KI außer Kontrolle

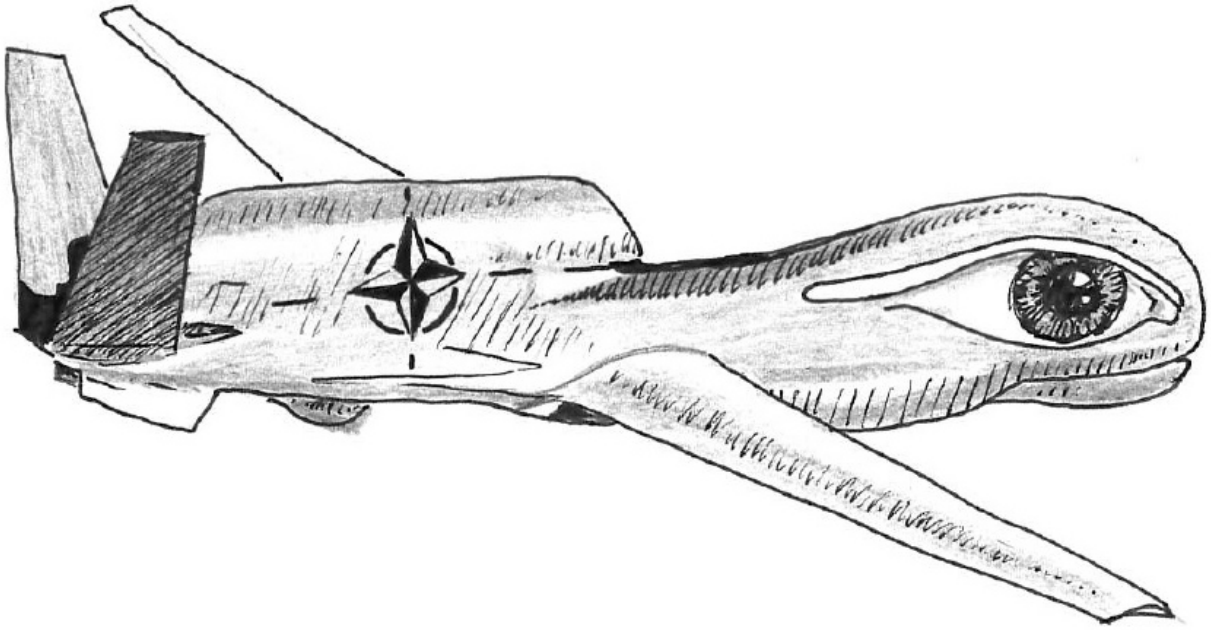
Autonome Waffensysteme und Meaningful Human Control

von Jens Hälterlein

Eines der dringendsten Probleme der internationalen Sicherheit und des Friedens ist die Entwicklung und Anwendung sog. Autonomer Waffensysteme (Aws). Der Begriff „autonom“ verweist darauf, dass diesen Systemen ein hoher Grad an Eigenständigkeit zugeschrieben wird. Die Steuerung unbemenschter militärischer Fahrzeuge (Drohnen, Panzer, U-Boote etc.) ebenso wie die Identifikation, Auswahl und das Angreifen von Zielen werden als Prozesse betrachtet, die solche Systeme „autonom“, das heißt ohne menschliches Zutun, vollziehen könnten. Befürworter*innen von Aws verbinden damit das Versprechen, militärische Operationen schneller und präziser (als der Gegner) ausführen zu können, ohne dabei das Leben der eigenen Soldat*innen unnötig zu gefährden. Kritiker*innen von Aws betonen hingegen, dass mit der Ausweitung KI-basierter, maschineller Autonomie ein Verlust menschlicher Kontrolle über diese Kampf-Maschinen droht.[1] Zahlreiche politische, zivilgesellschaftliche und wissenschaftliche Akteure (z.B. die Kampagne Stop Killer Robots) verbinden diese Kritik mit der politischen Forderung, eine hinreichende, effektive menschliche Kontrolle („Meaningful Human Control“) über Aws zu ermöglichen bzw. aufrechtzuerhalten. Der Begriff „meaningful“ besagt im Kern, dass es nicht ausreicht, wenn Menschen beim Einsatz von Aws formal die Letztentscheidung über zentrale Kriegshandlungen (insb. Tötungsentscheidungen) haben, sondern dass diese Entscheidung auf einem ausreichend tiefen Verständnis des von einer Maschine generierten Outputs, der als Entscheidungsgrundlage dienen soll, beruhen muss. Das setzt insb. die Möglichkeit zur Überprüfung und Bewertung dieses Outputs ohne übermäßigen Zeitdruck und sonstige einschränkende Faktoren voraus.[2]

Die Gewährleistung bzw. Aufrechterhaltung einer „Meaningful Human Control“ ist aus dieser kritischen Perspektive unerlässlich, da Aws – wie jedes KI-basierte System – notorisch vorurteilsbehaftet und daher fehleranfällig sind. So liegen bspw. bei Gesichtserkennungssystemen die Fehlerkennungsraten der zugrundeliegenden Algorithmen bei Gesichtern, die als weiblich, afroamerikanisch oder asiatisch klassifiziert werden, häufig deutlich höher als bei männlich und weiß gelesen Gesichtern, wobei mögliche Überschneidungen dieser Kategorien die höchsten Fehlerquoten aufweisen. Der Grund hierfür ist, dass in den Datensätzen, mit denen die Algorithmen trainiert werden, weiße männliche Gesichter zumeist deutlich überrepräsentiert sind und der Algorithmus daher besser darin ist, genau diese zu erkennen.[3] Während bereits der polizeiliche Einsatz von fehleranfälligen Gesichtserkennungssystemen schwerwiegende Konsequenzen haben kann, hat diese als „machine bias“ bezeichnete Eigenschaft von KI-basierten Systemen im militärischen Kontext zumeist sogar tödliche Folgen. Das Zielerkennungssystem eines Aws könnte z.B. Zivilist*innen als Kombattant*innen identifizieren, weil es bestimmte äußerliche Merkmale oder Verhaltensweisen fehlinterpretiert, oder weil es andere, für das richtige Verständnis der Situation relevante Faktoren übersieht. Die Unterscheidung von Kombattant*innen und Zivilist*innen ist allerdings nicht der einzige Punkt, in dem ein fehlerhaftes Aws völkerrechtlich betrachtet problematisch wäre. Ebenso zentral für das humanitäre Völkerrecht ist die auf dieser Unterscheidung basierende Abwägung und Wahrung der Verhältnismäßigkeit von potenziellen zivilen Opfern eines Angriffs auf ein legitimes militärisches Ziel (die sog. „Kollateralschäden“). In eine solche Abwägung müssen zahlreiche Aspekte und Faktoren miteinfließen. Vorab festgelegte quantitative Relationen genügen hier keineswegs. Daher gibt es auch ebenso viele Einfallstore für Vorurteile und Fehlleistungen einer „intelligenten“ Kampfmaschine, insofern dieser entsprechende Aufgaben übertragen werden. Nicht zuletzt deshalb betonen militärische Akteure, dass die Letztentscheidung über Kriegshandlungen mit potenziell tödlichem Ausgang weiterhin beim Menschen verbleibt – obwohl es rein technisch gesehen mittlerweile möglich sei, den kompletten Observation-Orientation-Decision-Action-(Ooda)-Zyklus einem Aws zu überlassen.

Dennoch gibt es ausreichend Grund zur Skepsis. Denn auch wenn die Zielidentifikation nicht unmittelbar zu einem Angriff auf das Ziel führt, sondern zunächst durch einen Menschen überprüft und bestätigt werden muss, ist



keinesfalls gewährleistet, dass ein fehlerhafter Output ohne Konsequenzen bleibt. Zahlreiche empirische Untersuchungen über den Einsatz algorithmischer Entscheidungsunterstützungssysteme zeigen, dass die Benutzer*innen den Output der Systeme kaum infrage stellen und sogar dazu neigen, diese als unfehlbar zu betrachten und somit einem automation bias unterliegen.[4] Ein solcher automation bias hat im militärischen Kontext bereits zu mehreren tödlichen Entscheidungen beigetragen, unter anderem bei einem Einsatz des Patriot-Raketensystems der US-Armee, das 2004 während des Irakkrieges einen britischen Tornado und eine amerikanische F/A-18 abschoss.[5] Ein automation bias führt zu zwei Arten von Fehlern: Bei einem „commission error" folgen die Benutzer*innen einer fehlerhaften Empfehlung eines automatisierten Entscheidungsunterstützungssystems. Übertragen auf das Szenario Aws würde dies bedeuten, dass das System Zivilist*innen fälschlicherweise als Kombattant*innen identifiziert, die Bediener*innen diesen Output jedoch nicht hinterfragen und dementsprechend, ohne es zu wissen, einen Angriff auf Zivilist*innen autorisieren. Bei einem „omission error" hingegen, übersehen die Benutzer*innen kritische Situationen, wenn diese nicht bereits vom System erkannt werden. Übertragen auf das Szenario Aws würde dies bedeuten, dass sie eine Gefahr, z.B. einen feindlichen Panzer oder Raketenwerfer, erst gar nicht zur Kenntnis nehmen, insofern diese vom Aws nicht als mögliche Ziele identifiziert wurden – was ebenfalls tödliche Folgen haben könnte (in diesem Fall für die eigenen Truppen).

Aber auch wenn menschliche Bediener*innen nicht einem automation bias unterliegen, sondern den Output eines Systems kritisch hinterfragen möchten, garantiert dies noch keine Meaningful Human Control. Denn es stellt sich die Frage, wie Menschen den Output eines auf KI-basierenden Systems überhaupt hinreichend verstehen könne, um eine informierte Entscheidung zu treffen. Denn dessen Verstehen würde letztlich eine detaillierte Erläuterung der verarbeiten Daten und der Datenverarbeitungsmethoden erfordern und hängt zugleich von der Komplexität der Algorithmen und dem Fachwissen der jeweiligen Bediener*innen ab. Bei Algorithmen aus dem Bereich des Maschinellen Lernens und insbesondere bei Künstlichen Neuronalen Netzen, die bei Bildverarbeitungsanwendungen wie der Zielerfassung besonders leistungsfähig sind, wären sogar Expert*innen aus dem Bereich der Informatik ggf. nicht in der Lage, die Funktionsweise des Systems im Detail zu verstehen. Daher wird ein KI-basiertes Entscheidungsunterstützungssystem selbst bei vollständiger Transparenz für den Anspruch einer Meaningful Human Control zum Problem, ganz zu schweigen von daraus resultierenden Fragen strafrechtlicher Verantwortung für militärische Handlungen. Denn diese liegt immer bei einem Menschen, unabhängig davon, wieviel Autonomie einem System, das in Entscheidungsprozesse involviert ist, zuerkannt wird.[6]

Schienen diese Fragen noch vor kurzem eher hypothetischer Natur zu sein und der Einsatz von Aws in realen Kriegskontexten ein Zukunftsszenario, so haben insbesondere die Kriege in der Ukraine und im Nahen Osten gezeigt, dass KI-basierte Kampfdrohnen und Zielfindungssystemen bereits zu einem zentralen Element gegenwärtiger Kriegsführung geworden sind – mit den uns bekannten katastrophalen Folgen. Umso wichtiger ist es, dass bereits seit 10 Jahren versucht wird, im Rahmen der UN-Rüstungskontrolle (der *UN-Konvention über Bestimmte Konventionelle Waffen*) verbindliche Regulierungen für *letale autonome Waffensysteme* (LAWS) zu erwirken – bisher allerdings ohne Erfolg. In erster Linie, da dies nicht im Interesse der Nationen ist, die solche Syste-

me entwickeln, bereits einsetzen oder es in Zukunft beabsichtigen. Das sind im Wesentlichen die USA, Großbritannien, Israel, Russland und China. Auf der anderen Seite stehen neben einigen internationalen NGOs insb. Akteure aus dem Globalen Süden, die sich für ein Ächtung von LAWS einsetzen. Deutschland hat sich zusammen mit Frankreich zwar für ein Verbot von sog. vollautonomen Waffensystemen eingesetzt, möchte aber gleichzeitig den Einsatz von sog. teilautonomen unter bestimmten Voraussetzungen ermöglichen. Diese Position muss in Zusammenhang mit der Entwicklung des *Future Combat Air Systems* (FCAS) betrachtet werden, das von beiden Nationen zusammen mit Spanien und Belgien entwickelt wird. FCAS soll hochgradig autonom agierende Drohnen und ein KI-basiertes Entscheidungsunterstützungssystem beinhalten und ab etwa 2040 den Kern der europäischen Luftstreitkräfte bilden. Als Alternative zu rechtlichen Regulierungen werden seit einigen Jahren freiwillige Selbstverpflichtungen präsentiert, in denen staatliche und militärische Akteure erklären, ausschließlich „vertrauenswürdige, erklärbare und transparente“ KI zum Einsatz zu bringen. So findet sich mittlerweile in einer Reihe von offiziellen militärischen Dokumenten ein freiwilliges Bekenntnis zum Leitbild der verantwortungsvollen militärischen KI („Responsible AI in the Military Domain“).[7] Letztlich ist es aber fragwürdig, ob diese freiwilligen Selbstverpflichtungen ein effektives Mittel der Rüstungskontrolle darstellen können. Insbesondere wenn sie rechtliche Regulierungen nicht ergänzen, sondern ersetzen, was gegenwärtig der Fall zu sein scheint. Die gegenwärtige globale Polarisierung und Militarisierung erzeugt ein Klima, in dem für ein Verbot oder zumindest ein vorläufiges Moratorium von AWS wenig Raum ist. Eher ist zu erwarten, dass die globale Verbreitung von AWS in Zukunft noch zunehmen wird. Denn KI-basierte Kampfdrohnen sind wesentlich preisgünstiger als Kampfjets. Gerade kleine Drohnen mit relativ kurzer Flugzeit und geringer Nutzlast sind in der Anschaffung und im Betrieb vergleichsweise preisgünstig. Der größte Kostenfaktor sind hier Softwarekomponenten. Aber auch diese Kosten sind kaum mit den Gesamtkosten für die Entwicklung der F-35 oder des FCAS vergleichbar. Es ist also davon auszugehen, dass Kosten einen zunehmend geringen Faktor bei der globalen Verbreitung KI-basierter Militärtechnologien darstellen. Das türkische Unternehmen Baykar hat bspw. seine KI-Kampfdrohne Bayraktar TB2 eben nicht nur in die Ukraine exportiert, sondern auch in zahlreiche afrikanische und asiatische Staaten, wo sie bereits in mehreren Kriegen zum Einsatz gekommen ist.

Ob sich beim Einsatz dieser Systeme eine Meaningful Human Control realisieren lässt und ob dies überhaupt von militärischer Seite gewollt ist, bleibt eine offene Frage. Grund zum Optimismus gibt es wenig. Umso wichtiger ist es, weiter auf ein Verbot von Waffensystemen hinzuarbeiten, die das Prinzip einer Meaningful Human Control konterkarieren.

Anmerkungen

1 Altmann, Jürgen (2019). Autonomous Weapon Systems – Dangers and Need for an International Prohibition. In Christoph Benzmüller & Heiner Stuckenschmidt (Hrsg.), KI 2019: Advances in Artificial Intelligence. 42nd German Conference on AI, Kassel, September 23-26, 2019: Proceedings, Cham: Springer, S. 1-17.

2 Article 36 (2016). Key elements of Meaningful Human Control. www.article36.org.

3 Hälterlein, Jens (2024) Biometrische Gesichtserkennung – Technologischer Solutionismus für mehr Sicherheit. In: CILIP Bürgerrechte &; Polizei 134. www.cilip.de.

4 Skitka, L. J., Mosier, K. L. und Burdick, M. (1999): Does automation bias decision-making? In: International Journal of Human-Computer Studies 51 (5), S. 991-1006.

5 Cummings, M. L. (2015): Automation Bias in Intelligent Time Critical Decision Support Systems. In: Harris, D. und Li, W.-C. (Hg.): Decision Making in Aviation. London, S. 289-294.

6 Barlag, Schirin & Beck, Susanne (2024). Menschlichkeit im Krieg? Die Bedeutung von „Meaningful Human Control“ für die Regulierung von autonomen Waffensystemen. *Ethik und Militär*, Heft 1, S. 60-67.

7 French Ministry of Armed Forces (2019): Artificial Intelligence in Support of Defence. Report of the AI Task Force. www.defense.gouv.fr. NATO (2021): An Artificial Intelligence Strategy for NATO. www.nato.int . U.S. Department of Defense (2020): DOD Adopts Ethical Principles for Artificial Intelligence. www.defense.gov.