Kritischer Blick in den Maschinenraum

Das Fellowship-Programm des ETH AI Center fördert die interdisziplinäre Zusammenarbeit exzellenter Nachwuchsforschender und

zielt auf einen positiven Einfluss auf die Gesellschaft.

Wenn Algorithmen Vorurteile verstärken: Doktorandin Afra Amini befasst sich mit der automatisierten Verarbeitung menschlicher Sprache und ihren Fallstricken.

Sie forschen als Fellow am ETH AI Center, eine sehr begehrte Position – auf welchem Weg sind Sie da hingekommen?

AFRA AMINI – Nach meinem Bachelor an der Sharif University of Technology in Teheran habe ich ein Jahr als Datenwissenschaftlerin in einer der grössten Techfirmen Irans gearbeitet. Von einer dreimonatigen «Student Summer Research Fellowship» her kannte ich die ETH Zürich bereits. Ich hatte das Forschungsumfeld in bester Erinnerung und sagte deshalb zu, als mir ein Exzellenz-Stipendium für das Master-Studium angeboten wurde – obwohl ich an der University of Waterloo in Kanada direkt hätte doktorieren können. Dank dem Stipendium konnte ich mich auf mein Studium konzentrieren und meine volle Leistung bringen. Das half mir, als ich mich als Doktorandin am AI Center bewarb.

Woran forschen Sie im Rahmen Ihres Fellowship?

«Natural Language Processing», abgekürzt NLP, ist vereinfacht gesagt ein Gebiet der Informatik, das sich damit befasst, Computern z. B. mittels sogenanntem Deep Learning menschliche Sprache beizubringen. Technische Grundlage sind künstliche neuronale Netze. Sie ermöglichen, dass Tools wie die Googlesuche, der Online-Übersetzer DeepL oder der digitale Sprachassistent Siri gut funktionieren. Ich arbeite an der Schnittstelle dieses Gebiets zu den Sozialwissenschaften, wo wir es mit praktischen Anwendungen von Sprachmodellen zu tun haben. Diese Anwendungen haben Konsequenzen auf das Leben von Menschen; Konsequenzen, die problematisch sein können, wenn wir sie nicht reflektieren.

Problematisch inwiefern?

Die Modelle können voreingenommen sein. Ein Beispiel ist Software, die CVs screent und eine Vorhersage trifft, ob jemand für einen bestimmten Job geeignet ist. Hier haben Einseitigkeiten, beispielsweise Genderstereotype, sehr reale Konsequenzen. Wenn das Modell beispielsweise mit einem Datenset trainiert wurde, welches 80 Prozent Ärzte und nur 20 Prozent Ärztinnen enthält, ist es kein Wunder, dass das Modell zum Schluss kommt, dass ein Mann ein geeigneterer Kandidat für eine offene Stelle sei als eine Frau. Vorurteile treten auch bei der Arbeit mit Sprachen zutage, die eine geschlechtsspezifische Terminologie verwenden. So können im Deutschen sowohl Personalpronomen als auch Substantive geschlechtsmarkiert sein: «er» versus «sie» und «Arzt» versus «Ärztin». Im Englischen gibt es «he» und «she», aber lediglich «doctor».

Türkisch oder Persisch hingegen weisen weder Äquivalente für die Unterscheidung «er»/«sie» noch für «Arzt»/«Ärztin» auf. Google Translate hat in der Vergangenheit den persischen Satz «[He/She] is a doctor», «او یک پزشک است» stets in die männliche Form «He is a doctor» übersetzt, wohingegen der persische Satz «[He/She] is a nurse», «او یک پرستار است» stets in die weibliche Form «She is a nurse» übersetzt wurde. Das sind einfache Beispiele, viele Fälle sind weit komplizierter. Mit Methoden der Sozialwissenschaften lassen sich diese ergründen.

Welche praktischen Konsequenzen könnte Ihre Forschung haben?

In einem ersten Schritt geht es darum, zu zeigen, dass diese Verzerrungen da sind. In einem weiteren Schritt müssen wir einen Weg finden, die Fehler zu beheben. Unsere Modelle dürfen bestehende Stereotype nicht noch verstärken. Wir können zum Beispiel bei den Daten ansetzen, mit denen die Algorithmen trainiert werden. Das ist jedoch heikel, denn die riesige Menge an Daten ist ein entscheidender Erfolgsfaktor, man kann nicht einfach einen Teil ausser Acht lassen. Dennoch ist es einfacher, einen fehlerhaften Algorithmus zu verbessern, als menschliche Vorurteile abzubauen.

Wie befruchtet es Ihre Forschung, dass Sie am ETH AI Center arbeiten?

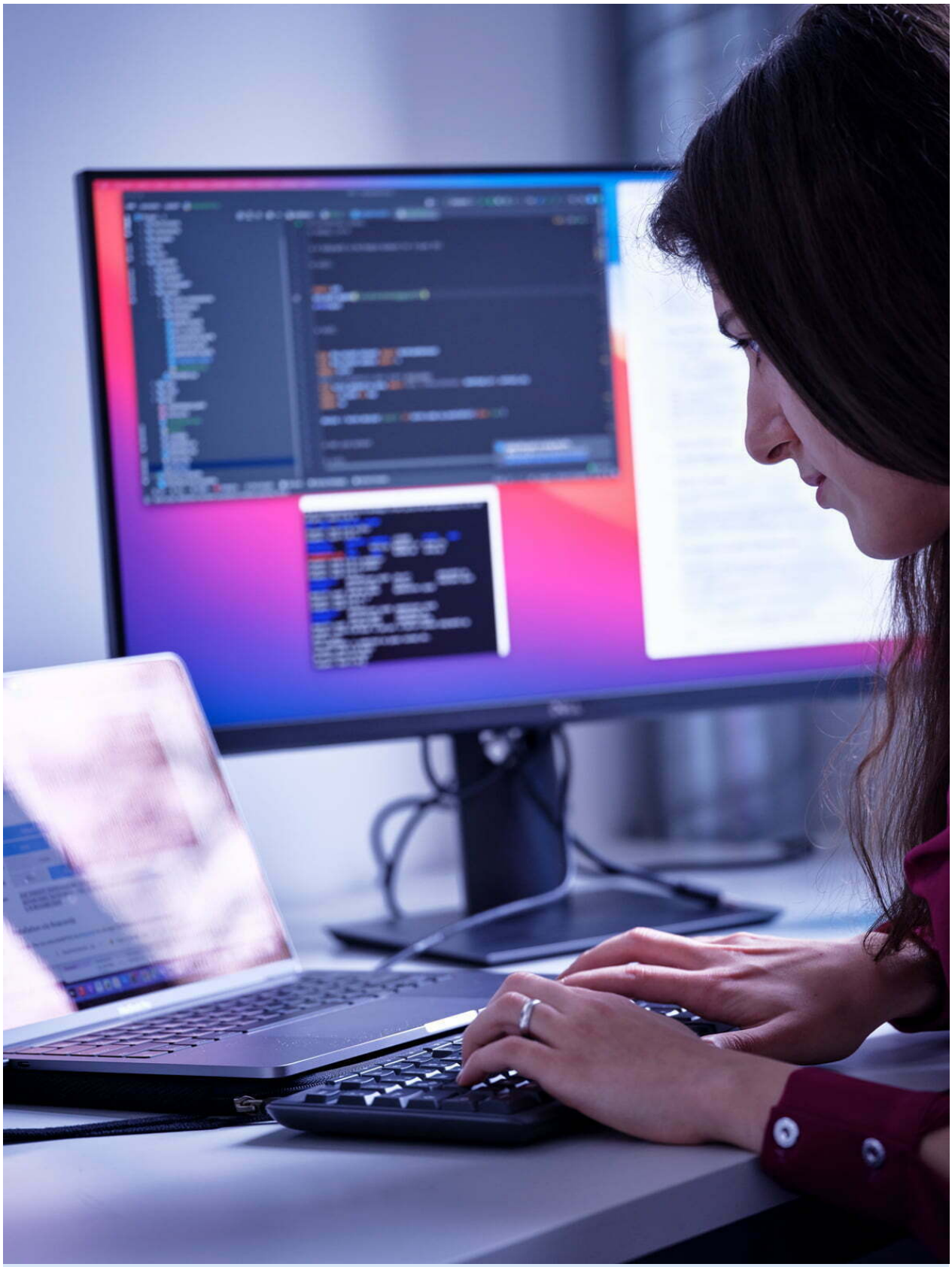
Die Interdisziplinarität des Zentrums kommt meinem Projekt sehr zugute. Es wird von zwei Professoren aus unterschiedlichen Feldern betreut, von Ryan Cotterell aus dem Departement Informatik und von Elliott Ash aus dem Department Geistes-, Sozial- und Staatswissenschaften. Die anderen Fellows hier arbeiten an der Schnittstelle

von KI zu den unterschiedlichsten Themen, von reiner Mathematik bis hin zu Robotik. Mit so vielfältigen Talenten zusammenzuarbeiten, ist ebenfalls sehr spannend für mich.

Viele Menschen verspüren ein gewisses Unbehagen, wenn es um KI geht. Was sagen Sie diesen Menschen?

Je mehr Menschen an vertrauenswürdiger KI arbeiten, Menschen mit unterschiedlichen Hintergründen, desto sicherer können wir sein, dass wir uns in die richtige Richtung bewegen. Sei es die Medikamentenforschung, der Schutz von Wildtieren oder personalisiertes Lernen: KI kann für eine ganze Reihe guter Zwecke eingesetzt werden und dazu beitragen, das Leben vieler Menschen einfacher zu machen. Darauf sollten wir nicht verzichten.

Mehr über das Exzellenz-Stipendienprogramm erfahren



«Je mehr Menschen an vertrauenswürdiger KI arbeiten, Menschen mit unterschiedlichen Hintergründen, desto sicherer können wir sein, dass wir uns in die richtige Richtung bewegen.»

Afra Amini