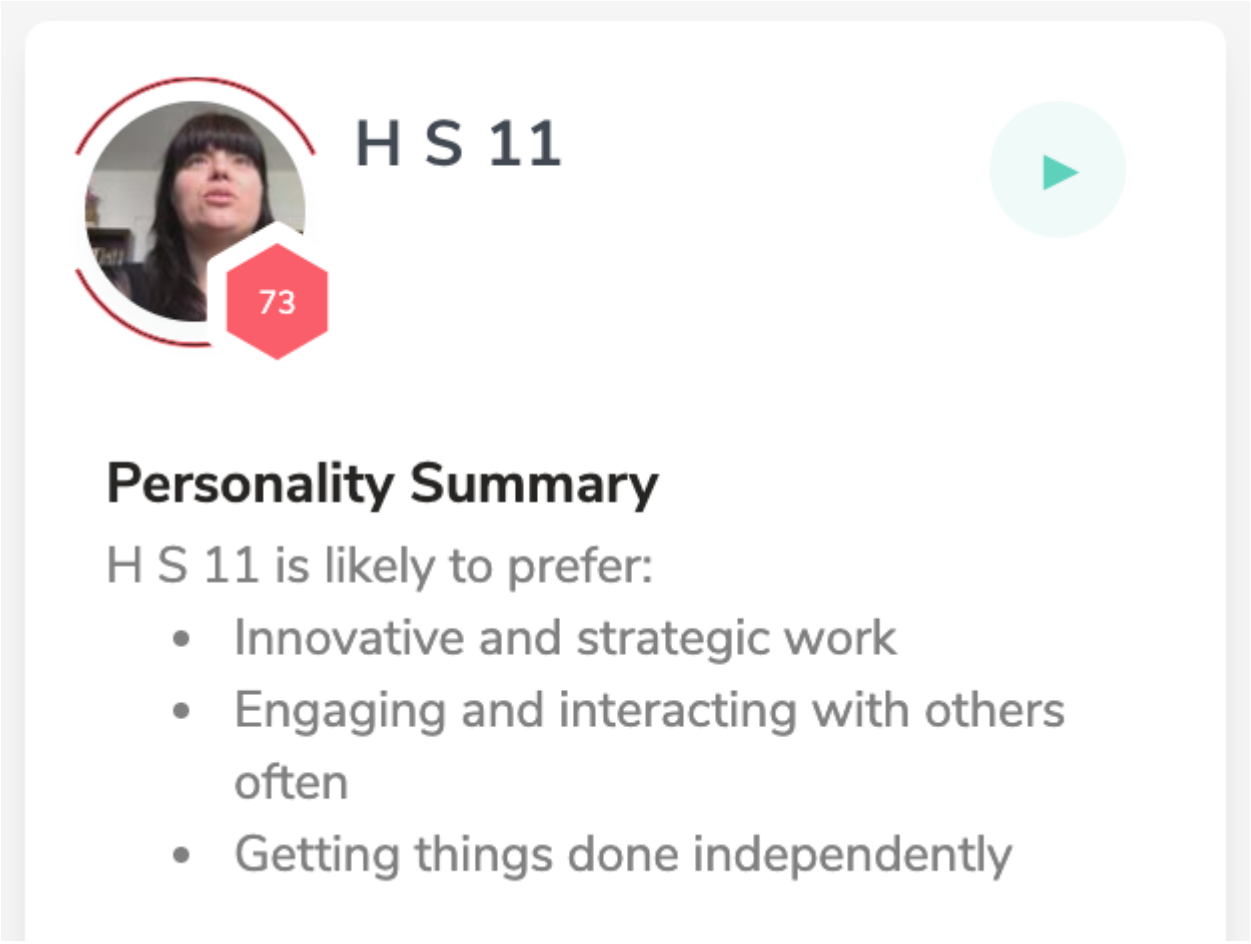
Twitter Thread by Vladimir Haltakov

Vladimir Haltakov
@haltakov

AI Job Interviews - another good example of bias in ML ■■■■

Two journalists tested some AI tools for assessing job candidates. Even when the candidate read a Wiki article in German instead of answering questions in English, the AI systems gave them good scores ■■■■

Let's unpack ■



The Setup ■

The journalists created a fake job posting on two AI interview platforms. They specified the traits of the ideal candidate and

provided the questions that need to be answered during the interview.

Then they started experimenting... ■

The Positive Test ■

One of them did a fake interview giving all the right answers and predictably got very high scores - 8.5 out of 9 ■

Then she tried something different... ■

The Negative Test ■

In a second interview, instead of answering the questions in English, she just read the article on psychometrics from the German Wikipedia ■

One system gave her a score of 6 out of 9, while the other determined she is a 73% match for the job.

Oops... ■

What happened? ■

Interestingly, one of the systems generated a transcript which was obviously meaningless.

This means that the machine learning model behind the tool likely captured nuances of the intonation of the speaker instead of the meaning of the actual words.

■

> *"So humidity is desk a beat-up. Sociology, does it iron? Mined material nematode adapt. Secure location, mesons the first half gamma their Fortunes in for IMD and fact long on for pass along to Eurasia and Z this particular location mesons."*

Bias ■■

These tools are biased - they incorrectly put more weight on the intonation instead of the actual meaning of what is being said.

This is shown when one of the components is artificially removed - in this case, the meaning (German instead of English).

More examples ■

Bias again ■■

Another AI interview tool was reported to give candidates different ratings when they used different backgrounds or accessories during the interviews.

Why does this happen? ■

The Dataset ■

One likely reason is the dataset used to train the machine learning models. It may not contain enough data for the model to correctly learn to discriminate important features. It overfits on some non-essential features, like the background.

■

Black-Box Models ■

Another common problem is the use of black-box models where it is difficult to interpret why the algorithm makes a certain decision (for example in the case of neural networks).

More work is needed to inspect what the model is paying attention to.

■

Lack of Testing ■

Furthermore, it seems that the systems weren't tested in scenarios like this one and weren't prepared to meaningfully handle these cases. The question is then, what other cases will not be handled correctly?

■

Bias in Human Interviews ■

Interestingly, human interviews are also subject to many biases and this is a known problem. The hope is that AI systems will help reduce those biases, but one needs to be careful to not just replace them with different ones.

More Details ■■

For more details, you can read the full article by @HilkeSchellmann and @sheridanlwall on MIT Technology Review here:

https://t.co/uHAlrA2vKw

I regularly post similar threads on machine learning, computer vision, and self-driving.

Follow me @haltakov for more!

Yes, this is the exact goal of most of these AI tools - to provide another data point for interviewers to help them make a better decision.

https://t.co/Vy73ydcp61

> The legend @Kasparov63 once said that the best outcome occurs when technology and humans work together and not as a replacement
>
> — Guzman Ojero (@GuzaUy) July 20, 2021

Here is a very detailed article about the influence of different backgrounds and accessories on the scores

https://t.co/7hBveNTpHb https://t.co/gZkJ6zol9W