Mehr Sicherheit durch fehlende Information in KI-Systemen?

Karl Hans Bläsius, 4.5.2021, https://www.hochschule-trier.de/informatik/blaesius/

Link zu diesem Dokument: www.fwes.info/fwes-KI-fi-21-1.pdf

English Version: www.fwes.info/fwes-KI-fi-21-1-en.pdf

Siehe auch: www.atomkrieg-aus-versehen.de

In militärischen Frühwarn- und Entscheidungssystemen zur Erkennung von Angriffen mit Atomwaffen werden zunehmend Techniken der Künstlichen Intelligenz (KI) eingesetzt, da es für Menschen immer schwieriger wird, solche Alarmmeldungen in der kurzen verfügbaren Zeit zu bewerten. Da die Datengrundlage bei diesen Anwendungen jedoch unsicher und unvollständig ist, können auch KI-Systeme nicht sicher entscheiden. Diese Zusammenhänge sind in https://www.fwes.info/fwes-ki-20-1.pdf beschrieben. Einige Passagen dieses Beitrags sind aus dem genannten Artikel entnommen. Hier geht es nun darum, ob die Sicherheit erhöht werden kann, wenn bestimmte Informationen nur dem Menschen vorbehalten und für die KI-Komponenten nicht verfügbar sind.

Frühwarnsysteme

Frühwarnsysteme basieren auf Sensoren und sehr komplexen Computer-Netzwerken und dienen der Vorhersage von Angriffen mit Atomraketen. Auf Grundlage einer solchen Erkennung könnten die eigenen Raketen zu einem Gegenschlag gestartet werden, bevor man selbst vernichtend getroffen und eine Gegenreaktion erschwert oder verhindert wird. In Frühwarnsystemen kann es zu Fehlalarmen, also falschen Angriffsmeldungen kommen, die ganz unterschiedliche Ursachen haben können. In Phasen politischer Entspannung sind die Risiken gering, dass die Bewertung einer Alarmmeldung zu einem atomaren Angriff führt. Die Situation kann sich drastisch ändern, wenn politische Krisensituationen vorliegen, eventuell mit gegenseitigen Drohungen oder wenn in zeitlichem Zusammenhang mit einem Fehlalarm weitere Ereignisse eintreten, die zur Alarmmeldung in Zusammenhang gesetzt werden könnten.

Automatische Entscheidungen

Die Weiterentwicklung von Waffensystemen mit höherer Treffsicherheit und immer kürzeren Flugzeiten (Hyperschallraketen) wird zunehmend Techniken der Künstlichen Intelligenz erforderlich machen, um für gewisse Teilaufgaben Entscheidungen automatisch zu treffen. Es gibt bereits Forderungen in Zusammenhang mit Frühwarnsystemen autonome KI-Systeme zu entwickeln, die vollautomatisch eine Alarmmeldung bewerten und gegebenenfalls einen Gegenschlag auslösen, da für menschliche Entscheidungen keine Zeit mehr bleibt. Die für eine Entscheidung verfügbaren Daten sind jedoch vage, unsicher und unvollständig. Deshalb

können auch KI-Systeme in solchen Situationen nicht zuverlässig entscheiden. Automatische Erkennungsergebnisse gelten deshalb nur mit einer gewissen Wahrscheinlichkeit und können falsch sein.

Kontextwissen

Aufgrund der unsicheren Datengrundlage stützen Menschen ihre Entscheidungen auch auf Kontextwissen über die politische Lage und die Einschätzung des Gegners. Wenn aufgrund der geringen verfügbaren Zeit zur Bewertung von Alarmmeldungen Entscheidungen weitgehend automatisch fallen sollen, müsste auch hierbei Kontextwissen zur weltpolitischen Lage einbezogen werden. Jedoch ist auch solches Wissen hochgradig vage, unsicher und unvollständig.

Die Bewertung der weltpolitischen Lage ist Gegenstand eines Projektes mit dem Namen „Preview", das die Bundeswehr im März 2018 gestartet hat, mit dem Ziel, auf der Basis von Methoden der Künstlichen Intelligenz Krisen und Kriege vorherzusagen. Dazu sollen große Datenmengen automatisch analysiert werden. Ausgewertet werden hierbei Internet-Quellen sowie militärische und wirtschaftliche Datenbanken und auch Geheimdienstinformationen. Die Art der verwendeten Daten umfasst ein großes Spektrum, wozu auch Handelsdaten, Marktpreise, demographische Entwicklungen, Kriminalitätsraten, Meinungen in sozialen Netzwerken oder Daten über politische Gewalt gehören. Auch in anderen Staaten (z.B. Schweden, USA) gibt es solche KI-basierten Systeme zur Vorhersage von Krisen und Kriegen.

Auch wenn solche Vorhaben wie das Projekt Preview sinnvoll zur frühzeitigen Erkennung von potentiellen Krisen, z.B. in Afrika sein können und es derzeit keine Hinweise auf einen Zusammenhang mit Frühwarnsystemen zur Erkennung von nuklearen Angriffen gibt, kann ein solcher Zusammenhang irgendwann eintreten: Wenn ein Frühwarnsystem einen Raketenangriff meldet und diese Situation über mehrere Alarmstufen hinweg in den entsprechenden Krisensitzungen bewertet wird, ist es durchaus möglich, dass Kommissionsmitglieder auch Zugriff auf ein solches System zur Kriegsvorhersage haben. Wenn dieses KI-System in einer solchen Situation einen Krieg vorhersagt, kann dies erheblichen Einfluss auf die Bewertung der Alarmmeldung durch die Kommissionsmitglieder haben.

Entscheidungsvorschläge

Eine fachliche Beurteilung der von einem KI-basierten System getroffenen Entscheidungen durch Menschen ist in der kurzen verfügbaren Zeit praktisch unmöglich. Dies gilt schon deshalb, weil die automatische Erkennung oft auf Hunderten von Merkmalen basiert. Die KI-Systeme können in der Regel keine einfachen nachvollziehbaren Begründungen liefern und selbst wenn Erkennungsmerkmale von einem KI-System ausgegeben werden, könnten diese nicht in der verfügbaren Zeit überprüft werden. Dem Menschen bleibt deshalb meist nur zu glauben, was ein KI-System liefert. Die zunehmende Verbreitung von KI-Systemen in unserer Alltagswelt fördert zudem das Vertrauen in die Entscheidungskompetenz von technischen

Systemen und es ist zu erwarten, dass automatisch bestimmte Entscheidungsvorlagen oder Lagebeurteilungen von Menschen als nur schwer zu ignorierende Faktoren bewertet werden.

Zitat aus einem Roman

In dem Roman Qualityland 2.0 von Marc Uwe Kling gibt es auf Seite 206 in Zusammenhang mit dem Auslösen des 3. Weltkrieges durch ein KI-System folgenden Dialog:

> Henryk überlegt. „Glaubst Du der Dritte Weltkrieg hätte verhindert werden können, wenn ein Mensch in die auslösende Entscheidungskette involviert gewesen wäre?"
>
> „Kommt auf den Menschen an", sagt Peter, „Außerdem … wenn man Entscheidungen trifft mithilfe einer K.I., auf Basis von Daten, die einem die K.I. anzeigt, wenn man also nur die Welt sieht, die die K.I. für einen aufbereitet hat, entscheidet man sich dann nicht fast zwangsläufig für das, was die K.I. vorschlägt? Verstehen Sie was ich meine?"
>
> „Du meinst, man braucht zusätzlichen Input. Input, den die K.I. nicht hat."
>
> „Ja."

Lösungsansatz: fehlende Information für KI-System

Vielleicht ist dieser Aspekt ja auch in Zusammenhang mit immer mehr KI in Frühwarnsystemen relevant. Ein Ansatz könnte z.B. sein, dass es feste Vereinbarungen gibt, dass Systeme wie Preview zur Vorhersage von Kriegen und Krisen definitiv nicht in Frühwarnsystemen zur Erkennung von Angriffen mit Atomwaffen angewendet werden dürfen. Dann ist klar, dass die Bewertung des Kontextes bezogen auf die aktuelle weltpolitische Lage nur von Menschen durchgeführt werden kann. Da das KI-System dieses Wissen nicht hat.

Wenn es den Menschen in Frühwarnsystemen bewusst ist, dass nur sie über bestimmtes Wissen verfügen, nicht aber das KI-System, dann ist es auch leichter, sich gegen eine Entscheidung der Maschine zu stellen.

Dies ist kein technischer, sondern ein psychologischer Aspekt. Das Selbstbewusstsein des Menschen gegenüber Entscheidungen einer Maschine kann so gestärkt werden. Der Mensch weiß, er kann sich gegen die Entscheidung der Maschine stellen, er kann dafür nicht verantwortlich gemacht werden, da er über zusätzliches Wissen verfügt, das die Maschine nicht hat.

Es sollte nicht sehr schwer sein, alle Atommächte von einer solchen Vereinbarung zu überzeugen, da dies allen Beteiligten als sinnvoll erscheinen müsste. Auch wenn das Einhalten einer solchen Vereinbarung kaum überprüft werden kann, spricht viel für das Einhalten, da man sich damit auch selbst schützt.

Natürlich kann eine solche Vereinbarung nicht dringend notwendige konkrete Abrüstungsvereinbarungen bzgl. Atomwaffen ersetzen, aber es könnte ein kleiner Schritt dazu sein, Vertrauen zwischen Atommächten aufzubauen und das Bewusstsein für mögliche Gefahren zu stärken.