JOHANNES KEPLER
UNIVERSITÄT LINZ

**Mag. Christian Savoy**
PR-Manager
Universitätskommunikation

T +43 732 2468 3012
christian.savoy@jku.at

**PRESSEMITTEILUNG**

Linz, 14. September 2022

# Feuer mit Feuer bekämpfen: Neue JKU KI verschleiert persönliche Merkmale vor Suchmaschinen-KI

**Sie wollten wissen, welche Sänger*innen die Charts anführen? Oder welchen Film Sie heute ansehen sollen? Moderne Suchmaschinen und Empfehlungssysteme arbeiten mit Künstlicher Intelligenz bzw. Deep-Learning-Methoden. Dadurch wird Ihre Frage zwar recht präzise beantwortet – gleichzeitig können aber auch Ihre persönlichen Merkmale erstaunlich genau vorhergesagt werden. Ihr Geschlecht zum Beispiel wird aus Ihrem Konsumverhalten von Musik oder Filmen zu 72 Prozent korrekt abgeleitet. An der Johannes Kepler Universität Linz wurde nun eine Methode entwickelt, die Ihre individuellen Merkmale verschleiert, ohne die Such- und Empfehlungs-Ergebnisse zu verschlechtern.**

Um keine Rückschlüsse der Empfehlungssysteme auf die privaten Eigenschaften von Anwender*innen zuzulassen, haben Prof. Markus Schedl und seine Kolleg*innen vom JKU Institut für Computational Perception eine eigene Deep-Learning-Architektur entwickelt: Adversarial Variational Auto-Encoder with Multinomial Likelihood, kurz *Adv-MultVAE*.

Dabei werden zwei Netzwerke aktiv, die separat arbeiten, aber eng gekoppelt sind. Ein Netzwerk löst die eigentliche Aufgabe, findet also Musik oder Filme, die dem Nutzer oder der Nutzerin gefallen könnten. Der große Unterschied ist das zweite Netzwerk. Dieses versucht, sensible persönliche Eigenschaften (wie Geschlecht oder Nationalität) möglichst genau vorherzusagen.

Im Gegensatz zu üblichen Empfehlungssystem-Algorithmen gehen die JKU Netzwerke ab dann in eine andere Richtung. Die Parameter, mit denen gearbeitet wird, werden nun schrittweise so angepasst, dass die Empfehlungen ähnlich bleiben, die Persönlichkeitsmerkmale aber weniger genau vorhergesagt werden können.
Im Endeffekt durchlaufen die Netzwerke eine Evolution, die am Ende gute Ergebnisse, aber schlechte Persönlichkeitsvorhersagen hervorbringt.

**Reduzierte Vorhersagegenauigkeit**
Getestet wurde das neue System mit Musik- und Filmdatensets. Dabei konnte die Vorhersagegenauigkeit für das Geschlecht der Anwender*innen auf 57 % für Filme und 62 % für Musik reduziert werden – die Empfehlungen selbst blieben qualitativ ähnlich.

Neben dem besseren Schutz der eigenen Daten hat diese Methode noch einen anderen Vorteil: „*Da unsere neuen Empfehlungssysteme nicht mehr so genau wissen, welches Geschlecht die Anwender*innen haben, wird ein breiteres Spektrum an Musik oder Filmen vorgeschlagen. Das heißt, Männern werden nun häufiger auch ‚weiblich' konnotierte Filme wie Casablanca empfohlen. Dadurch wirkt unser Ansatz auch Stereotypen und der Blasenbildung im Internet entgegen*", erklärt Schedl. Noch wirkt der JKU Algorithmus bei Frauen etwas weniger gut. *„Vermutlich, weil wir einfach mehr Daten von Männern vorliegen haben oder Frauen ein anderes Nutzungsverhalten als Männer haben"*, meint Schedl. Daher soll das System mit starkem Blick auf weibliche Merkmale weiterentwickelt werden.

**Frei verfügbarer Algorithmus**
Die Ergebnisse und der Algorithmus wurden vor kurzem auf der Top Information Retrieval Konferenz (ACM SIGIR) publiziert und sind frei verfügbar. „*Alles was man zur*

**JOHANNES KEPLER**
**UNIVERSITÄT LINZ**
Altenberger Straße 69
4040 Linz, Österreich
www.jku.at
DVR 0093696

JOHANNES KEPLER
UNIVERSITÄT LINZ

*Umsetzung benötigt, wie Source Code, Datensets etc., steht auf unserem Github-Account zur Verfügung. Mit ein bisschen Fachwissen kann man den Algorithmus also gerne selbst testen“*, lädt Schedl ein.

**Zur Publikation:** https://dl.acm.org/doi/10.1145/3477495.3531820
**Github-Account:** https://github.com/CPJKU/adv-multvae

**JOHANNES KEPLER**
**UNIVERSITÄT LINZ**
Altenberger Straße 69
4040 Linz, Österreich
www.jku.at
DVR 0093696