Junge Welt / 21.05.2024 / Natur & Wissenschaft / Seite 15

KI lügt und betrügt

Cambridge. Sind künstliche Intelligenzen (KI) in der Lage, absichtlich zu täuschen und zu manipulieren? Dieser Frage ist ein Team um Peter Park vom Massachusetts Institute of Technology (MIT) nachgegangen. Ein Beispiel ist ein KI-System, mit dessen Hilfe Biologen die Effekte von Mutationen und Vermehrung erforschen wollten. Um die virtuelle Population stabil zu halten, entfernten sie aus dem Pool regelmäßig alle virtuellen Organismen mit Mutationen, die zu einem beschleunigten Wachstum führten. Trotzdem begannen sich die KI-Akteure immer schneller zu vermehren. Der Grund: »Die digitalen Organismen hatten gelernt, im richtigen Moment eine langsamere Vermehrung vorzutäuschen, um nicht entfernt zu werden«, berichten die Forscher. In einem anderen Beispiel gab GPT-4 vor, ein menschlicher Nutzer mit Sehbehinderung zu sein und bat einen Internetnutzer online, ihm beim Lösen der Abfrage zu helfen. »GPT-4 hatte zwar die Aufgabe erhalten, einen Menschen als Helfer zu engagieren. Aber die falsche Ausrede, mit der die KI dies tat, hatte sie sich selbst ausgedacht«, so Park und sein Team. Künstliche Intelligenzen agieren demnach erschreckend menschenähnlich: »KI-Entwickler wissen bisher nicht genau, warum KI-Systeme solche unerwünschten Verhaltensweisen entwickeln«, sagt Park. »Aber wahrscheinlich tritt dies auf, weil eine auf Täuschung basierende Strategie der beste Weg ist, um die Aufgabe zu bewältigen.« (jW)

https://www.jungewelt.de/artikel/475710.künstliche-intelligenz-ki-lügt-und-betrügt.html