekutivdekreten und dem Einsatz loyaler, paramilitärischer Polizeikräfte durchregieren kann, sehr weitreichende legale Kompetenzen verleiht. Die Obersten Richter, die illegale Aktionen verhindern könnten, wurden vom amtierenden Präsidenten teilweise selbst ernannt und haben im Falle einer Verfassungskrise auch keine nennenswerte Kontrolle über den Polizeiapparat. Sollte die US-Regierung das Machtpotenzial ausschöpfen wollen, das die KI-Entwicklung verspricht (vgl. The White House 2025), könnte sie versuchen, sich unilateral Zugang zu den besten KI-Agenten zu verschaffen, sei es durch (erzwungene) Public–Private-Partnerships mit den KI-Firmen oder durch Verstaatlichungen (Cheng und Katzke 2024).

Wenn – wie im 2. Abschnitt ausgeführt – in den kommenden Jahren mit KI-Agenten zu rechnen ist, deren Zeithorizont sich in menschlichen Arbeitswochen oder gar -monaten bemisst, was könnte eine antidemokratische Regierung mit ihnen anstellen wollen? Der amerikanische Präsident stellte 2020 und insbesondere nach dem Sturm auf das Kapitol im Januar 2021 unter Beweis, dass er gewillt ist, Wahlen zu stehlen und Staatsstreiche zu billigen, wenn nicht selbst zu unternehmen. Sein Vizepräsident ist der Ziehsohn des rechtslibertären Tech-Milliardärs und Überwachungsunternehmers Peter Thiel, der seine Präferenz, die Demokratie abzuschaffen, öffentlich kundtut (Thiel 2009). Politische Akteure dieser Art können und wollen KI-Agenten nutzen, um ihre antidemokratischen Ziele zu verwirklichen und diese gegen jeden gegenwärtigen und künftigen Widerstand abzusichern.

3.2 KI-Agenten als Superspitzel

Ein entscheidender Weg dorthin führt über den Überwachungs- und Polizeiapparat des Staates. Historisch haben Autokratien und Diktaturen immer wieder versucht, ihre Bevölkerungen auszuspionieren und einzuschüchtern, um ihre Macht zu verstetigen. In der DDR beispielsweise kamen auf 16 Mio. Einwohner rund 91.000 Stasi-Angestellte und 180.000 Teilzeit-Informanten beziehungsweise „inoffizielle Mitarbeiter“ (Stasi-Unterlagen-Archiv, n.d.). Die Skalierung dieses sozialen Überwachungsapparats dauerte Jahrzehnte, war ressourcenraubend und am Ende nicht besonders stabil. KI-Agenten könnten wesentlich dazu beitragen, diese „Probleme“ zu lösen: Sie könnten so billig sein, dass der amerikanische Staat es sich gut leisten kann, seine Bevölkerung mit hunderten Millionen Agenten auszuspionieren – pro Einwohner ein KI-Überwachungsagent, wenn nicht mehrere. Diese KIs bräuchten weder Schlaf noch Pausen und könnten mit strikt loyalen Werten ausgestattet werden. Sie könnten neue Informationen rund um die Uhr miteinander abgleichen und so ein undurchdringliches Überwachungsnetz über die Bevölkerung legen. Sie würden aufgrund ihres noch beschränkten Zeithorizonts zwar von menschlichen Polizeikräften instruiert und supervidiert, könnten jedoch die insgesamt geleistete Überwachungsarbeit um Größenordnungen steigern. Die Skalierung dieses Systems würde nicht Jahrzehnte erfordern, sondern könnte in Monaten erfolgen. Eine lückenlose Überwachung wäre ökonomisch womöglich nicht nur tragbar, sondern sogar lukrativ: Die gesammelten und analysierten Personendaten könnten – im Zusammenspiel mit den überwachungskapitalistischen Firmen – monetarisiert werden.

Neben den Datenströmen aus den elektronischen Geräten, die wir rund um die Uhr nutzen oder auf uns tragen, würde die KI-Agentenarmee auch die Datenströme aus dem physischen Raum in die Analyse integrieren (Tokson 2025). Die Dichte an Überwachungskameras im öffentlichen Raum hat in den vergangenen Jahren massiv zugenommen (Lin und Purnell 2019). Nicht nur in China, sondern auch in westlichen Städten wird es immer schwieriger, sich in Innen- oder Außenräumen zu bewegen, ohne von ihnen erfasst zu werden. Überwachungsdrohnen, die unbemerkt über einer Stadt schweben, erkennen Personen an ihrem Gesicht und ihrer Gangart. Drohnen und andere Roboter werden auch die physische Polizeiarbeit effizienter und bedrohlicher machen: Mit Waffen bestückt können sie Menschen ferngesteuert oder autonom verfolgen und verhaften, Proteste niederschlagen oder Putschversuche unterstützen (wie den Sturm auf das Kapitol im Januar 2021). Nicht zuletzt könnten KI-Agenten auch Teile des Justizsystems automatisieren und speziell darauf angesetzt werden, Mitglieder oppositioneller Gruppen rechtlich zu belangen und einzuschüchtern.

Auch wenn die KI-Agenten in ihrem Zeithorizont und ihrer Zuverlässigkeit noch beschränkt sind und der menschlichen Supervision bedürfen, ist es gut möglich, dass sie übermenschliche Analysefähigkeiten mitbringen und das Predictive Policing auf eine neue Stufe heben werden. Und selbst wenn sie in ihren Analysefähigkeiten „nur“ an das menschliche Niveau heranreichen sollten, werden ihre weit übermenschliche Geschwindigkeit und Ausdauer wahrscheinlich bewirken, dass die prädiktiven Fähigkeiten des Polizeiapparats stark zunehmen. Einer autoritären Regierung könnte es dadurch gelingen, Widerstand im Keim zu ersticken und ihre Macht auf unbestimmte Zeit abzusichern (vgl. Risse 2018).

Auf diese Weise könnten Überwachungskapitalismus, Autoritarismus und agentische KI in den nächsten Jahren so zusammenwirken, dass insbesondere die US-amerikanische Demokratie existenziell und irreversibel beschädigt wird. Dieses dystopische Szenario braucht – wie eingangs erläutert – nicht besonders wahrscheinlich zu sein, damit es sich lohnt, es im Detail zu erforschen und gezielt zu bekämpfen. Es genügt aus risikoethischer Sicht, dass seine Bestandteile nicht als rein „spekulativ“ abgetan werden können. Im Lichte des oben Gesagten scheint uns dies gegeben. Sollten die USA – als

einzige Demokratie unter den Supermächten – einem solchen Szenario zum Opfer fallen, wäre die Zukunft der demokratischen Idee auch global höchst ungewiss: Gegen China, Russland und die USA als nukleare Superdiktaturen hätten die verbleibenden Demokratien einen schweren Stand, an sich und weil die USA und China über die mächtigsten KI-Systeme verfügen werden.

4 Wenn der Regierung die Macht über ihre KI-Agenten entgleitet

Autokraten können sich unter anderem deshalb oft an der Macht halten, weil sie über einen ausgeprägten Machtinstinkt verfügen. Sie ordnen alles dem Ziel unter, ihre eigene Machtposition zu festigen, und nutzen dazu das Misstrauen und die Angst ihrer Untergebenen strategisch aus. Eine machiavellistische Einstellung, wonach zur Machterlangung und -erhaltung jedes Mittel recht ist, lässt sich als Persönlichkeitsmerkmal bei Autokraten und Diktatoren oft feststellen (Nai und Toros 2020). Sie sind Teil dessen, was in der Persönlichkeitspsychologie als „dunkle Triade“ (Machiavellismus, Narzissmus, Psychopathie) oder „Tetrade“ (inklusive Sadismus) bezeichnet wird. Dass Autokraten machiavellistische Züge aufweisen, ist weniger durch den korrumpierenden Effekt von Macht zu erklären als durch die statistische Tatsache, dass sich Menschen mit machiavellistischen Zügen überproportional oft als Autokraten etablieren (Geng et al. 2017, Grosz et al. 2019, Peterson und Palmer 2022). Die entsprechenden Personen ignorieren soziale Werte und Normen, was es ihnen ermöglicht, ihre Ziele skrupelloser und effektiver zu verfolgen.

Ein ähnlicher Mechanismus könnte für einen ökonomischen und politischen Aufstieg besonders „machthungriger“ KI-Systeme sorgen („power-seeking AI“). Wie wir agentische KIs hinreichend zuverlässig mit unseren Werten und Normen ausstatten („alignment“), ist ein technisch und ethisch ungelöstes Problem. Insoweit die Ziele der KIs mit den Zielen kollidieren, die Menschen durch ihr Marktverhalten und ihre Staatsverfassungen verfolgen, werden superintelligente KI-Agenten einen Anreiz haben, uns zu entmachten (Carlsmith 2024, Kulveit et al. 2025). Ein solches Szenario würde zu einer realistischen Möglichkeit, wenn sich die Task-Dauern weiter ausdehnen und die KI-Agenten das menschliche Leistungsniveau in jeder relevanten Dimension approximieren oder übertreffen. Die gefährliche Machttendenz superintelligenter KI-Agenten, deren Zielfunktionen die Normen des demokratischen Rechtsstaats nicht strikt respektieren (politisches „misalignment“), liegt im Phänomen der instrumentellen Konvergenz begründet (Bostrom 2012). Die entsprechende These besagt, dass intelligente Systeme Ziele effektiver erreichen können, wenn sie Ressourcen und sozialen Einfluss unbegrenzt anhäufen, und zwar weitgehend unabhängig vom Inhalt der Ziele. Mit anderen Worten: Die „skrupellose“ Maximierung von Ressourcen und Einfluss ist ein effektives Mittel (Instrument) zum Zweck, und zwar für viele verschiedene Zweckbestimmungen. Insofern ist sie eine konvergente instrumentelle Strategie, die hinreichend intelligente KI-Agenten oft verfolgen werden, wenn die Alignment-Maßnahmen nicht greifen. Und selbst wenn nur eine Minderheit der Agenten „machiavellistische“ Tendenzen haben sollte, wird es gerade ihnen besonders häufig gelingen, Macht zu akkumulieren.

Superintelligenten KI-Agenten stünden die Mittel zur Verfügung, die auch die Autokraten der Gegenwart zu einer immer größeren Gefahr machen: die Datenmengen und Strukturen des Überwachungskapitalismus sowie Heerscharen untergeordneter KI-Agenten und Roboter. Entsprechende Risiken muten zum aktuellen Zeitpunkt noch spekulativ an. Doch aus entscheidungstheoretischer und risikoethischer Sicht lässt sich argumentieren, dass globale Katastrophenfälle auch dann ernst genommen werden sollten, wenn ihre Eintrittswahrscheinlichkeiten gering sind (vgl. Mukerji und Mannino 2020). Außerdem ist angesichts der rapide zunehmenden KI-Fähigkeiten keineswegs klar, dass die entsprechenden Wahrscheinlichkeiten gering bleiben werden.

5 Was tun?

KI-Systemrisiken haben wissenschaftlich-technische und soziopolitische Aspekte (Tabassi 2023). Beide bieten sich für gezielte Maßnahmen an. Wer über technische Fähigkeiten verfügt, kann beispielsweise an „privacy-preserving“ KI arbeiten (Feretzakis et al. 2024) oder versuchen, das Alignment-Problem auf eine Weise zu lösen oder einzudämmen, die Autokraten nicht gleichzeitig dazu ermächtigt, ihre Ziele effektiver zu verfolgen (Hellrigel-Holderbaum und Dung 2025). Wissenschaftlich-technische Arbeit allein wird jedoch bei Weitem nicht ausreichend sein: Spätestens dann, wenn die KI-Agenten einen Zeithorizont von menschlichen Wochen oder Monaten erreichen, wird die Politik die KI-Firmen verpflichten müssen, auf die Bremse zu treten und maximale Alignment-Vorkehrungen zu treffen. Wenn eine Regierung stattdessen auf Deregulierung und „Akzeleration“ setzt – in einem angeblichen Wettlauf gegen China (Ó hÉigeartaigh 2025) –, ist die Wahrscheinlichkeit weit geringer, dass das Alignment-Problem rechtzeitig hinreichend robust gelöst wird. Daraus ergibt sich das folgende Dilemma (vgl. Hellrigel-Holderbaum und Dung 2025): Bleibt das Alignment-Problem ungelöst und erreichen KI-Agenten in hinreichend vielen Dimensionen (über)menschliches Kompetenzniveau, droht uns ein Kontrollverlust. Wird das Alignment-Problem hingegen gelöst, können die entsprechenden KI-Agenten von einer autoritären Regierung effektiver eingesetzt werden, um antidemokratische Ziele zu verfolgen – insbesondere durch einen massiven Ausbau des Überwachungs- und Polizeiapparats, der alle weiteren Schritte absichert oder ermöglicht (Dresden et al. 2022; Shahbaz 2018). Mit anderen Worten: Um KI-Systemrisiken erfolgreich zu minimieren, müssen wir verhindern, dass autoritäre Kräfte an der Macht sind, wenn KI-Agenten sich dem menschlichen Leistungsniveau annähern und es zu übertreffen beginnen.

Dabei ist jene Regierung besonders relevant, deren Gewaltmonopol sich direkt auf die führenden KI-Firmen erstreckt: die US-amerikanische. Sie könnte sich unilateral (und möglicherweise auch geheim) Zugang zu den jeweils mächtigsten KI-Agenten verschaffen. Sollten diese zunächst sehr teuer sein, ist die finanzstarke US-Regierung bestens positioniert, sie trotzdem in riesiger Zahl zu erwerben oder im Rahmen eines neuen „Manhattan Project“ selbst zu erschaffen. Auch unabhängig von der aktuellen KI-Entwicklung haben die Verfechter der Demokratie jedoch guten Grund, sich gegenwärtig auf die USA zu konzentrieren: Die US-amerikanische Demokratie ist in ihrer Existenz derzeit viel gefährdeter als andere und fungiert auch als Sicherheitsgarant für die Existenz vieler, wenn nicht aller anderen Demokratien (s. unten). Man stelle sich vor, die USA verwandelten sich in eine genuine Autokratie und wären den Werten der verbleibenden Demokratien ähnlich feindlich gesinnt wie Russland und China. Die erwähnte Prognose, dass die Zukunft der demokratischen Idee auf diesem Planeten damit höchst ungewiss wäre – wirtschaftlich, kulturell und natürlich militärisch – erfordert dann leider kaum Spekulation.

Viele Verfechter der Demokratie verspüren den verständlichen, oft wohlbegründeten Impuls, sich auf ihren eigenen Nationalstaat zu konzentrieren. So werden in Deutschland„kurze Anleitungen zur Verteidigung der Demokratie“ (Polenz 2024) verfasst, die sich ganz selbstverständlich auf die deutsche und ausschließlich die deutsche Demokratie beziehen. Die guten Gründe dafür liegen auf der Hand: Wer in Deutschland lebt und die deutsche Politik und Gesellschaft gut kennt, hat epistemische und praktische Vorteile, wenn er primär die deutsche Demokratie stärken will. Und wer dies als deutscher Staatsbürger tut, vermeidet die ethischen und rechtlichen Komplikationen, die sich durch politisches Engagement – das heißt durch den Einsatz von Zeit und Geld – in anderen Nationalstaaten ergeben. Auf der anderen Seite der Waagschale jedoch finden sich auch gewichtige Gründe: (i) Man kann sich natürlich auch als Nicht-Amerikaner über die US-amerikanische Politik kundig machen und es ist möglich, sich ethisch und rechtlich einwandfrei für Demokratien zu engagieren, denen man nicht als Staatsbürger angehört (s.

unten); (ii) die europäischen und viele weitere Demokratien sind in ihrer Existenz derzeit nicht akut gefährdet, die amerikanische schon; und (iii) von der amerikanischen Demokratie hängt geo- und technologiepolitisch viel mehr ab, weil sie als nukleare Supermacht die Existenz anderer Demokratien garantiert und weil die führenden KI-Firmen ihrer direkten Jurisdiktionsgewalt unterstehen.

Wer Zeit oder Geld spenden will, um einen Beitrag zum Schutz der US-amerikanischen Demokratie zu leisten, kann auf spezialisierte Organisationen wie Power for Democracies zurückgreifen (www.powerfordemocracies.org), die akademische Studien, Datensätze und NGO-Berichte zur Frage auswerten, mit welchen konkreten Maßnahmen Demokratien am effektivsten geschützt werden können. Darunter finden sich auch einige Maßnahmen, die Ausländern rechtlich offenstehen. Weil die amerikanischen Wahlen 2026 und 2028 – um das Repräsentantenhaus und den Senat beziehungsweise die Präsidentschaft – durch wenige Stimmen entschieden werden könnten, haben strategisch kluge Zeit- und Geldspenden das außerordentliche Potenzial, über Weltgeschichte zu entscheiden.

Was die politische Ethik des Engagements für Demokratien betrifft, denen man nicht als Staatsbürger oder Einwohner angehört, ist unter anderem nach den Parteilichkeitspflichten zu fragen, die man der eigenen Demokratie gegenüber hat. Es ist plausibel, dass solche Pflichten insbesondere dann bestehen, wenn man von den Normen und Leistungen eines demokratischen Staats direkt profitiert. Diese Pflichten sind jedoch nicht absolut: Die Bewahrung anderer Demokratien ist auch wichtig, denn die Menschen, deren Grundrechte andernfalls verletzt würden, haben einen Anspruch auf unseren Beistand; sie ist doppelt wichtig, wenn über die Zukunft der eigenen Demokratie auch in den USA mitentschieden wird, sodass man die Parteilichkeitspflichten womöglich sogar besser erfüllt, wenn man sich in den USA engagiert; und nicht zuletzt kann die Bewahrung anderer Demokratien viel dringlicher sein, wenn die eigene Demokratie nicht akut gefährdet ist.

Das Erstarken antidemokratischer Kräfte in Deutschland und anderswo ist besorgniserregend. Dennoch dürfen die folgenden Unterschiede zur amerikanischen Situation nicht vergessen werden: Antidemokratische Kräfte sind in Deutschland bis auf Weiteres nicht an der Macht. Sollten sie einer künftigen Bundesregierung angehören, wären sie vielleicht nur Juniorpartner in einer Koalition und hätten auch als Seniorpartner weniger Macht als der US-Präsident, der alle Minister ernennt, per Exekutivdekret durchregieren und aktuell auf Mehrheiten in beiden Parlamentskammern zählen kann. Zudem ist zu erwarten, dass die deutschen Verfassungsrichter viel deutlicher auf der Seite des demokratischen Rechtsstaats stünden als die amerikanischen, die zu einem Drittel vom Präsidenten selbst nominiert wurden und zu zwei Dritteln auf der Seite seiner Partei stehen.

Insgesamt scheint uns ein Übergewicht der Gründe dafür zu sprechen, sich mit knappen Ressourcen zur Verteidigung der Demokratie derzeit in den USA zu engagieren. Dieses Urteil ist auch unabhängig von der KI-Entwicklung plausibel. Bedenkt man zusätzlich, dass sich eine autoritäre US-Regierung in den kommenden Jahren privilegierten Zugang zu KI-Agenten verschaffen könnte, die in mehr und mehr Dimensionen (über)menschliches Leistungsniveau erreichen, drängt sich das Urteil erst recht auf.

Literatur

Acemoglu, D., & Restrepo, P. (2018). Artificial intelligence, automation and work. *NBER Working Paper Series*. https://doi.org/10.3386/w24196.

Bengio, Y., Hinton, G., Yao, A., et al. (2024). Managing extreme AI risks amid rapid progress. *Science*, 384, 842–845.

Bessen, J. (2015). Toil and technology: Innovative technology is displacing workers to new jobs rather than replacing them entirely. *Finance & Development*, 52, 16.

Bostrom, N. (2012). The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22, 71–85.

Bureau of Labor Statistics, U.S. Department of Labor. (n.d.). Tellers. In *Occupational Outlook Handbook*. Suitland.

Butlin, P. (2024). The agency in language agents. *Inquiry*, 0, 1–21.

Carlsmith, J. (2024). Is power-seeking AI an existential risk? *Arxiv preprint*. https://doi.org/10.48550/arXiv.2206.13353.

Cheng, D., & Katzke, C. (2024). Soft Nationalization: How the US Government Will Control AI Labs. *SuperIntelligence – Robotics – Safety & Alignment*, 1(1).

Dresden, J., Bair, A., & Raderstorf, B. (2022). The authoritarian playbook. *Protect Democracy*.

Eth, D., & Davidson, T. (2025). Will AI R&D automation cause a software intelligence explosion?

Feretzakis, G., et al. (2024). Privacy-preserving techniques in generative AI and large language models: A narrative review. *Information*, 15, 697.

Ferrag, M. A., Tihanyi, N., & Debbah, M. (2025). From LLM reasoning to autonomous AI agents: A comprehensive review. *Arxiv preprint*.

Friederich, S., & Dung, L. (2025). Against the Manhattan project framing of AI alignment. *Mind & Language*, 1–18.

Geng, Y., et al. (2017). Relations between Machiavellianism, internalizing and externalizing behavior problems in adolescents. *Personality and Individual Differences*, 119, 296–300.

Grace, K., et al. (2024). Thousands of AI authors on the future of AI. https://doi.org/10.48550/arXiv.2401.02843.

Grosz, M. P., et al. (2019). The development of narcissistic admiration and Machiavellianism in early adulthood. *Journal of Personality and Social Psychology*, 116, 467–482.

Hellrigel-Holderbaum, M., & Dung, L. (2025). Misalignment or misuse? The AGI alignment trade-off. Forthcoming in *Philosophical Studies*.

Herre, B., Rodés-Guirao, L., & Ortiz-Ospina, E. (2013). Democracy. *Our World in Data*.

Jain, L. C., et al. (2014). A review of online learning in supervised neural networks. *Neural Computing and Applications*, 25, 491–509.

Korinek, A. (2023). Scenario planning for an AGI future. *IMF*.

Korinek, A., & Suh, D. (2024). Scenarios for the transition to AGI. *Arxiv preprint*.

Königs, P. (2024). In defense of 'surveillance capitalism'. *Philosophy & Technology*, 37, 122.

Kulveit, J., et al. (2025). Gradual disempowerment: Systemic existential risks from incremental AI development. *Arxiv preprint*.

Kwa, T., et al. (2025). Measuring AI ability to complete long tasks. *Arxiv preprint*.

Lin, L., & Purnell, N. (2019). A world with a billion cameras watching you is just around the corner. *The Wall Street Journal*.

Little, A. T., & Meng, A. (2024). Measuring democratic backsliding. *PS: Political Science & Politics*, 57, 149–161.

Lührmann, A., Tannenberg, M., & Lindberg, S. I. (2018). Regimes of the world (RoW). *Politics and Governance*, 6, 60–77.

METR (2025a). Measuring AI Ability to Complete Long Tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.

METR (2025b). How Does Time Horizon Vary Across Domains? https://metr.org/blog/2025-07-14-how-does-time-horizon-vary-across-domains/.

Mukerji, N. & Mannino, A. (2020). *Covid-19: Was in der Krise zählt. Über Philosophie in Echtzeit*. Reclam.

Nai, A. & Toros, E. (2020). The peculiar personality of strongmen. *Political Research Exchange* 2(1).

Nord, M., et al. (2025). 25 years of autocratization – Democracy trumped? *Democracy Report 2025*. V-Dem Institute.

Novikov, A., et al. (2025). AlphaEvolve: A coding agent for scientific and algorithmic discovery. *Arxiv preprint*.

Ó hÉigeartaigh, S. (2025). The most dangerous fiction: The rhetoric and reality of the AI race. *Available at SSRN*.

Park, J. S., et al. (2023). Generative agents: Interactive simulacra of human behavior. *Arxiv preprint*.

Peterson, R. D., & Palmer, C. L. (2022). The dark triad and nascent political ambition. *Journal of Elections, Public Opinion and Parties*, 32, 275–296.

Polenz, R. (2024). *Tu was!: Kurze Anleitung zur Verteidigung der Demokratie*. C.H. Beck.

Reworr, & Volkov, D. (2025). LLM agent honeypot: Monitoring AI hacking agents in the wild. *Arxiv preprint*.

Risse, M. (2018). Human rights and artificial intelligence: An urgently needed agenda. *HKS Working Paper* No. RWP18–015.

Russell, S. (2025). CEPR webinar series on the economics of artificial intelligence.

Schneier, B. (2015). *Data and Goliath: The hidden battles to collect your data and control your world*. W. W. Norton.

Sevilla, J., et al. (2022). Compute trends across three eras of machine learning. *IJCNN 2022*.

Shahbaz, A. (2018). The rise of digital authoritarianism. *Freedom House*.

Stasi-Unterlagen-Archiv (n.d.). Was war die Stasi? https://www.bundesarchiv.de/.

Susskind, D. (2020). *A world without work*. Penguin UK.

Susskind, D. (2024). Technological unemployment. In *The Oxford handbook of AI governance* (pp. 641–659). OUP.

Susskind, D. (2025). What will remain for people to do? *Knight First Amendment Institute*.

Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0). *NIST*.

Tetlock, P. E., & Gardner, D. (2015). *Superforecasting: The art and science of prediction*. Crown Publishers.

The White House (2025). Winning the race: America's AI action plan.

Thiel, P. (2009). The education of a libertarian. *Cato Unbound*.

Todd, B. (2025a). When do experts expect AGI to arrive? *80,000 Hours*.

Todd, B. (2025b). The case for AGI by 2030. *Benjamin Todd substack*.

Tokson, M. (2025). Artificial intelligence and the anti-authoritarian Fourth Amendment. *University of Utah College of Law Research Paper* No. 635.

Villalobos, P., et al. (2024). Will we run out of data? *Arxiv preprint*.

Vries, A. de. (2023). The growing energy footprint of artificial intelligence. *Joule*, 7, 2191–2194.

Zuboff, S. (2018). *Das Zeitalter des Überwachungskapitalismus*. Campus Verlag.

Zuboff, S. (2022). Surveillance capitalism or democracy? *Organization Theory*.

Adriano Mannino ist Postdoctoral Fellow an der University of California, Berkeley (Kavli Center for Ethics, Science, and the Public) und Technology & Human Rights Fellow an der Harvard University (Carr-Ryan Center for Human Rights). Mannino verfügt über einen interdisziplinären Hintergrund in Philosophie, Rechtswissenschaft, Unternehmertum und Politik. Er erwarb einen BA an der Universität Bern und promovierte an der LMU München. Seine Forschungsinteressen liegen in der normativen und angewandten Ethik, Entscheidungstheorie und politischen Theorie. Aktuell befasst er sich schwerpunktmäßig mit ethisch-politischen Herausforderungen der aufkommenden KI-, Neuro- und Biotechnologien. Zudem berät er die Denkfabrik Power for Democracies, die evaluiert, mit welchen Maßnahmen Demokratien weltweit am besten geschützt werden können.

Nils Althaus ist unabhängiger Wissenschaftler, freier Journalist und Publizist. Er veröffentlicht unter anderem in der Süddeutschen Zeitung, der Neuen Zürcher Zeitung, dem Tagesspiegel, Gehirn & Geist (Spektrum) und dem Philosophie Magazin. Althaus erwarb einen MSc in Biochemie und Molekularbiologie an der ETH Zürich. Seine Schwerpunkte liegen auf den ethischen und gesellschaftlichen Implikationen neuer Technologien wie der KI und Gehirn-Computer-Schnittstellen sowie auf empirischen und philosophischen Fragen rund um Bewusstsein, Empfindungsfähigkeit, Schmerz und Leid.