Kann KI übermenschlich sein? Mängel im Top-Spielerbot werfen Zweifel auf

Forschung zeigt, dass selbst Superintelligenzen im Go-Spiel anfällig sind. Erfahren Sie, wie KI-Systeme wie KataGo gegen Angriffe kämpfen und welche Auswirkungen dies auf die Entwicklung von KI-Systemen haben könnte.

Gespräche über **übermenschliche künstliche Intelligenz (KI)** nehmen zu. Doch Forschungsergebnisse haben Schwächen in einem der erfolgreichsten KI-Systeme aufgedeckt – ein Bot, der das Brettspiel Go spielt und die besten menschlichen Spieler der Welt schlagen kann – was zeigt, dass eine solche Überlegenheit zerbrechlich sein kann. Die Studie wirft Fragen auf, ob allgemeinere KI-Systeme anfällig für Schwachstellen sein könnten, die ihre Sicherheit und Zuverlässigkeit gefährden könnten, und sogar ihren Anspruch, ‚übermenschlich' zu sein.

„Das Papier hinterlässt ein großes Fragezeichen darüber, wie das ehrgeizige Ziel erreicht werden kann, robuste KI-Agenten der realen Welt aufzubauen, denen die Menschen vertrauen

können“, sagt Huan Zhang, Informatiker an der University of Illinois Urbana-Champaign. Stephen Casper, ein Informatiker am Massachusetts Institute of Technology in Cambridge, fügt hinzu: „Es liefert einige der stärksten Beweise bisher dafür, dass es schwierig ist, fortgeschrittene Modelle so zuverlässig zu implementieren, wie man es sich wünscht.“

Die Analyse, die im Juni **online als Preprint veröffentlicht**[1] wurde und noch nicht peer-reviewed wurde, nutzt sogenannte adversarielle Angriffe – die AI-Systemen Eingaben **zuführen, die darauf abzielen, die Systeme zu Fehlern zu verleiten**, sei es für Forschungszwecke oder für böswillige Zwecke. Bestimmte Eingaben können zum Beispiel Chatbots ‚jailbreaken‘, indem sie schädliche Informationen ausgeben, die sie normalerweise unterdrücken sollten.

Beim Go platzieren zwei Spieler abwechselnd schwarze und weiße Steine auf einem Gitter, um die Steine des anderen Spielers zu umgeben und einzufangen. Im Jahr 2022 berichteten Forscher über **die Ausbildung von adversariellen AI-Bots zum Besiegen von KataGo**[2], dem besten Open-Source-Go-spielenden KI-System, das normalerweise die besten Menschen problemlos schlägt (und handslos). Ihre Bots fanden Schwachstellen, die regelmäßig KataGo besiegten, obwohl die Bots ansonsten nicht sehr gut waren – menschliche Amateure konnten sie besiegen. Außerdem konnten Menschen die Tricks der Bots verstehen und anwenden, um KataGo zu besiegen.

Ausnutzung von KataGo

War das eine einmalige Sache, oder wies diese Arbeit auf eine fundamentale Schwäche in KataGo hin – und, in Erweiterung, auf andere KI-Systeme mit scheinbar übermenschlichen Fähigkeiten? Um dies zu untersuchen, nutzen die Forscher unter der Leitung von Adam Gleave, Geschäftsführer von FAR AI, einer gemeinnützigen Forschungsorganisation in Berkeley, Kalifornien und Co-Autor des Papers von 2022[2], adversarielle Bots, um drei Möglichkeiten zu testen, Go-KIs gegen solche Angriffe zu

verteidigen[1].

Die erste Verteidigung war eine, die die KataGo-Entwickler bereits nach den Angriffen von 2022 eingesetzt hatten: KataGo Beispiele für Spielsituationen, die bei den Angriffen beteiligt waren, zu geben und es spielen zu lassen, um zu lernen, wie man gegen diese Situationen spielt. Das ähnelt dem, wie es sich generell das Go-Spielen beigebracht hat. Die Autoren des neuesten Papers fanden jedoch heraus, dass ein adversarieller Bot lernte, selbst diese aktualisierte Version von KataGo zu schlagen, und 91 % der Zeit gewann.

Die zweite Verteidigungsstrategie, die das Team von Gleave ausprobierte, war iterativ: eine Version von KataGo gegen adversarielle Bots zu trainieren, dann Angreifer gegen das aktualisierte KataGo zu trainieren und so weiter, neun Runden lang. Aber auch das führte nicht zu einer unbesiegbaren Version von KataGo. Die Angreifer fanden weiterhin Schwachstellen, wobei der letzte Angriff KataGo 81 % der Zeit besiegte.

Als dritte Verteidigungsstrategie trainierten die Forscher ein neues Go-spielendes KI-System von Grund auf. KataGo basiert auf einem Berechnungsmodell, das als convolutional neural network (CNN) bekannt ist. Die Forscher vermuteten, dass CNNs sich zu sehr auf lokale Details konzentrieren könnten und globale Muster übersehen. Deshalb bauten sie einen Go-Spieler mit einem alternativen **neuronalen Netzwerk** namens vision transformer (ViT). Aber ihr adversarieller Bot fand einen neuen Angriff, der ihm half, 78 % der Zeit gegen das ViT-System zu gewinnen.

Schwache Gegner

In all diesen Fällen waren die adversariellen Bots – obwohl sie in der Lage waren, KataGo und andere führende Go-spielende Systeme zu schlagen – darauf trainiert, versteckte Schwachstellen in anderen KIs zu entdecken, und nicht, vielseitige Strategen zu sein. „Die Gegner sind immer noch

ziemlich schwach – wir haben sie recht leicht besiegt“, sagt Gleave.

Und da Menschen in der Lage sind, die Taktiken der adversariellen Bots zu nutzen, um führende Go-KIs zu besiegen, macht es dann noch Sinn, diese Systeme übermenschlich zu nennen? „Das ist eine großartige Frage, mit der ich definitiv gerungen habe“, sagt Gleave. „Wir haben begonnen zu sagen, ‚typischerweise übermenschlich‘.“ David Wu, ein Informatiker in New York, der KataGo zuerst entwickelte, sagt, dass starke Go-KIs „im Durchschnitt übermenschlich“ sind, aber nicht „in den schlechtesten Fällen“.

Gleave sagt, dass die Ergebnisse weitreichende Auswirkungen auf KI-Systeme haben könnten, einschließlich der **großen Sprachmodelle, die Chatbots wie ChatGPT zugrunde liegen**. „Die wichtigste Erkenntnis für KI ist, dass diese Schwachstellen schwer zu beseitigen sein werden“, sagt Gleave. „Wenn wir das Problem in einem einfachen Bereich wie Go nicht lösen können, dann scheint es in naher Zukunft wenig Aussicht darauf zu geben, ähnliche Probleme wie Jailbreaks in ChatGPT zu beheben.“

Was die Ergebnisse für die Möglichkeit bedeuten, eine KI zu schaffen, die menschliche Fähigkeiten umfassend übertrifft, ist weniger klar, sagt Zhang. „Obwohl dies oberflächlich betrachtet darauf hindeutet, dass Menschen möglicherweise noch einige Zeit wichtige kognitive Vorteile gegenüber KI behalten“, sagt er, „glaube ich, dass die entscheidende Erkenntnis darin besteht, dass **wir die KI-Systeme, die wir heute bauen, noch nicht vollständig verstehen**.“

1. Tseng, T., McLean, E., Pelrine, K., Wang, T. T. & Gleave, A. Preprint at arXiv

https://doi.org/10.48550/arXiv.2406.12843 (2024).

2. Wang, T. T. *et al.* Preprint at arXiv **https://doi.org/10.48550/arXiv.2211.00241** (2022).