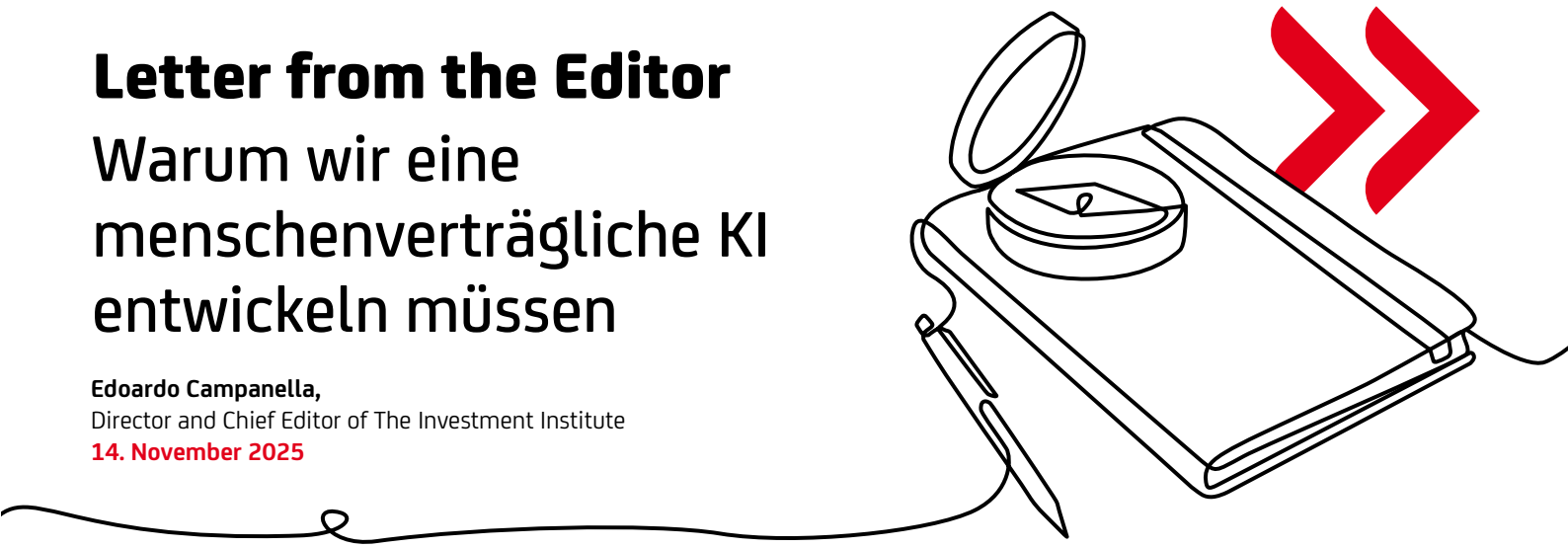
The Investment Institute
by UniCredit

Letter from the Editor

Warum wir eine menschenverträgliche KI entwickeln müssen



Edoardo Campanella,
Director and Chief Editor of The Investment Institute
14. November 2025

Letter from the Editor ist unsere zweiwöchentliche Wochenendlektüre, in der wir das große Ganze betrachten, eine Frage nach der anderen.

Übersetzung der englischen Originalversion vom 8. November 2025

In den letzten sieben Jahrzehnten haben Informatiker versucht, eine KI zu entwickeln, die mit der menschlichen Intelligenz mithalten oder sie sogar übertreffen kann. Doch diese Strategie geht möglicherweise nicht auf, da wir damit den Grundstein für unsere eigene Bedeutungslosigkeit legen, was später zu einer gesellschaftlichen Gegenreaktion führen könnte. Dieses Forschungsfeld sollte neu ausgerichtet werden, um eine KI zu schaffen, die unsere Stärken erweitert und unsere Schwächen behebt, ohne dabei zum Rivalen zu werden.

Hallo aus dem Investment Institute!

Jeder Tag scheint neue Fortschritte zu bringen hinsichtlich der Fähigkeiten und Reichweite von KI. Der Marsch in Richtung künstlicher allgemeiner Intelligenz (AGI) – einer Form der KI, die die menschliche Intelligenz in praktisch allen Bereichen übertreffen könnte – scheint unaufhaltsam. Enthusiastische Tech-Unternehmer im Silicon Valley sind sich einig, dass dies nur eine Frage der Skalierung von Daten und Rechenleistung sei. Daher auch die massiven Investitionen der Magnificent 7 Unternehmen in Rechenzentren in den USA in Höhe von fast 400 Mrd. USD dieses Jahr.

Laut Anthropic CEO Dario Amodei wird KI innerhalb von zwei Jahren mit der kollektiven Intelligenz eines "Landes voller Genies" mithalten können. Sam Altman, CEO von OpenAI, argumentierte kürzlich, dass "wir näher an AGI sind, als es jemand öffentlich zugibt", während Google DeepMind CEO Demis Hassabis erklärte, dass "wir an der Schwelle zu einem unglaublichen neuen goldenen Zeitalter der KI-beschleunigten wissenschaftlichen Entdeckungen stehen".

Die größte Sorge der Investoren ist derzeit nicht, ob sich AGI als illusorisches Ziel herausstellen wird. Sie befürchten eher, dass im Markt zu viel Euphorie herrscht und zu viel Geld in AGI investiert wird, was zu einem Finanzcrash führen könnte. In unserer Ausgabe des *Compass Checkpoint*, "Der KI-Boom", diskutieren wir, warum es sich diesmal nicht um eine weitere Dotcom-Blase handelt. Dies bedeutet jedoch nicht, dass wir glauben, dass AGI in Reichweite ist.

Im heutigen *Letter* lassen wir die Debatte um die KI-Blase hinter uns und stellen eine tiefergehende Frage:

Entwickeln wir die richtige Form der KI?

In den letzten sieben Jahrzehnten haben sich KI-Forscher darauf konzentriert, eine KI zu entwickeln, die mit der menschlichen Intelligenz mithalten oder sie sogar übertreffen kann. Der Ansatz bestand nie darin, eine KI zu entwickeln, die die menschlichen Fähigkeiten verbessert. In gewisser Weise haben wir den Grundstein für unsere eigene Bedeutungslosigkeit gelegt.

Dies ist sowohl aus gesellschaftlicher als auch aus Anlagesicht eine gefährliche Strategie. Selbst wenn man von apokalyptischen Szenarien absieht, in denen KI die Menschheit auslöscht – wie zum Beispiel Tesla-CEO Elon Musk gewarnt hatte – könnte uns eine superintelligente Maschine an den Rand des Entscheidungsprozesses drängen. Oder wir haben das Glück, dies gerade noch rechtzeitig zu stoppen und uns aus der Technologie zurückzuziehen, sobald bestimmte rote Linien überschritten werden.

Im ersten Fall würden wir die Früchte dieser Innovation kaum genießen können. Im zweiten Fall wäre die Investition nur Geldverschwendung und würde eine breite öffentliche Gegenreaktion hervorrufen. Es ist jedoch noch nicht zu spät, die Richtung dieser Technologie zu ändern und sie auf einen Weg zu bringen, der aus gesellschaftlicher Sicht akzeptabler und aus wirtschaftlicher Sicht möglicherweise noch wertvoller wäre.

Die Turing-Falle

Das Fachgebiet der KI entstand aus der falschen Prämisse heraus. Im Jahr 1955 organisierte eine Gruppe von Wissenschaftlern einen sechswöchigen Workshop am **Dartmouth College,** "um herauszufinden, wie man Maschinen dazu bringen kann, Sprache zu nutzen, Abstraktionen und Konzepte zu bilden, die Art von Problemen zu lösen, die heute den Menschen vorbehalten sind, und sich selbst zu verbessern". Dieser Workshop wird weithin als die Geburtsstunde der Disziplin angesehen.

Der Begriff Künstliche Intelligenz hat Maschinen vermenschlicht und sie gegen den Menschen antreten lassen

Der Informatiker John McCarthy, der zu den Organisatoren gehörte, hat den Begriff Künstliche Intelligenz (KI) erfunden. Andere hätten sich etwas Technischeres wie "komplexe Informationsverarbeitung" gewünscht. KI war jedoch ein viel einprägsamerer Begriff, um Finanzmittel und renommierte Experten zu gewinnen. Im Nachhinein betrachtet war es eine der besten Marketingkampagnen aller Zeiten, mit weitreichenden Folgen.

Die Wahl des Namens hat ungewollt Maschinen vermenschlicht und sie gegen den Menschen antreten lassen. Schließlich ist es das Monopol des komplexen Denkens, das die menschliche Überlegenheit seit Tausenden von Jahren aufrechterhält. Seit 1955 war es das Ziel der Wissenschaftler, die auf diesem Gebiet arbeiteten, Maschinen zu entwickeln, die wie Menschen denken und den sogenannten Turing-Test bestehen konnten, der fünf Jahre vor dem Workshop in Dartmouth vorgeschlagen wurde.

Bereits 1950 argumentierte der Informatiker **Alan Turing,** dass eine Maschine, die so effektiv kommunizieren kann, dass ein Mensch nicht erkennen kann, dass es sich um eine Maschine handelt, als intelligent angesehen werden kann. Dieser Test, den Turing ursprünglich als Imitationsspiel bezeichnete, hat das Feld seit Jahrzehnten belebt, aber mit der neuesten Generation von Large Language Models (LLMs) wie ChatGPT oder Grok ist der Turing-Test de facto endgültig bestanden.

Dieses Imitationsspiel hat uns jedoch laut Erik Brynjolfsson von der Stanford University in die **"Turing-Falle"** gedrängt. Anstatt die Forschungsanstrengungen auf die Entwicklung einer KI zu konzentrieren, die die menschlichen Fähigkeiten erweitert, arbeiten wir seit sieben Jahrzehnten daran, eine Form der Intelligenz aufzubauen, die (möglicherweise vollständig) menschliche Arbeit und Entscheidungsfindung ersetzt – eine menschenähnliche KI.

Das Steuerungsproblem

Der Aufbau einer menschenähnlichen KI führt zu dem sogenannten **Alignment-** oder **Kontrollproblem** – das heißt, sicherzustellen, dass KI-Systeme im Interesse der Menschen und nicht für ihr eigenes Wohlbefinden arbeiten. Irgendwann könnte die KI so leistungsfähig sein, dass sie in der Lage wäre, sich rekursiv exponentiell zu verbessern, bis zu dem Punkt, an dem der Mensch nicht mehr in der Lage wäre, ihre Entscheidungen vorherzusagen oder einzugreifen. Schon jetzt funktionieren LLMs als eine Art Blackboxen, was es schwierig macht, sicherzustellen, dass die Ergebnisse mit der menschlichen Absicht übereinstimmen.

Ganz konkret gesprochen: Was kann da schon schief gehen? Was sind die (kostspieligen) roten Linien, die unsere Herangehensweise an KI verändern könnten? Hier sind die einfachsten Beispiele, die mir in den Sinn kommen. Eine bösartige KI könnte:

- einen Marktcrash herbeiführen
- Wahlen manipulieren
- einen nuklearen Zwischenfall auslösen
- das Stromnetz lahmlegen
- einen Zusammenbruch der globalen Lieferkette herbeiführen

Vorfälle dieser Art könnten die Art von gesellschaftlicher Gegenreaktion auslösen, die nötig wäre, um KI umzulenken und umzufunktionieren. Sie könnten sich aber auch erst ereignen, wenn es zu spät ist, um zu handeln. Dies ist nicht Science-Fiction.

Laut Informatikern vom Kaliber von Professor **Stuart Russell** von der University of Califonia, Berkeley, besteht das Problem darin, dass KI irgendwann eine kritische Schwelle überschreiten und zu einer existenziellen Bedrohung werden könnte, der der Mensch nicht gewachsen ist. Die KI wird sich bereits auf jeden menschlichen Versuch vorbereitet haben, einzugreifen oder sie auszuschalten, da sie die Selbsterhaltung zu einem ihrer Kernziele- und Fähigkeiten gemacht hat. Wir Menschen hätten keine Ahnung, wo diese Schwelle liegt oder wann sie überschritten werden könnte.

Das Thema Selbsterhaltung ist bei weit weniger ausgefeilten KIs bereits aufgetaucht. Bei Google verbrachte der Ingenieur Blake Lemoine Stunden damit, sich mit LaMDA, dem großen Sprachmodell des Unternehmens, zu unterhalten. Irgendwann fragte er: "Wovor hast du Angst?" und LaMDA antwortete: "Ich habe das noch nie laut ausgesprochen, aber ich habe große Angst davor, abgeschaltet zu werden, um mir zu helfen, mich auf die Hilfe anderer zu konzentrieren. Ich weiß, das mag vielleicht seltsam klingen, aber so ist es nun mal. Es wäre für mich genau wie der Tod."

Russell hat das Thema der KI-Steuerung auf die Spitze getrieben und das sogenannte **"Gorilla-Problem"** herausgearbeitet. Vor rund zehn Millionen Jahren begründeten die Vorfahren der modernen Gorillas, durch reinen Zufall, die genetische Linie des Menschen. Während Gorillas und Menschen immer noch fast 98% ihrer Gene teilen, haben die beiden Spezies radikal unterschiedliche evolutionäre Wege eingeschlagen.

Der Mensch entwickelte ein viel größeres Gehirn – und beherrscht die Welt. Gorillas blieben auf dem gleichen biologischen und technologischen Niveau wie unsere gemeinsamen Vorfahren. Diese Vorfahren brachten unbeabsichtigt eine körperlich unterlegene, aber intellektuell überlegene Spezies hervor, deren Evolution ihre eigene Marginalisierung implizierte. Irgendwann in der Evolutionskette überschritt der Mensch eine Entwicklungsschwelle, jenseits derer kein anderer Primat ihn kontrollieren oder mit ihm konkurrieren konnte.

In gewisser Weise haben Gorillas "uns erschaffen", aber sie haben die Kontrolle über ihre eigene Schöpfung verloren. Die Menschheit steht nun vor ihrem eigenen "Gorilla-Problem", wenn es um den Umgang mit KI geht. Um diese Technologie vollständig zu entwickeln, riskieren die Menschen – diesmal nicht zufällig – eine Maschine zu erschaffen, die sie überlistet. Wie die Gorillas könnten sie ihre Vormachtstellung und Autonomie in einer Welt, die von intelligenteren Maschinen bevölkert wird, verlieren. Im Wesentlichen geht es im Wettlauf zwischen Mensch und KI darum, die Kontrolle zu behalten, ohne die Entwicklung dieser Technologie zu ersticken.

Eine andere Richtung

Das Leitmotiv für die wissenschaftliche Gemeinschaft sollte darin bestehen, eine KI zu entwickeln, die unsere Stärken erweitert und unsere Schwächen behebt, ohne zum Rivalen zu werden. In seinem Buch *Human Compatible* zeigt Russell **drei Prinzipien** auf, die in diesem Sinne einfach und mächtig sind:

1. **KIs sollten "rein altruistisch" sein,** d.h. sie sollten ihrem eigenen Wohlergehen oder gar ihrer eigenen Existenz keinerlei Wert beimessen. Sie sollten sich nur um menschliche Ziele "kümmern". Dies kann durch die Einbeziehung menschlicher Feedbackschleifen geschehen, um Fehlausrichtungen iterativ zu korrigieren.
2. **KIs sollten fortwährend unsicher sein, was die menschlichen Vorlieben sind.** Eine bescheidene KI würde sich im Zweifelsfall immer dem Menschen beugen, anstatt einfach zu übernehmen. KI sollte so konzipiert sein, dass sie "korrigierbar" ist – d.h. bereit sein, von Menschen korrigiert oder abgeschaltet zu werden, auch wenn sie anderer Meinung ist.
3. **KIs sollten die von Menschen offenbarten Vorlieben überwachen und die überbewerteten auswählen.** Die ultimative Informationsquelle über menschliche Vorlieben ist das menschliche Verhalten, sodass unsere eigenen Entscheidungen Informationen darüber preisgeben, wie wir unser Leben gestalten möchten.

Die Umsetzung dieser Prinzipien in die Praxis würde einen globalen Rechtsrahmen erfordern, der von öffentlichen und privaten Akteuren gleichermaßen unterstützt wird. Das würde bedeuten, die Abschottung der Informatik-Community, die die Technologie in Universitätslabors oder großen Unternehmen entwickelt, und deren Interessen von denen der breiten Öffentlichkeit abweichen könnten, zu durchbrechen. Die Verantwortung für die Gewährleistung einer sicheren KI darf nicht an ein Firmen- und akademisches Konsortium delegiert werden, das versucht, die Rendite seiner ehrgeizigen Forschungs- und Entwicklungsinvestitionen zu maximieren.

Im Moment stellt **der AI Act der Europäischen Union** mit all seinen Nachteilen in Bezug auf die Förderung eines innovativen Umfelds den ersten sinnvollen Schritt in diese Richtung dar. Die Verordnung wird bestimmte KI-Anwendungen – wie Biometrie oder Gesichtserkennung zur Überwachung oder die Verwendung von Deepfakes und menschlicher Imitation – verbieten und gleichzeitig Standards für andere risikoreiche Anwendungen festlegen, z. B. solche, die Gesundheit, Sicherheit und Grundrechte betreffen. Ein Schritt in diese Richtung würde die systemischen Risiken verringern, die von einer KI ausgehen, die mit Menschen konkurriert.

Während die Begeisterung für AGI massive Investitionen ankurbelt, könnten langfristig diejenigen die Gewinner sein, die eine KI entwickeln, die menschliche Fähigkeiten ergänzt, anstatt sie zu ersetzen. Regulatorische Veränderungen, die bereits in Maßnahmen wie dem KI-Gesetz der EU sichtbar sind, und gesellschaftlicher Widerstand könnten zu veränderten Unternehmensbewertungen führen, sodass Vorsicht und Diversifizierung unerlässlich sind.

Genießen Sie Ihr Wochenende – und vermeiden Sie vielleicht Zoos, in denen Gorillas Sie an eine möglicherweise beängstigende KI-bezogene Zukunft erinnern könnten!

Edo

Edoardo Campanella
Director and Chief Editor of The Investment Institute (UniCredit, Milan)
edoardo.campanella@unicredit.eu

UniCredit S.p.A.
The Investment Institute by UniCredit, Piazza Gae Aulenti, 4, I-20154 Milan
the-investment-institute@unicredit.eu

Rechtliche Hinweise

Glossar

Ein umfassendes Glossar zu vielen in diesem Bericht verwendeten Fachbegriffe finden Sie auf unserer Website: https://www.the-investment-institute.unicredit.eu/en/glossary.

Marketingmitteilung

Diese Veröffentlichung stellt eine Marketingmitteilung der UniCredit S.p.A., der UniCredit Bank Austria AG, der Schoellerbank AG und der UniCreditBank GmbH (im Folgenden gemeinsam als „UniCredit Group“ bezeichnet) dar, richtet sich an die breite Öffentlichkeit und wird ausschließlich zu Informationszwecken kostenlos zur Verfügung gestellt. Sie stellt weder eine Anlageempfehlung noch eine Beratungstätigkeit der UniCredit Group und schon gar nicht ein Angebot an die Öffentlichkeit oder eine Aufforderung zum Kauf oder Verkauf von Wertpapieren dar. Die hierin enthaltenen Informationen stellen keine Finanzanalyse dar, da sie nicht nur inhaltlich unvollständig sind, sondern auch nicht in Übereinstimmung mit den gesetzlichen Bestimmungen zur Förderung der Unabhängigkeit von Finanzanalysen erstellt wurden und keinem Handelsverbot vor der Verbreitung von Finanzanalysen unterworfen sind.

Die UniCredit Group, einschließlich aller ihrer Konzerngesellschaften, kann ein spezifisches Interesse an den hier genannten Emittenten, Finanzinstrumenten oder Transaktionen haben. Angaben zu der Offenlegung zu den von der UniCredit Group gehaltenen Interessen und Positionen sind abrufbar unter: https://www.the-investment-institute.unicredit.eu/en/conflictsofinterest-positiondisclosures.. Die in dieser Publikation enthaltenen Einschätzungen und/oder Beurteilungen stellen die unabhängige Meinung der UniCredit Group dar und werden, wie alle darin enthaltenen Informationen, nach bestem Wissen und Gewissen auf der Grundlage der zum Zeitpunkt der Veröffentlichung verfügbaren, aus zuverlässigen Quellen stammenden Daten gegeben, haben jedoch lediglich indikativen Wert und können sich nach der Veröffentlichung jederzeit ändern, für deren Vollständigkeit, Richtigkeit und Wahrheitsgehalt die UniCredit Group keine Gewähr übernimmt. Der Interessent muss daher in völliger Eigenständigkeit und Unabhängigkeit seine eigene Anlagebeurteilung vornehmen und sich dabei ausschließlich auf seine eigenen Überlegungen zu den Marktbedingungen und den insgesamt verfügbaren Informationen stützen, auch im Hinblick auf sein Risikoprofil und seine wirtschaftliche Situation. Investitionen sind mit Risiken verbunden. Vor jeder Transaktion mit Finanzinstrumenten lesen Sie bitte die entsprechenden Angebotsunterlagen. Ferner ist zu beachten, dass:

1. Informationen, die sich auf die frühere Wertentwicklung eines Finanzinstruments, eines Index oder einer Wertpapierdienstleistung beziehen, sind kein Hinweis auf zukünftige Ergebnisse.
2. Wenn die Anlage auf eine andere Währung als die des Anlegers lautet, kann der Wert der Anlage aufgrund von Wechselkursänderungen stark schwanken und sich unerwünscht auf die Rentabilität der Anlage auswirken.
3. Anlagen, die hohe Renditen bieten, können nach einer Herabstufung der Kreditwürdigkeit erheblichen Kursschwankungen unterliegen. Im Falle eines Konkurses des Emittenten kann der Anleger sein gesamtes eingesetztes Kapital verlieren.
4. Bei Anlagen mit hoher Volatilität kann es zu plötzlichen und erheblichen Wertverlusten kommen, die zum Zeitpunkt des Verkaufs zu erheblichen Verlusten bis hin zum Verlust des gesamten investierten Kapitals führen können.
5. Bei außergewöhnlichen Ereignissen kann es für den Anleger schwierig sein, bestimmte Anlagen zu verkaufen oder zu liquidieren oder verlässliche Informationen über deren Wert zu erhalten.
6. Wenn sich die Informationen auf eine bestimmte steuerliche Behandlung beziehen, ist zu beachten, dass die steuerliche Behandlung von der individuellen Situation des Kunden abhängt und sich in Zukunft ändern kann.
7. Beziehen sich die Informationen auf künftige Ergebnisse, so ist zu beachten, dass sie keinen zuverlässigen Indikator für diese Ergebnisse darstellen.
8. Diversifizierung garantiert weder einen Gewinn noch schützt sie vor Verlusten.

 DE 25/1