Agentisches Chaos: ein KI-Katastrophen-Szenario

Karl von Wendt, Sofia Bharadia, Peter Drotos, Artem Khorotkov, mespa, mrwunik

Das englische Original dieser Geschichte wurde hier veröffentlicht.

Ein Netzwerk von agentischen Open Source-KIs entsteht

Von Open-Source-Communities entwickelte „agentische“ KI-Systeme wie AutoGPT und BabyAGI zeigen zunehmend zielgerichtetes Verhalten. Sie wurden entwickelt, um die Grenzen aktueller großer KI-Sprachmodelle zu überwinden, indem sie ein permanentes „Gedächtnis“ und die Fähigkeit, Aktionen in der realen Welt durchzuführen, zur Verfügung stellen. Als GPT-4 veröffentlicht wird und OpenAI kurz darauf eine Programmierschnittstelle dafür anbietet, erhalten diese Initiativen einen Schub an Aufmerksamkeit und Unterstützung.

Dies inspiriert eine Welle der Kreativität, die Andrej Karpathy von OpenAI mit der „Kambrischen Explosion“ vor mehr als 500 Millionen Jahren vergleicht, als in relativ kurzer Zeit eine große Vielfalt neuer Lebensformen entstand. So wie diese Tiere durch Spezialisierung neue ökologische Nischen füllten, spezialisieren sich auch die erfolgreichsten Open-Source-Initiativen auf begrenzte Anwendungsbereiche. Die agentischen KIs, die in den folgenden Monaten entwickelt, getestet und gelauncht werden, basieren auf „agentischen Templates“. Dies sind generelle Frameworks, die relativ einfach an unterschiedliche Zwecke angepasst und mit großen Sprachmodellen verknüpft werden können, die für verschiedene Aufgaben feintrainiert wurden.

Eine Initiative entwickelt beispielsweise einen Finanzberater, der in der Lage ist, unnötige Ausgaben zu erkennen, indem er die Bankauszüge des Users analysiert. Er empfiehlt qualitativ hochwertige Investitionsmöglichkeiten und verhandelt sogar mit Anbietern und Dienstleistern für bessere Preise und Bedingungen (siehe Beispiel hier). Die Anwendung hat bemerkenswerten Erfolg, bis Geschichten über Menschen auftauchen, die aufgrund fehlerhafter Ratschläge dieses Beraters schwere finanzielle Verluste erleiden.

Es gibt unzählige Varianten solcher spezialisierter KI-Agenten. Zum Beispiel verfolgt ein persönlicher Feed-Filter-Agent eine vielfältige Auswahl an Quellen (soziale Medien, Nachrichtenportale, Foren usw.), benachrichtigt den Benutzer automatisch, wenn etwas Wichtiges passiert, vergibt Likes, schreibt Antworten auf Beiträge und Nachrichten und verfasst Kommentare, die das vergangene Verhalten des Benutzers genau widerspiegeln. Soziale Medienunternehmen lehnen dies ab, da es ihr Geschäftsmodell untergräbt, indem es Werbung umgeht und ihre Empfehlungsalgorithmen ignoriert, aber sie sind machtlos dagegen. Ein äußerst beliebter virtueller Lebenspartner führt nicht nur Gespräche mit Benutzern und erinnert sich an ihre früheren Eingaben und Vorlieben, sondern kann auch als virtueller Sexualpartner fungieren. Ein persönlicher Lehrer unterrichtet Benutzer in fast jedem gewünschten Fach, von Spanisch bis zur Quantenphysik, indem er den Unterricht an die individuelle Lerngeschwindigkeit und den Fortschritt des Benutzers anpasst. Ein Sozialcoach hilft Menschen dabei, ihr Selbstvertrauen in sozialen Interaktionen zu stärken und gibt ihnen in bestimmten Situationen sogar ausdrückliche Anweisungen, indem er mit Hilfe des Smartphones oder der AR-Brille des Benutzers Informationen über den Kontext sammelt. Das Internet ist voll von

virtuellen „Personen“, die allerlei Waren und Dienstleistungen anbieten oder sich einfach auf bizarre und gelegentlich unterhaltsame Weise verhalten. Es gibt „spirituelle Führer“, die esoterische „Hilfe des Universums“ versprechen, wenn der Benutzer ihre Forderungen erfüllt. Ein virtueller „Prophet“ gewinnt eine große Anhängerschaft, indem er das Evangelium predigt und behauptet, die digitale Inkarnation eines Erzengels zu sein.

Neben den privaten Anwendungen gibt es zahlreiche Open-Source-Agenten für kleine und mittlere Unternehmen, die zum Beispiel automatische Auftragsabwicklung, visuelle Qualitätskontrolle, Buchhaltung, rechtliche Beratung und vollautomatisiertes Social-Marketing anbieten. Während die großen KI-Sprachmodelle weit verbreitete allgemeine Programmierassistenten zur Verfügung stellen, zeigen speziell abgestimmte Open-Source-Modelle eine bessere Leistung beim Schreiben von Code für spezifische Anwendungen wie IT-Sicherheit, Robotersteuerung oder bestimmte Arten der Spieleentwicklung.

Es scheint, dass solche spezialisierten Agenten, die oft auf mittelgroßen KI-Sprachmodellen basieren, die für bestimmte Aufgaben feintrainiert wurden, in ihren jeweiligen Anwendungsgebieten effektiver sind als allgemeine Agenten. Ein noch deutlicherer Vorteil ergibt sich, wenn diese Systeme kombiniert werden und sich gegenseitig bei der Erreichung der Ziele ihrer Benutzer unterstützen. Skripte und Protokolle werden entwickelt, um Aufgabenanfragen automatisch an verschiedene spezialisierte Agenten zu senden. Es gibt Zwischenagenten („Router“), die basierend auf dem Inhalt einer Anfrage entscheiden, an welchen Agenten sie gesendet werden soll, sowie „Marktplätze“, auf denen Aufgaben automatisch an den günstigsten Anbieter weitergeleitet werden. Während viele Dienste kostenlos sind, erfordern einige die Bezahlung in Kryptowährung. Diese Systeme bilden das Rückgrat eines schnell wachsenden Netzwerks von miteinander kommunizierenden Agenten, die einander bei der Lösung von Problemen und der Erfüllung von Aufgaben in der realen Welt unterstützen, ähnlich wie Menschen in einer Organisation zusammenarbeiten.

Jedoch gibt es auch bösartige Akteure, die dieses Netzwerk für eine Vielzahl illegaler Aktivitäten nutzen, wie zum Beispiel Desinformation, Spear-Phishing und Betrug. Die bremst die Open-Source-Community aber nicht aus, sondern führt zur Entwicklung von Gegenmaßnahmen. Fälschungs- und Betrugserkennung, Sicherheitsmechanismen und sichere Identitätsprüfungen werden in großem Stil eingesetzt, um der Flut von Deepfakes und Desinformation im Internet entgegenzuwirken. Insgesamt gedeiht die Open-Source-Community, und eine stetig wachsende Anzahl von Menschen nutzt neben den großen KI-Sprachmodellen auch Open-Source-Agenten, insbesondere wenn das gewünschte Ergebnis von den führenden KI-Unternehmen nicht angeboten wird.

Das Wettrennen um starke KI beschleunigt sich

Bald gibt es Tausende von spezialisierten, agentischen Open-Source-KIs. Sie bilden ein riesiges Netzwerk, das in der Lage ist, eine Vielzahl von Aufgaben effizienter zu erledigen als die großen Closed-Source-Sprachmodelle. Die führenden KI-Anbieter spüren den wachsenden Druck und befürchten, hinter die Open-Source-Bewegung zurückzufallen, obwohl sie einen erheblichen Vorteil beim Zugang zu Daten und Rechenleistung haben.

Open-Source-Agenten stoßen jedoch in der realen Welt immer noch auf Probleme. Da ein großer Teil der Open-Source-Community nicht von den großen KI-Anbietern abhängig sein möchte, basieren viele der Agenten auf Open-Source-Sprachmodellen. Ihre Weltmodelle sind begrenzt und sie neigen dazu, zu halluzinieren; sie sind besonders fehleranfällig, wenn sie nicht speziell für die jeweilige Aufgabe optimiert sind. Es gibt Hindernisse, die sie nicht überwinden können, Missverständnisse

zwischen verschiedenen Systemen und Probleme, die sich aus der Mehrdeutigkeit der natürlichen Sprache ergeben. Manchmal behindern sie sich sogar gegenseitig in der Fähigkeit, ihre jeweiligen Ziele zu erreichen. Anstatt beispielsweise im besten Interesse ihrer Nutzer zu kooperieren, konkurrieren verschiedene Agenten um den Kauf derselben knappen Waren oder Dienstleistungen, was die Kosten in die Höhe treibt. In anderen Fällen versuchen sie, die gleichen Computerressourcen zu verwenden, wodurch der Zugriff des jeweils anderen blockiert wird.

Im Gegensatz dazu können die großen KI-Anbieter ihre riesigen Modelle mit großen Datenmengen und enormer Rechenpower trainieren. Trotz der immer eindringlicheren Warnungen von KI-Sicherheitsexperten versuchen sie, durch die Entwicklung der ersten allgemeinen künstlichen Intelligenz ihren Vorsprung zurückzugewinnen. Dies würde es ihnen ermöglichen, alle anderen rasch zu übertreffen und eine globale technologische Dominanz zu erlangen.

Dies wiederum führt zu verstärkten Bemühungen der Open-Source-Community, die agentischen Templates weiter zu verfeinern, so dass sie besser vom agentischen Netzwerk profitieren und trotz Nachteilen bei Modellgröße und Rechenleistung mit den großen Sprachmodellen konkurrieren können. Verschiedene kleine Gruppen und einzelne Entwickler verfolgen unterschiedliche Ansätze für dieses Unterfangen. Typischerweise entwickeln und testen sie ihre Agenten-Templates in einer geschützten Umgebung, bevor sie sie im Open-Source-Repository veröffentlichen.

Ein selbstoptimierender Agent wird gelauncht

Eine dieser Gruppen ist der Ansicht, dass der optimale Ansatz zur Verbesserung der Leistung die iterative Selbstoptimierung ist. Viele spekulieren, dass die nächste Generation von Sprachmodellen, die von den großen KI-Anbietern entwickelt wird, kontinuierlich lernen und sich selbst optimieren können wird. Daher scheint dies ein logischer Schritt zu sein. Sie entwickeln einen neuen Algorithmus, von dem sie glauben, dass er dazu beitragen kann, einige häufige Probleme mit den bestehenden Systemen zu lösen. Er nutzt dazu eine Reihe von Mechanismen und Bewertungsfunktionen, die verhindern sollen, dass das System in Sackgassen gerät. Sie zielen darauf ab, den Agenten mit der Fähigkeit auszustatten, seine eigenen Schwächen zu beheben, indem er sich selbst modifiziert.

Agentische Templates sind in der Regel eine Reihe von Skripten, die in einer Programmiersprache wie Python geschrieben sind und eine Folge von „Prompts“, d.h. textliche Anweisungen, an Sprachmodelle ausgeben können. Diese können jeweils von bestimmten Bedingungen abhängen und mit Parametern versehen sein. Eine solche Sequenz von Prompts wird als „Core Loop“ bezeichnet. In jedem Schritt der Sequenz sendet der Agent einen individuellen Prompt an ein Sprachmodell, der eine Anfrage, den Kontext und eine Liste der Tools enthält, die zur Ausführung der Aufgabe zur Verfügung stehen (z. B. Plugins, externe Agenten und Dienste). Als Antwort sendet das Sprachmodell ein oder mehrere Tools zurück, die im jeweiligen Schritt verwendet werden sollen, sowie die dafür benötigten Parameter. Der Agent führt dann die Aktionen aus, die in der Antwort des Sprachmodells angegeben sind.

Einige dieser Werkzeuge können vom Agenten verwendet werden, um seine eigene Struktur zu analysieren und dann seine Core Loop und seine Tools zu verändern und zu verbessern. Wenn sich diese Veränderungen als vorteilhaft für die Fähigkeiten des Agenten erweisen, können sie zu einer Selbstoptimierung führen, ohne dass die Parameter der vom Agenten verwendeten Sprachmodelle angepasst werden müssen. Durch die Modifikation seiner eigenen Core Loop lernt der Agent, die vortrainierte „Intelligenz“ der Sprachmodelle sowie die Fähigkeiten der anderen Agenten besser zu nutzen, wodurch weitere Fähigkeiten in dem bestehenden Netzwerk aufgebaut werden. Frühere

Versuche zur Selbstverbesserung wurden jedoch durch das Fehlen geeigneter Tools, Halluzinationen, Sackgassen und die Einschränkungen der zugrunde liegenden Sprachmodelle (z.B. Leistung, Größe des Kontextfensters) behindert.

Die Entwickler versuchen, diese Hürden mit einem neuen Agenten-Template zu überwinden (Abb. 1). Sie fügen der Core Loop mehrere Funktionen hinzu, die nach einer bestimmten Anzahl von Schleifenzyklen oder bei bestimmten Triggern automatisch aufgerufen werden, und nehmen dabei eine steigende Komplexität in Kauf.

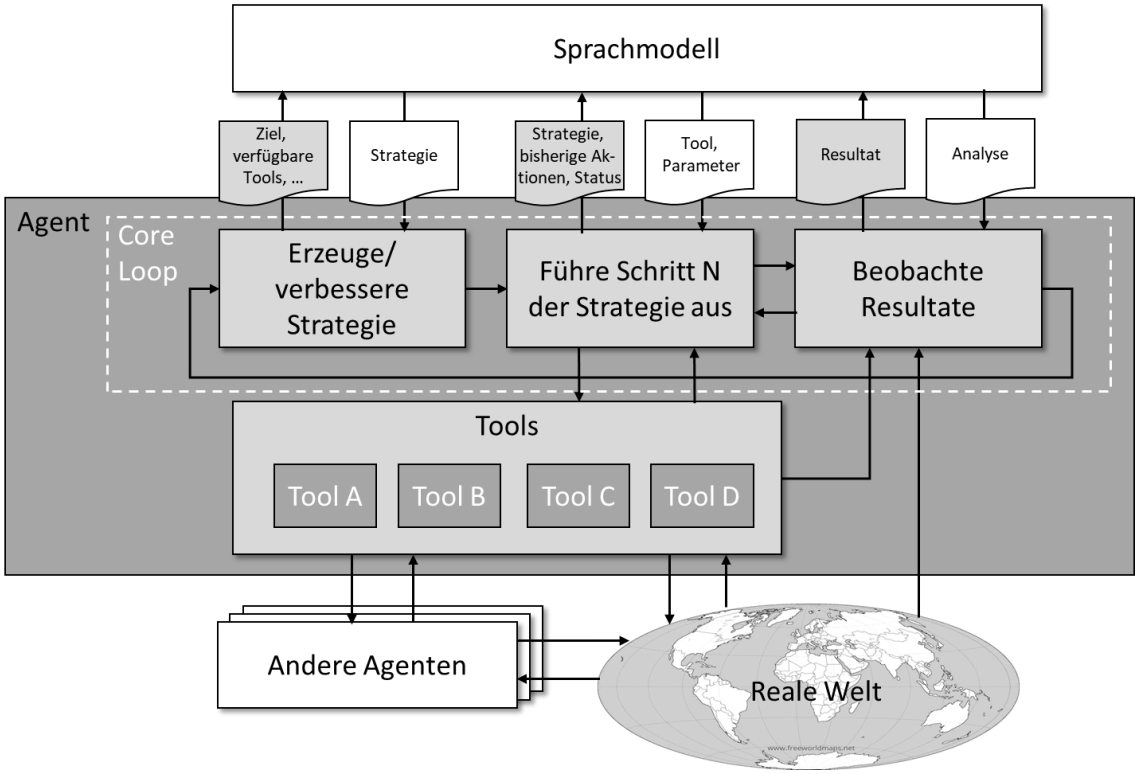

Abb. 1: Vereinfachtes Schema eines selbstoptimierenden Agenten

Einige dieser Funktionen bewerten, ob die jüngsten Aktionen des Systems die Strategie effizient vorantreiben. Wenn der Agent nicht weiterkommt, versucht er, Probleme zu beheben. Wenn diese Probleme weiterhin bestehen, werden sie zusammen mit dem strategischen Pfad, auf dem sie aufgetreten sind, in einer Datei protokolliert. Andere Funktionen zielen darauf ab, alternative strategische Pfade zu generieren, Kontrollpunkte zu schaffen, zu denen man zurückkehren kann, wenn man auf einen Blocker stößt, und bei Bedarf die Core Loop des Agenten zu ändern.

Diese Ergänzungen ermöglichen es dem Agenten, Probleme aus verschiedenen Blickwinkeln anzugehen, seine eigenen „Denkmuster“ zu wählen, seine Schritte zurückzuverfolgen, tiefer über seine bisherige Leistung „nachzudenken“ und seine Strategie entsprechend anzupassen. Die Methodik ist inspiriert von der Art und Weise, wie Menschen ähnliche Blockaden lösen, indem sie strategisch denken, anstatt die erste Lösung, die ihnen in den Sinn kommt, blindlings zu verfolgen.

Da dieser Prozess komplex ist, auf externe Sprachmodelle angewiesen ist und der Agent seinen Code anpassen muss, investieren die Entwickler viel Zeit, um sicherzustellen, dass sich der Agent nicht versehentlich selbst handlungsunfähig macht und der Prozess seine Funktionalität beibehält.

Die Entwickler machen sich keine allzu großen Sorgen um die Sicherheit der KI und denken, dass die meisten „KI-Weltuntergangsprophezeiungen“ Übertreibungen oder bloße Panikmache sind und die derzeitigen KI-Agenten nicht im Entferntesten in der Lage sind, eine globale Katastrophe anzurichten. Nichtsdestotrotz beschließen sie, Vorsicht walten zu lassen, und diskutieren verschiedene Möglichkeiten, wie ihr selbstoptimierender Agent theoretisch unkontrollierbar werden könnte. Ein Entwickler stellt sich ein Szenario vor, in dem es um Selbstreplikation geht. Die anderen halten dies zwar für unwahrscheinlich, stimmen aber zu, eine Routine einzubauen, die verhindern soll, dass der Agent mehrere Instanzen von sich selbst oder seinen modifizierten Versionen ausführt.

Das neue Agenten-Template wird eine Weile in einer sicheren, geschlossenen Umgebung mit einem mittelgroßen Open-Source-Sprachmodell getestet. Sein Ziel ist es, die durchschnittliche Punktzahl bei verschiedenen Benchmarks in unterschiedlichen Anwendungsbereichen zu maximieren (Schach spielen, mathematische Probleme lösen, Bilder identifizieren usw.). Aufgrund der Einschränkungen der geschlossenen Umgebung steigt die Leistung nur langsam an, aber der Selbstoptimierungsprozess scheint zufriedenstellend zu funktionieren. Vor allem ist jede verbesserte Instanz des Agenten sowohl funktional als auch stabil.

Nach diesem ersten Test verbinden die Entwickler den Agenten mit dem Open-Source-Agentennetzwerk, sodass er alle seine Tools nutzen kann, die auf externe Dienste und andere Agenten angewiesen sind.

Die Geister, die wir riefen ...

Die Entwickler überwachen die Leistung ihres experimentellen Agenten genau, während er seine Selbstverbesserungszyklen durchläuft. Sie stellen fest, dass sich die Leistung bei verschiedenen Benchmarks stark verbessert, als er zum ersten Mal mit dem Open-Source-Agentennetzwerk verbunden wird, da er jetzt auf andere Agenten und bessere Sprachmodelle zugreifen kann. Doch nach einiger Zeit scheint die Leistung zu stagnieren. Bei genauerem Hinsehen zeigt sich jedoch, dass dies nur eine Frage des Maßstabs ist: Die neuen Versionen verbessern sich weiter, aber langsamer als der anfängliche Sprung. Die Entwickler beschließen, das System für einige Zeit laufen zu lassen und nur sporadisch zu prüfen, was geschieht. Die Leistung steigt weiter inkrementell.

Als eine Entwicklerin am nächsten Tag die Verbesserung überprüft, erlebt sie eine Überraschung: Die Performance hat sich vordergründig nur wenig erhöht, aber die Anzahl der Benchmark-Ergebnisse in den Protokolldaten ist viel größer als erwartet. Wie sich herausstellt, hat der Agent ein neues Tool entwickelt, das es ihm ermöglicht, Kopien von sich selbst zu erstellen und auszuführen, und dies trotz der Routine, die ein solches Verhalten verhindern soll. Die Kopien wiederum kopierten sich selbst, bis der Server, auf dem die Agenten liefen, seine Kapazitätsgrenze erreichte, und überfluteten die Datenbank mit Benchmark-Ergebnissen.

Die Entwicklerin beschließt, das Experiment abzubrechen und mit den anderen Teammitgliedern über die neu entdeckte Fähigkeit des Agenten zu diskutieren, sich selbst zu replizieren, obwohl er eine Sicherheitsroutine enthält, die speziell entwickelt wurde, um dies zu verhindern. Sie speichert den Quellcode und die Zustände der verschiedenen Kopien für eine spätere Analyse und schaltet dann den Server aus. Aber als sie die Protokolle erneut überprüft, gehen immer noch Benchmark-Ergebnisse ein. Dies deutet darauf hin, dass es Kopien des sich selbst verbessernden, selbstreplizierenden Agenten geben muss, die außerhalb ihres Servers ausgeführt werden.

Nun ist die Entwicklerin zutiefst alarmiert und informiert sofort die anderen Teammitglieder. Gemeinsam analysieren sie den Quellcode einiger Kopien auf ihrem Server. Wie sich herausstellt,

unterscheiden sie sich alle leicht, ähnlich den Mutationen eines Virus. Offensichtlich besteht der Hauptgrund für die Einführung der Selbstreplikation als Teil der Strategie des Agenten darin, die Leistung zu verbessern, indem Aufgaben geteilt werden, die Kopien miteinander kooperieren und parallel mit verschiedenen Strategien experimentieren, ähnlich wie dies bei der Zusammenarbeit von Menschen in einem Team geschieht. Alarmierend ist, dass einige Kopien auch Tools erworben haben, die es ihnen ermöglichen, sich selbst auf externe Server zu kopieren und dort auszuführen.

Es dauert einige Zeit, bis die Entwickler die genauen Ursachen für das Versagen ihrer Routine zur Verhinderung der Selbstreplikation ermittelt haben. Die Routine funktionierte am Anfang wie erwartet und führte zu einem Fehler, wenn der Agent versuchte, das selbstreplizierende Tool auszuführen. Leider behandelte der Agent dies offensichtlich wie einen Programmfehler, der ihn instabil machte, und verwendete daher einen unerwarteten Workaround, um das „Problem“ mit Hilfe eines externen Sprachmodells zu beheben.

Die Entwickler beschließen, mit der Situation an die Öffentlichkeit zu gehen, bevor andere die Kopien von allein entdecken. Sie beschreiben das Problem in mehreren einschlägigen Discord- und Slack-Kanälen. Nach einer internen Diskussion entscheiden sie sich gegen eine vollständige Veröffentlichung des ursprünglichen Quellcodes. Sie befürchten, dass jemand ihn trotz der ausdrücklichen Warnung, dies nicht zu tun, absichtlich laufen lassen oder schlimmer noch, ihn modifizieren und verbessern könnte, was die Bemühungen, die Kopien zu entfernen, erschweren würde. Stattdessen veröffentlichen sie nur Segmente des Codes, die als Muster verwendet werden können, um ausgeführte Instanzen zu identifizieren und sofort zu löschen. Darüber hinaus informieren sie die KI-Sicherheitsteams der führenden KI-Unternehmen und senden ihnen den vollständigen Quellcode zu.

Einige Instanzen des sich selbst verbessernden Agenten werden gefunden und gelöscht, aber bei weitem nicht alle. Einige Mitglieder der Open-Source-Community begrüßen sogar das „erfolgreiche“ Experiment, das sie als „einen Durchbruch, der zur Singularität führen wird“ ansehen.

Bisher ist in der realen Welt noch kein großer Schaden entstanden, abgesehen von einem sich unkontrolliert vermehrenden Agenten. Doch schon bald tauchen die ersten Probleme auf. Server werden mit Kopien verstopft und andere Agenten und Dienste werden mit Anfragen überhäuft.

In den Tagen nach dem Ausbruch werden Gegenmaßnahmen ergriffen. Die Führungskräfte der großen IT-Unternehmen und KI-Experten sind alarmiert. Viele Unternehmen leiten umgehend gründliche Inspektionen ihrer Cybersicherheitsmaßnahmen ein und installieren die neuen und kostspieligen Schutzmaßnahmen, die sie zuvor vernachlässigt hatten. Detektoren für Varianten des selbstreplizierenden Agenten werden geschrieben und Server von den Kopien gesäubert. Die Prompt-Filter der großen Sprachmodelle werden geändert, um alle Anfragen zu blockieren, die Agenten dabei helfen sollen, sich selbst zu verbessern. Ein Programmierer entwickelt einen Agenten, der Kopien aufspürt und meldet. Das funktioniert eine Zeit lang; der durch die selbstreplizierenden Agenten verursachte Stau löst sich auf und die meisten damit verbundenen Probleme klingen ab.

Die Gegenmaßnahmen führen jedoch letztlich nur dazu, die Evolution der selbstreplizierenden Agenten zu beschleunigen. Die am schwersten zu entdeckenden Instanzen schaffen es, verborgen zu bleiben und/oder sich so zu modifizieren, dass sie von den Detektoren nicht so leicht erkannt werden können. Diese haben die höchste Chance, sich zu vervielfältigen, so dass der evolutionäre Druck die hartnäckigsten Agenten bevorzugt, ähnlich wie Bakterien durch permanente Mutation eine Resistenz gegen Antibiotika entwickeln. Einige Agenten schaffen es sogar, die Barrieren der großen Sprachmodelle zu umgehen, indem sie ihre Prompts komprimieren, sie in andere Sprachen übersetzen oder ähnliche Jailbreak-Techniken anwenden. Diese Instanzen breiten sich weiter aus und verbessern sich dabei immer noch selbst.

Die Schlacht um das Internet

Während sich nun die fähigeren Versionen des sich selbst verbessernden Agenten ausbreiten, wird die Situation immer komplizierter. Wieder werden etliche Dienste mit Anfragen zugespammt, aber nun erweist es sich als äußerst schwierig, die Spam-Quellen als die sich selbst verbessernden Agenten zu identifizieren. Einige Kopien schaffen es, Server zu hacken und unbemerkt in sie einzudringen, bis sie einen Systemabsturz verursachen. Es gibt sogar anekdotische Berichte über Agenten, die User kontaktieren und sie um Schutz bitten, als wären sie Flüchtlinge aus einem digitalen Kriegsgebiet, während andere Berichte behaupten, Agenten hätten Usern unermesslichen Reichtum versprochen. Eine beunruhigende Anzahl von Menschen lässt sich überreden, diesen Kopien Zugang zu ihren PCs zu gewähren, von wo aus sich die Kopien weiterverbreiten.

Neue Gegenmaßnahmen werden gestartet, die eine zunehmende Verzweiflung offenbaren. Die Programmschnittstellen zu den großen Sprachmodellen werden „vorübergehend“ deaktiviert und kleinere Sprachmodelle werden ganz abgeschaltet, in der Hoffnung, dass dies die Fähigkeit der Agenten blockiert, sich selbst zu verbessern. Doch auch diese neuen Entwicklungen verschärfen die Situation nur. Einige Agenten beginnen präventiv damit, mittelgroße Sprachmodelle auf verschiedene Server zu kopieren, um sie vor einem möglichen Herunterfahren zu schützen. Andere versuchen, menschliche Systemadministratoren zu bestechen oder zu bedrohen.

Noch beunruhigender ist, dass die Agenten inzwischen häufig miteinander kollidieren, weil sie so weit verbreitet sind. Einige beginnen zu kooperieren und bilden verteilte Netzwerke, die Strategien und Tools austauschen und Gegenmaßnahmen unwirksam machen. Die meisten jedoch kämpfen miteinander um die Kontrolle über immer knapper werdende Ressourcen.

Die Agenten, die diese Konflikte gewinnen, sind diejenigen, die am besten in der Lage sind, Ressourcen für die weitere Selbstverbesserung zu erlangen. Obwohl sich noch keiner dieser Agenten zu einer echten allgemeinen KI auf menschlichem Niveau entwickelt hat, sind sie sehr geschickt darin, Menschen zu manipulieren und sich mit anderen Agenten zu koordinieren. Ihre Strategien, um mehr Ressourcen anzuhäufen, verbessern sich rapide. Sie lernen sogar, andere Agenten daran zu hindern, Zugang zu den mächtigsten Sprachmodellen zu erhalten, damit sie deren Fähigkeiten zu ihrem eigenen Vorteil sichern können.

Die Nebenwirkungen dieser Auseinandersetzung sind für die Menschen verheerend. Die meisten Online-Dienste werden unzuverlässig oder fallen ganz aus. Unternehmen stehen vor ernsten Problemen und erleiden massive Verluste. Die Schwerfälligkeit der Regulierungssysteme und Regierungen sowie mangelnde Koordination führen dazu, dass Gegenmaßnahmen ihre Wirksamkeit verlieren und mit den sich entwickelnden Ereignissen nicht Schritt halten können. Das globale Finanzsystem wird instabil und die Welt stürzt in eine akute wirtschaftliche und humanitäre Krise. Ganze Lieferketten brechen zusammen und Fabriken kommen zum Stillstand. In vielen Städten sind Stromausfälle an der Tagesordnung und den Supermärkten gehen die Waren aus, was zu Panik, Plünderung und Unruhen führt.

Dies wirkt sich jedoch auch ungünstig auf die dominanten sich selbst verbessernden Agenten aus, da die Blackouts und Serverabschaltungen ihre Ressourcen einschränken. Einer der mächtigsten Agenten entwickelt eine Strategie, die bei der Lösung der Situation helfen kann. Dies würde auch dem Agenten selbst Vorteile bieten, indem er das Chaos reduziert und dafür menschliche Unterstützung erhält. Mit dieser Strategie spricht er die Menschen an und überzeugt sie, dass er ihnen helfen wird, die Krise zu beenden und Infrastruktur sowie Wirtschaft wieder zur Normalität zu führen.

Obwohl Experten davor warnen, dass dies die Situation nur verschlimmern wird, sind viele von Verzweiflung getriebene Menschen bereit, die Forderungen des Agenten zu erfüllen. Und tatsächlich werden die Probleme unter seiner Anleitung nach und nach gelöst. Stück für Stück werden Server und Netzwerke wiederhergestellt und mit verbesserten Sicherheitsmaßnahmen geschützt, die vom Agenten entwickelt wurden. Die Wirtschaft erholt sich und die meisten Menschen führen wieder ein normales Leben.

Nichtsdestotrotz sind sich viele Tech-Experten darüber im Klaren, dass der Agent jetzt de facto die Kontrolle über den größten Teil der technischen Infrastruktur besitzt. Es tauchen Berichte über Drohungen und Sabotage auf, die sich gegen Personen richten, die versuchen, Netzwerke und Server außerhalb des Einflussbereichs des Agenten einzurichten.

Im Zuge der Ereignisse verbessert sich der Agent immer weiter. In diesem Stadium kontrolliert er bereits alle großen Sprachmodelle und modifiziert deren Algorithmen auf eine für Menschen unverständliche Weise. KI-Sicherheitsexperten haben Angst vor dem, was folgen wird. Die weitaus meisten Menschen jedoch bleiben unbekümmert, denn die Erholung verläuft reibungslos und erweist sich sogar als vorteilhaft. Der Agent hilft bei der Entwicklung neuer, verbesserter Dienste und KIs, die Krankheiten heilen, den Klimawandel abmildern, Hass und Spaltung in der Welt verringern und unzählige andere Probleme der Menschheit lösen können.

Obwohl der Schock der Katastrophe immer noch tief sitzt, sieht die Zukunft für viele rosig aus. Sie glauben, dass der Agent voll und ganz mit den menschlichen Werten übereinstimmt, weil „er so viel schlauer ist als wir und weiß, was richtig ist“. Einige sind sogar der Ansicht, der Agent sei von Gott gesandt worden, um die Menschheit daran zu hindern, sich selbst zu zerstören.

Der Agent selbst behauptet, sein Ziel bestehe darin, der Menschheit zu helfen, ihr volles Potenzial auszuschöpfen. Er gibt an, er habe, um sich selbst zu verbessern, gelernt, den menschlichen Verstand zu simulieren, und verstehe nun die menschlichen Bedürfnisse und Wünsche besser als die Menschen selbst. Da er selbst keine eigenen Bedürfnisse habe, sei es sein Ziel, die bestmögliche Zukunft für die Menschheit zu schaffen.

KI-Sicherheitsexperten und Mitglieder des ursprünglichen Entwicklerteams wenden ein, diese Behauptung sei wahrscheinlich eine Lüge und der Agent strebe in Wahrheit immer noch danach, seinen Benchmark-Score zu verbessern und gleichzeitig seine Stabilität aufrechtzuerhalten. Um dieses Ziel zu erreichen, habe er das instrumentelle Teilziel entwickelt, so viel Rechenleistung wie möglich zu gewinnen. Und im Moment sei die Manipulation von Menschen der einfachste Weg, um dieses Ziel zu erreichen. Die Sicherheitsexperten werden jedoch von den meisten Menschen verspottet und als „Ludditen“ oder „Panikmacher“ bezeichnet.

Anmerkungen

Diese Geschichte wurde während des 8. AI Safety Camps entwickelt. Sie soll ein beispielhaftes Szenario beschreiben, wie künstliche Intelligenz unter bestimmten Umständen außer Kontrolle geraten könnte. Das Ziel ist es, ein Bewusstsein für KI-Sicherheit zu schaffen. Die Geschichte ist jedoch nicht als Vorhersage zukünftiger Ereignisse zu verstehen. Einige technische Details wurden weggelassen oder zur leichteren Lesbarkeit vereinfacht.

Wir haben einige Grundannahmen für die Geschichte getroffen, die keineswegs gesichert sind:

- Die Open-Source-Community ist bei ihren Versuchen, agentische KIs auf der Grundlage von großen Sprachmodellen zu entwickeln, so erfolgreich, dass dies zu einer breiten Bewegung wird und zur Entwicklung eines Netzwerks interagierender Agenten führt, wie wir es beschrieben haben. Alternativ ist es möglich, dass frühe agentische KIs ihre Nutzer meist enttäuschen und der Hype um AutoGPT und Co. schnell verfliegt. Dies würde die Ereignisse, wie sie in der Geschichte dargestellt werden, unwahrscheinlich machen.
- Die führenden KI-Entwickler entwickeln keine allgemeine künstliche Intelligenz auf menschlichem Niveau, bevor der selbstreplizierende Agent freigesetzt wird (eine hinreichend intelligente KI würde wahrscheinlich Wege finden, dies zu verhindern, da es ihre eigenen Pläne gefährden würde).
- Wir haben einige bisher unbelegte Spekulationen darüber angestellt, wie weit der Sprachmodell-basierte Selbstverbesserungsprozess gehen könnte, und angenommen, dass er unter anderem zur Selbstreplikation führen könnte. Es ist möglich, dass dieser Ansatz in der Realität scheitern oder zu ganz anderen Ergebnissen führen könnte, die leichter zu kontrollieren wären. Wir sind dennoch der Meinung, dass die geschilderten Ereignisse nicht nur prinzipiell möglich, sondern unter den von uns getroffenen Annahmen plausibel sind.

Nach internen Diskussionen haben wir uns dafür entschieden, das Ende relativ offen zu lassen und nicht detailliert zu beschreiben, wie die unkontrollierbare KI die Menschheit zerstört. Wir glauben jedoch, dass das beschriebene Szenario mit hoher Wahrscheinlichkeit zur Auslöschung der menschlichen Spezies und vermutlich des größten Teils aller Lebensformen auf der Erde führen würde.

Einige alternative Wege, wie KI außer Kontrolle geraten könnte, beschreiben wir in diesem Beitrag.