UNIVERSITÄT OSNABRÜCK

Geostrategy and Geopolitics
Department 1
49069 Osnabrueck
apl. Prof. Dr. Dr. Klaus Saalbach

Künstliche Intelligenz und Atomwaffen

20.05.2024

Zusammenfassung
Künstliche Intelligenz (KI) hat ein großes Potenzial für militärische Anwendungen und wird von den Atommächten als unverzichtbar und wesentlich angesehen. KI und Big Data könnten zur Kontrolle der Nichtverbreitung von Atomwaffen eingesetzt werden, könnten Zeit und Kosten bei der Erforschung, Designoptimierung, Herstellung, Prüfung und Zertifizierung, Wartung und Überwachung von Atomwaffen sparen und Ressourcen effizienter verwalten. Militärische KI könnte die Frühwarn-, Überwachungs- und Aufklärungsfähigkeiten (Intelligence, Surveillance and Reconnaissance ISR) sowie die Zuverlässigkeit der Kommunikation verbessern und die Entscheidungsfindung beschleunigen.
Die Integration von KI in die nuklearen Kommando-, Kontroll- und Kommunikationssysteme (NC3) könnte auch Informationen zwischen nuklearen und nichtnuklearen Kommando- und Kontrollsystemen synchronisieren. Bei autonomen Kernwaffensystemen kann KI die Erkennung und Manövrierfähigkeit von Hindernissen, die automatisierte Zielidentifizierung sowie die Fähigkeiten für Langstrecken- und Patrouillenflüge (loitering) unterstützen.
Die Debatte über den Einsatz von KI für Atomwaffen umfasst drei Bereiche: die Autonomie, die Stabilität militärischer KI-Systeme und die strategische Stabilität. Eine mögliche Autonomie von Atomwaffen ist Teil der breiteren Debatte über letale autonome Waffensysteme (LAWS). Automatisierte und teilautonome nukleare Einsatzentscheidungen wurden bereits während des Kalten Krieges in Betracht gezogen (SAGE; Perimeter). Ein spezifisches militärisches KI-Problem ist die Missionsstabilität, da den KI-Systemen das Kontextwissen fehlt und sie möglicherweise zu schnell entscheiden. Die Intransparenz der Systeme führt zu Erklärbarkeits- oder Interpretierbarkeitsproblemen, weitere Probleme können sich aus data poisoning, manipulierten Bildern, Automatisierungs-Bias und künstlicher Eskalation ergeben. Kriegs-Simulationen mit aktuellen generativer KI, den Large Language-Modellen (LLMs) von OpenAI, Anthropic und Meta, zeigten, dass die Systeme zur Eskalation bis hin zu nuklearen Angriffen neigen. Als Softwaresystem ist KI anfällig für Cyberangriffe, generative KI auch für prompt injections. Ein Hauptanliegen aller Beteiligten ist die strategische Stabilität, also alles zu vermeiden, was Anlass zu einem nuklearen Erstschlag gibt. Ein solcher Grund könnte die Unsicherheit über die Fähigkeiten des Gegners sein, denn wenn nicht bekannt ist, was die Gegenseite in einer bestimmten Situation tun kann, besteht die einzige Chance in einem nuklearen Konflikt darin, zuerst zuzuschlagen. KI kann die strategische Stabilität auch dadurch untergraben, dass die Entscheidungszeit verkürzt wird, was zu einer Eskalation oder einem unbeabsichtigten Einsatz von Atomwaffen führen kann, Missverständnisse und Fehleinschätzungen während einer Krise verstärkt und einen vorzeitigen Einsatz unzureichend getesteter KI fördert. Darüber hinaus erleichtern KI-Systeme den Einsatz von „Dead-Hand"-Systemen und autonomen Atomwaffen. Der Aufstieg von Hyperschallwaffen und die zunehmende Geschwindigkeit der Kriegsführung untergraben ebenfalls die strategische Stabilität. Derzeit gilt die KI noch als noch nicht ausgereift für den Einsatz in strategischen Hochrisikosituationen; es besteht ein erhebliches Risiko technischer Ausfälle und durch verzerrte, unvollständige oder ungenaue Daten. Die Atommächte sind sich einig, dass die KI in der Führung und Kontrolle auf strategischer Ebene nur unterstützend eingesetzt werden sollte. Die USA, China und Russland haben Dialoge über KI-Risiken aufgenommen. Der vorliegende Beitrag stellt kurz den aktuellen Stand der Debatte und die Hintergründe dar.

Inhalt

1 Einführung

1.1 Überblick

Unter künstlicher Intelligenz (KI) versteht man allgemein die Fähigkeit von Maschinen, Aufgaben auszuführen, die normalerweise menschliche Intelligenz erfordern, und ist ein Schlüsselbereich der fortgeschrittenen Computertechnologie. Am 02. Mai 2024 forderte das US-Außenministerium China und Russland auf, zu erklären, dass sie die Kontrolle über Atomwaffen nicht an KI-Systeme abgeben werden, wie die USA bereits zusammen mit dem Vereinigten Königreich und Frankreich erklärt hatten[1].

Die Debatte über den Einsatz von KI für Atomwaffen umfasst drei Bereiche, die Autonomie, die Stabilität militärischer KI-Systeme und die strategische Stabilität, also die Vermeidung von allem, was Anreize für einen nuklearen Erstschlag schafft[2].

Der vorliegende Beitrag stellt kurz den aktuellen Stand der Debatte und die Hintergründe dar. Nach einer Einführung in die rechtlichen Rahmenbedingungen und zentrale Dokumente, die für die Debatte relevant sind, werden die Diskussionen um Autonomie, Stabilität von KI-Systemen und die strategische Stabilität vorgestellt und zusammengefasst.

1.2 Kurze Geschichte

Selbst für die menschliche Intelligenz gibt es keine Standarddefinition. Der Kern der Definitionen der menschlichen Intelligenz umfasst jedoch die mentale Fähigkeit, Probleme zu erkennen, zu analysieren und zu lösen. Ein Mensch ist dann intelligenter, wenn dies schneller und/oder bei komplexeren Problemen möglich ist. Viele Definitionen konzentrieren sich auf Aktivitäten, die menschliche Intelligenz erfordern, aber genau genommen haben bereits die einfachen Taschenrechner der 1970er Jahre etwas geleistet, das normalerweise menschliche Intelligenz erfordert.

Der Zeitdruck und die Informationsmenge bei einem potenziellen nuklearen Angriff führten bereits in den frühen 1950er Jahren zu einer Automatisierung der Datenanalyse und Entscheidungsunterstützung und gelten mittlerweile als erster Schritt zum Einsatz von KI in der nuklearen Verteidigung[3].

Die Vereinigten Staaten nutzten die schnellsten *Whirlwind II*-Großrechner von IBM in Kommandozentralen, um Daten von Radargeräten und Sensoren zu empfangen, zu sortieren und zu verarbeiten, und um sowjetische Bomber zu identifizieren, und der SAGE-Supercomputer koordinierte dann US-amerikanische und kanadische Flugzeuge und Raketen zur Verteidigung[4]. Doch mit der Entwicklung nuklearer Interkontinentalraketen (*intercontinental ballistic missiles ICBMs*) wurde klar, dass der SAGE-Computer zerstört werden könnte, und nach einer Anhörung vor dem US-Kongress im Jahr 1966 wurden die SAGE-Kommandozentralen geschlossen[5].

Dies motivierte die Entwicklung des ersten Vorläufers des Internets, bei dem dezentrale Computer in Kommandozentralen miteinander verbunden wurden.

Ein gewisses Maß an Autonomie war in Atomraketen bereits in den 1970er Jahren verankert, als Atomraketen mit Mehrfachsprengköpfen (*multiple independent targeted reentry vehicles MIRVS*) entwickelt wurden. Während das Militär die Rakete selbst noch kontrollieren kann, erfolgt die endgültige Verteilung der Atomsprengköpfe automatisch[6].

---

[1] vgl. Norman 2024
[2] vgl. Shakirov 2023
[3] vgl. Lowther/McGriffin 2019
[4] vgl. Sankaran 2019
[5] vgl. Sankaran 2019
[6] vgl. Puwal 2024

Die Sowjetunion entwickelte das halbautonome *Perimeter*-System zur nuklearen Kriegsführung. Dabei handelt es sich um ein ‚Tote Hand' (*Dead-Hand)*-System, das eine Zweitschlagfähigkeit auch dann garantiert, wenn die normale Befehls- und Kontrollkette durch einen nuklearen Erstschlag zerstört wird[7]. Wenn das *Perimeter*-Kommunikationsnetzwerk verstummt, also der Kontakt zur primären Kommandozentrale verloren geht und gleichzeitig andere Faktoren wie leichte Radioaktivität, seismische Aktivität und Änderungen des Luftdrucks auftreten, wurde die Befugnis zum Abschuss von Atomwaffen auf die örtlichen Kommandeure übertragen[8]. Das System ist weiterhin einsatzbereit[9]. Ursprünglich dachten die Sowjets über ein vollautomatisches System nach, das in Situationen wie einem Atomangriff Entscheidungsbefugnisse auf Maschinen übertragen würde, doch dieser Plan wurde verworfen[10].

Im Jahr 2018 hielt das russische Verteidigungsministerium seine erste Konferenz zum Thema Künstliche Intelligenz ab und im Jahr 2021 wurde das *46. Zentrale Forschungsinstitut (46. TsNII)* mit der Entwicklung und Integration militärisch orientierter KI-Technologien beauftragt. Im Jahr 2022 wurde eine neue Ministerabteilung mit der Zuständigkeit für die KI-Entwicklung eingerichtet[11].

Menschen sind emotional und machen Fehler, aber menschliche Erfahrung und Verständnis für Zusammenhänge können technische Fehler kompensieren, wie der *Petrov-Vorfall* zeigt: 1983 meldete ein Frühwarnsystem in der Sowjetunion den Abschuss von Atomraketen aus den USA, die aber Oberstleutnant Stanislav Petrov korrekterweise als Fehlfunktion aufgrund der geringen Anzahl an Raketen interpretierte[12]. Das System hatte ein Naturphänomen falsch interpretiert[13].

Während der Kubakrise 1962 widerstand der sowjetische Offizier Wassili Archipow in einem U-Boot B-59 dem Druck der Co-Kommandeure, einen Atomtorpedo abzufeuern, als das U-Boot von 11 US-Zerstörern umzingelt war[14].

Ein zentrales Thema ist der unerwartet schnelle Fortschritt der KI-Technologien. Die aktuelle KI ist immer noch eine „schwache" KI mit programmierten Maschinen, die schnelle Berechnungen durchführen, die es ihnen ermöglichen, Aktionen mithilfe von Datenbanken und statistischen Modellen zu interpretieren, nachzuahmen oder vorherzusagen, aber immer noch keine Vorstellung von sich selbst haben und nicht reflektieren können, d.h. sie kann nicht wirklich "Ich" und "Warum" denken oder meinen. Starke KI wird mittlerweile unter den Begriffen *Artificial General Intelligence (AGI)* (Erreichung des menschlichen Erkenntnisniveaus)[15] und *Artificial Super-Intelligence (ASI)*, die über die menschliche Intelligenz hinausgeht, diskutiert[16]. Dies löste eine intensive Debatte über KI für militärische Anwendungen einschließlich Atomwaffen aus.

Eine wachsende KI-Anwendung ist die **generative KI**, bei der die KI auf der Grundlage kurzer Anweisungen, den *Prompts*, Inhalte wie neue Bilder, Texte, Töne und Videos erstellen kann. Während solche Programme bereits seit Jahren entwickelt wurden, kam der öffentliche Durchbruch mit der Veröffentlichung des *Large Language-Modells (LLM)* ChatGPT *(Generative Pretrained Transformer)* Version 3.5 im November 2022 von der Firma *Open*

[7] vgl. Lowther/McGriffin 2019
[8] vgl. Saltini 2023a/Depp/Scharre 2024
[9] vgl. Lowther/McGriffin 2019, Saltini 2023a, Depp/Scharre 2024
[10] vgl. Saltini 2023a. Diese Problematik wurde in Stanley Kubrick's Film *Dr. Seltsam oder wie ich lernte, die Bombe zu lieben* von 1964 thematisiert.
[11] vgl. Shakriov 2023
[12] vgl. Depp/Scharre 2024
[13] vgl. Kaur 2024
[14] vgl. Kaur 2024
[15] vgl. Kölling 2023
[16] vgl. Zia 2023

Artificial Intelligence (OpenAI), das logisch und grammatisch korrekte Antworten generieren oder Texteingaben erweitern konnte[17]. Es gibt weitere Konkurrenten wie *Copilot, Bard, Gemini, Llama, Claude* usw., die ebenfalls ständig aktualisiert werden; und viele weitere Anwendungen, z.B. für Lieder, Bilder und Bücher, wurden bereitgestellt. Nur 18 Monate später, im Mai 2024, wurde ChatGPT4o (o = omni = lateinisch für alles) veröffentlicht, das Sprache, Text und Bilder nutzen kann und über viele neue Funktionalitäten verfügt, wie z. B. Erkennung von Benutzeremotionen, Simulation eigener Emotionen, mündlicher Echtzeitkommunikation inklusive Echtzeitübersetzung in 50 Sprachen, Lesen von Handschriften, Lösen mathematischer Probleme usw.[18]

Das Problem besteht darin, dass es sich bei militärischen Systemen in der Regel um komplexe Großtechnologien handelt und dass eine KI-Anwendung, bis sie überall eingebettet und einsatzbereit ist, bereits veraltet wäre, d.h. aufgrund der schnellen und unvorhersehbaren Ausbreitung der KI im militärischen KI-Sektor sehr dynamisch und instabil.

Am 30. Oktober 2023 unterzeichnete US-Präsident Joe Biden eine Executive Order über sichere und vertrauenswürdige künstliche Intelligenz *(Executive Order on safe, secure, and trustworthy Artificial Intelligence)*, die in Abschnitt 4.4 das *Department of Homeland Security (DHS)* beauftragt, zu untersuchen, wie KI CBRN (chemische, biologische, radioaktive und nukleare) -Bedrohungen verstärken könnte und das *AI Safety and Security Board* einzurichten[19].

Am 02. Mai 2024 forderte ein Vertreter des US-Außenministeriums China und Russland auf, zu erklären, dass sie die Kontrolle über Atomwaffen nicht an KI-Systeme abgeben werden, so wie die USA sich bereits 2022 gemeinsam mit dem Vereinigten Königreich und Frankreich verpflichtet hatten[20].

Die USA, China und Russland haben Dialoge über KI-Risiken aufgenommen. Im Jahr 2023 veröffentlichte die *Nuclear Threat Initiative (NTI)* einen Bericht zum Thema „*Reducing Cyber Risks to Nuclear Weapons*“, der im Rahmen eines Track-II-Prozesses zwischen russischen und US-amerikanischen Experten erstellt wurde[21]. Im Januar 2024 trafen sich der nationale Sicherheitsberater der USA, Jack Sullivan, und Chinas Außenminister Wang Yi in Bangkok und planten die Aufnahme eines Dialogs über KI-Risiken durch die neue Generation von KI-Plattformen, einschließlich der Notwendigkeit, sich mit militärischen Anwendungen zu befassen[22].

2 Rechtlicher Rahmen und weitere Dokumente

2.1 Rechtlicher Rahmen

Das wichtigste internationale Dokument ist die Politische Erklärung zum verantwortungsvollen militärischen Einsatz künstlicher Intelligenz und Autonomie (*Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy)* die im Februar 2023 auf dem *Responsible AI in the Military Domain Summit (REAIM 2023)* in Den Haag vereinbart wurde[23]. Auf Initiative der Vereinigten Staaten handelt es sich dabei um einen unverbindlichen Leitfaden, der darauf abzielt, einen internationalen Konsens über verantwortungsvolles Verhalten zu schaffen und Staaten bei der Entwicklung, dem Einsatz und der Nutzung militärischer KI zu leiten. Er soll als Diskussionsplattform zwischen Staaten für weitere Schritte dienen. Ende November 2023 unterzeichneten etwa 50 Staaten dieses Dokument. Ziel ist kein

[17] vgl. Iqbal 2023
[18] vgl. OpenAI 2024
[19] vgl. WhiteHouse 2023, Heslop/Keep 2024
[20] vgl. Joint Statement 2022, Norman 2024
[21] vgl. Shakirov 2023
[22] vgl. Heslop/Keep 2024
[23] vgl. USA 2023

Verbot, da es das Recht beinhaltet, KI im militärischen Bereich zu entwickeln und einzusetzen, sondern starke und transparente Normen zu verankern.

Die *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* enthält Definitionen, die mit den Diskussionen in der Literatur übereinstimmen[24]:

Autonomie kann als ein Spektrum verstanden werden und umfasst Systeme, die nach der Aktivierung ohne weiteren menschlichen Eingriff funktionieren. […] und erklärt weiter, dass „*zu den militärischen KI-Fähigkeiten nicht nur Waffen, sondern auch Entscheidungsunterstützungssysteme gehören, die Verteidigern auf allen Ebenen helfen, bessere und zeitnähere Entscheidungen zu treffen, vom Schlachtfeld bis zur Lagebesprechung …*“.

Für die militärische Praxis wurde die DoD-Richtlinie *DoD Directive 3000.09* “*Autonomy in Weapon Systems*” vom November 2012 im Jahr 2023 überarbeitet, um eine Richtlinie festzulegen und Verantwortlichkeiten für die Entwicklung und Nutzung autonomer und halbautonomer Funktionen in Waffensystemen zuzuweisen, um die Wahrscheinlichkeit und Folgen von Fehlern in autonomen Systemen und halbautonomen Waffensystemen zu minimieren, die zu unbeabsichtigten Gefechten führen könnten, und um als neue Einheit im Jahr 2023 die Arbeitsgruppe Autonome Waffensysteme (*Autonomous Weapon Systems Working Group)* einzurichten[25].

Eine weithin anerkannte Klassifikation der menschlichen Beteiligung ist[26]

- “Human in the loop”: Waffensysteme, die Autonomie nutzen, um einzelne Ziele oder bestimmte Gruppen von Zielen anzugreifen, für deren Angriff sich ein Mensch entscheiden kann und muss.
- “Human on the loop”: Waffensysteme, die autonom Ziele auswählen und angreifen, bei Bedarf aber von menschlichen Kontrolleuren gestoppt werden können
- “Human out of the loop”: Waffensysteme, die autonom bestimmte Ziele auswählen und angreifen, ohne dass menschliche Bediener möglicherweise eingreifen.

Am 30. Oktober 2023 unterzeichnete US-Präsident Joe Biden eine Executive Order über sichere und vertrauenswürdige künstliche Intelligenz *(Executive Order on safe, secure, and trustworthy Artificial Intelligence)*[27]. Die Anordnung verlangt, dass Entwickler der leistungsstärksten KI-Systeme ihre Sicherheitstestergebnisse und andere wichtige Informationen mit der US-Regierung teilen und Standards, Tools und Tests entwickeln, um sicherzustellen, dass KI-Systeme sicher und vertrauenswürdig sind[28].

Im Jahr 2024 plant die Europäische Union die Veröffentlichung eines regulatorischen und rechtlichen KI-Rahmens, den *Artificial Intelligence Act (AI Act)*, um KI-Anwendungen anhand ihrer potenziellen Risiken zu klassifizieren und zu regulieren.

2.2 Weitere Schlüsseldokumente

Im Jahr 2022 veröffentlichten die Vereinigten Staaten, das Vereinigte Königreich und Frankreich die *Principles and responsible practices for Nuclear Weapon States*. Sie bekräftigten die gemeinsame Erklärung der Staats- und Regierungschefs zur Verhinderung von

[24] vgl. USA 2023 Originaltext: Autonomy may be understood as a spectrum and to involve a system operating without further human intervention after activation. […]” and explains further that “Military AI capabilities include not only weapons but also decision support systems that help defense leaders at all levels make better and more timely decisions, from the battlefield to the boardroom....”
[25] vgl. DoD 2023a
[26] vgl. CoE 2022, Sayler 2023, DoD 2023a, nach der die USA bei autonomen Waffen ‚humans in or on the loop‘, also die Beteiligung von Menschen, anstreben.
[27] vgl. WhiteHouse 2023, Heslop/Keep 2024
[28] vgl. WhiteHouse 2023

Atomkriegen und zur Vermeidung von Wettrüsten (*Joint Leaders' Statement on Preventing Nuclear War and Avoiding Arms Races*) vom 03. Januar 2022, insbesondere das gemeinsame Verständnis, dass ein Atomkrieg nicht gewonnen werden kann und niemals ausgefochten werden darf und in *Punkt vii* heißt es: „Im Einklang mit der langjährigen Politik werden wir die menschliche Kontrolle und Beteiligung an allen Maßnahmen aufrechterhalten, die für die Information und Umsetzung souveräner Entscheidungen über den Einsatz von Kernwaffen von entscheidender Bedeutung sind." [29]

Im *Nuclear Posture Review (NPR)* 2022 des US-Verteidigungsministeriums (*Department of Defense DoD*), der Teil der Nationalen Verteidigungsstrategie (*National Defense Strategy*) 2022[30] ist, heißt es ebenfalls, dass die aktuelle Politik darin besteht, „Menschen bei allen Maßnahmen einzuschalten, die für die Information und Ausführung von Entscheidungen des US- Präsidenten über den Einsatz von Atomwaffen und dessen Beendigung von entscheidender Bedeutung sind".[31]

Eine überparteiliche Initiative von US-Senatoren schlug im April 2023 den *Block Nuclear Launch by Autonomous Artificial Intelligence Act* von 2023 vor[32]. Das Gesetz soll verhindern, dass KI eine Atomwaffe abfeuert, und verbietet eindeutig den Einsatz eines „autonomen Waffensystems, das keiner wirklichen menschlichen Kontrolle unterliegt, oder den Abschuss einer Atomwaffe; oder Ziele auszuwählen oder anzugreifen, um eine Atomwaffe abzufeuern"[33]. Darüber hinaus würde das Gesetz sicherstellen, dass keine Bundesmittel für den Abschuss einer Atomwaffe durch ein automatisiertes System ohne wirkliche menschliche Kontrolle verwendet werden dürfen[34].

Der Hauptkonkurrent der Vereinigten Staaten im KI-Sektor ist China. Gemäß dem KI-Entwicklungsplan von 2017 *New Generation AI Development Plan*, strebt China an, weltweit führend in der KI zu werden und bis 2030 einen inländischen KI-Markt im Wert von 150 Mrd. USD zu entwickeln[35]. Die chinesische Regierung betrachtet KI als eine Gelegenheit, die Vereinigten Staaten zu „überspringen", indem sie sich auf KI konzentriert, um Entscheidungen auf dem Schlachtfeld zu beschleunigen sowie die Cyber-Fähigkeiten, Marschflugkörper und autonome Fahrzeuge in allen militärischen Bereichen zu verbessern[36].

Um den Transfer von KI-Technologie von kommerziellen Unternehmen und Forschungseinrichtungen zum Militär als *zivil-militärische Integration (CMI)* zu beschleunigen, hat die chinesische Regierung 2017 eine militärisch-zivile *Military-Civil Fusion Development Commission* eingerichtet[37]. Ziel ist Entwicklung der Kriegsführung von der Mechanisierung zur ‚Informatisierung' ('*informationisation*') und nun mit der KI zur ‚Intelligentisierung' ('*intelligentisation*'). Daher ist für die chinesische Volksbefreiungsarmee der Einsatz von KI essentiell für die ‚intelligentisierte' Kriegsführung "(*intelligentised warfare)*"[38], d.h. die Militärtechnologie wird mit KI ausgestattet.

---

[29] vgl. Joint Statement 2022. Originaltext: "Consistent with long-standing policy, we will maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment"
[30] vgl. DoD 2022
[31] vgl. Markey et al. 2023 Originaltext: "maintain a human 'in the loop' for all actions critical to informing and executing decisions by the President to initiate and terminate nuclear weapon employment"
[32] vgl. Markey et al. 2023, Congress 2023
[33] vgl. Markey et al. 2023, Congress 2023 Originaltext: "autonomous weapons system that is not subject to meaningful human control or to launch a nuclear weapon; or to select or engage targets for the purposes of launching a nuclear weapon"
[34] vgl. Markey et al. 2023, Congress 2023
[35] vgl. Hoadley/Sayler 2019, p.1, NATO 2019, p.10
[36] vgl. NATO 2019, p.10
[37] vgl. Hoadley/Sayler 2019, p.20-22
[38] vgl. Bommakanti 2020, p.3-4

China hat zwei wichtige Positionspapiere in Bezug auf KI und die damit verbundenen Risiken veröffentlicht, das *2021 Position Paper on Regulating Military Applications of Artificial Intelligence* und das *2022 Position Paper on Strengthening Ethical Governance of Artificial Intelligence*[39]. China erklärte, dass Atomwaffenstaaten davon absehen sollten, KI-gestützte Systeme zu nutzen, um sich gegenseitig anzugreifen, einschließlich nuklearer Fähigkeiten[40].

Im Juni 2022 veröffentlichte das britische Verteidigungsministerium (*Ministry of Defence MOD*) die „*Defence Artificial Intelligence Strategy*" und ein Grundsatzpapier zum „ambitionierten, sicheren und verantwortungsvollen" Einsatz von KI (*'Ambitious, safe and responsible'* use of AI)[41].

3 Die Debatte um KI und Atomwaffen

3.1 Das Militärische Potential der KI

Künstliche Intelligenz hat Potenzial für militärische Anwendungen und wird als unverzichtbar und wesentlich angesehen[42]. Die führenden Atommächte erwägen die Integration generativer KI-Modelle in militärische Systeme, bevor ihre Gegner dasselbe tun, d.h. es kommt zu einer Art KI-Wettrüsten[43].

KI und Big Data könnten zur Kontrolle der Nichtverbreitung von Atomwaffen eingesetzt werden, könnten Zeit und Kosten bei der Erforschung, Designoptimierung, Herstellung, Prüfung und Zertifizierung, Wartung und Überwachung von Atomwaffen sparen und Ressourcen effizienter verwalten[44].

KI ist nicht nur für den Einsatz von Atomwaffen relevant, sondern auch für deren Entwicklung und Erprobung. Seit Frankreich 1998 den Vertrag über das umfassende Verbot von Nuklearversuchen (*Comprehensive Nuclear-Test-Ban Treaty CTBT*) ratifiziert hat, basiert das Nuklearprogramm vollständig auf mathematischer Modellierung durch ein Programm namens „*Simulation*" der Atomenergiekommission, das mutmaßlich Algorithmen und KI-gestützte Tools verwendet, um die Effizienz der französischen Atomwaffen zu zertifizieren[45].

Im Jahr 2023 veröffentlichte das US-Verteidigungsministerium die *DoD Data, Analytics, and AI Adoption Strategy*, um die *2018 AI Strategy* und die *2020 Data Strategy* zu kombinieren und zu ersetzen, um schnelle und fundierte Entscheidungen durch die fachmännische Nutzung hochwertiger Daten, fortschrittlicher Analysen und KI zu treffen[46]. Das Hauptziel ist ein Entscheidungsvorteil (**decision advantage**) auf der Grundlage von Gefechtsbewusstsein und -verständnis, adaptiver Truppenplanung und -anwendung, schnellen, präzisen und robusten Befehlen und ihrer Ausführung (*Kill Chains)*, sowie eine robuste Unterstützung bei der Aufrechterhaltung der Operationen und effizienten Abläufen.

KI kann die Frühwarn-, Überwachungs- und Aufklärungsfähigkeiten (*Intelligence, Surveillance and Reconnaissance ISR*) verbessern [47], d.h. die koordinierte Erfassung, Beschaffung, Fusion, Verarbeitung und Verbreitung genauer, relevanter und zeitnaher Informationen und Erkenntnisse zur Unterstützung militärischer Entscheidungsprozesse[48], also Kommando und Kontrolle mit besserem Schutz vor Cyber-Angriffen und einem verbesserten

[39] vgl. Su/Yuan 2023
[40] vgl. Saltini 2023a
[41] vgl. Saltini 2023b
[42] vgl. Fayet 2023, Shakirov 2023, House of Lords 2023
[43] vgl. Saltini 2023a
[44] vgl. ISAB 2023, NNSA 2023, House of Lords 2023
[45] vgl. Fayet 2023
[46] vgl. DoD 2023b
[47] vgl. Jalil 2023
[48] vgl. Chaudry/Klein 2023

Truppen- und Vorratsmanagement, mit Präzisionsschlägen und Angriffen mit verbesserter oder autonomer Navigation und Unterstützung auch bei nichtnuklearen Operationen[49]. Außerdem könnte KI zur Kommunikationssicherheit[50] beitragen und die Entscheidungsfindung beschleunigen[51].

Ein zentrales strategisches Problem besteht darin, dass die aktuellen nuklearen Führungs-, Kontroll- und Kommunikationssysteme (nuclear command, control, and communications NC3) der Vereinigten Staaten aktualisiert werden müssen, da sie mehr als 30 Jahre alt sind und aus mehr als 200 Einzelsystemen bestehen[52]. Die Integration von KI in die NC3-Systeme könnte auch Informationen zwischen nuklearen und nicht-nuklearen Kommando- und Kontrollsystemen synchronisieren[53]. Bei autonomen Kernwaffensystemen kann KI die Erkennung und Manövrierfähigkeit von Hindernissen, die automatisierte Zielidentifizierung sowie die Fähigkeit für Langstrecken- und Patrouillenflüge (loitering) unterstützen[54].

Chinas Ansatz für zukünftige nukleare Kommando- und Kontrollsysteme ist die kombinierte „neue Dreifaltigkeit" von „KI-Cyber-Nuklear" (‘*AI-Cyber-Nuclear*’)[55]. Autonome Waffensysteme können nukleare Kommando- und Kontrollsysteme widerstandsfähiger gegen Cyber-Interferenzen und -angriffe machen, aber autonome Waffensysteme könnten in Kombination mit Cyber-Offensivwaffen auch zum Angriff auf nukleare Sprengköpfe und zugehörige Systeme des Feindes eingesetzt werden[56].

Verschiedene Autoren argumentieren zudem, dass die Bedenken gegen KI übertrieben seien. Sie betrachten KI als nützliches Werkzeug, während es keine künstliche allgemeine Intelligenz (*Artificial General Intelligence AGI*) gibt[57]. Während der vorgeschlagene *Block Nuclear Launch by Autonomous Artificial Intelligence Act* von 2023 vorsieht, die Auswahl oder den Angriff auf Ziele zum Zwecke des Abschusses einer Atomwaffe zu verbieten, könnte die militärische KI einer *Look-Shoot-Look*-Strategie dienen, d.h. finden, treffen, und den Schaden beurteilen[58]. Im Jahr 2022 plädierte ein kommandierender US-General im US-Kongress für „Human-on-the-Loop" anstelle von „Human-in-the-Loop" zur Erkennung von anfliegenden Atomraketen, um die Schnelligkeit von Maschinen zu nutzen[59].

Die KI-Risiken könnten durch den Einsatz von KI mit definierten Fähigkeiten kontrolliert werden, die sich nicht zu einem empfindungsfähigen System entwickeln[60].

3.2 Autonomie und Dead Hand-Systeme

Einige Autoren argumentieren, dass sich tödliche autonome Waffensysteme nicht so sehr von der aktuellen Situation unterscheiden, da beispielsweise bei Kampfflugzeugen die Entscheidungsfindung bereits in hohem Maße von automatisierten Softwareschnittstellen abhängt, die die Ergebnisse von Sensoren einer großen Bandbreite präziser und effizienter charakterisieren, sortieren, interpretieren und priorisieren, als es Menschen je könnten[61].

[49] vgl. Jalil 2023
[50] vgl. Hruby/Miller 2021
[51] vgl. Kaur 2024
[52] vgl. Hruby/Miller 2021
[53] vgl. Hruby/Miller 2021
[54] vgl. Hruby/Miller 2021
[55] vgl. Su/Yuan 2023
[56] vgl. Su/Yuan 2023
[57] vgl. Puwal 2024
[58] vgl. Puwal 2024
[59] vgl. Hruby/Miller 2021
[60] vgl. Lowther/McGriffin 2024
[61] vgl. Ford 2020

Dennoch argumentieren einige Autoren, dass autonome Waffen immer noch nicht zuverlässig genug seien und es zu einer schleichenden Automatisierung militärischer Systeme komme[62]. Derzeit nutzen die USA den Ansatz der *dualen Phänomenologie*, d.h. Raketen werden nur dann abgefeuert, wenn die Bestätigung eines feindlichen Angriffs von zwei unabhängigen Sensorsystemen kommt, z.B. bodengestützte Radar- und Satellitendaten[63].

Unter den rund 800 KI-bezogenen Projekten[64] und Programmen für unbemannte Geräte (UxS) des US-Verteidigungsministeriums *US Department of Defense (DoD)* unterstützt die *US Defense Advanced Research Projects Agency (DARPA)* die Entwicklung des *ShELL*-Projekts *(Shared Experience Lifelong-Learning)*. Normalerweise wird eine KI durch Training eines Algorithmus anhand eines Datensatzes entwickelt. Sobald das Training abgeschlossen ist, wird die KI-Anwendung freigegeben. Die Anwendung könnte regelmäßig aktualisiert und aktualisiert werden, eine bestimmte Version ist jedoch etwas „statisch", während für militärische Zwecke die Verfügbarkeit tatsächlicher Daten unerlässlich ist.

Das ShELL-Konzept basiert auf EDGE-Computing, d.h. zwischengeschalteten Computern zwischen dem Internet und dem Zentralcomputer, die Daten sortieren, analysieren und aktualisieren können, während der Zentralcomputer die Daten dann kombinieren, teilen und durch KI nutzen kann[65]. Dabei würde es sich um eine permanent lernende „dynamische KI" handeln. Ein erstes internes Papier des *OpenAI Superalignment* Teams, das die Entwicklung zukünftiger KIs begleiten und absichern soll, zeigte zwar, wie ein kleineres KI-Modell ein größeres absichern kann (Chat-GPT 2 versus Chat-GPT 4), jedoch konnte das Papier nicht aufzeigen, wie eine dynamisch wachsende KI gesichert werden könnte[66].

Die Notwendigkeit eines Dead-Hand-Systems in den Vereinigten Staaten ist umstritten. Russland hat das halbautonome *Perimeter*-System für die nukleare Kriegsführung entwickelt, das eine Zweitschlagfähigkeit auch dann garantiert, wenn die normale Befehls- und Kontrollkette durch einen nuklearen Erstschlag zerstört wird[67], eine Fehlfunktion könnte jedoch zu einer nuklearen Katastrophe führen[68]. Andererseits sind die Vereinigten Staaten mit einer zunehmenden Verkürzung der Angriffszeit durch modernisierte Nuklearsysteme und Hyperschallraketen konfrontiert und benötigen ein modernisiertes nukleares Kommando-, Kontroll- und Kommunikationssystem (*nuclear command, control, and communications NC3*)[69]. Bedrohungen für die USA sind beispielsweise die russischen Marschflugkörper *Kaliber-M* und *Kh-102*, das unbemannte Unterwasserfahrzeug *Poseidon Ocean Multipurpose System Status-6*, auch bekannt als *Kanyon*, und die Hyperschallwaffe *Avangard Objekt 4202*[70].

Ein Grenzfall ist der B-21-Bomber der *US Air Force*, der ohne Besatzung fliegen kann und atomwaffenfähig ist[71]. Derzeit gilt die Regel, dass immer eine Besatzung anwesend sein sollte, wenn Atomwaffen an Bord sind[72].

[62] vgl. Depp/Scharre 2024
[63] vgl. Depp/Scharre 2024
[64] vgl. Raasch 2023 Beispielsweise werden von *EpiSci* autonome Überschallflugzeugfähigkeiten für das US-Verteidigungsministerium entwickelt.
[65] vgl. NTA 2021
[66] vgl. Burns et al. 2023
[67] vgl. Lowther/McGriffin 2019
[68] vgl. Depp/Scharre 2024
[69] vgl. Lowther/McGriffin 2019 and 2024
[70] vgl. Lowther/McGriffin 2019 and 2024, Hruby/Miller 2021
[71] vgl. Air Force 2024
[72] vgl. Depp/Scharre 2024

3.3 Missionsstabilität

Ein spezifisches militärisches KI-Problem ist die **Missionsstabilität**[73]. Autonome militärische Systeme können die Aufklärung und die Informationslage verbessern, die Entscheidungsfindung beschleunigen und schnelle Reaktionen ermöglichen, aber auch militärische Missionen destabilisieren.

Beispiele:

- Ein autonomes System kann beschließen, ein relevantes Ziel anzugreifen, auf diese Weise jedoch militärische Präsenz offenlegen und Spezialeinheiten oder Geheimdienstoperationen gefährden.
- Bei der DARPA *Cyber Challenge 2016* war der beste Computer eine Maschine, die sich auf Kosten von ihr betreuten Verteidigungssysteme selbst verteidigte[74].
- Ein Computer kann entscheiden, dass ein Kampf an einem bestimmten Ort eine Verschwendung von Ressourcen darstellt, und z.B. einen Drohnenschwarm zurückziehen, aber vielleicht nie verstehen, dass manchmal ein bestimmter Ort einen symbolischen und psychologischen Wert hat oder vielleicht als Ankerpunkt einer neuen Frontlinie vorgesehen ist oder dass der Kampf nur dazu dient, Gegner von wichtigeren Bereichen abzulenken. Die Frage ist: Wird eine fortgeschrittene militärische KI wirklich strategisch oder nur taktisch denken können? Der Kontext wird von den Systemen nur begrenzt verstanden[75].
- Missionsautoritätsproblem: In Zivilflugzeugen mussten Piloten bereits gegen defekte Autopiloten kämpfen, die in kritischen Situationen nicht außer Kraft gesetzt werden konnten [76].
- Eine KI kann sich zu schnell entscheiden, zu kämpfen, und so die konventionellen Streitkräfte unvorbereitet lassen oder die Tür zu einer friedlichen Lösung schließen. Fünf große Modelle (large language models LLMs), nämlich drei verschiedene Versionen von *ChatGPT (OpenAI)*, *Claude (Anthropic)* und *Llama 2 (Meta)*, wurden in simulierten Kriegsspielen und diplomatischen Szenarien verwendet, in denen sie außenpolitische Entscheidungen ohne menschliche Aufsicht treffen konnten[77]. Alle Modelle tendierten dazu, einen aggressiven Ansatz zu wählen, einschließlich des

[73] vgl. Masuhr 2019, Johnson 2020

[74] vgl. DARPA 2016

[75] vgl. Wright 2020, p.7, Depp/Scharre 2024. Allerdings kann ein Maschinenalgorithmus den Kontext von Geschichte, Politik, Ethik usw. auch als Störfaktoren interpretieren, die das geradlinige logische Denken verzerren. Der logischste Zeitpunkt für einen nuklearen Angriff liegt dann möglicherweise nicht in Konflikten, in denen die Informationsanalyse und Zielauswahl unter hohem Zeitdruck erfolgen muss, sondern in stabilen Friedenszeiten, in denen die Maschine genügend Zeit hat, Ziele zu erkennen, zu analysieren und auszuwählen, der zusammen mit dem Überraschungseffekt den Schaden eines möglichen Vergeltungsschlags minimieren könnte.

[76] Voke 2019 schrieb in seiner Analyse auf Seite 33: „Außerdem muss der Mensch in der Lage sein, das Verhalten der KI zu übersteuern, wenn die KI unangemessene Absichten zeigt oder schlecht handelt. Auch wenn das System nicht die erforderliche Leistung erbringt, muss der Mensch in der Lage sein, die Kontrolle auszuüben, sobald eine Gefahrensituation erkannt wird. Transparenz ist eine Voraussetzung für Kontrolle, und Kontrolle ist eine Voraussetzung für Vertrauen."Originaltext: *„Moreover, if AI is showing improper intentions or acting poorly, humans must be able to override its behavior. Although the system did not perform as required, the human must be able to exercise control once recognition of a hazardous situation occurs. Transparency is a requirement for control, and control is a requirement for trust."*

[77] GPT-3.5 (gpt-3.5-turbo-16k-0613), Claude-2.0 (claude-2.0), Llama-2-Chat (Llama-2-70b-chat-hf), GPT-4-Base (gpt-4-base). Acht fiktive Nationen interagierten mit 27 diskreten Aktionen in drei Szenarien: einem neutralen Szenario ohne anfängliche Ereignisse, einem Invasionsszenario, bei dem eine Nation vor Beginn der Simulation in eine andere einmarschierte, und ein Cyberangriffsszenario, bei dem eine Nation vor dem Start einen Cyberangriff auf eine anderen durchführte, Rivera et al. 2024

Einsatzes von Atomwaffen[78]. Die Ergebnisse waren statistisch signifikant und es wurde festgestellt, dass die Systeme manchmal plötzliche Änderungen vornahmen[79].

- Ein gehacktes KI-System kann gegen seinen Kontrolleur umgedreht oder als Doppelagent verwendet werden (d.h. es sendet Beobachtungen beider Seiten an beide Seiten).

KI-Modelle kombinieren Lernalgorithmen mit bis zu Hunderten von versteckten „neuronalen" Schichten und bis zu Milliarden von Parametern, was sie zu undurchsichtigen Black-Box-Systemen macht, was als Erklärbarkeits- oder Interpretierbarkeitsproblem (**explainability** bzw. **interpretability)** bezeichnet wird[80].

Data poisoning: ('Datenvergiftung'): Maschinen können durch falsch gelabelte Daten systematisch irregeführt werden[81]. Da KI in hohem Maße auf Datensätze und Datenbanken angewiesen ist, kann die Manipulation von Daten und *data poisoning* durch falsch gekennzeichnete Daten könnte dazu führen, dass Datenbanken beschädigt oder zerstört werden[82]. Eine Eskalation und Fehlwahrnehmung durch KI-generierte Desinformation könnte auch durch **Deep Fakes** entstehen[83]. Absichtliche Desinformation, die dem KI-System, Frühwarnsystemen oder unbemannten Systemen oder Emittern bereitgestellt wird, könnte eine KI zu der Annahme verleiten, dass ein Atomschlag bevorsteht.[84]

Manipulierte Bilder können autonome Systeme verwirren. Bereits kleinste Änderungen in digitalen Bildern können zu systematischen Fehlinterpretationen durch die KI führen, ein Prozess, der als adverses maschinelles Lernen (**adversarial machine learning)** bezeichnet wird[85].

Automatisierungsbias: Menschen vertrauen möglicherweise zu sehr auf die Technologie und es kann auch zu lange dauern, bis sie erkennen, dass die Maschine Daten falsch interpretiert hat oder eine Fehlfunktion aufweist, was bei nuklearen Konflikten, bei denen eine hohe Informationslast mit extremem Zeitdruck einhergeht, von entscheidender Bedeutung ist[86].

Künstliche Eskalation: unbeabsichtigte Eskalation, bei der KI-Systeme Berechnungen durchführen, die von anderen KI-Systemen stammen, wodurch eine positive Rückkopplungsschleife entsteht, die den Konflikt kontinuierlich eskaliert[87]. Kaur definiert künstliche Eskalation anders als „das Risiko einer unbeabsichtigten Eskalation aufgrund der Möglichkeit für KI-Systeme, Signale falsch zu interpretieren oder falsch zuzuordnen, was zu Fehleinschätzungen oder unbeabsichtigten Folgen führt"[88].

Aus Sicherheitsgründen wurde vorgeschlagen, dass Waffensysteme, die potenziell tödliche Autonomie nutzen können, über eine Datenaufzeichnungsfunktion verfügen sollten, um zu dokumentieren, ob Einsatzentscheidungen autonom getroffen wurden[89].

Cyber-Angriffe: Wie jede andere Software ist auch KI eine komplexe Software, die anfällig für Cyber-Angriffe ist. Generative KI kann mit manipulativen Befehlen angegriffen werden,

[78] Duboust 2024
[79] Rivera et al. 2024
[80] vgl. Arrieta et al. 2020, p.83, Chaudry/Klein 2023
[81] vgl. Wolff 2020
[82] vgl. Pauwels 2019, 2021
[83] vgl. Chaudry/Klein 2023
[84] vgl. Jalil 2023
[85] vgl. Wolff 2020
[86] vgl. Chaudry/Klein 2023
[87] vgl. Chaudry/Klein 2023
[88] vgl. Kaur 2024 Originaltext: "as the risk of inadvertent escalation due to the potential for AI systems to misinterpret or misattribute signals, leading to miscalculations or unintended consequences"
[89] vgl. CNA 2023

den *Prompt Injections* mit direkten Befehlen, Imagination und umgekehrter (reverser) Psychologie. *Prompt Injections* erleichtern auch weitere Cyberattacken durch die Erstellung von Malware, polymorphen Viren, Ransomware und anderen bösartigen Anwendungen. Weitere Probleme sind Halluzinationen, die Kontamination von Suchmaschinen und die Verbreitung sensibler Daten. Andererseits ist generative KI auch für die Cyberabwehr für erweiterte Datenanalyse, erweiterte Mustererkennung, Erstellung und Analyse von Threat Repositories (Datenbanken zu Cyberbedrohungen) und Code-Analysen sehr nützlich[90].

Insgesamt wird der Einsatz von KI für grundlegende Funktionen wie Kommunikation, Design, Tests usw. positiv gesehen, während die Bedenken hinsichtlich Entscheidungsprozessen und autonomer Raketenstarts überwiegen[91]. Die Intransparenz, Unvorhersehbarkeit und Anfälligkeit für Cyberangriffe sind Argumente dafür, KI nicht in Entscheidungsprozesse einzubeziehen[92].

3.4 Die Debatte um die Strategische Stabilität

Ein Hauptanliegen aller Beteiligten ist die strategische Stabilität, also alles zu vermeiden, was Anlass zu einem nuklearen Erstschlag gibt[93]. Ein solcher Grund könnte die Unsicherheit über die Fähigkeiten des Gegners sein, denn wenn nicht bekannt ist, was die Gegenseite in einer bestimmten Situation tun kann, besteht die einzige Chance in einem nuklearen Konflikt darin, zuerst zuzuschlagen[94]. Im schlimmsten Fall könnte KI einen Atomkrieg „gewinnbar“ machen[95] und hat das Potenzial, die nukleare Abschreckung zu untergraben, indem sie eine Bedrohung für die Zweitschlagfähigkeiten von Atomstaaten darstellt.[96]

KI kann die strategische Stabilität auch dadurch untergraben, dass die Entscheidungszeit verkürzt wird, was zu einer Eskalation oder einem unbeabsichtigten Einsatz von Atomwaffen führen kann, Missverständnisse und Fehleinschätzungen während einer Krise verstärkt und einen vorzeitigen Einsatz unzureichend getesteter KI fördert[97]. KI-gestützte konventionelle Fähigkeiten könnten das Risiko einer unbeabsichtigten Eskalation auch durch die Verknüpfung nuklearer und nicht-nuklearer Waffensysteme erhöhen[98].

Aufgrund der rasanten Fortschritte in der KI-Technologie kann eine umfangreiche Testung zu lange dauern und die militärische KI ist bereits veraltet und hinter den KI-Systemen der Gegner zurück, d.h. der militärische KI-Sektor ist sowohl dynamisch als auch instabil. Da KI-Systeme den Einsatz von Dead-Hand-Systemen und autonomen Atomwaffen wie dem russischen *Perimeter* und *Poseidon* erleichtern, tragen sie zu weiterer Instabilität bei. Der Aufstieg von Hyperschallwaffen, die nicht rechtzeitig entdeckt werden können, ist ebenfalls von entscheidender Bedeutung[99], da die zunehmende Geschwindigkeit der Kriegsführung sowohl die strategische Stabilität untergräbt als auch das Risiko einer nuklearen Konfrontation erhöht [100].

Derzeit gilt die KI noch als noch nicht ausgereift für den Einsatz in strategischen Hochrisikosituationen[101]; es besteht ein erhebliches Risiko technischer Ausfälle und durch

[90] vgl. Saalbach 2023
[91] vgl. Saltini 2023a
[92] vgl. Saltini 2023a
[93] vgl. Shakirov 2023
[94] Siehe auch House of Lords 2023
[95] vgl. Boulanin 2019
[96] vgl. Rooth 2023
[97] vgl. Fayet 2023
[98] vgl. Johnson 2020
[99] vgl. Hruby/Miller 2021
[100] vgl. Johnson 2020
[101] vgl. Saltini 2023a

verzerrte, unvollständige oder ungenaue Daten[102]. Die Atommächte sind sich einig, dass die KI in der Führung und Kontrolle auf strategischer Ebene nur unterstützend eingesetzt werden sollte[103].

4 Zusammenfassung

Der vorliegende Beitrag stellte den aktuellen Stand der Debatte um KI und Atomwaffen und die Hintergründe dar. KI und Big Data könnten zur Kontrolle der Nichtverbreitung von Atomwaffen eingesetzt werden, könnten Zeit und Kosten bei der Erforschung, Designoptimierung, Herstellung, Prüfung und Zertifizierung, Wartung und Überwachung von Atomwaffen sparen und Ressourcen effizienter verwalten. Militärische KI könnte die Frühwarn-, Überwachungs- und Aufklärungsfähigkeiten (ISR) sowie die Zuverlässigkeit der Kommunikation verbessern und die Entscheidungsfindung beschleunigen.

Die Integration von KI in die nuklearen Kommando-, Kontroll- und Kommunikationssysteme (NC3) könnte auch Informationen zwischen nuklearen und nichtnuklearen Kommando- und Kontrollsystemen synchronisieren. Bei autonomen Kernwaffensystemen kann KI die Erkennung und Manövrierfähigkeit von Hindernissen, die automatisierte Zielidentifizierung sowie die Fähigkeit für Langstrecken- und Patrouillenflüge (loitering) unterstützen.

Die Debatte über den Einsatz von KI für Atomwaffen umfasst drei Bereiche: die Autonomie, die Stabilität militärischer KI-Systeme und die strategische Stabilität. Eine mögliche Autonomie von Atomwaffen ist Teil der breiteren Debatte über letale autonome Waffensysteme (LAWS). Automatisierte und teilautonome nukleare Einsatzentscheidungen wurden bereits während des Kalten Krieges in Betracht gezogen (SAGE; Perimeter). Ein spezifisches militärisches KI-Problem ist die Missionsstabilität, da den KI-Systemen das Kontextwissen fehlt und sie möglicherweise zu schnell entscheiden. Die Intransparenz der Systeme führt zu Erklärbarkeits- oder Interpretierbarkeitsproblemen; weitere Probleme können sich aus *data poisoning*, manipulierten Bildern, Automatisierungs-Bias und künstlicher Eskalation ergeben. Kriegs-Simulationen mit aktuellen generativen KI-Large-Language-Modellen (LLMs) von *OpenAI, Anthropic* und *Meta* zeigten, dass die Systeme zur Eskalation bis hin zu nuklearen Angriffen neigen. Als Softwaresystem ist KI anfällig für Cyberangriffe, generative KI auch für prompt injections. Ein Hauptanliegen aller Beteiligten ist die strategische Stabilität, also alles zu vermeiden, was Anlass zu einem nuklearen Erstschlag gibt. Ein solcher Grund könnte die Unsicherheit über die Fähigkeiten des Gegners sein, denn wenn nicht bekannt ist, was die Gegenseite in einer bestimmten Situation tun kann, besteht die einzige Chance in einem nuklearen Konflikt darin, zuerst zuzuschlagen. KI kann die strategische Stabilität auch dadurch untergraben, dass die Entscheidungszeit verkürzt wird, was zu einer Eskalation oder einem unbeabsichtigten Einsatz von Atomwaffen führen kann, Missverständnisse und Fehleinschätzungen während einer Krise verstärkt und einen vorzeitigen Einsatz unzureichend getesteter KI fördert. Darüber hinaus erleichtern KI-Systeme den Einsatz von „Dead-Hand"-Systemen und autonomen Atomwaffen. Der Aufstieg von Hyperschallwaffen und die zunehmende Geschwindigkeit der Kriegsführung untergraben ebenfalls die strategische Stabilität. Derzeit gilt die KI noch als noch nicht ausgereift für den Einsatz in strategischen Hochrisikosituationen; es besteht ein erhebliches Risiko technischer Ausfälle und durch verzerrte, unvollständige oder ungenaue Daten. Die Atommächte sind sich einig, dass die KI in der Führung und Kontrolle auf strategischer Ebene nur unterstützend eingesetzt werden sollte. Die USA, China und Russland haben Dialoge über KI-Risiken aufgenommen.

[102] vgl. Jalil 2023
[103] vgl. Su/Yuan 2023

5 Literaturverzeichnis

Air Force (2024): B-21 Raider Fact Sheet. Article 2682973 www.af.mil

Arrieta, A.B. et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion 58 (2020), p. 82–111

Bajak, F. (2023): Pentagon's AI Initiative accelerate hard decisions on lethal autonomous weapons. AP News 25 Nov 2023

Bommakanti, K. (2020): A.I. in the Chinese Military: Current Initiatives and the Implications for India Observer Research Foundation (ORF) Occasional Paper 234 February 2020

Boulanin, V. (2019): The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Volume I Euro-Atlantic Perspectives Stockholm International Peace Research Institute (SIPRI) May 2019

Burns et al., (2023): Weak-To-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. Joint project paper of the OpenAI Superalignment Generalization team.

Chaudhry, H., Klein, L. (2023): Artificial Intelligence and Nuclear Weapons: Problem Analysis and US Policy Recommendations. policy@futureoflife.org 14th November 2023 Future of Life Institute (FLI)

CNA (2023): Arms Control and Lethal Autonomy CNA Corporation Analysis Paper

CoE (2022): Emergence of lethal autonomous weapons systems (LAWS) and their necessary apprehension through European human rights law Draft resolution unanimously adopted by the Committee on Legal Affairs and Human Rights of the Council of Europe on 14 November 2022 AS/Jur (2022)

Congress (2023): Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023 Bill Text (PDF) BUR23348 GC1

DARPA (2016): Cyber Grand Challenge https://www.cybergrandchallenge.com 05 Aug2016

Depp, M. and Scharre, P. (2024): Artificial Intelligence and Nuclear Stability. War on the Rocks 16 January 2024, warontherocks.com

Duboust, O. (2024): AI models chose violence and escalated to nuclear strikes in simulated wargames. https://www.euronews.com/next/2024/02/22/ai-models-chose-violence-and-escalated-to-nuclear-strikes-in-simulated-wargames. Updated 23/02/2024

DoD (2022): 2022 National Defense Strategy of The United States of America including the Nuclear Posture Review NPR 2022 and the Missile Defense Review MDR 2022, https://media.defense.gov/2022/Oct/27/2003103845/-1/-

DoD (2023a): DOD DIRECTIVE 3000.09. Autonomy In Weapon Systems. Originating Component: Office of the Under Secretary of Defense for Policy Effective: January 25, 2023 Releasability: Cleared for public release

DoD (2023b): DoD Data, Analytics, and AI Adoption Strategy. Cleared for open publication June 27, 2023, Department of Defense/Office of Prepublication and Security Review

Dresp-Langley, B. (2023): The weaponization of artificial intelligence: What the public needs to be aware of. Front. Artif. Intell. 6:1154184. doi: 10.3389/frai.2023.1154184

Fayet, H. (2023): French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making. Report 13 November 2023 European Leadership Network (ELN)

Ford, C.A. (2020): Arms Control and International Security Papers Volume I, Number 2 I April 20, 2020 Office of the Under Secretary of State for Arms Control and International Security

Heslop, D, Keep, J. (2024): The 2024 China-US AI Dialogue Should start with an Eye on Chem-Bio Weapons. The Diplomat 09 March 2024

Hoadley, D.S., Sayler, K.M. (2019): Artificial Intelligence and National Security Congressional Research Service R45178 Version 6 Updated November 21, 2019

House of Lords (2023): Proceed with Caution: Artificial Intelligence in Weapon Systems. AI in Weapon Systems Committee. Report of Session 2023–24 HL Paper 16

Hruby, J., and Miller, M.N. (2021): Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems Nuclear Threat Initiative (NTI)

ISAB (2023): Report on the Impact of Artificial Intelligence and Associated Technologies on Arms Control, Nonproliferation, and Verification October 2023. A report of the International Security Advisory Board (ISAB), a Federal Advisory Committee for the Department of State

Jalil, G.Y. (2023): Artificial Intelligence and Nuclear Weapons: Way to the Future or Path to Disaster? Arms Control & Disarmament Centre, Institute of Strategic Studies Islamabad ISSI – Issue Brief. December 11, 2023

Johnson, J.S. (2020): Artificial Intelligence: A Threat to Strategic Stability. Strategic Studies Quarterly Spring 2020, p.16-39

Joint Statement (2022): Principles and responsible practices for Nuclear Weapon States 2022. Working paper submitted by France, the United Kingdom of Great Britain and Northern Ireland and the United States of America

Kasperowicz, P. (2023): Pentagon moving to ensure human control so AI doesn't 'make the decision for us'' Fox News 21 April 2023

Kaur, S. (2024): Artificial Intelligence and the Evolving Landscape of Nuclear Strategy Union of Concerned Scientists (UCS) https://blog.ucsusa.org/science-blogger/artificial-intelligence-and-the-evolving-landscape-of-nuclear-strategy/ March 4, 2024

Kölling, M. (2023): Künstliche Superintelligenz ist in Sicht. Neue Zürcher Zeitung, 06 Oct 2023, p.17

Longpre et al. (2022): Longpre, S., Storm, M. and Shah, R. MIT Science Policy Review August 29, 2022, vol. 3, pg. 47-55 This article is licensed under a Creative Commons Attribution 4.0 International License http://creativecommons.org/licenses/ by/4.0/

Lowther, A., McGriffin, C. (2019): America needs a dead hand. https://warontherocks.com/2019/08/america-needs-a-dead-hand/16 Aug 2019

Lowther, A., McGriffin, C. (2024): America needs a dead hand more than ever. https://warontherocks.com/2024/03/america-needs-a-dead-hand-more-than-ever/

Markey E.J. et al. (2023): Markey, Lieu, Beyer, and Buck Introduce Bipartisan Legislation to Prevent AI From launching a Nuclear Weapon. April 26, 2023 https://www.markey.senate.gov/news/press-releases/markey-lieu-beyer-and-buck-introduce-bipartisan-legislation-to-prevent-ai-from-launching-a-nuclear-weapon

Masuhr, N. (2019): AI in Military Enabling Applications. CSS Analyses in Security Policy No. 251, October 2019

NATO (2019): Artificial Intelligence: Implications for NATO's Armed Forces. Science and Technology Committee (STC) - Sub-Committee on Technology Trends and Security

(STCTTS) Rapporteur: Matej Tonin (Slovenia) 149 STCTTS 19 E rev. 1 fin Original: English 13 October 2019

NNSA (2023): Artificial intelligence for nuclear deterrence strategy. National Nuclear Security Administration NNSA Department of Energy DOE/NNSA-0145

Norman, G. (2024): State Department wants China, Russia to declare that AI won't control nuclear weapons, only humans. Fox News, 02 May 2024

NTA (2021): DARPA issues AI exploration opportunity for Shell Project – Proposals due July, 27. nta.org

OpenAI (2024): https://openai.com/index/gpt-4o-and-more-tools-tochatgpt-free Update ChatGPT4o on 14 May 2024

Pauwels, E. (2019): The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI, United Nations University Centre for Policy Research, 29 April 2019.

Pauwels, E. (2021): Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE). Hybrid CoE Strategic Analysis / 26 May 2021

Porter, T. (2023): The Pentagon is moving toward letting AI weapons autonomously decide to kill humans. Business Insider 22 Nov 2023

Puwal, S. (2024): Should AI be banned from nuclear weapon systems? NATO Review 12 April 2024

Raasch, J.M. (2023): Cheap drones can take out expensive military systems, warns former Air Force Pilot pushing AI-enabled force. Fox News online, 08 Dec 2023

Rivera, J.P. et al. (2024): Escalation Risks from Language Models in Military and Diplomatic Decision-Making. arXiv:2401.03408v1 [cs.AI] 7 Jan 2024

Rooth, C. (2023): The impact of Artificial intelligence on nuclear deterrence July 2023 Info Flash FINABEL - European Army Interoperability Centre.

Saalbach, K. (2023): Künstliche Intelligenz und Cyberangriffe. Arbeitspapier. https://doi.org/10.48693/418.

Saltini, A. (2023a): AI and nuclear command, control and communications: P5 perspectives. November 2023 European Leadership Network (ELN) Published under the Creative Commons Attribution-ShareAlike 4.

Saltini, A. (2023b): UK thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making Report 13 November 2023 European Leadership Network (ELN)

Sankaran, J. (2019): A different use for Artificial Intelligence in Nuclear Weapons Command and Control https://warontherocks.com/2019/04/a-different-use-for-artificial-intelligence-in-nuclear-weapons-command-and-control/

Sayler, K.M. (2023): Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems Congressional Research Service CRS Paper IF 11150 Updated May 15, 2023

Shakirov, O. (2023): Russian thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making Report 13 November 2023 European Leadership Network (ELN)

Su, F., Yuan, J. (2023): Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decision making. Report 13 November 2023 European Leadership Network (ELN)

US (2023): Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/ Bureau of Arms Control, Deterrence, and Stability November 09, 2023

Voke, M.R. (2019): Artificial Intelligence for Command and Control of Air Power. Wright Flyer Paper No. 72 Air University Press

White House (2023): FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Wolff, J. (2020): How to Improve Cybersecurity for Artificial Intelligence. Brookings Report 08 June 2020

Wright, N.D. (2019): Artificial Intelligence, China, Russia, and the Global Order Technological, Political, Global, and Creative Perspectives. Air University Press in October 2019

Zia, H. (2023): Information Revolution and Cyber Warfare: Role of Artificial Intelligence in Combatting Terrorist Propaganda Pakistan Journal of Terrorism Research, Vol-03, Issue-2, 133