Künstliche Intelligenz für die Cybersicherheit

Künstliche Intelligenz (KI) wird die Cybersicherheit in den kommenden Jahren voraussichtlich verändern. KI wird sowohl Angriff als auch Verteidigung im Cyberraum weiterentwickeln und die Landschaft von Cyberbedrohungen mitprägen. Der Umgang mit diesen Veränderungen fordert vor allem staatsnahe Akteure heraus, die einen angemessenen politischen und normativen Rahmen schaffen müssen.

Von Matteo E. Bonfanti
und Kevin Kohler

KI ist ein Sammelbegriff, der 1955 vom Computerwissenschaftler John McCarthy geprägt und als «Wissenschaft und Technik intelligenter Maschinen» definiert wurde. Heute bezeichnet KI wegbereitende Systeme und ein Forschungsgebiet. Als solches ist KI eine wissenschaftliche Disziplin mit dem Ziel, künstliche Systeme zu Tätigkeiten zu befähigen, von denen man annimmt, dass Menschen für deren Ausführung ein gewisses Mass an rationalem Denken oder Intelligenz benötigen. Zu diesem Zweck gibt es verschiedene Ansätze. Einer davon ist maschinelles Lernen mit den Kernbestandteilen Lernalgorithmen, Daten und Rechenleistung für das Trainieren von Algorithmen. Die meisten Erfolge der jüngeren Zeit wurden in einem Teilgebiet des maschinellen Lernens erzielt: *Deep Learning* (tiefes Lernen). Dabei werden tiefe neuronale Netze bestehend aus zahlreichen Schichten mit künstlichen Neuronen verwendet, die eingegebene Daten verarbeiten. Als Inspiration für die neuronalen Netzwerke dient das menschliche Gehirn. Aufgrund ihrer zunehmenden Lernkapazität und Entscheidungsfähigkeit dürften künstliche Systeme mit der Zeit immer autonomer werden.

Durch maschinelles Lernen können auf Basis von Fotos echter Menschen Bilder erstellt werden.
*Raten Sie: Welches dieser Kinder gibt es und welches ist KI-generiert? (*Lösung auf der letzten Seite)*

KI wird zudem als Befähigungstechnologie bezeichnet, da sie in zahlreichen Gebieten Anwendung findet, wie etwa im Bevölkerungsschutz (siehe CSS Analyse Nr. 260) und zu militärischen Zwecken (siehe CSS Analyse Nr. 251). Dabei kann sie für Gutes eingesetzt werden, aber auch, um Schaden anzurichten. Wenig überraschend ist, dass sie auch im Bereich Cybersicherheit zielführend genutzt werden kann, etwa als sogenannte «KI für die Cybersicherheit». Dieser Ausdruck bezieht sich auf technologische Lösungen: Strategien und Fähigkeiten des maschinellen Lernens werden für die Verarbeitung grosser Informationsmengen eingesetzt. Die daraus abgeleiteten Erkenntnisse können das Vorgehen zur Erfüllung bestimmter Ziele im Cyberraum beeinflussen.

Sicherheitsbedenken zu KI

Die Forschungsgemeinschaft rund um KI pflegt eine Vorliebe für Offenheit – der Glaubenssatz der Forschenden lautet, dass

Wissen frei verfügbar sein sollte. Sie streben ausserdem nach öffentlichen Zeitstempeln für die Forschung, um in einem hochdynamischen Gebiet professionelles Ansehen zu erlangen. Deshalb veröffentlichen sie oft nicht nur vage Beschreibungen ihrer Ergebnisse, sondern verbreiten Quellcodes, trainierte Modelle, Tutorials und sogar Datensätze offen im Internet. Auch der Zugang zum letzten Hauptbestandteil von KI – Rechenleistung – ist durch Anbieter von auf Anfrage verfügbaren *Cloud*-Diensten einfacher geworden. Deshalb breiten sich Forschungsaktivitäten und Entwicklungen zu KI schnell aus. Dieses Phänomen wird manchmal auch Demokratisierung der KI genannt.

KI-Forschende untersuchen neue Ansätze, um die aus der Verbreitung immer leistungsstärkerer, vielfach einsetzbarer Tools entstehenden Schäden – etwa durch böswillige Akteure – einzudämmen. Dazu gehören Lizenzen für verantwortungsvolle KI oder die stufenweise Veröffentlichung trainierter Modelle. Letzteres basiert auf der verantwortungsvollen Veröffentlichung von Zero-Day-Sicherheitsrisiken in der Cybersicherheit und wurde erstmals 2019 bei der Veröffentlichung des Sprachmodells GPT-2 von OpenAI umgesetzt. Auf Anwendungsebene herrscht weniger Offenheit, da Unternehmensdatensätze und damit trainierte Modelle aus wirtschaftlicher Sicht als immaterielle Vermögenswerte gelten, deren Vertraulichkeit gegen Diebstahl und Spionage geschützt werden muss. In einigen Ländern gibt es Exportbeschränkungen für besonders sensible Algorithmen oder Datensätze wie etwa genomische Informationen über die Bevölkerung. Ausserdem wird die stärkere Einschränkung von KI-Hardware und zugehöriger Tools diskutiert.

KI ist noch immer nicht robust genug und scheitert vor allem auf merklich nichtmenschliche Art und Weise. Beispielsweise könnte KI Bilder aufgrund zufälliger Hintergrundkorrelationen im Trainingsdatensatz, ungewöhnlicher Blickwinkel oder Manipulationen von untergeordneten Elementen, die Menschen nur unbewusst wahrnehmen, falsch klassifizieren. Daraus entstehen zahlreiche neue, noch nicht ausgemerzte und oft unbekannte Schwachstellen, die gegnerische Akteure ausnutzen können, um die Entscheidungsqualität KI-gestützter Systeme zu beeinträchtigen. Eine Ausnutzung könnte etwa «Datenvergiftungs»-Angriffe beinhalten, wobei Trainingsdaten durch Einspeisung manipuliert werden, damit der Lernalgorithmus Fehler macht. Des Weiteren könnten «feindliche Beispiele» eingesetzt werden: digitale Inputs sowie Artefakte aus dem echten Leben, die so konzipiert wurden, dass Anwendungen, die maschinelles Lernen nutzen, diese falsch klassifizieren. Das ist vor allem dann effektiv, wenn die KI-Modell-Parameter bekannt sind – bei sogenannten White-Box-Angriffen. Dennoch können solche Angriffe auch ohne dieses Wissen, als sogenannte Black-Box-Angriffe, erfolgreich sein.

> KI kann Cyberbedrohungen vermehren, qualitativ verändern und neue Bedrohungen hervorbringen.

KI und Cyberbedrohungen

Die Anwendung von KI im Cyberraum kann die Landschaft der Cyberbedrohung auf drei Arten beeinflussen: Vorausgesetzt, es bestehen keine nennenswerten Präventionsmassnahmen, könnte KI bestehende Cyberbedrohungen verstärken (Menge), die typischen Merkmale dieser Bedrohungen verändern (Qualität) sowie neue, unbekannte Bedrohungen hervorbringen (Menge und Qualität).

Durch KI könnten noch mehr Akteure zur Ausführung böswilliger Cyberaktivitäten fähig sein. Diese Akteure könnten ihre Aktivitäten häufiger und gegen mehr mögliche Ziele durchführen. Dies lässt sich aus der Effizienz, Skalierbarkeit und Anpassungsfähigkeit von KI sowie der «Demokratisierung» von Forschung und Entwicklung auf diesem Gebiet folgern. Speziell die Verbreitung von KI-Komponenten unter den traditionellen Akteuren der Cyberbedrohung – Staaten, Kriminelle, HacktivistInnen und Terrorgruppen – könnte dazu führen, dass die Durchführung von Angriffen für eine steigende Zahl von Organisationen bezahlbar wird. Da KI-Anwendungen ausserdem skalierbar sind, könnten Akteure mit genügend Ressourcen für Angriffe solche auch häufiger durchführen. Neue Ziele könnten sich als lohnenswert erweisen.

In qualitativer Hinsicht könnten KI-gestützte Cyberangriffe auch effektivere, zielgerichtetere und komplexere Aktionen und Angriffe ermöglichen. Dabei leitet sich der höhere Wirkungsgrad aus der Effizienz, Skalierbarkeit und Anpassungsfähigkeit dieser Lösungen ab. Potenzielle Ziele können schneller identifiziert und geprüft werden.

Durch KI könnte zudem eine neue Art böswilliger Aktivitäten entstehen, die jene Schwachstellen ausnutzen, die durch die Nutzung ebendieser Technologien in die Cybersysteme gelangt sind. In diesem Fall wird Cybersicherheit selbst zum Gegenstand von KI-Forschung und -Entwicklung. KI-gestützte Cybersysteme benötigen Schutz vor Cyberstörungen oder Angriffen, damit sie weiterhin korrekt funktionieren, verlässlich und integer sind. Praktiken für die Cybersicherheit sowie die Förderung grösserer Cyberhygiene-Programme mit spezifischen Anforderungen an die KI-Forschung, -Entwicklung und -Anwendung können unter dem Begriff «Cybersicherheit für KI» zusammengefasst werden.

Defensiver und offensiver Einsatz

Viele Eigenschaften, aufgrund derer sich KI für Anwendungen zur Cyberverteidi-

KI-gestützte Erkennung von Malware

Durch das Training mit einem grossen Datensatz, der als Goodware oder Malware gekennzeichnete Dateien beinhaltet, kann ein neuronales Netzwerk in annehmbarem Umfang lernen, eine neue Datei ohne Rückgriff auf manuell aktualisierte Listen als bösartig einzustufen. Das wird wohl dazu beitragen, die **Erkennung moderner, neuer Malware zu verbessern**, die automatisch neue Varianten generieren kann, um die traditionellen regelbasierten Ansätze zur Identifizierung zu umgehen. Diese Varianten können mithilfe von KI-gestützter Malware-Erkennung der richtigen Malware-Gruppe zugeordnet werden. Gleichzeitig ist diese binäre Klassifizierung keineswegs einfach. In der Masse all dieser Dateien ist die Häufigkeit von Malware sehr gering, weshalb die **Klassifizierung oft falsch positive Resultate ergibt** und ausführbare Dateien seriöser Softwareprodukte blockiert. Einige Unternehmen haben als Übergangslösung eine «weisse Liste» für gefahrlose Dateigruppen angelegt, jedoch zeigt Forschung, dass eine Malware dann einfach um Dateien von weissen Listen erweitert werden kann und so unentdeckt bleibt. Daher ist in absehbarer Zukunft die **KI-gestützte Erkennung von Malware kein Ersatz für traditionelle Methoden, sondern eine ergänzende Massnahme** dazu.

gung eignet, machen sie gleichzeitig nützlich für Cyberangriffe. In den nächsten drei bis fünf Jahren ist deshalb zu erwarten, dass Unternehmen KI-basierte Cyberverteidigungs-Mechanismen zum Schutz ihrer Vermögenswerte wie Netzwerke, Informationen und Mitarbeitende einführen werden, um Gegner abzuwehren, die sowohl KI-gestützte als auch Nicht-KI-gestützte Angriffe starten könnten. In ähnlicher Weise wird es Akteure geben, die KI-gestützte Mechanismen für Cyberangriffe einsetzen, um Ziele mit oder ohne KI-gestützter Cyberverteidigung zu kompromittieren. Besonders bei Handlungen, die auf den Schutz vor oder die Durchführung von Computer-Netzwerk-Operationen, ob Angriffe oder Ausnutzung, abzielen, könnten KI-basierte Cyberfähigkeiten eine Unterstützung sein. KI wird wohl auch den Schutz vor und die Durchführung von sogenannten Cyberinformations- und Beeinflussungsoperationen begünstigen.

Der Einsatz von KI zur Generierung von *Cyberintelligence*, sprich verwertbarem Wissen zur Unterstützung der Entscheidungsfindung im Zusammenhang mit dem Cyberraum, kann sowohl in Sachen Verteidigung als auch Angriff ein Vorteil sein. Tatsächlich kann KI verschiedene Funktionen eines *Cyberintelligence*-Prozesses übernehmen, insbesondere die Sammlung, Verarbeitung und Analyse von Informationen. Sie kann die Informationserfassung vorantreiben und deren Reichweite auf eine Vielzahl von Quellen und Endpunkte ausweiten. Auch die Auswahl und Gegenprüfung von Informationen mittels zusätzlicher Daten aus anderen Quellen könnte verbessert werden. KI kann ausserdem Analyseprozesse durch das Auffinden versteckter Muster und Korrelationen in verarbeiteten Daten unterstützen. Durch die Nutzung von KI-Fähigkeiten für diese Funktionen wird der *Cyberintelligence*-Prozess hinsichtlich Automation und Geschwindigkeit wahrscheinlich voranschreiten.

Computer-Netzwerk-Operationen

Die Fähigkeit von KI-Komponenten zur Generierung von *Cyberintelligence* übersetzt sich auf spezifische Verteidigungsanwendungen auf taktischer und technischer sowie operativer Ebene der Cybersicherheit. Im operativen Kontext könnte KI für die Auslesung und Verarbeitung von Daten aus Programmen für die Netzwerksicherheitsanalyse genutzt werden. KI könnte diese Daten dann mit anderen verfügbaren Informationen abgleichen. Auf taktischer Ebene wird KI zunehmend als Hilfe bei der Suche nach, Analyse von und, falls möglich, Prävention von Cyberbedrohungen dienen. Insbesondere wird KI zur Aufwertung von Angriffserkennungssystemen *(Intrustion Detection Systems,* IDS*)* mit dem Ziel der Auffindung illegaler Aktivitäten in Computern oder Netzwerken beitragen. Dasselbe gilt für Spam- oder Phishing-Erkennungssysteme sowie Tools für die Malware-Erkennung und Analyse (siehe Box). KI-Komponenten werden künftig auch für Multi-Faktoren-Authentifizierungs- oder Identifikationssysteme eingesetzt werden und so helfen, die Verhaltensmuster eines bestimmten Benutzers oder einer Benutzerin aufzuzeichnen und mögliche Musterveränderungen zu erkennen. Ein weiteres vielversprechendes Gebiet für den taktischen Einsatz von KI zur Verteidigung ist das automatisierte Testen auf Schwachstellen, auch Fuzzing genannt.

KI-Anwendungen werden auch für Cyberangriffe eingesetzt werden, um beispielsweise ein Zielunternehmen oder Benutzer, deren Netzwerke und die verarbeiteten Daten zu kompromittieren. Sie werden noch zahlreichere und komplexere Cyberangriffe ermöglichen. Wie schon im Hinblick auf die Verteidigung erwähnt, können KI-Anwendungen *Cyberintelligence* generieren – solche kann jedoch auch zur Vorbereitung und Umsetzung von Angriffen verwendet werden. Sie könnten die Zielauswahl und priorisierung für *Social-Engineering*-Cyberangriffe verbessern. Bei dieser Art von Angriffen werden Zielpersonen für illegitime Zwecke psychologisch manipuliert, bestimmte Informationen offenzulegen oder eine bestimmte Aktion auszuführen. Sämtliche online verfügbaren Informationen des potenziellen Opfers können dank KI gesammelt, verarbeitet und zur automatischen Erstellung individueller böswilliger Webseiten, E-Mails sowie Links verwendet werden.

KI-generierte *Synthetic Media* können für Erpressung, Betrug, Sabotage und politische Propaganda benutzt werden.

KI-Komponenten werden auch zur verbesserten Erkennung und Ausnutzung von gegnerischen Sicherheitslücken beitragen. Malware wird durch KI mit komplexeren Designs und Funktionen ausgestattet und noch besser getarnt werden können. KI-gestützte Malware kann verhindern, dass sie entdeckt wird, und bei einer Verhaltensänderung des Ziels auf kreative Art reagieren. Als autonomes und anpassungsfähiges Implantat lernt sie von ihrem Wirt und bleibt so unentdeckt. Sie sucht und klassifiziert interessante Inhalte für die Datenexfiltration, sucht und infiziert neue Ziele und findet neue Wege oder Methoden, um sich durch ein Netzwerk zu bewegen und dabei das eigentliche Hauptziel des Angriffs – die Schlüsseldaten – ausfindig zu machen. Bereits 2018 entwickelten IBM-Forschende eine solche Malware, genannt Deeplocker. Schliesslich wird KI auch zur Täuschung von Identifizierungs- und Authentifizierungssystemen, beispielsweise mittels biometrischer Identifizierung, genutzt werden.

Informationen und Beeinflussung

Zu erwarten ist, dass KI künftig die Planung und Durchführung von Informations- und Beeinflussungsaktionen im Cyberraum beeinflusst. Sie wird die digitale Informationserfassung sowie die Überwachung des Onlineverhaltens von Zielobjekten durch Automation vorantreiben. Durch KI entstehen mehr Werkzeuge für die Informationsversorgung und die Beeinflussung des Gegners durch den Cyberraum und innerhalb desselben. KI könnte das Management von Bots und Social Bots in den sozialen Medien verbessern und Nachrichten generieren, die auf die jeweils empfänglichsten Nutzerinnen und Nutzer abzielen. Gemäss einem anhaltenden Trend werden vor allem KI-gestützte Lösungen, die auf *Generative Adversarial*

Weiterführende Literatur

Matteo E. Bonfanti, Artificial Intelligence and the Offence-Defence Balance in Cyber Security, in: Dunn Cavelty, M. & Wenger A. (Hrsg.), *Cyber Security Politics: Socio-Technological Transformations and Political Fragmentation* (Routledge, 2020, i. E.).

Matteo E. Bonfanti, Cyber Intelligence: In Pursuit of a Better Understanding for an Emerging Practice, in: *INSS Cyber, Intelligence, and Security* 2:1 (2018), S. 105–121.

Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation,* Februar 2018.

Ben Buchanan, *A National Security Research Agenda for Cybersecurity and Artificial Intelligence. Center for Security and Emerging Technology,* CSET Issue Brief, Mai 2020.

Networks (GAN) basieren, die Generierung manipulierter digitaler Inhalte unterstützen. Diese Inhalte sind bekannt als *Synthetic Media* oder *Deepfakes*. Dabei handelt es sich um hyperrealistische Video-, Audio-, Bild- oder Textinhalte, die durch manuelle oder andere konventionelle forensische Methoden nur schwer als Fälschung entlarvt werden können. *Synthetic Media* könnte sehr schnell missbraucht werden. Bereits heute gibt es viele in den Medien dokumentierte Fälle von schädigenden *Deepfakes*. Meistens handelt es sich dabei um Videos, die mittels KI für gezieltes Onlinemobbing, Stalking und Verleumdung im Cyberraum manipuliert wurden. In der nahen Zukunft werden *Synthetic Media* wahrscheinlich als Online-Waffe für Erpressung, Betrug und Unternehmenssabotage durch Marktvorgänge oder andere manipulative Handlungen sowie politische Propaganda-Aktionen dienen.

Regierungen können eine wichtige Rolle spielen, die durch KI ausgelöste Transformation der der Cybersicherheit zu steuern.

Obwohl KI für die obengenannten Zwecke missbraucht und solche Aktionen ermöglichen wird, wird sie ebenfalls zu deren Bekämpfung beitragen. Aus Verteidigungssicht kann KI die Entdeckung und die Reaktion auf Beeinflussungen und Informationsaktionen im Cyberraum unterstützen. So kann sie zur Überwachung der Online-Umgebung wie etwa Plattformen sozialer Medien beitragen, frühe Anzeichen böswilliger Handlungen wie eine steigende Bot-Anzahl oder vermehrte Social-Bot-Aktivitäten sowie veränderte digitale Inhalte erkennen, darunter auch *Synthetic Media*.

Eine Gouvernanzfrage

KI wird die Cybersicherheit in den kommenden Jahren beeinflussen und die Landschaft der Cyberbedrohung sowohl quantitativ als auch qualitativ bereichern. Durch KI wird die Anzahl der Akteure im Bereich Cyberbedrohung höchstwahrscheinlich ansteigen. Diese werden zusätzliche Schwachstellen und Ziele ausnutzen und ihre böswilligen Aktionen vorantreiben können. Handkehrum wird KI auch die Abwehr solcher Bedrohungen stärken, indem sie die Entdeckung unbekannter Schwachstellen und böswilliger Cyberaktivitäten sowie die Implementierung von Gegenmassnahmen ermöglicht. KI wird also sowohl die Cyberverteidigung als auch Cyberangriffe unterstützen. Ob defensive oder offensive Anwendungen stärker profitieren werden, ist schwer zu sagen. Wahrscheinlich hängt dies davon ab, inwiefern öffentliche oder private Interessengruppen der Cybersicherheit KI-Anwendungen beherrschen und nutzen können. Die allgemeine Fähigkeit der Interessengruppen, die aus dem Einsatz dieser Technologien entstehenden Risiken, Bedrohungen und Chancen zu identifizieren, zu verstehen und damit umzugehen, ist ebenfalls ein entscheidender Faktor.

Durch das Management und die Steuerung des durch KI ausgelösten Wandels der Cybersicherheit können Regierungen den Umgang mit den Risiken und Chancen stark beeinflussen. Bislang haben sie Innovationen im KI-Bereich durch verschiedene politische Mechanismen unterstützt. So haben sie in KI-Infrastrukturen investiert, die akademische Ausbildung und fachliche Schulung gefördert, Geld für die wissenschaftliche Forschung vergeben, Anreize für öffentlich-private Partnerschaften und Zusammenarbeit geschaffen sowie die Normierung durch beschaffungspolitische Beschlüsse gefördert. Nach Rücksprache mit dem Privatsektor und der Zivilgesellschaft haben sie Leitlinien oder Grundnormen verabschiedet, darunter Grundrechte und Datenschutzgesetze, um verantwortungsvolle und vertrauenswürdige Innovationen auf diesem Technologiegebiet zu fördern.

In vielen Ländern richten die Regierungen ihr Handeln auf den Erwerb von KI-Fähigkeiten nach breit angelegten nationalen KI-Strategien aus, von denen die meisten die Cybersicherheit als ein vielversprechendes Anwendungsgebiet behandeln. Diese Strategien werden anschliessend durch sektorielle Politikinstrumente oder andere technische Unterlagen ergänzt. Im Allgemeinen verfolgen Regierungen das Ziel, KI-Fähigkeiten für die relevanten nationalen Cybersicherheitsakteure zugänglich zu machen und sicherzustellen, dass diese KI als Vorteil gegenüber ihren Konkurrenten nutzen können.

Um den durch KI ausgelösten Wandel der Cybersicherheit zu beeinflussen, können Regierungen auch dynamische Standards für das Testen, Überprüfen und die Zertifizierung von KI-Tools für Cyber-Anwendungen festlegen. Auf internationaler Ebene können sie gemeinsame Normen für die KI-Forschung und Entwicklung anstreben sowie durchdachte Beschränkungen für die Verbreitung von Wissen und Fähigkeiten in diesem technischen Gebiet erwägen. Des Weiteren können sie durch die Operationalisierung wichtiger Grundsätze zu vertrauenswürdiger KI, wie sie etwa die EU und die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (OECD) verabschiedet haben, eine positive und inklusive KI-Gouvernanz fördern.

Für mehr zu Cybersicherheitspolitik, siehe CSS Themenseite.

Matteo E. Bonfanti ist Senior Researcher im Team Risiko und Resilienz am CSS.

Kevin Kohler ist Researcher im Team Risiko und Resilienz am CSS.

* *Lösung des Bildrätsels auf Seite 1: Tatsächlich sind beide Fotos KI-generiert. Diese «Kinder» existieren nicht.*

Die **CSS Analysen zur Sicherheitspolitik** werden herausgegeben vom Center for Security Studies (CSS) der ETH Zürich. Das CSS ist ein Kompetenzzentrum für schweizerische und internationale Sicherheitspolitik. Jeden Monat erscheinen zwei Analysen auf Deutsch, Französisch und Englisch.

Herausgeber: Benno Zogg
Lektorat: Julian Kamasa
Layout: Rosa Guggenheim

Feedback und Kommentare: analysen@sipo.gess.ethz.ch
Weitere Ausgaben und Abonnement: www.css.ethz.ch/cssanalysen

ISSN: 2296-0236; DOI: 10.3929/ethz-b-000417506