A critical view of the engine room

The ETH AI Center fellowship programme promotes interdisciplinary collaboration between excellent young researchers, and aims

to have a positive impact on society.

When algorithms reinforce biases: doctoral student Afra Amini is studying automated human language processing and its pitfalls.

You are researching as a fellow at the ETH AI Center, a very coveted position – how did you get there?

AFRA AMINI – Following my Bachelor's degree at Sharif University of Technology in Tehran, I spent a year working as a data scientist at one of Iran's biggest tech companies. I already knew ETH Zurich from a three-month Student Summer Research Fellowship. I had great memories of the research environment and therefore accepted when I was offered an Excellence Scholarship for the Master's programme – even though I could have gone straight to the University of Waterloo in Canada for my doctorate. Thanks to the scholarship, I was able to focus on my studies and perform at my best, and this helped when I applied to be a doctoral student at the AI Center.

What are you researching in the context of your fellowship?

Simply put, "natural language processing", or NLP for short, is a field of computer science that focuses on teaching computers to learn human language, for example by means of so-called "deep learning". This approach is built on the technical basis of artificial neural networks. They enable tools such as Google Search, the online translator DeepL or the digital voice assistant Siri to work well. I work at the interface of this field with the social sciences, where we deal with practical applications of language models. These applications have consequences for people's lives – consequences that can be problematic if we fail to reflect on them.

Problematic in what way?

The models can be biased. One example is software that screens CVs and predicts whether someone is suitable for a particular job. Here, biases such as gender stereotypes have very real consequences. If the model was trained with a dataset containing 80 percent male doctors and only 20 percent female, for example, it would be unsurprising to see the model conclude that a man would be a more suitable candidate for a vacancy than a woman. Gender bias also emerges when working with languages that use gender-specific terminology. In German, for example, both personal pronouns and nouns can be gender-specific: "er" versus "sie", and "Arzt" for a male doctor versus "Ärztin" for a female doctor. English has "he" and "she", but only "doctor".

Turkish and Persian, on the other hand, do not have equivalents for the distinction between "he" and "she" or for "male doctor" and "female doctor". In the past, Google Translate has always translated the Persian phrase "[He/She] is a doctor",
"او یک پزشک است" into the masculine form "He is a doctor", while the Persian phrase "[He/She] is a nurse", "او یک پرستار است", has always been translated into the feminine form "She is a nurse". These are simple examples, but many cases are far more complicated and can be explored using methods from the social sciences.

What might be the practical implications of your research?

The first step is to show that these distortions exist, and then we have to find a way to fix the errors. Our models must not reinforce existing stereotypes. We can start with the data used to train the algorithms, for example. This is tricky, however, because the huge amount of data is a crucial factor for success, and you can't just disregard one part. But it's certainly easier to improve a flawed algorithm than to reduce human bias.

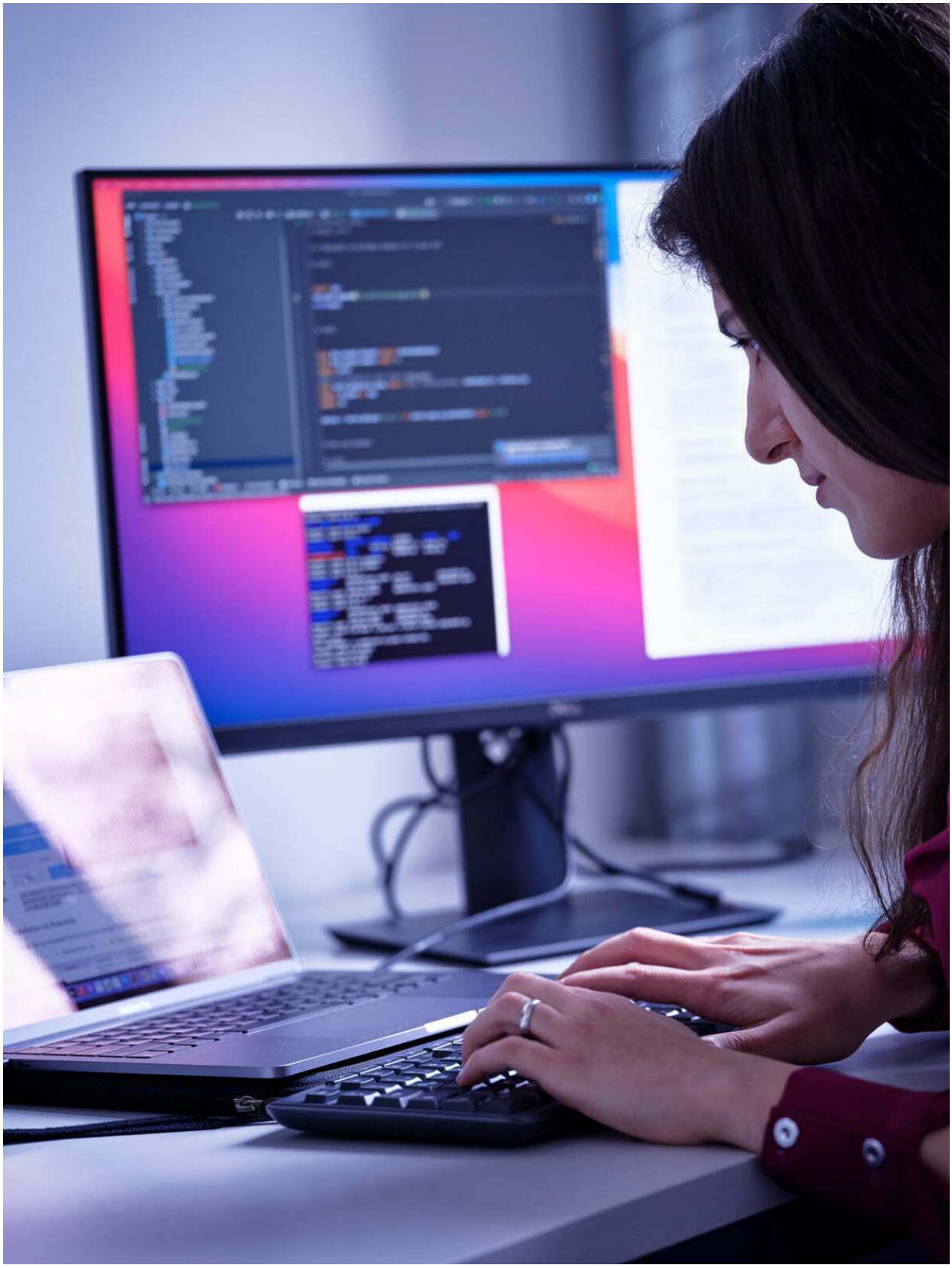
How does your research benefit from your work at the ETH AI Center?

The interdisciplinary approach of the Center is great for my project, which is supervised by two professors from different fields – Ryan Cotterell from the Department of Computer Science, and Elliott Ash from the Department of Humanities, Social and Political Sciences. The other fellows here work at the interface of AI on a wide variety of topics, from pure mathematics to robotics. Working with such a diverse range of talent is also really exciting for me.

Lots of people feel a sense of unease when it comes to AI. What do you say to them?

The more people work on trustworthy AI, people with different backgrounds, the more certain we can be that we're moving in the right direction. Drug research, wildlife conservation or personalised learning: AI can be put to a whole range of good uses and help make life easier for many people. We shouldn't miss out on that.

More about the Excellence Scholarships programme



> “The more people work on trustworthy AI, people with different backgrounds, the more certain we can be that we’re moving in the right direction.”
>
> **Afra Amini**