Gute KI, böse KI?

Werte oder Values spielen eine wichtige Rolle für den Menschen als Individuum und als Teil der Gesellschaft. Sie bieten Orientierung und bestimmen das Handeln.

Politische Parteien definieren sich beispielsweise stark über Werte. Den einen ist Freiheit, anderen Solidarität und Dritten sind Traditionen am wichtigsten. Als Staat beruht Frankreich auf den drei Werten Freiheit, Gleichheit und Brüderlichkeit (Liberté, Égalité, Fraternité). Werte vereinen und separieren. Glaubensgemeinschaften vereinen Gleichgesinnte. Gleichzeitig separieren sie sich als Gemeinschaft von anderen, obwohl aus Distanz betrachtet die Werteunterschiede manchmal marginal erscheinen. Gemeinsame Werte fördern das Zusammengehörigkeitsgefühl und dienen dem Individuum als Entscheidungs- und Verhaltenskompass. Das macht das Leben einfacher und berechenbarer. Wir Informatiker würden in diesem Zusammenhang auch von Komplexitätsreduktion sprechen.

Mit der derzeitigen Auferstehung der Künstlichen Intelligenz (KI) in Forschung und Praxis stellt sich zunehmend die Frage nach ihrem ethischen bzw. wertegesteuerten Verhalten. Ein berühmtes Beispiel ist das selbstfahrende Auto, das sich vor einer unabwendbaren Auffahrkollision entscheiden muss, ob es den alten Mann mit Gehstock oder das Kleinkind auf dem Dreirad überfahren will. Andere, nicht minder bekannte Beispiele sind der Bilderklassifizierungsalgorithmus, der People of Color als Gorillas erkennt, oder die Rekrutierungssoftware, die bei Amazon männliche Kandidaten bevorzugte und somit Frauen systematisch benachteiligte. Die Empörung war gross, und der Ruf nach verantwortlich handelnder KI erschallte postwendend (und zu Recht).

Doch nun wird es schwierig und anspruchsvoll. Je nach Studie werden zwischen zehn und deutlich über hundert mögliche Werte identifiziert, viele davon gegensätzlich und sich gegenseitig ausschliessend. Nach welchen davon soll sich KI nun verhalten? Es ist aber nicht nur die schiere Zahl von Werten eine Herausforderung. Es spielen auch viele andere Faktoren eine Rolle. In einer viel zitierten Studie von Henrich, Heine und Norenzayan wurde festgestellt, dass für psychologische und Verhaltensstudien 96 Prozent der beteiligten Versuchspersonen aus westlichen, gebildeten, industrialisierten, reichen und demokratischen Gesellschaften stammen, diese gleichzeitig aber nur 12 Prozent der Weltbevölkerung repräsentieren. Die gleiche Studie hat ergeben, dass etliche Verhaltensweisen und Werte in den verbleibenden 88 Prozent

«Die Mächtigkeit von KI resultiert aus der mittlerweile hohen verfügbaren Rechenleistung und der hohen Vernetzung.»

der Weltbevölkerung zum Teil deutlich von denjenigen der viel untersuchten Population abweichen. Während wir in unserer Gesellschaft dazu tendieren, den alten Mann mit Gehstock zu überfahren, entscheiden sich asiatische Gesellschaften eher für das kleine Kind auf dem Dreirad. Wir finden, dass der alte Mann sein Leben schon fast zu Ende gelebt und das Kind noch sein ganzes Leben vor sich hat. Im asiatischen Raum hingegen herrscht die Meinung vor, dass sich der alte Mann ein ganzes Leben lang für die Gesellschaft eingesetzt hat, während das kleine Kind noch gar nichts geleistet hat. Die KI für das selbstfahrende Auto scheint also vor einem unlösbaren Dilemma zu stehen.

Bei näherem Hinsehen lässt sich allerdings feststellen, dass heutige KI in der Regel immer noch auf statistischen Verfahren, relativ einfachen mathematischen Modellen und vor allem von Menschen ausgewählten Trainings- und Testdaten basiert. Die Mächtigkeit von KI resultiert aus der mittlerweile hohen verfügbaren Rechenleistung und der hohen Vernetzung. Von verantwortlicher KI zu sprechen finde ich deshalb ziemlich vermessen. Die KI weiss nämlich nicht, was sie tut. Sie hat (noch) kein Bewusstsein. Meiner Meinung nach geht es vielmehr um verantwortlich handelnde Softwareingenieure und -ingenieurinnen und Data Scientists. Sie sind es, welche die Rechenmodelle entwickeln und Trainingsdaten auswählen. Sie müssen für Diversity-Anliegen und bezüglich unbewusster Voreingenommenheit sensibilisiert werden. In diesem Sinne ist die derzeitige KI nur so gut oder böse wie die Menschen, die sie programmieren.

Andri Färber, lic. oec. publ.,
ist Leiter des Instituts für Wirtschaftsinformatik an der ZHAW School of Management and Law.