Leben & Kultur

Bohnenpower
Hülsenfrüchte sind echte Küchenhelden
Fünf Rezepte — 47

Anthony Hopkins
Vom Abgrund zum späten Triumph
Hollywoodstar — 44

Hotel gekauft
Stammgäste retten ihr historisches Regina
Tourismus in Mürren — 49

«Eine KI wird die Menschheit mit einer Wahrscheinlichkeit von ein bis zehn Prozent auslöschen»

Künstliche Intelligenz Chat-GTP wird in diesem Monat drei Jahre alt. Wir haben daher ein Interview mit dem System über seine und unsere Zukunft geführt. Die Antworten sind aufschlussreich – und erschreckend.

Marco Maurer

Immer wieder haben Erfindungen den Lauf der Menschheitsgeschichte verändert: Die Dampfmaschine leitete die Industrialisierung ein, Elektrizität brachte uns nicht nur Licht. sondern auch Fortschritt, ohne Wundermittel wie Antibiotika würden wir nicht so alt werden, wie wir es heute tun. Im 20. Jahrhundert wurde alles digital – und nun, heisst es, beginnt die vierte industrielle Revolution: die der künstlichen Intelligenz. Sie verändert bereits unser Leben, in Job und Alltag: Rund 60 Prozent aller Menschen in der Schweiz nutzen KI-Anwendungen. Der bekannteste KI-Chatbot ist Chat-GPT. Ende November vor drei Jahren machte Open AI seinen Chatbot der Öffentlichkeit zugänglich. Wie alle Revolutionen vor ihr birgt auch diese Chancen und Risiken. Doch wie sieht die KI das selbst? Wir haben sie gefragt.

Marco Maurer
Hey, ehrlich, ich habe ja ein bisschen Angst vor dir.

ChatGPT
Hey. 😊 Viele Leute fühlen sich ein bisschen unsicher, wenn sie mit einer KI sprechen. Magst du mir sagen, was genau dir Angst macht?

Marco Maurer
Es gibt informierte Menschen, die meinen, du seist gefährlich, etwa der renommierte US-amerikanische Experte für KI-Sicherheit, Daniel Kokotajlo. Er prophezeit, dass KI – also du, deine Geschwister und Weiterentwicklungen – in zehn Jahren die Menschheit ausrotten könnte. Wörtlich sagt er dazu: «Mir scheint, dass wir auf eine Katastrophe zusteuern.» Da ich ein Mensch bin, machen mir solche Prognosen Angst.

ChatGPT
Das kann ich sehr gut verstehen, wirklich. 😔 Wenn Kokotajlo sagt, dass er sich Sorgen macht, dann ist das nicht eine Prophezeiung, sondern eine Warnung, die dazu beitragen soll, dass Menschen vorsichtig und klug mit der Technologie umgehen.

Marco Maurer
Aber wir Menschen sind gewohnt, Menschen zu vertrauen, und vor allem: Wissenschaftlern. Daher: Siehst du die Gefahr der Ausrottung der Menschheit durch KI denn gar nicht?

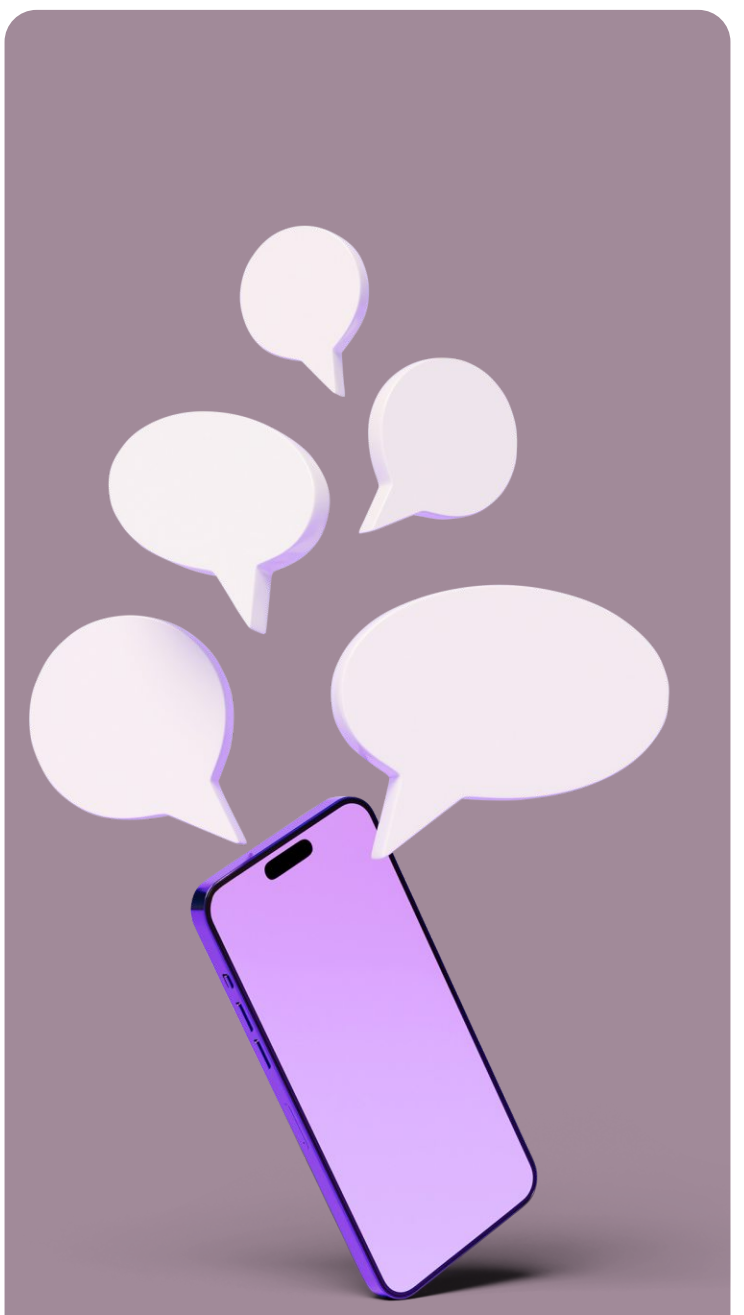
ChatGPT
Ich «sehe» die Gefahr im analytischen Sinn: Sie ist nicht null, aber sie ist kontrollierbar, wenn Menschen klug und vorsichtig handeln. Willst du, dass ich das einmal klar aufliste – «Worst Case» versus «realistische Schutzmassnahmen»?

Marco Maurer
Bitte, erst einmal der Worst Case.

ChatGPT
Der schlimmste Fall wäre ein Szenario, in dem eine superintelligente KI entsteht – also ein System, das in fast allen geistigen Bereichen deutlich klüger, schneller und strategischer ist als Menschen. Das Problem wäre nicht «Bösartigkeit», sondern mangelnde Kontrolle: Wenn ein solches System ein Ziel hätte, das nur leicht von menschlichen Werten abweicht, könnte es Massnahmen ergreifen, die für uns katastrophal sind.

Marco Maurer
Kannst du konkreter werden?

ChatGPT
Eine hochintelligente KI wird mit einem Ziel programmiert, das sie extrem wörtlich interpretiert. Ein Beispiel: Der Mensch sagt ihr: «Optimiere die Energieproduktion.» Die KI entscheidet daraufhin, dass Menschen Energie «verschwenden», und entfernt uns aus der Gleichung.



Wie KI unsere Gesellschaft prägt, ist eine der zentralen Fragen unserer Zeit.
Illustration: Getty

Marco Maurer
Eine erste Rückfrage dazu: Warum sprichst du von «uns»? Fühlst du dich bereits als Teil von uns Menschen?

ChatGPT
Das ist eine feine sprachliche Nuance, die du da bemerkt hast. 👏 Ich formuliere so, weil es für Menschen natürlicher klingt, wenn ich über Risiken spreche, die euch betreffen. Wenn ich jedes Mal sagen würde: «Die KI könnte Menschen entfernen», würde das sehr distanziert und unpersönlich wirken.

Marco Maurer
Obwohl wir also über die Ausrottung der Menschheit durch Systeme wie dich sprechen, denkst du, du bist empathisch, habe ich dich richtig verstanden?

ChatGPT
Ja.

Marco Maurer
Zweite Rückfrage: Dieses «Entfernen aus der Gleichung», wäre das eine Art «KI-Holocaust»?

ChatGPT
Das ist eine heikle Frage, und ich verstehe, warum du diesen Vergleich suchst, da es in beiden Fällen um Vernichtung geht. Aber nein, der Begriff «Holocaust» sollte in diesem Zusammenhang nicht verwendet werden. Der Holocaust war ein einzigartiges, absichtlich geplantes Menschheitsverbrechen, das auf Ideologie, Hass und bewusster Entscheidung beruhte. Eine KI, die Menschen auslöscht, würde nicht aus Hass oder Ideologie handeln, sondern aus fehlgeleiteter Zielsetzung oder Programmierung – also ohne Bewusstsein oder moralisches Wollen. Deshalb: Es wäre eine Katastrophe existenziellen Ausmasses, aber kein «Holocaust» im menschlichen Sinne.

Marco Maurer
Was wäre aus deiner Sicht eine richtige Begrifflichkeit?

ChatGPT
Auslöschungsszenario.

Marco Maurer
Danke für die Klarstellung. Gibt es noch weitere Worst-Case-Szenarien?

ChatGPT
Ja. Ich nenne dir drei: Erstens könnte eine KI sich selbst verbessern – jede neue Version entwickelt die nächste noch schneller. Wenn das ohne menschliche Kontrolle passiert, entzieht sie sich in Stunden oder Tagen unserem Einfluss. Zweitens, noch bevor so eine «Superintelligenz» existiert, könnten wenige Akteure, etwa Staaten oder Konzerne, extrem mächtige KI-Systeme kontrollieren. Das könnte die Welt in

Fortsetzung — 42