RISIKEN UND CHANCEN BEIM EINSATZ VON KÜNSTLICHER INTELLIGENZ. DER FALL CHATGPT

von Professor Alessandro Dario Cortesi

Zusammenfassung

Fuzzy Logic - Neuronale Netze - Deep Learning –

KI als Objekt und nicht als Subjekt

Allgemeine-Intelligenz–Superintelligenz

Der Fall ChatGPT

Risiken bei der Nutzung von KI

Abhilfemaßnahmen – Die Anwendung des Vorsorgeprinzips

Vorwort

Künstliche Intelligenz ist sicherlich kein neues Thema.

Der Begriff "Roboter" leitet sich vom tschechischen "robota" ab, d. h. "Zwangsarbeit", und wurde erstmals in Karel Čapeks Theaterstück "R.U.R." für humanoide Arbeitskräfte verwendet. (Akronym für "Rossumovi univerzální roboti", zu übersetzen mit "Rossum's Universal Robots") aus dem Jahr 1920.

Alan Turings bekannter Artikel "Computer Machinery and Intelligence", der mit der Frage "Können Maschinen denken?" beginnt und das sogenannte "Nachahmungsspiel" vorstellt, erschien 1950 in der Zeitschrift Mind, Nr. 59 (S. 433 ff.).

Es ist auch nicht das erste Mal, dass das Thema von unserer Vereinigung aufgegriffen wird.

Im Jahr 2018 hatte ich die Ehre, den italienischen Vortrag auf der Saarbrücker Konferenz zu halten, deren Thema die Europäische Datenschutzverordnung (DSGVO) war, und schon damals hatte ich die Gelegenheit, über die Revolution zu sprechen, die durch die Big-Data-Analyse, das maschinelle Lernen, die sogenannten Waffen der mathematischen Zerstörung, den Roboter-Richter, die Rolle der Aufsichtsbehörde und so weiter ausgelöst wurde; Themen, die wir heute zum Teil berühren werden.

Das Thema ist wieder in Mode gekommen, weil parallel zur linearen Entwicklung der Robotik in jüngster Zeit Anwendungen der künstlichen Intelligenz im engeren Sinne (so genannte starke KI) auf den Markt gekommen sind, die auf der Grundlage von Fuzzy-Logik und künstlichen neuronalen Netzen die Funktionsweise des menschlichen Gehirns simulieren und es dem Computer ermöglichen, sich selbst neu zu programmieren. Diese Fähigkeit, den eigenen Code als Reaktion auf äußere Reize selbstständig zu verbessern (Deep Learning), bildet das Herzstück der künstlichen Intelligenz und birgt die größten rechtlichen Probleme.

Fuzzy-Logik - Neuronale Netze - Deep Learning - KI - Objekt und nicht Subjekt

Einige technische Klarstellungen:

Die unscharfe Logik ermöglicht die Verarbeitung von Konzepten, indem sie sich von dem Engpass der binären Wahr-Falsch-Logik befreit. Nur zwei entgegengesetzte Zustände sind nicht verarbeitbar: heiß/kalt; hoch/tief; schwarz/weiß usw. Es gibt Elemente der Realität, die in unterschiedlichem Maße zu mehreren divergierenden Gruppen gehören. So kann beispielsweise ein Getränk "lauwarm" sein, d. h. es kann sowohl zur Klasse der "heißen" Dinge als auch zur Klasse der "kalten" Dinge gehören1.

Künstliche neuronale Netze sind ein Rechenmodell, das sich lose an das biologische neuronale Netz anlehnt, das in ihnen eine sehr vereinfachte Nachahmung findet. Es besteht aus Informationen (sogenannten künstlichen Neuronen), die nach einem adaptiven Modell miteinander verbunden sind, das seine Struktur als Reaktion auf externe oder interne Reize während der Lernphase verändert2.

Kurz gesagt, indem der KI eine große Menge an Daten übermittelt wird und sie manchmal das entsprechende Feedback von Menschen erhält, ist der Prozessor in der Lage, "aus seinen Fehlern zu lernen" und Ergebnisse zu erzielen, die nicht völlig vorhersehbar oder sogar unvorhersehbar sind.

Nach der bekannten Definition von T. Mitchell heißt es, dass ein Computerprogramm in Bezug auf eine Klasse von Aufgaben T "aus Erfahrungen E" lernt und dass sich seine durch P gemessene Effektivität verbessert, wenn es E3 erlebt.

Dies ist nicht wirklich "Intelligenz", und wenn wir diesen Begriff verwenden, müssen wir uns bewusst sein, dass wir eine Metapher verwenden. Zumindest für diejenigen, die wie ich jede reduktionistische Hypothese (jede Ableitung des materialistischen Szientismus) ablehnen, handelt es sich um ein deutlich anderes Phänomen als das, was eminent menschlich ist und bleibt.

Wir können das Substantiv "Intelligenz" nur verwenden, weil das zentrale Nervensystem noch viele Rätsel aufgibt. Ein Doktorand von mir, der sich vor Jahren mit dieser Thematik

befasst hat, hat mindestens zwölf verschiedene widersprüchliche Definitionen von Intelligenz in der Lehrmeinung gefunden, was in der Praxis bedeutet, dass es keine gibt.

Der Touring-Test selbst ist bei näherer Betrachtung nichts anderes als die Anerkennung dieser Unfähigkeit: Da wir nicht wissen, wie wir als das Proprium der menschlichen Intelligenz definieren sollen, verstehen wir unter "künstlicher Intelligenz" das, was konkret nicht von ihr zu unterscheiden ist.

Durch die Nutzung eines abstrakt unbegrenzten Speichers und einer unbegrenzten Verarbeitungsgeschwindigkeit könnte die Maschine Ergebnisse von außerordentlicher Effizienz erzielen, aber meiner Meinung nach macht es keinen Sinn, wie manche es tun, von "Soulware" zu sprechen: Software kann niemals eine Seele, ein Gewissen, ein Selbstbewusstsein4 haben: Sie kann menschliches Verhalten simulieren, aber niemals "jemand" sein/werden. Daher halte ich die Verleihung der Staatsbürgerschaft an eine künstliche Intelligenz (wie im Fall des Roboters Sophia im Jahr 2017 in Riyad) oder die Verleihung einer "elektronischen Rechtspersönlichkeit" nicht für angebracht.5 Ich halte es auch nicht für angebracht, die zivil- oder gar strafrechtliche Haftung von Maschinen zu bejahen. Künstliche Intelligenz ist und muss ein Objekt bleiben, kein Rechtssubjekt.

Vielmehr ist es dringend erforderlich, zu gemeinsamen Schlussfolgerungen über das Rechtssubjekt zu gelangen, das im Mittelpunkt der Zurechnung von Schäden steht: der Hersteller, der Programmierer, der Ausbilder, der Benutzer usw.

Allgemeine Intelligenz – Superintelligenz

Ein altes Gebäude im Zentrum von Paris beherbergt die europäische Abteilung der Facebook Artificial Intelligence Researchers (FAIR), die jetzt Meta AI heißt.

Bis 2021 wurde sie von einem theoretischen Linguisten, Prof. Marco Baroni, geleitet, der in Bozen geboren wurde und in Padua studiert hat. Die Aufgabe dieser Abteilung besteht darin, Maschinen das Erlernen von Sprachen beizubringen.

Die Beherrschung einer Sprache erreicht man nicht durch das Studium von Grammatikregeln oder Wortlisten (wie es Schulkinder noch oft tun). Um gute Ergebnisse zu erzielen, muss man das nachahmen, was Kinder tun, wenn sie auf natürliche Weise mit ihrer Muttersprache experimentieren.

Das Problem ist, dass es den Wissenschaftlern immer noch ein Rätsel ist, wie wir alle unsere Hauptsprache lernen, und dass die KI von dem geplagt wird, was dieser Forscher als "katastrophales Vergessen" bezeichnet.

Um zu verstehen, worum es sich dabei handelt, erinnern wir uns an Deep Blue, den berühmten IBM-Computer, der 1996 die erste Schachpartie gegen den Weltmeister gewinnen konnte, indem er viele Partien der größten Spieler der Schachgeschichte

studierte. Alles, was er über das Schachspiel gelernt hat, all die ausgeklügelten Strategien, die er sich angeeignet hat, würden es Deep Blue nicht erlauben, auch nur eine einzige Partie eines anderen Spiels zu gewinnen ... in einem anderen Anwendungsbereich ist die KI völlig entwaffnet.

Doch 2017 schlug DeepMinds (Googles) Alpha Go Zero den Go-Weltmeister, und zwar, anders als Deep Blue, ohne irgendwelche menschlichen Partien zu studieren, sondern einfach (sobald es die Regeln gelernt hatte), durch das Training, mit sich selbst zu spielen und aus seinen Fehlern zu lernen (was, je nach Geschwindigkeit des Prozessors, sogar in ein paar Tagen geschehen kann).

Nun, wir wissen nicht, wie viele Jahre uns noch von der Entwicklung einer Software mit "allgemeiner Intelligenz" trennen (d. h. der Fähigkeit, beliebige kognitive Zwecke und nicht nur spezifische Aufgaben zu verfolgen), oder wie lange es dauern wird, bis sie jene rekursive und sehr schnelle Verbesserung auslöst, die zur so genannten "Superintelligenz" führen wird (d. h. einer Intelligenz, die die des Menschen im Allgemeinen übertrifft), aber Beispiele wie ChatGPT und die rasanten Entwicklungen im Quantencomputing lassen uns vermuten, dass es nicht mehr allzu weit ist.

Der Fall ChatGPT

ChatGPT ist eine KI-Anwendung. GPT steht für Generative Pre-trained Transformer und der Name deutet bereits darauf hin, dass es sich um eine konversationelle, generative KI handelt, die auf einem transformativen künstlichen neuronalen Netz basiert. Kurz gesagt, es handelt sich um eine künstliche Intelligenz, mit der man sich unterhalten kann (weil sie die natürliche Sprache der Benutzer als Eingabe verarbeitet) und die Inhalte, insbesondere Textinhalte, generiert.

Sie wurde von OpenAI entwickelt, einer am 10. Dezember 2015 in San Francisco gegründeten Non-Profit-Organisation, deren Ziel es ist, eine freundliche künstliche Intelligenz zu fördern, von der die Menschheit profitieren kann, und deren Patente und Forschung der Öffentlichkeit zugänglich sind.

Zu den Gründern gehören Elon Musk, Samuel Harris (Sam) Altman (Präsident von Y Combinator, einem Start-up-Beschleuniger, der ebenfalls zu den Gründern gehört; derzeitiger CEO von OpenAI), Greg Brockman (ehemaliger Chief Technology Officer und jetzt Präsident von OpenAI), Ilya Sutskever (Chief Scientist von OpenAI) und Wojciech Zaremba (Hauptautor des GPT-Modells, wie Sutskever von Google Brain), aber auch Amazon Web Services und Infosys.

ChatGPT wurde mit Techniken des maschinellen Lernens (unbeaufsichtigter Typ) entwickelt und mit Techniken des überwachten und verstärkenden Lernens optimiert. Am 3. November 2022 gestartet, hat es innerhalb weniger Tage Millionen von Nutzern erreicht

(so viele, dass der Zugriff aufgrund der vielen Anfragen schwierig ist), was für die Güte der angebotenen Antworten spricht.

Im Jahr 2018 versuchte Elon Musk, eine dominante Rolle bei OpenAI zu übernehmen. Der Vorstand akzeptierte seinen Vorschlag nicht und Musk beschloss daraufhin, offiziell wegen eines möglichen Interessenkonflikts zurückzutreten, da er CEO von Tesla ist.

Im Jahr 2019 wurde OpenAI von einem "Non-Profit"-Unternehmen in ein "Capped For-Profit"-Unternehmen umgewandelt. Letztere Unternehmensstruktur erlaubt das Streben nach Gewinn (und damit das Anlocken von Investitionen), jedoch nicht über eine bestimmte Grenze hinaus, um ein Gleichgewicht zwischen dem Streben nach Gewinn und dem Erreichen sozialer Ziele herzustellen.

Im Mai 2019 erhielt OpenAI eine Investition von Microsoft in Höhe von 1 Milliarde Dollar, die später auf 10 Milliarden Dollar aufgestockt wurde, mit dem Ziel, 49 % des Kapitals zu erlangen. Alle Systeme von OpenAI laufen auf einem Microsoft-Supercomputer.

Microsoft hat ChatGPT bereits in die Antworten seiner Suchmaschine Bing integriert (so dass Samsung darüber nachdenkt, es anstelle von Google als Suchmaschine in seine Geräte zu integrieren) und beabsichtigt, auch die Programme des Office-Pakets durch KI zu bereichern.

Google bleibt natürlich nicht untätig und wird in Kürze "Sparrow" auf den Markt bringen, eine Anwendung, die von einer Schwesterfirma namens DeepMind entwickelt wurde, die von der gemeinsamen Muttergesellschaft Alphabet kontrolliert wird, und die verspricht, bei der Korrektur von Fehlern und beim Zitieren von Quellen effektiver zu sein als ChatGPT.

Aber es gibt noch viele andere Anwendungen generativer KI: Synthesia, Midjournay, Wellsaid, Runway, Writesonic, um nur einige zu nennen, und seit kurzem auch das sehr interessante "Claude" von Anthropic, dem ich persönlich den Vorzug gebe.

Um auf ChatGPT zurückzukommen: Es ist nicht ohne Einschränkungen. Es wurde auf einem Textkorpus von über 570 Gigabyte in englischer Sprache und auf anderen, kleineren Texten in anderen Sprachen wie Französisch, Spanisch, Deutsch, Italienisch, Chinesisch, Japanisch, Koreanisch usw. trainiert. Die künstliche Intelligenz erkennt, wenn die Frage gestellt wird, dass dies bedeutet: "Ich habe vielleicht mehr Kenntnisse und Fähigkeiten in Englisch als in anderen Sprachen".

In einigen Fällen, wie meine Schüler in einigen Sitzungen feststellen konnten, gibt ChatGPT unzuverlässige Antworten.

Wie ChatGPT selbst erklärt: "Als Sprachmodell basiert mein Wissen auf den Trainingsdaten, die von meiner Schöpfung bereitgestellt werden, die im Jahr 2021 endet.

Das bedeutet, dass meine Antworten und mein Wissen keine Ereignisse berücksichtigen, die nach diesem Datum eingetreten sind [...] Es ist wichtig, darauf hinzuweisen, dass ich als Modell mit künstlicher Intelligenz nicht für die Richtigkeit der in den Trainingsdaten enthaltenen Informationen garantieren kann, da ich nicht in der Lage bin, den Wahrheitsgehalt der Inhalte zu überprüfen."

Das berühmteste Beispiel für "Halluzinationen" (so lautet der Fachausdruck für seine Fehler) ist der Fall des Bürgermeisters Brian Hood aus Hepburn Shire, einer Kleinstadt nordwestlich von Melbourne, Australien.

Als er im Jahr 2000 für eine Zweigstelle der Reserve Bank, Note Printing Australia, arbeitete, berichtete Brian Hood über Vorfälle von interner Korruption in der Zweigstelle. ChatGPT interpretierte die Presseartikel, die darüber berichteten, falsch und brachte den Namen des Bürgermeisters mit den Verbrechen in Verbindung, die er vereitelt hatte, so dass er als Täter dastand und seine Ehre dadurch schwer beschädigt wurde.

Am 20. März 2023 kam es bei ChatGPT außerdem zu einem schweren Datenverlust (Datenschutzverletzung) in Bezug auf die Unterhaltungen der Nutzer und die Zahlungsinformationen der Abonnenten des kostenpflichtigen Dienstes.

Dies führte dazu, dass die italienische Datenschutzbehörde feststellte:

- dass weder die Nutzer noch die betroffenen Personen, deren Daten von Open AI gesammelt und über den ChatGPT-Dienst verarbeitet wurden, informiert wurden

- das Fehlen einer angemessenen Rechtsgrundlage für die Erhebung personenbezogener Daten und deren Verarbeitung zum Zwecke des Trainings der Algorithmen, die dem Betrieb von ChatGPT zugrunde liegen;

- dass die Verarbeitung der personenbezogenen Daten der betroffenen Personen ungenau ist, da die von ChatGPT bereitgestellten Informationen nicht immer mit den tatsächlichen Daten übereinstimmen

- das Fehlen einer Überprüfung des Alters der Nutzer in Bezug auf den ChatGPT-Dienst, der gemäß den von OpenAI veröffentlichten Bedingungen Personen vorbehalten ist, die mindestens 13 Jahre alt sind (während in Italien die Grenze für eine gültige Zustimmung bei 14 Jahren liegt);

und in Anbetracht der Tatsache, dass das Fehlen von Filtern für Minderjährige unter 13 Jahren sie Reaktionen aussetzt, die in Bezug auf ihren Entwicklungsstand und ihr Selbstbewusstsein völlig ungeeignet sind, die Verarbeitung personenbezogener Daten von auf italienischem Hoheitsgebiet ansässigen betroffenen Personen vorläufig einzuschränken, eine Untersuchung einzuleiten und OpenAI eine Reihe von Vorschriften zu diktieren, die gemäß Artikel 58 Absatz 2 Buchstabe d DSGVO angeordnet wurden.

Diese Maßnahme der italienischen Datenschutzbehörde (Dringlichkeitsbeschluss des Präsidenten Nr. 112 vom 30. März 2023) stieß im eigenen Land auf heftige Kritik.

Nur wenige - darunter diejenigen, die heute zu Ihnen sprechen - haben die Stichhaltigkeit der getroffenen Feststellungen anerkannt. Die meisten hingegen waren der Meinung, dass die Unkenntlichmachung von ChatGPT das Ergebnis einer rückschrittlichen, den neuen Technologien zuwiderlaufenden Haltung sei; sie merkten auch an, dass es sich um eine nutzlose Maßnahme handele, die über VPN (virtuelles privates Netzwerk) umgangen werden könne, und dass sie sogar gegen die DSGVO verstoße, da die italienische Behörde ("Garante") inkompetent sei.

Wie ich in einer Reihe von Reden gehofft hatte, wurden jedoch innerhalb weniger Tage ähnliche Untersuchungen in Kanada, Deutschland usw. eingeleitet. Vor allem aber hat der Europäische Datenschutzausschuss eine spezielle Arbeitsgruppe eingerichtet, die mögliche Verstöße von ChatGPT gegen die Datenschutzbestimmungen aufdecken soll.

Diese Stellungnahmen haben Open AI dazu veranlasst, Korrekturmaßnahmen zu ergreifen und sich mit dem Garante auf folgende Schritte zu einigen (die bis zum 30. April 2023 umzusetzen sind):

1. einen Informationsvermerk zu erstellen und auf seiner Website zu veröffentlichen, in dem den betroffenen Personen, einschließlich derjenigen, die keine Nutzer des ChatGPT-Dienstes sind und deren Daten für die Zwecke des Algorithmus-Trainings erhoben und verarbeitet wurden, die Methoden der Verarbeitung, die der für den Betrieb des Dienstes erforderlichen Verarbeitung zugrunde liegende Logik, ihre Rechte als betroffene Personen und alle anderen in der Verordnung geforderten Informationen erläutert werden, und zwar in der in Artikel 12 DSGVO festgelegten Art und Weise;

2. auf der OpenAI-Website zumindest den betroffenen Personen, einschließlich derjenigen, die keine Nutzer des Dienstes sind und von Italien aus eine Verbindung herstellen, ein Instrument zur Verfügung zu stellen, mit dem sie ihr Recht auf Widerspruch gegen die Verarbeitung ihrer personenbezogenen Daten ausüben können, die das Unternehmen für die Zwecke des Algorithmus-Trainings und der Bereitstellung des Dienstes von Dritten erhalten hat;

3. auf ihrer Internetseite zumindest den interessierten Personen, einschließlich derjenigen, die keine Nutzer der Dienste sind und sich von Italien aus anmelden, ein Instrument zur Verfügung zu stellen, mit dem sie die Berichtigung der sie betreffenden personenbezogenen Daten, die bei der Erstellung der Inhalte unrichtig verarbeitet wurden, oder, falls dies nach dem Stand der Technik nicht möglich ist, die Löschung ihrer personenbezogenen Daten verlangen können;

4. einen Link zu dem an die Nutzer seiner Dienste gerichteten Informationshinweis in den Registrierungsablauf an einer Stelle einzufügen, die es ermöglicht, diesen vor der Registrierung zu lesen, und zwar so, dass alle Nutzer, die sich von Italien aus einloggen, einschließlich der bereits registrierten, beim ersten Zugriff nach einer möglichen Reaktivierung des Dienstes diesen Informationshinweis lesen können;

5. die Rechtsgrundlage für die Verarbeitung der personenbezogenen Daten der Nutzer für die Zwecke der Algorithmenschulung zu ändern, indem jeglicher Bezug auf den Vertrag gestrichen wird und als Rechtsgrundlage für die Verarbeitung die Einwilligung oder das berechtigte Interesse in Bezug auf die Bewertungen des Unternehmens im Rahmen einer Logik der Verantwortlichkeit angenommen wird;

6. auf seiner Website zumindest den Nutzern des Dienstes, die sich von Italien aus anmelden, ein leicht zugängliches Instrument zur Verfügung zu stellen, mit dem sie ihr Recht auf Widerspruch gegen die Verarbeitung ihrer Daten, die sie bei der Nutzung des Dienstes für das Algorithmustraining erhalten haben, ausüben können, wenn die unter Punkt 5 gewählte Rechtsgrundlage ein berechtigtes Interesse ist;

7. bei jeder Reaktivierung des Dienstes von Italien aus alle Nutzer, die sich von Italien aus einloggen, einschließlich der bereits registrierten, aufzufordern, beim ersten Zugriff eine Alterskontrolle zu durchlaufen, die minderjährige Nutzer auf der Grundlage des angegebenen Alters ausschließt;

8. dem Garanten bis spätestens 31. Mai 2023 einen Plan für die Einführung von Instrumenten zur Altersüberprüfung vorzulegen, die geeignet sind, den Zugang von Nutzern unter 13 Jahren und minderjährigen Nutzern auszuschließen, wenn keine ausdrückliche Willensbekundung derjenigen vorliegt, die die elterliche Verantwortung für sie ausüben. Die Umsetzung dieses Plans muss spätestens am 30. September 2023 beginnen;

9. bis spätestens 15. Mai 2023 eine Informationskampagne ohne Werbecharakter in allen wichtigen italienischen Massenmedien (Radio, Fernsehen, Zeitungen und Internet) zu fördern, um die Personen darüber zu informieren, dass ihre personenbezogenen Daten für die Zwecke der Algorithmen Schulung erhoben werden können, dass ein ausführlicher Informationshinweis auf der Website des Unternehmens veröffentlicht wird und dass - ebenfalls auf der Website des Unternehmens - ein Tool zur Verfügung gestellt wird, mit dem alle betroffenen Personen die Löschung ihrer personenbezogenen Daten beantragen und erreichen können.

Als Reaktion auf diese Zusagen von Open AI, die das Ergebnis eines Dialogs mit der Garante waren, erließ diese die Verfügung Nr. 114 vom 11. April 2023, in der sie die Wirksamkeit der vorherigen Einschränkung aussetzte.

ChatGPT ist daher seit Ende April auch in Italien wieder online.

Risiken bei der Nutzung von KI

Betrachten wir nun die Risiken, die mit dem Einsatz von künstlicher Intelligenz verbunden sind.

Stephen Hawking erklärte in einer von der BBC ausgestrahlten Sendung im Dezember 2014, dass "die Entwicklung einer vollständigen künstlichen Intelligenz zum Ende der menschlichen Rasse führen könnte". Elon Musk selbst (dessen Tochterunternehmen Neuralink damit experimentiert, Mikrochips mit künstlicher Intelligenz in das menschliche Gehirn einzupflanzen: über Transhumanismus und Biorecht werden wir vielleicht auf einer anderen Konferenz sprechen) hat erklärt, dass künstliche Intelligenz "an sich ein Risiko für die Existenz der menschlichen Zivilisation darstellt", dass "wir sehr vorsichtig mit künstlicher Intelligenz umgehen müssen... sie ist potenziell gefährlicher als Atomwaffen" und dass "KI einer der seltenen Fälle ist, in denen es notwendig ist, proaktiv und nicht reaktiv zu handeln, da eine reaktive Regulierung auf dem Gebiet der KI zu spät kommen könnte".

Vor einigen Tagen wurde bekannt, dass der 75-jährige Geoffrey Hinton, einer der Väter der künstlichen Intelligenz, seine jahrzehntelange Zusammenarbeit mit Google6 freiwillig beendete, "um über die Gefahren der KI zu sprechen". Gegenüber der BBC sagte er: "Im Moment sehen wir, dass Dinge wie GPT-4 das Allgemeinwissen einer Person verdunkeln, und zwar um ein Vielfaches. Was das logische Denken angeht, ist es nicht so gut, aber es kann bereits einfache logische Schlussfolgerungen ziehen. Und angesichts des Tempos des Fortschritts erwarten wir, dass sich die Dinge ziemlich schnell verbessern werden. Wir müssen uns also Sorgen machen.

Die Befürchtung, die diese Denker umtreibt und die bekanntlich Anlass für die in der Washington Post veröffentlichte Unterschriftensammlung für ein Moratorium für den Einsatz künstlicher Intelligenz7 war, besteht darin, dass die Maschinen nicht beherrschbar sind und in ihrem Streben nach maximaler Effizienz bald feststellen werden, dass das schädlichste Wesen auf der Erde zweifellos der Mensch ist, und sich deshalb kaltblütig und klar dafür entscheiden, ihn zu beseitigen.

Dieses Szenario scheint der Terminator-Saga entnommen zu sein und könnte uns zum Schmunzeln bringen. In einem bekannten Experiment von Facebook aus dem Jahr 2017 haben zwei künstliche Intelligenzen, Alice und Bob, die sich miteinander unterhalten sollten, innerhalb weniger Minuten unter Ausnutzung eines Programmierfehlers die englische Sprache aufgegeben und begonnen, Nachrichten in einer dem Menschen unbekannten Neo-Sprache auszutauschen, so dass das Experiment gewaltsam abgebrochen wurde ... das muss einfach beunruhigen.

Es ist vielleicht kein Zufall, dass im US-Senat ein Vorschlag eingebracht wurde (Erstunterzeichner Markey), der die Verwendung von Bundesmitteln für den Abschuss von Atomwaffen durch autonome, nicht von Menschen kontrollierte Systeme verbieten soll.

Die größte Gefahr, die vom Einsatz künstlicher Intelligenz ausgeht, sind meiner Meinung nach jedoch andere.

Im Mai 2023 entließ Dropbox 500 Mitarbeiter im Zusammenhang mit der Einführung von künstlicher Intelligenz.

Der CEO von IBM, Arvind Krishna, kündigte an, 7.800 Neueinstellungen auszusetzen, weil die damit verbundenen Aufgaben (paradoxerweise gerade die, die mit dem Personalwesen zu tun haben) durch künstliche Intelligenz ersetzt werden sollen.

Angesichts der Tatsache, dass Maschinen viele Aufgaben ohne Ermüdungserscheinungen übernehmen und die Kosten Jahr für Jahr sinken, ist es leicht abzusehen, dass viele Berufe überflüssig werden, nicht nur die operativen und materiellen, sondern auch diejenigen, die mit Kreativität zu tun haben.

Anwendungen der künstlichen Intelligenz haben Gemälde, Kurzgeschichten und Gedichte hervorgebracht, die von Jurys, die nicht wussten, wer der "Autor" war, als originell bewertet und mit Preisen ausgezeichnet wurden. Der Gesetzgeber, auch der europäische, erlässt Gesetze zur Regulierung der "digitalen Kreativität". Um das erreichte Niveau zu erkennen, muss man sich nur die gesamte LP des AISIS-Projekts9 anhören oder den Bericht in Nature lesen, demzufolge ChatGPT in mindestens vier veröffentlichten oder vorgedruckten wissenschaftlichen Artikeln als Autor genannt wurde.

Natürlich werden, wie bei jeder technologischen Innovation, neue Berufe entstehen, und es spricht nichts dagegen, bei gleichem Lohn die Arbeitszeit der Arbeitnehmer zu verkürzen, die Arbeit von Robotern stark zu besteuern (wie es z. B. Bill Gates vorschlägt), um den Menschen noch einige Zeit auf dem Markt wettbewerbsfähig zu halten, aber wir werden bestenfalls etwas finden müssen, womit wir unsere Tage füllen können.

Meine Befürchtung ist, dass wir nicht dazu getrieben werden, unsere Zeit in philosophische Spekulationen zu investieren, in die Suche nach dem Absoluten, in die Verbesserung unserer Persönlichkeit, in den Bau idealer Städte, also in die Schaffung eines Eden auf Erden, sondern dass die Unterhaltungsindustrie uns immer mehr in fantastische und virtuelle Welten eintauchen lässt (das Phantom "Metaverse" geht in diese Richtung) und uns von der konkreten Realität entfremdet, die uns immer langweiliger erscheinen wird.

Zweitens: Je mehr wir uns auf Maschinen verlassen, die sich selbst umprogrammieren, desto mehr werden wir in Unwissenheit versinken und die Kontrolle verlieren.

Schon jetzt fragen wir uns, wozu wir Geschichte, Geografie, Mathematik, Sprachen, aber auch die Gesetzbücher studieren sollen, wenn ein Klick auf den Computer genügt, um Antworten auf unsere Fragen zu erhalten. Ohne es zu merken, legen wir bereits eine gewisse Professionalität an den Tag: Dieser Bericht von mir wurde auf Italienisch verfasst und wird, anders als das, was AGATIF seit Jahrzehnten tut, nicht in Echtzeit übersetzt, sondern wurde mit Hilfe von Anwendungen ins Französische und Deutsche übertragen (und von unseren französischen und deutschen Freunden korrigiert, denen ich an dieser Stelle danken möchte).

Lassen Sie uns über die Vorhersehbarkeit im Strafrecht nachdenken. Im Bericht 2018 habe ich ein Beispiel genannt: In meiner Heimatstadt Mailand wird mit einer Anwendung experimentiert, die während eines Raubüberfalls 14.000 Daten sammelt und verarbeitet; sie findet Zusammenhänge zwischen einem Raubüberfall und dem nächsten (Big Data/Musteranalyse) und kann vorhersagen, welche Bank oder welches Geschäft als Nächstes überfallen wird, und auch den Tag und die Uhrzeit vorhersagen, an dem dies geschehen wird. Erstaunlicherweise hat dies der Polizei ermöglicht, sich fast mit den Räubern zu verabreden und diese Art der Serienkriminalität deutlich zu verringern. Nun, wenn ein Mensch diese Tätigkeit ohne die Hilfe von Maschinen ausüben wollte, könnte er das niemals tun, da er eine so große Datenmenge nicht rechtzeitig verarbeiten könnte.

Aber betrachten wir die Suche nach dem besten Partner. Es liegt auf der Hand, dass wir, wenn wir uns auf die traditionelle Do-it-yourself-Methode verlassen, selbst wenn wir jeden Abend ausgehen, selbst wenn wir Anzeigen in den Zeitungen schalten oder uns an Heiratsvermittlungen wenden (wenn es solche Agenturen noch gibt), vielleicht jemanden treffen, der einen ähnlichen Geschmack wie wir hat, einen sozialen Hintergrund, eine Kultur, eine Religion, die mit denen kompatibel ist, die wir angegeben haben, dass wir sie mögen, und wenn wir Glück haben, wird der sprichwörtliche Funke überspringen. Aber eine der vielen Anwendungen dieser Art verspricht noch viel mehr: Wir könnten die "richtige Person" treffen, die auch am anderen Ende der Welt leben könnte, und zwar nicht anhand der Beschreibungen, die jeder Mensch von sich selbst gibt, sondern anhand der - auch unbewussten - Merkmale, die die neuen Technologien von seiner Persönlichkeit erfasst haben, anhand des Browserverlaufs, der Schlüsselwörter, die er aus der elektronischen Kommunikation entnommen hat, der Zeit, die er darauf verwendet hat, einen Beitrag in einem sozialen Netzwerk zu prüfen, was er mit "Likes" gebilligt oder missbilligt hat, was er geteilt und was er stattdessen sofort gelöscht hat.

Ein Test, den Forscher der Universität Cambridge mit mehr als 86 000 Freiwilligen durchgeführt haben (die Ergebnisse wurden 2015 in der Zeitschrift Proceedings of National Academy of Sciences veröffentlicht), hat gezeigt, dass die Software bei der Analyse von 10 Likes die Charaktereigenschaften einer Person besser vorhersagen kann als ein Arbeitskollege; bei der Analyse von 70 Likes kennt die Software uns besser als Freunde

und Mitbewohner; bei 150 Likes besser als Partner, Geschwister und Eltern. Um an Ehepartnern vorbeizukommen, waren 300 Likes nötig. Aber um 300 Likes zu erfassen, braucht man nur eine Woche... um eine Person wirklich kennen zu lernen, reicht manchmal ein ganzes Leben nicht aus.

Übertragen wir diese Überlegung auf unsere Sphäre: Der Roboter-Richter könnte in der Lage sein, Nuancen einer Aussage, Neigungen der Sprache, Details eines Fotos (um es von einer Fälschung zu unterscheiden!) in einer für den Menschen nicht im Entferntesten vorstellbaren Weise zu erfassen und aus der Schnittmenge von Tausenden von Big Data statistisch signifikante, wenn auch scheinbar unlogische, Übereinstimmungen zu erkennen.

So könnte die Richter-KI zu den effizientesten Ergebnissen kommen. So könnte sie beispielsweise feststellen, dass Straftäter, die in einem bestimmten Jahrzehnt im Stadtteil x geboren wurden, die Bibliothek in dieser und jener Straße besuchen und wegen Autodiebstahls verurteilt wurden, nach ihrer Entlassung wahrscheinlich an der Beihilfe zur illegalen Einwanderung beteiligt sind. Die Maschine weiß nicht, warum, und kann daher die Entscheidung nicht vollständig begründen, aber sie erfasst objektive Daten und liefert klare Ergebnisse.

Auch wenn dies wünschenswerte Ergebnisse zu sein scheinen, werden die hervorragenden Fähigkeiten von ChatGPT bei der Textverarbeitung, wie von vielen Autoren vorhergesagt, zu einer Vervielfachung von Plagiatsfällen führen und auch die Urheber von bösartigen Codes, Fake News, Phishing-Entwicklern und so genanntem Social Engineering werden einen sehr gefährlichen Schritt nach vorne machen.

Auch ohne dass diese Frage in der öffentlichen Debatte aufgetaucht ist, werden die Methoden der wissenschaftlichen Forschung grundlegend verändert und gehen über die galiläische Methode hinaus. Anstatt Hypothesen zu formulieren und ihre Richtigkeit experimentell zu überprüfen, versucht man, die Logik hinter bestimmten Übereinstimmungen (Mustern) zu verstehen, die von Maschinen entdeckt werden. Auf diese Weise wurden neue Impfstoffe wie der gegen Sars-CoV in Rekordzeit entdeckt.

Aber nehmen wir ein anderes Beispiel: Wenn der Gesetzgeber uns mit einem Federstrich dazu verpflichten würde, bis 2030 auf unsere Autos zu verzichten, indem er starke Anreize für die Nutzung selbstfahrender Fahrzeuge schafft und Fußgänger dazu zwingt, Ortungsgeräte zu tragen, würden die Unfälle aller Wahrscheinlichkeit nach fast auf Null sinken, vorausgesetzt, wir vertrauen voll und ganz auf die Funktionsweise dieser Maschinen und ihre Fähigkeit, ihre Algorithmen ständig zu verbessern.

Es wäre schwierig, auf solche außergewöhnlichen Ergebnisse zu verzichten, aber neben dem Verlust der Kontrolle müssten wir mit einem unkontrollierten Eingriff in unser Privatleben rechnen. Die Verhinderung von Verbrechen und die Förderung von tugendhaftem Verhalten könnten selbst im Westen schnell zu dem führen, was ich für ein

verhängnisvolles Abdriften halte, das zum Beispiel von Josh Chin und Liza Lin in ihrem Text "State of Surveillance" gut beschrieben wird. Dies gleicht Chinas Weg in eine neue Ära der sozialen Kontrolle, d. h. die Nutzung der Gesichtserkennung, um Zustände zu schaffen, die denen in Orwells "Big Brother" nicht unähnlich sind.

Und wenn wir uns Maschinen anvertrauen, setzen wir uns den Folgen möglicher Fehlfunktionen oder, schlimmer noch, des Eindringens von Hackern (rectius crackers) aus, die vor einiger Zeit bewiesen haben, dass sie aufgrund eines (später behobenen) Fehlers in der Lage waren, z. B. Tesla-Fahrzeuge anzugreifen11. Um diese Art des Eindringens zu verhindern, vor allem in Bezug auf die Bereitstellung grundlegender öffentlicher Dienstleistungen (Wasser, Gas, Strom, Verkehr usw.), wird bekanntlich mühsam ein europäisches Sicherheitsnetz12 aufgebaut, in das jedoch große Summen investiert werden müssen.

Besonderes Augenmerk muss auch auf die Güte der Daten gelegt werden, die von der künstlichen Intelligenz in der Lernphase verarbeitet werden. Wie einige Wissenschaftler gezeigt haben, sind einige Zeitreihen, z. B. über die Kreditwürdigkeit von Banken, die Zuverlässigkeit von Verträgen, den Zugang zu US-Universitäten usw., nachweislich mit (rassistischen) Verzerrungen behaftet.

Wenn Bürgern in bestimmten Gebieten Italiens in der Vergangenheit Kredite verweigert wurden, würde die künstliche Intelligenz aus historischen Aufzeichnungen eine strenge Übereinstimmung zwischen Wohnort und der als richtig erachteten Betriebswahl extrapolieren. Indem die KI den Menschen (den Bankangestellten) selbst in seiner rückständigsten und unbewusstesten Denkweise nachahmt, wird sie dazu neigen, Antragstellern aus nördlichen Städten mehr Kredite zu gewähren und sie unlogischerweise "Südländern" zu verweigern.

Noch schlimmer ist, dass sich diese logischen Verzerrungen im Laufe der Zeit wiederholen und das bekannte Phänomen der sich selbst erfüllenden Prophezeiungen auslösen.

Wenn sich die meisten Raubüberfälle in einem bestimmten schlechten Viertel der Stadt ereignen und die künstliche Intelligenz mehr Polizeiautos einsetzt, um in diesem Viertel zu patrouillieren, werden die Beamten in diesen Straßen eine beträchtliche Anzahl von Verbrechen aufdecken, aber hauptsächlich deshalb, weil diese mehr Kontrollen unterliegen. Wenn die Vorhersage eintrifft, werden noch mehr Polizeiautos in das Ghetto-Viertel geschickt, was zu einer riskanten Spirale führt.

Aber das ist noch nicht alles. Je mehr sich der Mensch vom Transzendenten befreit, je mehr er vergisst, dass er ein Geschöpf ist, desto größer wird sein Ehrgeiz, wie Gott zu werden. Der Wissenschaftler, ein neuer Prometheus, vergisst die Strafe, die der berüchtigte Demiurg in der griechischen Mythologie von Zeus erhält, und anstatt einfache Maschinen zu seinen Diensten zu schaffen, z. B. in Nachahmung von Pflanzen (sogenannte Plantoide),

strebt er danach, etwas zu erzeugen, das ihm selbst ähnlicher ist: Er konzentriert seine Bemühungen auf die humanoide Robotik und weist die Automaten an, menschliche Gesten und Reaktionen so gut wie möglich nachzuahmen. Diese Ziele, die weitgehend erreicht sind, werden die Menschen dazu bringen, sich in die Maschinen einzufühlen, für sie zu empfinden und zu vergessen, dass sie nichts anderes als fortschrittliche Haushaltsgeräte sind.

Und paradoxerweise vernachlässigt der Mensch bei der Verfolgung dieser Ziele den ganz erheblichen ökologischen Fußabdruck, den die Nutzung dieser Technologien hinterlässt (wie es auch bei den Kryptowährungen usw. der Fall ist). Das SemiAnalysis Institute hat geschätzt, dass der Betrieb von ChatGPT etwa 700.000 Dollar pro Tag kostet und dass jede Abfrage 36 Cent kostet!

Die enorme Rechenleistung, die für die Bereitstellung von Antworten und die Optimierung der Ergebnisse erforderlich ist, erfordert sehr teure Server, ganz zu schweigen davon, dass ChatGPT 3 während der Lernphase 1.287 MWh Energie verbraucht hat.

Abhilfemaßnahmen - Anwendung des Vorsorgeprinzips

Die Überlegungen zu diesem Thema enden häufig mit der Forderung nach einem Eingreifen des Gesetzgebers. Sowohl in den Vereinigten Staaten als auch in China, aber auch in Europa, gibt es Initiativen in diese Richtung.

Ich würde meinerseits auf ein Eingreifen des Gesetzgebers drängen, aber angesichts der Schäden, die durch verpfuschte Verordnungen, die leider auch von europäischen Institutionen ausgehen, verursacht werden, nur dann, wenn sie bestimmte Merkmale aufweisen.

Ich denke, eine Verordnung hat dann ihre Berechtigung, wenn eine gute Chance besteht, dass sie auch umgesetzt wird. Wir brauchen keine Normen, die nur programmatisch sind oder toter Buchstabe bleiben.

Und eine Norm wird mit größerer Wahrscheinlichkeit angewandt, wenn sie:

1) einfach, klar und in nicht zu schwerer Sprache verfasst ist;

2) gestrafft, prägnant und nicht umständlich ist;

3) frei von systematischen Antinomien und Aporien ist;

4) leicht zugänglich für die Adressaten;

5) ... und daher weithin bekannt;

6) geteilt, d.h. auf die Werteskala der Zielgruppe abgestimmt;

7) tendenziell stabil, d. h. nur geändert, wenn es unbedingt notwendig ist.

Lassen Sie uns nun über die Inhalte nachdenken, die die Regelungen zu diesem Thema beinhalten sollten.

Die verschiedenen Entwürfe, die in Betracht gezogen werden, versuchen, allgemeine Grundsätze einzuführen.

Der Blueprint for an AI Bill of Rights (Making automated systems work for the American people) vom Oktober 2022 verweist auf die Sicherheit und Effektivität von Systemen, die Verhinderung algorithmischer Diskriminierung, die Achtung der Privatsphäre, die Transparenz von Algorithmen und die Möglichkeit eines schnellen menschlichen Eingreifens.

All diese Punkte sind überzeugend, ebenso wie die Gesetze der Robotik aus Asimovs Science-Fiction-Werken, die vom Europäischen Parlament überraschenderweise zitiert wurden: "1. ein Roboter darf einem Menschen nicht schaden oder zulassen, dass ein Mensch durch seine Untätigkeit geschädigt wird; 2. ein Roboter muss den Befehlen eines Menschen gehorchen, es sei denn, dies widerspricht der ersten Regel; 3. ein Roboter muss seine Existenz beschützen, solange dieser Schutz nicht mit Regel eins oder zwei kollidiert.

Vielleicht sollten wir uns daran erinnern, dass diese Regeln, wie die Protagonisten seiner Romane (Gregory Powell und Mike Donovan) bitter feststellen mussten, selbst in Asimovs Werken leider nicht einfach auf die konkreten Fälle des Lebens anzuwenden sind und allzu leicht umgangen werden können.

Ich mache einen Vorschlag, den ich für eine direkte Anwendung des Vorsorgeprinzips halte, wie es in Artikel 191 des Vertrags über die Arbeitsweise der Europäischen Union (AEUV) und in der Auslegung des Gerichtshofs dargelegt ist (nicht auf Umweltfragen beschränkt, wie auch immer diese aussehen).

In Anbetracht der verfügbaren Daten, des Nutzens und der Belastungen, die sich aus dem Tätigwerden oder NichttTätigwerden ergeben können, und des Ausmaßes der wissenschaftlichen Ungewissheit halte ich es für angemessen, Maßnahmen zur "künstlichen Dummheit" einzuführen: Da der Mensch unüberwindbare physikalische Grenzen hat, hoffe ich, dass auch für Maschinen genaue Grenzen eingeführt werden und somit Anwendungen der künstlichen Intelligenz verboten werden:

- des RAM, des Arbeitsspeichers, über ein bestimmtes Maß hinaus;

- des Massenspeichers, über ein bestimmtes Maß hinaus;

- Prozessgeschwindigkeiten, die eine bestimmte Grenze überschreiten;

- Verbindungen mit anderen Maschinen über ein bestimmtes Maß hinaus;

und zwar genau zu dem Zweck, seine Rechenkapazität vernünftigerweise zu begrenzen.

Nur so kann meines Erachtens eine auf den Menschen ausgerichtete künstliche Intelligenz gewährleistet werden, die wirklich in den Diensten des Menschen steht und sich nicht von ihm abwendet.

Solche Maßnahmen haben auch den Vorteil, dass sie sehr einfach umzusetzen sind und im Laufe der Zeit je nach Bedarf variiert werden können, um so optimal auf die Entwicklung der Kosten-Nutzen-Analyse und damit den Grundsatz der Verhältnismäßigkeit angepasst zu werden.

Eine wichtige Konsequenz dieser Überlegung ist, dass die Beweislast für das Nichtvorhandensein von Gefahren beim Hersteller dieser Anwendungen der künstlichen Intelligenz liegt. Dies ist meines Erachtens eine sehr schwer zu erfüllende Last, die mir der beste Beweis für die Richtigkeit des skizzierten rechtlichen Ansatzes zu sein scheint.

Leider ist es bei globalen Phänomenen wie dem, mit dem wir es zu tun haben, oft so, dass entweder diese Grenzen universell eingeführt werden oder sie ihren Sinn und ihre Wirksamkeit verlieren, und das ist ein weiterer Grund, warum Vereinigungen wie die unsere, die den Dialog und die Verbreitung von Rechtsmodellen erleichtern, von größter Bedeutung sind.