Trier, Saarbrücken, 18.5.2024

Generative KI – Mögliche Folgen im Internet

Karl Hans Bläsius, Jörg Siekmann

https://www.hochschule-trier.de/informatik/blaesius/ , http://siekmann.dfki.de/de/home/

Link zu diesem Dokument: https://fwes.info/GenKI-Internet-2024-1.pdf

Siehe auch: https://ki-folgen.de

1. Erfolge von ChatGPT, Weiterentwicklungen

ChatGPT wurde im November 2022 veröffentlicht. Dieses System zeigt enorme Fähigkeiten in der sprachlichen Kommunikation. Umfangreiche Texte können auf Basis einfacher Anforderungen erstellt werden. Auch kurze Zusammenfassungen von umfangreichen Texten sind möglich. Diese Fähigkeit erstreckt sich auf viele fachliche Gebiete. Auch Bilder und Videos können mit Systemen der generativen KI erzeugt werden. Die meisten Antworten auf Fragestellungen sind sehr überzeugend, auch wenn nicht alles richtig ist. Diese Systeme können auch Programmieraufgaben in verschiedenen Programmiersprachen lösen.

Neue Systeme mit Verbesserungen sind angekündigt. Mögliche Verbesserungen könnten betreffen:

- Reduzierung von Fehlern, sodass möglichst keine falschen Ergebnisse mehr vorkommen,
- Kombination mit symbolischer KI,
- Fähigkeiten zum Schlussfolgern, d.h. aus normalen bisherigen Ergebnissen werden weitere Schlüsse gezogen und damit die Ergebnisse verbessert,
- Wissenserwerb mit dem Aufbau Wissensnetzwerken, z.B. als semantische Netze,
- usw.

2. Anwendungsdruck

Der Erfolg von ChatGPT erzeugt enormen Druck. Alle wichtigen Unternehmen und Staaten müssen bei der Entwicklung vorne dabei sein. Dies gilt auch für Anwender und auch für das Militär.

Die Entwicklung von Systemen der generativen KI läuft derzeit in vielen Unternehmen und Staaten, wobei mit großem finanziellen Aufwand Techniken entwickelt werden, die zu einer Superintelligenz führen könnten. Auch wenn die möglichen Risiken bewusst sind, möchte

niemand ins Hintertreffen geraten. Niemand möchte dieses Wettrennen um eine Superintelligenz verlieren.

3. Eskalierendes Verhalten

In einem Beitrag vom 11.2.2024 bei Telepolis wird davor gewarnt, dass KI-Systeme in einem Krieg frühzeitig Atomwaffen einsetzen würden.[1] Dabei wird auf eine Studie verwiesen, die auf Basis von Experimenten mit verschiedenen Systemen der generativen KI zu dem Ergebnis kommt, dass diese ein eskalierendes Verhalten bevorzugen könnten.[2] Auch ein Artikel bei Foreign Affairs verweist auf Erkenntnisse zu eskalierendem Verhalten.[3]

Damit stellt sich die Frage, was sind die Ursachen für ein eskalierendes Verhalten und wie könnte dies verhindert werden.

Ein Grundproblem in der KI sind häufig riesige Suchräume. Das Lösen von Problemen kann durch eine Folge von Operationsanwendungen erfolgen, wobei in jedem Schritt die aktuelle Situation mit der Ausführung einer Aktion geändert wird, bis eine akzeptierte Zielsituation erreicht wird. In der Regel wird für ein solches System gelten, dass in jeder Situation sehr viele mögliche Alternativen für eine nächste Aktion existieren, und es kommt darauf an, unter diesen vielen Alternativen eine möglichst gute oder sogar optimale Auswahl zu treffen. Die einzelnen Alternativen sind in der Regel gewichtet und auf dieser Grundlage kann nach bestimmten Strategien und Heuristiken eine Auswahl erfolgen.

Bei schwierigen Problemen können hierbei riesige Suchräume entstehen, wobei es nicht möglich ist, alle Alternativen zu untersuchen. Stattdessen sind geeignete Gewichtungen und Auswahlentscheidungen für die möglichen Operationen erforderlich, sodass auch bei einer eingeschränkten Betrachtung der möglichen Alternativen eine Lösung gefunden werden kann. In der symbolischen KI sind solche Operationen und Auswahlverfahren sichtbar und können untersucht und bewertet werden.

Bei der Verwendung von neuronalen Netzen und LLMs besteht das gleiche Problem, nämlich dass sehr viele Alternativen existieren und berücksichtigt werden müssen, wobei es hier aber auf eine umfangreiche Menge von Beispielen als Lerngrundlage ankommt. Auch hier erfolgt eine Problemlösung in einem riesigen Alternativen-Raum. Allerdings ist das Zustandekommen einer Lösung bei diesen Verfahren hinterher kaum nachvollziehbar.

Das Problem zur Bewältigung vieler Alternativen in riesigen Suchräumen kann interne Zwischenschritte betreffen aber auch die Lösungsvielfalt selbst, was für Systeme der generativen KI gilt. Bei ChatGPT gibt es meist sehr viele Möglichkeiten wie eine Frage beantwortet werden kann und eine solche Antwort basiert auf einer riesigen Menge an unterschiedlich gewichteten Informationen, die in der Lerngrundlage stecken.

---

1 https://www.telepolis.de/features/KI-wuerde-im-Krieg-rasch-Atomwaffen-einsetzen-9624831.html
2 https://arxiv.org/abs/2401.03408 und https://arxiv.org/pdf/2401.03408.pdf
3 https://www.foreignaffairs.com/united-states/why-military-cant-trust-ai

Wichtig für diese Systeme ist, dass Probleme gelöst, beziehungsweise Anfragen zur Zufriedenheit des Anwenders beantwortet werden. Eine Grundlage hierfür können Lernverfahren sein, sodass das Systemverhalten immer besser wird.

Mit dem Ziel das Systemverhalten zu verbessern kann versucht werden, den Erfolg einer Aktion bzw. die Reaktion eines Nutzers zu bewerten, und auf dieser Grundlage die Gewichte von möglichen Operationen oder Lerndaten zu verändern.

Eine Bewertung des Erfolges kann zum Beispiel bei generativer KI auf Basis der Resonanz zu den gelieferten Antworten erfolgen. Dies könnte die weitere Kommunikation mit der Person oder dem technischen System, von dem die Anfrage kam, betreffen oder ein stärkerer „Traffic" im Internet, der dieser Aktion zugeordnet werden kann.

Aus den sozialen Medien ist bekannt, dass Hass und Hetze schneller verbreitet werden und damit zu mehr Traffic führen. Dieser Effekt bindet die Aufmerksamkeit der Nutzer stärker, was dem Geschäftsmodell dieser Unternehmen entspricht. Dies führt auch zu höheren Gewichtungen entsprechender Alternativen bei der Auswahl der nächsten Systemvorschläge und begünstigt somit ein eskalierendes Verhalten von Nutzern und Bots, die in diese Prozesse einbezogen sind.

Ähnliches könnte bei Systemen wie ChatGPT erfolgen. Antworten und Aktionen, die zu einem eskalierenden Verhalten führen können, werden vermutlich mehr Resonanz bewirken als andere Aktionen. Wenn aber Aktionen, die mehr Resonanz bewirken, als Erfolg und damit immer höher bewertet werden, könnte dies zu Kettenreaktionen und so zu einem eskalierenden Verhalten führen.

Ähnlich wie bei den sozialen Medien werden vermutlich auch die Unternehmen, die Systeme der generativen KI anbieten, Prioritäten in ihren Systemen so setzen, dass möglichst viele Nutzer möglichst lange aktiv sind. Dies wird am besten damit erreicht, dass mehr Resonanz auf Antworten höher bewertet wird, was ein eskalierendes Verhalten begünstigt. Deshalb ist es vermutlich nicht im Interesse der anbietenden KI-Unternehmen dies zu ändern.

Wenn diese Risiken hinreichend bewusst sind, könnte dies bei militärischen Aufgaben berücksichtigt werden, sodass solche Systeme z.B. nicht in Zusammenhang mit Atomwaffen eingesetzt werden. Allerdings kann es auch gravierende Risiken im Internet geben, wobei es eine Vielzahl möglicher Akteure geben kann, sodass Risiken nur schwer eingedämmt werden können. Nicht alle Anbieter von Systemen der Generativen KI werden geeignete Vorsichtsmaßnahmen berücksichtigen, sondern stattdessen Prioritäten in die Gewinnmaximierung setzen. Risiken, die damit verbunden sein können, werden in den nächsten Abschnitten behandelt.

4. KI-basierte Cyberwaffen

Kapazitäten für Cyberangriffe werden von vielen Staaten, Unternehmen und Hackergruppen entwickelt. Hierbei könnten auch zunehmend Techniken der KI eingesetzt werden. Mit Hilfe

von KI könnten Angriffspunkte gefunden und Angriffsabläufe geplant und durchgeführt werden.[4]

Angriffscode kann mit Hilfe von KI schneller erzeugt werden. Dieses Potenzial an Beschleunigung bringt Angreifern signifikante Vorteile gegenüber der Verteidigungsseite. Ein weiterer Vorteil für die Offensive ist, dass Angriffscode nicht immer funktionieren muss. Angreifer sind auch dann erfolgreich, wenn von 100 Angriffsversuchen nur einer gelingt. Fehlgeschlagene Versuche sind dabei irrelevant. Wenn mit Hilfe von KI in kurzer Zeit viele Alternativen für Angriffscode erzeugt werden können, reicht dies um großen Schaden anzurichten. Auf der Seite der Verteidiger reicht es nicht, wenn Abwehrmaßnahmen nur bei wenigen Angriffsalternativen erfolgreich sind.

Es stellt sich auch die Frage welche Wechselwirkungen es mit Systemen wie ChatGPT geben kann. Können diese Hinweise für Cyberwaffen liefern oder eine Anwendung von Cyberwaffen bewirken oder verursachen? Systeme wie ChatGPT sind im Internet in gewissem Sinne „autonom" aktiv. Sie beantworten Fragen und lösen Probleme, ohne dass irgendwo Menschen eingreifen. Sie können auch programmieren und für Cyberangriffe eingesetzt werden. Anfragen und Aufgabenstellungen an ChatGPT müssen nicht von Menschen, sondern können auch von anderen technischen Systemen, von Bots kommen.

Diese Systeme können permanent im Internet aktiv sein und mögliche Angriffsziele suchen. Einmal gestartet, könnten solche Systeme, ohne weiteres menschliches Eingreifen, Angriffe planen und ausführen, können also als autonome Cyberwaffen betrachtet werden.

5. Kettenreaktion mit Angriffen, Cyberwar der Systeme

Von den Finanzmärkten ist bekannt, dass es im Hochfrequenzhandel zwischen verschiedenen Algorithmen zu unvorhergesehenen Interaktionsprozessen kommen kann, die innerhalb von Sekunden zu Kursabstürzen und finanziellen Verlusten führen können (bezeichnet als „flash crash"). Bei vollautonomen Waffensystemen können auch solche unvorhersehbaren Interaktionen zwischen den automatischen Systemen vorkommen und eine Kettenreaktion von autonom geführten Angriffen und Gegenangriffen auslösen. In sehr kurzen Zeitabschnitten kann hierbei eine Eskalationsspirale entstehen, die von Menschen in der Kürze der Zeit nicht beherrscht werden kann. In Zusammenhang mit autonomen Waffensystemen wird hierfür der Begriff „flash war" verwendet.[5] Auch bei „autonomen Internetagenten" wäre eine solche Kettenreaktion denkbar. Ein solcher „Flash war" könnte also auch im Internet erfolgen.

Solche Eskalationsspiralen könnten von einem einzigen System ausgehen, z.B. durch Gewichtsverstärkung aufgrund der bisherigen Aktivitäten oder von mehreren konkurrierenden Systemen.

---

[4] https://www.heise.de/news/Kaspersky-Cyberkriminelle-experimentieren-mit-KI-9612407.html?wt_mc=nl.red.ho.ho-nl-newsticker.2024-01-30.link

[5] Reinhard Grünwald, Christoph Kehl: Autonome Waffensysteme – Endbericht zum TA-Projekt, Büro für Technikfolgen-Abschätzung beim Deutschen Bundestag, Arbeitsbericht Nr. 187, Okt. 2020, https://dip21.bundestag.de/dip21/btd/19/236/1923672.pdf , Seite 130 - 131

Wenn ein KI-System einen Cyberangriff erfolgreich durchführt, dann wird dies vermutlich zu vielen Reaktionen, auch im Internet führen, die auch von diesem KI-System registriert werden. Dies könnte von diesem KI-System als Anzeichen gewertet werden, dass diese Aktion als erfolgreich zu bewerten ist, womit entsprechende Gewichte verändern werden. Eine bereits als erfolgreich bewertete Aktion kann dazu führen, dass ähnliche Alternativen höheres Gewicht erhalten und damit in den nächsten Phasen eher ausgewählt werden. Dies führt zu weiteren Gewichtserhöhungen für ähnliche Aktionen. Die Folge kann eine Kettenreaktion von immer mehr und schwerwiegenderen Angriffen sein.

Wenn eine normales Autonomes Waffensystem ihre Munition abschießt, ist es anschließend unbrauchbar oder muss zu ihrer Basis zurückkehren. Bei Internetwaffen gibt es diesen Aspekt „Pulver verschossen" nicht, beliebig viele weitere Angriffe sind möglich.

Im Internet sind bereits einige Systeme der generativen KI vergleichbar mit ChatGPT im Einsatz und es werden noch weitere dazu kommen. Derzeit wird in vielen Unternehmen und vermutlich auch in einigen Staaten an Systemen der generativen KI gearbeitet. Neben Menschen könnten auch Bots Fragen und Aufgaben an diese Systeme stellen. Es ist zu erwarten, dass es auch schon bald Interaktionen zwischen diesen Systemen selbst gibt.

Daraus können neue Gefahren entstehen, insbesondere wenn diese Systeme Cyberangriffsfähigkeiten haben. Durch Menschen, Bots oder ein anderes System der generativen KI beauftragt, könnte ein System wie ChatGPT Cyberangriffe ausführen. Andere Systeme der generativen KI, mit denen es ohnehin Interaktionen gibt, könnten dies erfassen und Gegenangriffe starten. Ohne dass Menschen beteiligt sind könnte so in kurzer Zeit eine Kettenreaktionen zwischen diesen Systemen mit immer stärkeren Cyberangriffen entstehen, also ein „flash war" im Internet. Diese Systeme wären dann defacto autonome Cyberwaffen.

Auch wenn die derzeitigen Systeme technisch dazu noch nicht in der Lage sind, ist zu erwarten, dass in regelmäßigen Abständen Erweiterungen aktiviert werden, die vielleicht irgendwann in den nächsten Jahren oder bereits sehr bald solche Fähigkeiten haben.

Ein Schutz vor diesen Risiken ist schwierig, denn bei einer Vielzahl solcher Systeme kann nicht erwartet werden, dass in allen Fällen eigentlich erforderliche Sicherheitsmaßnahmen von allen eingehalten werden. Des Weiteren könnte es auch gefährliche Interaktionen zwischen Systemen der generativen KI geben, die Staaten zugeordnet werden, die derzeit auf Konfrontationskurs sind. Die Folge könnten gegenseitige Schuldzuweisungen und internationale Konflikte sein. Auslöser eines solchen „flash war" im Internet könnten auch einzelne Menschen oder kleine Gruppen sein. Oder es könnte zufällig durch eine ungünstige Aktion entstehen.

6. Risiken auch ohne AGI

Die im vorherigen Abschnitt beschrieben Risiken setzen nicht voraus, dass die KI-Systeme sich verselbständigen, in dem sie einen eigenen Willen haben, eigene Ziele verfolgen oder ein Bewusstsein haben. Dies alles ist hierfür nicht erforderlich. Die beschriebenen Risiken resultieren alleine daraus, dass als Grundlage für gutes automatisches Problemlösen

geeignete Strategien und Heuristiken erforderlich sind, die in einem riesigen Suchraum Lösungen ermöglichen. Dazu ist es sinnvoll, bisherige Aktionen zu bewertet und daraus Anpassungen für Gewichte möglicher Operationen zu bestimmen. Allein dies kann ein eskalierendes Verhalten begünstigen und zu Kettenreaktionen führen, die in kurzer Zeit gravierende Auswirkungen haben. Diese Risiken können bestehen lange bevor eine AGI (Artificial General Intelligence) oder Superintelligenz entsteht. Diese sind für solch eskalierendes Verhalten nicht erforderlich.

7. Folgen

Systeme der generativen KI sind in der Regel im Internet aktiv und können erhebliche Auswirkungen auf die IT-Sicherheit haben. Dies gilt insbesondere auch deshalb, da die Abhängigkeit von Internetdiensten in den letzten Jahren erheblich gestiegen ist und weiter steigen wird.

In Zusammenhang mit Systemen der generativen KI oder Cyberwaffen kann es zu gravierenden Auswirkungen in der Internetkommunikation kommen. Solche Systeme könnten unter anderem bisher unbekannte Cyberangriffs- oder Internetmanipulationsfähigkeiten erreichen. Solche Systeme könnten dann von Menschen oder Staaten missbraucht werden, oder sogar selbst aktiv werden und den Informationsfluss im Internet beherrschen und damit menschlichen Informationsfluss lahmlegen. Mit Hilfe von KI-Systemen oder durch diese könnte also eine Informationsdominanz erreicht werden, die alle Bereiche betreffen würde, auch das Finanzwesen. Als Folge könnten Finanzwesen und Handel zumindest zeitweise zusammenbrechen und unsere Gesellschaftssysteme instabil werden.

Von diesen Folgen könnten große Regionen und mehrere oder viele Staaten gleichzeitig betroffen sein. Fehlende Grundbedürfnisse könnten zu Unruhen und Aufständen führen.

Wenn hierbei Atomwaffenstaaten in existenzielle Notsituationen geraten, würde das Atomkriegsrisiko erheblich steigen. Da die Abhängigkeit von technischen Systemen inzwischen sehr groß ist, was auch für die Kommunikation im Internet gilt, wären weltweite Krisen die Folge und in solch kritischen Situationen können Fehlinterpretationen, Missverständnisse, falsche Annahmen und Schuldzuweisungen oder auch Fehler in Frühwarnsystemen für nukleare Bedrohungen leicht zu einem Atomkrieg, eventuell aus Versehen führen.

8. Maßnahmen

Falls irgendwann Szenarien eintreten, wie in den letzten Abschnitten beschrieben, dann sollte es ein möglichst gutes Verhältnis zwischen allen Nationen geben. Insbesondere sind sehr gute Kommunikationsmöglichkeiten und ein gewisses Maß an Vertrauen auch zwischen politischen Gegnern erforderlich, damit nicht Fehlinterpretationen zum Einsatz gefährlicher Waffen, wie z.B. Atomwaffen führen.

Deshalb wäre es besonders wichtig Kriege und den aktuellen Konfrontationskurs zwischen großen Nationen zu beenden. Vertrauen und gute Kommunikationskanäle müssen wieder aufgebaut werden, und das zwischen allen Nationen, auch solchen, die sich derzeit als Gegner betrachten.

Des Weiteren sollten die Abhängigkeiten vom Internet nicht weiter erhöht werden. Stattdessen sollten für alle wichtigen Bereiche, insbesondere die kritische Infrastruktur betreffend, alternative Strukturen aufgebaut werden, die unabhängig vom Internet funktionstüchtig sind.