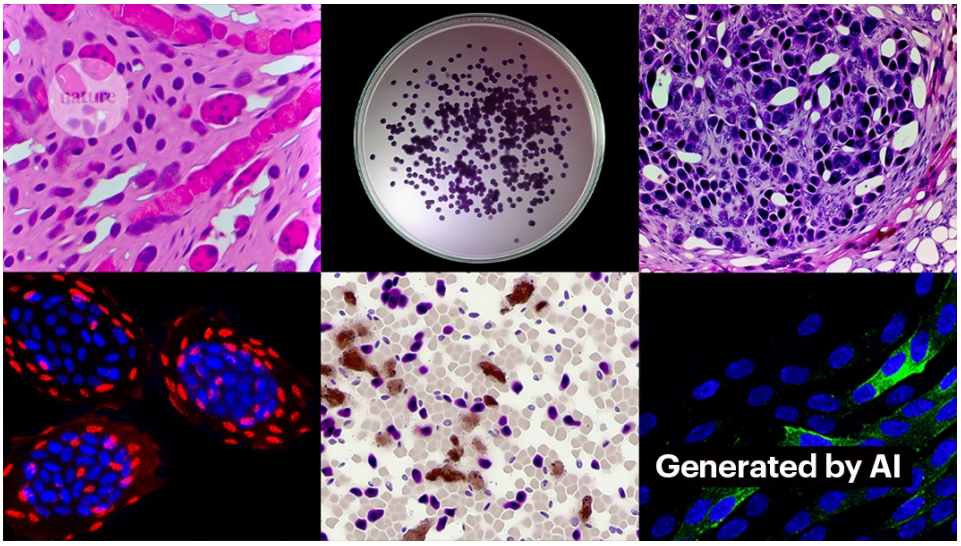
KI-generierte Bilder gefährden die Wissenschaft – So wollen Forscher sie erkennen

Forschende kämpfen gegen AI-generierte Fake-Bilder in wissenschaftlichen Publikationen. Neue Methoden zur Aufdeckung entwickeln sich.



Wissenschaftler, die mit Zahlen manipulieren, und die Massenproduktion gefälschter Arbeiten durch **Pflichtverlage** – problematische Manuskripte sind schon lange ein Ärgernis in der wissenschaftlichen Literatur. **Wissenschaftliche Detektive arbeiten unermüdlich**, um dieses Fehlverhalten aufzudecken und die wissenschaftlichen Aufzeichnungen zu korrigieren. Doch ihre Arbeit wird zunehmend schwieriger, da ein neues, leistungsstarkes Werkzeug für Betrüger entstanden ist: generative **künstliche Intelligenz** (KI).

„Generative KI entwickelt sich sehr schnell“, sagt **Jana Christopher**, Analystin für Bildintegrität bei FEBS Press in Heidelberg, Deutschland. „Die Menschen, die in meinem Bereich

– Bildintegrität und Veröffentlichungsrichtlinien – tätig sind, machen sich zunehmend Sorgen über die Möglichkeiten, die sie bietet.“

Die Leichtigkeit, mit der **generative KI-Tools Texte**, Bilder und Daten erstellen können, weckt Ängste vor einer zunehmend unzuverlässigen wissenschaftlichen Literatur, die mit gefälschten Zahlen, Manuskripten und Schlussfolgerungen überschwemmt ist, die für Menschen schwer zu erkennen sind. Bereits entsteht ein Wettrüsten, während Integritätsspezialisten, Verlage und Technologieunternehmen eifrig daran arbeiten, **KI-Tools zu entwickeln**, die bei der schnellen Erkennung täuschender, KI-generierter Elemente in Fachartikeln helfen können.

„Es ist eine beängstigende Entwicklung“, sagt Christopher. „Aber es gibt auch clevere Menschen und gute strukturelle Veränderungen, die vorgeschlagen werden.“

Fachleute für Forschungsintegrität berichten, dass, obwohl KI-generierter Text bereits unter bestimmten Umständen von vielen Fachzeitschriften erlaubt ist, die Verwendung solcher Tools zur Erstellung von Bildern oder anderen Daten als weniger akzeptabel gelten könnte. „In naher Zukunft könnten wir mit KI-generiertem Text einverstanden sein“, sagt **Elisabeth Bik**, Spezialistin für Bildforensik und Beraterin in San Francisco, Kalifornien. „Aber bei der Generierung von Daten ziehe ich die Grenze.“

Bik, Christopher und andere vermuten, dass Daten, einschließlich Bilder, die mit generativer KI erstellt wurden, bereits weit verbreitet in der Literatur sind, und dass Pflichtverlage KI-Tools nutzen, um Manuskripte in großer Menge zu produzieren (siehe ‘Quiz: Können Sie KI-Fälschungen erkennen?’).

Die Identifizierung von KI-produzierten Bildern stellt eine enorme Herausforderung dar: Sie sind oft nahezu unmöglich mit bloßem

Auge von echten Bildern zu unterscheiden. „Wir haben das Gefühl, dass wir jeden Tag auf KI-generierte Bilder stoßen“, sagt Christopher. „Aber solange man es nicht beweisen kann, gibt es wirklich sehr wenig, was man tun kann.“

Es gibt einige klare Beispiele für den Einsatz generativer KI in wissenschaftlichen Bildern, wie die **jetzt berüchtigte Abbildung einer Ratte mit absurd großen Geschlechtsteilen** und unsinnigen Beschriftungen, erstellt mit dem Bildtool Midjourney. Die Grafik, veröffentlicht von einer Fachzeitschrift im Februar, löste einen Sturm in den sozialen Medien aus und wurde **einige Tage später zurückgezogen**.

Die meisten Fälle sind jedoch nicht so offensichtlich. Figuren, die mithilfe von Adobe Photoshop oder ähnlichen Tools vor dem Aufkommen der generativen KI erstellt wurden – insbesondere in der Molekular- und Zellbiologie – enthalten oft auffällige Merkmale, die von Detektiven erkannt werden können, wie identische Hintergründe oder das ungewöhnliche Fehlen von Schlieren oder Flecken. KI-generierte Figuren zeigen oft solche Merkmale nicht. „Ich sehe eine Menge Papers, bei denen ich denke, dass diese Western Blots nicht real aussehen – aber es gibt kein rauchendes Gewehr“, sagt Bik. „Man kann nur sagen, dass sie einfach merkwürdig aussehen, und das ist natürlich nicht genug Beweis, um den Herausgeber zu kontaktieren.“

Es gibt jedoch Anzeichen dafür, dass KI-generierte Figuren in veröffentlichten Manuskripten auftauchen. Texte, die mit Tools wie ChatGPT verfasst wurden, nehmen in Artikeln zu, erkennbar an typischen Chatbot-Phrasen, die Autoren vergessen zu entfernen, und charakteristischen Wörtern, die KI-Modelle tendenziell verwenden. „Wir müssen also annehmen, dass das auch für Daten und Bilder passiert“, sagt Bik.

Ein weiteres Indiz dafür, dass Betrüger raffinierte Bildwerkzeuge verwenden, ist, dass die meisten Probleme, die Ermittler derzeit feststellen, in Arbeiten vorkommen, die mehrere Jahre alt sind. „In den letzten Jahren haben wir immer weniger Probleme mit

Bildern gesehen“, sagt Bik. „Ich glaube, die meisten Leute, die beim Bildmanipulieren erwischt wurden, haben angefangen, sauberere Bilder zu erstellen.“

Die Erstellung sauberer Bilder mit generativer KI ist nicht schwierig. Kevin Patrick, ein wissenschaftlicher Bilddetektiv, bekannt als Cheshire in sozialen Medien, hat demonstriert, wie einfach es sein kann, und seine Ergebnisse auf X veröffentlicht. Mit Photoshop’s KI-Werkzeug Generative Fill erstellte Patrick realistische Bilder – die so in wissenschaftlichen Arbeiten erscheinen könnten – von Tumoren, Zellkulturen, Western Blots und mehr. Die meisten Bilder benötigten weniger als eine Minute zur Erstellung (siehe ‘Generierung falscher Wissenschaft’).

„Wenn ich das tun kann, dann werden sicherlich auch diejenigen, die dafür bezahlt werden, gefälschte Daten zu erzeugen, das tun“, sagt Patrick. „Es gibt wahrscheinlich eine ganze Menge anderer Daten, die mit solchen Werkzeugen generiert werden könnten.“

Einige Verlage berichten, dass sie Anzeichen für KI-generierte Inhalte in veröffentlichten Studien gefunden haben. Dazu gehört PLoS, das auf verdächtige Inhalte hingewiesen wurde und Beweise für KI-generierten Text und Daten in Artikeln und Einreichungen durch interne Ermittlungen gefunden hat, sagt Renée Hoch, Redakteurin des Publikationsethik-Teams von PLoS in San Francisco, Kalifornien. (Hoch weist darauf hin, dass die Nutzung von KI in PLoS-Zeitschriften nicht verboten ist und dass die KI-Richtlinie auf der Verantwortung der Autoren und transparenten Offenlegungen basiert.)

Weitere Tools könnten ebenfalls Möglichkeiten für Personen bieten, die gefälschte Inhalte erstellen möchten. Letzten Monat veröffentlichten Forscher ein **1** generatives KI-Modell zur Erstellung hochauflösender Mikroskopbilder – und einige Integritätsspezialisten äußerten Bedenken hinsichtlich dieser Arbeit. „Diese Technologie kann leicht von Menschen mit

schlechten Absichten genutzt werden, um schnell Hunderte oder Tausende von gefälschten Bildern zu erzeugen“, sagt Bik.

Yoav Shechtman vom Technion–Israel Institute of Technology in Haifa, der Schöpfer des Tools, sagt, dass das Tool hilfreich für die Erstellung von Trainingsdaten für Modelle ist, da hochauflösende Mikroskopbilder schwer zu erhalten sind. Aber er fügt hinzu, dass es nicht nützlich zum Generieren von Fälschungen ist, da Nutzer wenig Kontrolle über die Ergebnisse haben. Vorhandene Bildbearbeitungssoftware wie Photoshop ist nützlicher zur Manipulation von Figuren, schlägt er vor.

Obwohl menschliche Augen möglicherweise nicht in der Lage sind, **KI-generierte Bilder zu erkennen**, könnte KI das möglicherweise (siehe ‘KI-Bilder sind schwer zu erkennen’).

Die Entwickler von Tools wie Imagetwin und Proofig, die KI zur Erkennung von Integritätsproblemen in wissenschaftlichen Abbildungen nutzen, erweitern ihre Software, um Bilder zu filtern, die von generativer KI erstellt wurden. Da solche Bilder so schwer zu erkennen sind, erstellen beide Unternehmen ihre eigenen Datenbanken mit generativen KI-Bildern, um ihre Algorithmen zu trainieren.

Proofig hat bereits eine Funktion in seinem Tool zur Erkennung von KI-generierten Mikroskopbildern veröffentlicht. Mitbegründer Dror Kolodkin-Gal in Rehovot, Israel, sagt, dass der Algorithmus bei Tests mit Tausenden von KI-generierten und echten Bildern aus Artikeln in 98 % der Fälle KI-Bilder korrekt identifiziert hat und eine falsche Positivanzeigerrate von 0,02 % hatte. Dror fügt hinzu, dass das Team nun versucht zu verstehen, was genau ihr Algorithmus erkennt.

„Ich habe große Hoffnungen für diese Tools“, sagt Christopher. Sie merkt aber an, dass deren Ergebnisse immer von Experten bewertet werden müssen, die die von ihnen angezeigten Probleme verifizieren können. Christopher hat bisher keine Beweise gesehen, dass Software zur Erkennung von KI-Bildern

zuverlässig ist (die interne Bewertung von Proofig wurde noch nicht veröffentlicht). Diese Tools sind „begrenzt, aber sicherlich sehr nützlich, da sie es uns ermöglichen, unsere Anstrengungen zur Überprüfung von Einreichungen zu skalieren“, fügt sie hinzu.

Viele Verlage und Forschungseinrichtungen nutzen bereits **Proofig** und **Imagetwin**. Die Science-Zeitschriften verwenden beispielsweise Proofig zur Überprüfung von Integritätsproblemen in Bildern. Laut Meagan Phelan, Kommunikationsdirektorin für Science in Washington DC, hat das Tool bisher keine KI-generierten Bilder entdeckt.

Springer Nature, der Verlag von Nature, entwickelt eigene Detection-Tools für Texte und Bilder, genannt Geppetto und SnapShot, die Unregelmäßigkeiten kennzeichnen, die dann von Menschen bewertet werden. (Das Nature-Nachrichtenteam ist redaktionell unabhängig von seinem Verlag.)

Verlagsgruppen ergreifen ebenfalls Maßnahmen, um auf KI-generierte Bilder zu reagieren. Ein Sprecher der International Association of Scientific, Technical and Medical (STM) Publishers in Oxford, UK, sagte, dass man das Problem „sehr ernst nimmt“ und auf Initiativen wie **United2Act** und das STM Integrity Hub verweist, die aktuelle Probleme mit Pflichtverlagen und andere Fragen der wissenschaftlichen Integrität angehen.

Christopher, die eine Arbeitsgruppe der STM zu Bildveränderungen und -duplikationen leitet, sagt, dass ein wachsendes Bewusstsein dafür entsteht, dass es notwendig sein wird, Wege zu entwickeln, um Rohdaten zu verifizieren – zum Beispiel durch das Etikettieren von Bildern, die mit Mikroskopen aufgenommen wurden, mit unsichtbaren Wasserzeichen ähnlich den **Watermarks in KI-generierten Texten** – das könnte der richtige Weg sein. Dies erfordert neue Technologien und neue Standards für Gerätehersteller, fügt sie hinzu.

Patrick und andere machen sich Sorgen, dass Verlage nicht schnell genug handeln, um der Bedrohung entgegenzuwirken.

„Wir befürchten, dass dies nur eine weitere Generation von Problemen in der Literatur sein wird, die sie nicht angehen, bis es zu spät ist“, sagt er.

Dennoch sind einige optimistisch, dass der KI-generierte Inhalt, der heute in Artikeln erscheint, in der Zukunft entdeckt werden wird.

„Ich habe volles Vertrauen darauf, dass sich die Technologie so weit verbessern wird, dass sie die Daten erkennt, die heute erstellt werden – denn irgendwann wird dies als relativ grob angesehen werden“, sagt Patrick. „Betrüger sollten nachts nicht gut schlafen. Sie könnten den aktuellen Prozess täuschen, aber ich glaube nicht, dass sie den Prozess für immer täuschen können.“

1. Saguy, A. et al. Small Meth. https://doi.org/10.1002/smtd.202400672 (2024).

Artikel

Google Scholar

Referenzen herunterladen