FRA-Pressemitteilung
Wien, 8. Dezember 2022

SPERRFRIST: 8. Dezember 2022, 6.00 Uhr MEZ

Algorithmen auf Verzerrungen testen, um Diskriminierung zu vermeiden

Künstliche Intelligenz ist überall und betrifft jeden. In einem neuen Bericht der Agentur der Europäischen Union für Grundrechte (FRA) wird der Einsatz künstlicher Intelligenz bei der vorausschauenden Polizeiarbeit und der Erkennung von beleidigenden Äußerungen untersucht. Sie zeigt, wie Verzerrungen in Algorithmen entstehen und wie sie sich auf das Leben der Menschen auswirken können. Dies ist das erste Mal, dass die FRA praktische Erkenntnisse darüber liefert, wie Verzerrungen entstehen. Die Agentur fordert die politischen Entscheidungsträger auf, dafür zu sorgen, dass KI auf Verzerrungen getestet wird, die zu Diskriminierung führen könnte.

„Gut entwickelte und getestete Algorithmen können für viele Verbesserungen sorgen. *Aber ohne angemessene Kontrollen laufen Entwickler und Nutzer in hohem Maße Gefahr, das Leben von Menschen zu beeinträchtigen*", so der Direktor der FRA, Michael O'Flaherty. „*Es gibt keine schnelle Lösung. Wir brauchen aber ein System, mit dem Verzerrungen vor und bei der Verwendung von Algorithmen bewertet und abgemildert werden können, um Menschen vor Diskriminierung zu schützen.*"

Für ihren neuen Bericht „Verzerrungen in Algorithmen – künstliche Intelligenz und Diskriminierung" (Bias in algorithms – Artificial Intelligence and discrimination) hat die FRA zwei Fallstudien entwickelt, um mögliche Verzerrungen in Algorithmen zu testen:

1) **Vorausschauende Polizeiarbeit** zeigt, wie Verzerrungen im Laufe der Zeit verstärkt werden können, was möglicherweise zu diskriminierender Polizeiarbeit führt. Wenn die Polizei in ein bestimmtes Gebiet geht und sich dabei nur auf Vorhersagen stützt, die von verzerrten Verbrechensaufzeichnungen beeinflusst sind, wird sie hauptsächlich in diesem Gebiet Verbrechen feststellen. Dadurch entsteht eine sogenannte Feedbackschleife. In diesem Fall beeinflussen Algorithmen andere Algorithmen, indem sie diskriminierende Praktiken verstärken oder bewirken, die möglicherweise unverhältnismäßig auf ethnische Minderheiten ausgerichtet sind.
2) **Erkennung von beleidigenden Äußerungen** analysiert ethnische und geschlechtsspezifische Verzerrungen in Systemen zur Erkennung beleidigender Rede. Die Analyse zeigt, dass Instrumente zur Erkennung von Hetze im Internet zu verzerrten Ergebnissen führen können. Algorithmen können sogar harmlose Formulierungen wie „Ich bin Muslim" oder „Ich bin jüdisch" als beleidigend kennzeichnen. Es gibt auch geschlechtsspezifische Verzerrungen in geschlechtssensiblen Sprachen wie Deutsch oder Italienisch. Das kann zu einem ungleichen Zugang zu Online-Diensten aus potenziell diskriminierenden Gründen führen.

Diese Ergebnisse erfordern eine umfassende Bewertung von Algorithmen. Die FRA fordert daher die EU-Organe und -Länder zu Folgendem auf:

- **Durchführung von Tests auf Verzerrungen** – Algorithmen können verzerrt sein oder im Laufe der Zeit Verzerrungen entwickeln, was zu Diskriminierung führen kann. Die Durchführung von Tests auf Verzerrungen vor und während der Nutzung,

insbesondere bei der automatisierten Entscheidungsfindung, verringert dieses Risiko.

- **Bereitstellung von Leitlinien zu sensiblen Daten** – zur Bewertung potenzieller Diskriminierung können Daten über geschützte Merkmale (z. B. ethnische Zugehörigkeit, Geschlecht) benötigt werden. Dies erfordert eine Anleitung dazu, wann eine solche Datenerhebung zulässig ist. Sie muss gerechtfertigt, notwendig und mit wirksamen Garantien versehen sein.
- **Bewertung ethnischer und geschlechtsspezifischer Verzerrungen** – ethnische und geschlechtsspezifische Verzerrungen bei Spracherkennungs- und -Vorhersagemodellen sind stark ausgeprägt. Sie müssen von Fall zu Fall geprüft werden. Solche Bewertungen müssen faktengestützt sein und den Aufsichtsgremien sowie der Öffentlichkeit zugänglich gemacht werden.
- **Berücksichtigung aller Diskriminierungsgründe** – Vorurteile sind breit gefächert. Daher müssen alle verbotenen Diskriminierungsgründe wie Geschlecht, Religion oder ethnische Herkunft geprüft werden. Zur Bekämpfung der Diskriminierung durch Algorithmen sind verschiedene bestehende und vorgeschlagene EU-Rechtsvorschriften erforderlich, einschließlich des Vorschlags für eine Gleichbehandlungsrichtlinie.
- **Streben nach mehr Sprachenvielfalt** – Spracherkennungsmodelle konzentrieren sich tendenziell auf Englisch. Es besteht die Notwendigkeit, die Forschung zu anderen Sprachen zu fördern und zu finanzieren, um die Verwendung von ordnungsgemäß getesteten, dokumentierten und gewarteten Sprachwerkzeugen für alle EU-Amtssprachen zu fördern.
- **Verbesserung des Zugangs für faktengestützte Aufsicht** – was sich hinter KI-Systemen verbirgt, kann weitgehend unbekannt sein. Eine wirksame Aufsicht erfordert einen verbesserten Zugang zu den Daten und Dateninfrastrukturen, um das Risiko von Verzerrungen in Algorithmen zu erkennen und zu bekämpfen.

Diese Erkenntnisse sollen zu laufenden regulatorischen Entwicklungen beitragen, indem politische Entscheidungsträger, Fachleute für Menschenrechte, die Technologieindustrie und die Öffentlichkeit über das Risiko von Verzerrungen in der KI informiert werden.

Sie sind Teil des Projekts der FRA zu künstlicher Intelligenz und Big Data. In früheren Forschungsarbeiten wurden Probleme bei der Nutzung von KI festgestellt und die EU und die Mitgliedstaaten aufgefordert, dafür zu sorgen, dass KI alle Grundrechte schützt.

Weitere Auskünfte finden Sie unter: media@fra.europa.eu / Tel.: +43 1 580 30 653