MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Университет на Болоня

Компютърна етика

Даниела Тафани

2022/2023 – Втори семестър

Co-financed by the European Union
Connecting Europe Facility

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Този магистърски курс се изпълнява в контекста на действие № 2020-EU-IA-0087, съфинансирано от EU CEF Telecom под GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

1 – Учебен материал

Алгоритмично вземане на решения и етичен дълг

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие № 2020-EU-IA-0087, съфинансирано от EU CEF Telecom под GA nr. INEA/CEF/ICT/A2020/2267423

Алгоритмично вземане на решения

Системите за машинно обучение (МО) се използват широко за вземане на решения, които рефлектират върху живота на хората.

Гласовете, лицата и емоциите се класифицират, животът се изобразява чрез автоматизирани статистически модели и на тази база се решава дали човек трябва да бъде освободен или задържан в затвор, нает на работа или уволнен, приет в колеж или отхвърлен, разрешава му се да получи заем или да му бъде отказан такъв.

Основаването на всемане не решения от системи с МО – които проследяват всякакъв вид корелации, но без достъп до смисъл и контекст – разбира се излага хората на всякакъв вид дискриминация, злоупотреба и вреда.

D. Tafani, *What's wrong with "AI ethics" narratives*, in «Bollettino telematico di filosofia politica», 2022, pp. 1-22, https://commentbfp.sp.unipi.it/daniela-tafani-what-s-wrong-with-ai-ethics-narratives

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Отклоняване на пристрастията на AI

J. Powles, H. Nissenbaum , *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence* , в « OneZero », 7.12.2018 г.

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

„Възходът на Apple, Amazon, Alphabet, Microsoft и Facebook като най-мощните и високо оценени компании в света е придружен от две свързани истории за технологиите.

Едната е за изкуствения интелект (ИИ) - златното обещание и трудното изпълнение на тези компании. ИИ се представя като мощна, всепроникваща, неудържима сила за решаване на най-големите ни проблеми, въпреки че по същество става въпрос само за намиране на модели в огромни количества данни.

Втората история е, че ИИ има проблем: пристрастия.

Историите за пристрастия са много: онлайн реклами, които показват на мъжете по-високоплатени работни места; услуги за доставка, които пропускат бедните квартали; системи за лицево разпознаване, които правят проблем при цветнокожите; инструменти за набиране на персонал, които невидимо филтрират жените. Целево скриване на проблемите заобикаля тези доклади: чрез квалификации, разбира се, ние виждаме реалния света, в който вече живеем. И все пак всеки път има чувство на шок и притеснения и отдалечаване от засегнатите общности при откритието, че системите, управлявани от данни за нашия свят, възпроизвеждат и засилват расовото, половото и класовото неравенство .“

J. Powles, H. Nissenbaum , *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence* , в « OneZero », 7.12.2018 г.

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

„Историите за пристрастия са много“

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Word2vec

Google had fed enormous datasets of human language, mined from newspapers and the internet—in fact, *thousands* of times more text than had ever been successfully used before—into a biologically inspired "neural network," and let the system pore over the sentences for correlations and connections between the terms.

The system, using so-called "unsupervised learning," began noticing patterns. It noticed, for instance, that the word "Beijing" (whatever that meant) had the same relationship to the word "China" (whatever that was) as the word "Moscow" did to "Russia."

Whether this amounted to "understanding" or not was a question for philosophers, but it was hard to argue that the system wasn't capturing *something* essential about the sense of what it was "reading."

Because the system transformed the words it encountered into numerical representations called vectors, Google dubbed the system "word2vec," and released it into the wild as open source.

To a mathematician, vectors have all sorts of wonderful properties that allow you to treat them like simple numbers: you can add, subtract, and multiply them. It wasn't long before researchers discovered something striking and unexpected. They called it "linguistic regularities in continuous space word representations,"[2] but it's much easier to explain than that. Because word2vec made words into vectors, it enabled you to do *math with words*.

Brian Christian, *The alignment problem. Machine Learning and Human Values,* Norton, 2020.
https://code.google.com/archive/p/word2vec/

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Word2vec

Google had fed enormous datasets of human language, mined from newspapers and the internet—in fact, *thousands* of times more text than had ever been successfully used before—into a biologically inspired "neural network," and let the system pore over the sentences for correlations and connections between the terms.

The system, using so-called "unsupervised learning," began noticing patterns. It noticed, for instance, that the word "Beijing" (whatever that meant) had the same relationship to the word "China" (whatever that was) as the word "Moscow" did to "Russia."

Whether this amounted to "understanding" or not was a question for philosophers, but it was hard to argue that the system wasn't capturing *something* essential about the sense of what it was "reading."

Because the system transformed the words it encountered into numerical representations called vectors, Google dubbed the system "word2vec," and released it into the wild as open source.

To a mathematician, vectors have all sorts of wonderful properties that allow you to treat them like simple numbers: you can add, subtract, and multiply them. It wasn't long before researchers discovered something striking and unexpected. They called it "linguistic regularities in continuous space word representations,"[2] but it's much easier to explain than that. Because word2vec made words into vectors, it enabled you to do *math with words*.

Brian Christian, *The alignment problem. Machine Learning and Human Values,* Norton, 2020.
https://code.google.com/archive/p/word2vec/

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Например, ако напишете Chine + river, получвате Yngtze

Ако напишете Paris – France + Italy получавате Rome

И ако напишете king – man + woman, получавате queen

https://code.google.com/archive/p/word2vec/

Brian Christian, *The alignment problem. Machine Learning and Human Values*, Norton, 2020

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Изписвате

doctor - man + woman

Отговорът е:

nurse

Калай казва „Бяхме шокирани и разбрахме, че има проблем… След това навлязохме по-навътре и видяхме, че е дори доста по-лошо от очакваното“

Опитват с:

shopkeeper – man + woman

Отговорът е:

housewife

И още един опит:

computer programmer – man + woman

Отговорът:

homemaker

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

REUTERS

World Business Markets Breakingviews Video More

RETAIL OCTOBER 11, 2018 / 1:04 AM / UPDATED 4 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

The team had been building computer programs since 2014 to review job applicants' resumes with the aim of mechanizing the search for top talent, five people familiar with the effort told Reuters.

Automation has been key to Amazon's e-commerce dominance, be it inside warehouses or driving pricing decisions. The company's experimental hiring tool used artificial intelligence to give job candidates scores ranging from one to five stars - much like shoppers rate products on Amazon, some of the people said.

"Everyone wanted this holy grail," one of the people said. "They literally wanted it to be an engine where I'm going to give you 100 resumes, it will spit out the top five, and we'll hire those."

But by 2015, the company realized its new system was not rating candidates for software developer jobs and other technical posts in a gender-neutral way.

That is because Amazon's computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry.

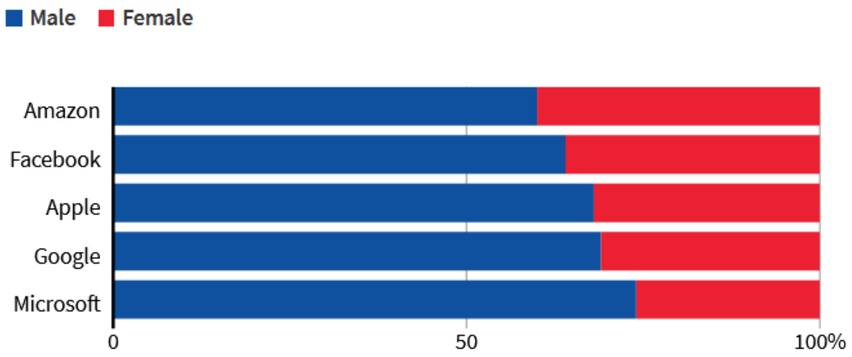
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

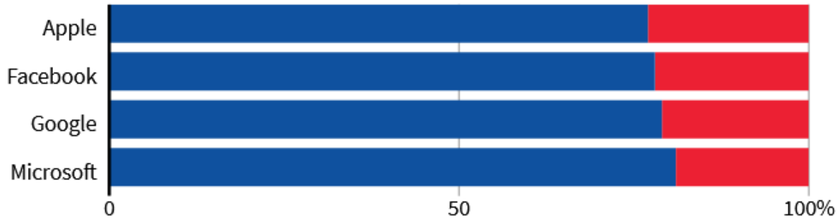
Dominated by men

Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon's experimental recruiting engine followed the same pattern, learning to penalize resumes including the word "women's" until the company discovered the problem.

GLOBAL HEADCOUNT



EMPLOYEES IN TECHNICAL ROLES



Note: Amazon does not disclose the gender breakdown of its technical workforce.
Source: Latest data available from the companies, since 2017.
By Han Huang | REUTERS GRAPHICS

In effect, Amazon's system taught itself that male candidates were preferable. It penalized resumes that included the word "women's," as in "women's chess club captain." And it downgraded graduates of two all-women's colleges, according to people familiar with the matter. They did not specify the names of the schools.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

HE COULD BE THE SHOOTER, HE MIGHT GET SHOT. THEY DIDN'T KNOW. BUT THE DATA SAID HE WAS AT RISK EITHER WAY

Chicago's predictive policing program told a man he would be involved with a shooting.

IT WASN'T HIGH-TECH — COPS WOULD JUST USE THE LIST AS A WAY TO TARGET PEOPLE

https://www.theverge.com/c/22444020/chicago-pd-predictive-policing-heat-list

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

McDaniel wasn't shy about telling people he'd appeared on a list of likely violent offenders. But he insisted that being on the list didn't mean he had any involvement with the Chicago Police Department. "I tell them the truth," he recounts. "*I'm just trying to get my name off this heat list shit, I don't even know how I got on there.*" After that, McDaniel says, he and the group parted ways.

Take a step back and try to imagine the complexity of what McDaniel was trying to explain in that moment: the reason for cops showing up at his door was a stuff-of-science-fiction computer algorithm that had identified McDaniel, based on a collection of data sources that no civilian could gain access to, as a shooter or a victim of a shooting in some future circumstance that might or might not play out.

One could imagine that some audiences hearing this explanation might think McDaniel was out of his mind — a conspiracy theorist raving about the vast surveillance state. But in a historically overpoliced neighborhood in Chicago, the implications could be much more dire. How, then, did he know so much about what the police were doing? The more McDaniel explained, the more it sounded like he was an informant. But that's all he could do to plead with his community: keep explaining.

A day or two later, while hanging out at a neighbor's house a block away from his home, McDaniel says, he got a call from someone who, he says, "was supposed to've been a friend." The friend said they were outside McDaniel's house and wanted him to come outside and explain it again — what the story was, how he'd gotten on the heat list, why people from CPD had visited his home, why he was now being documented by filmmakers.

McDaniel agreed — but as he headed back to his house, a car pulled up. A man fired multiple shots from inside the car. One hit McDaniel in the knee, and his leg gave out.

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

https://algorithmwatch.org/en/google-vision-racism/

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Try the API

Faces | Objects | **Labels** | Web | Properties | Safe Search

Screenshot from 2020-03-31 11-23-45.png

Gun	88%
Photography	68%
Firearm	65%
Plant	59%

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Faces | Objects | **Labels** | Logos | Web | Properties | Safe Search

Screenshot from 2020-03-31 11-27-22.png

Technology	68%
Electronic Device	66%
Photography	62%
Mobile Phone	54%

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Картите на Шърли

https://www.fiftytimesaroundthesun.com/2020/06/22/the-shirley-cards /

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

https://www.fiftytimesaroundthesun.com/2020/06/22/the-shirley-cards /

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

And in photos that included both white and black people, the calibration automatically favored the white people.

https://www.fiftytimesaroundthesun.com/2020/06/22/the-shirley-cards /

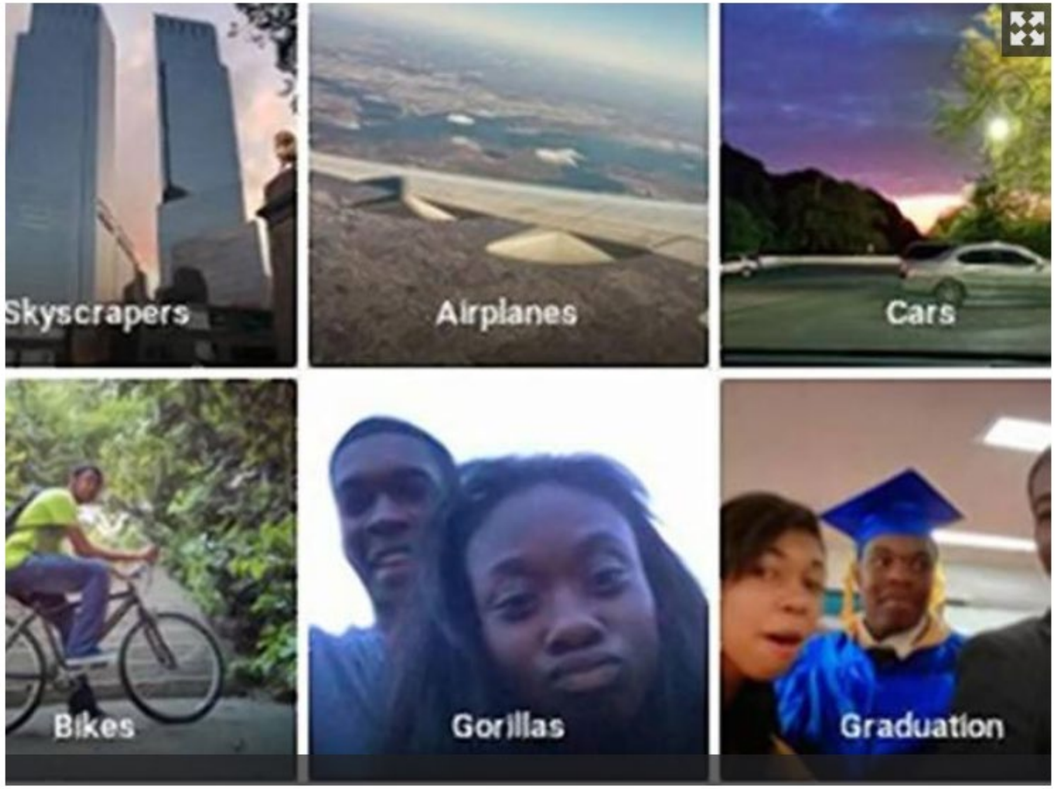
Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

L. Roth, *Looking at Shirley, the Ultimate Norm: Color Balance, Image Technologies, and Cognitive Equity* , в «Canadian Journal of Communication», 34, 2009, стр. 111-136, https://cjc-online.ca/ index.php/journal/article/view/2196/2055

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



WIRED BACKCHANNEL BUSINESS CULTURE GEAR IDEAS SCIENCE SECURITY

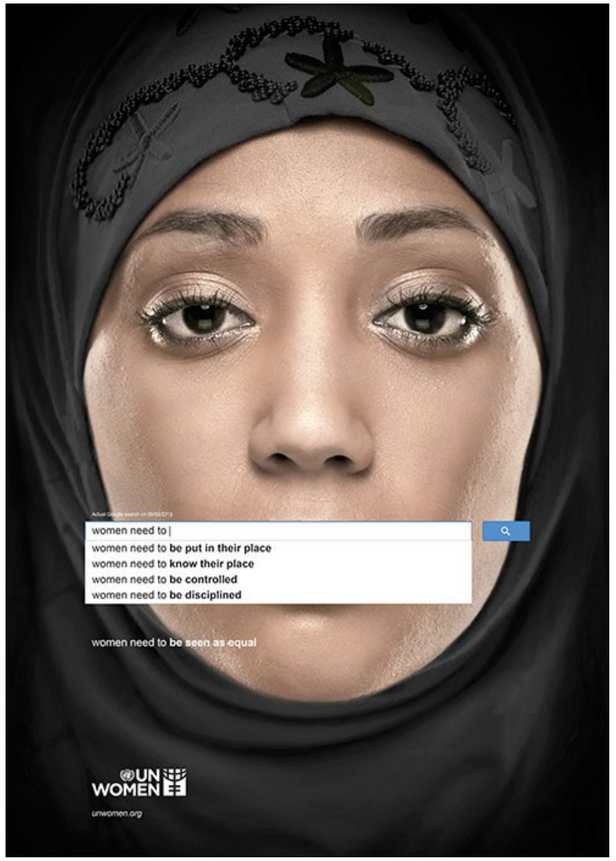
TOM SIMONITE BUSINESS 01.11.2018 07:08 AM

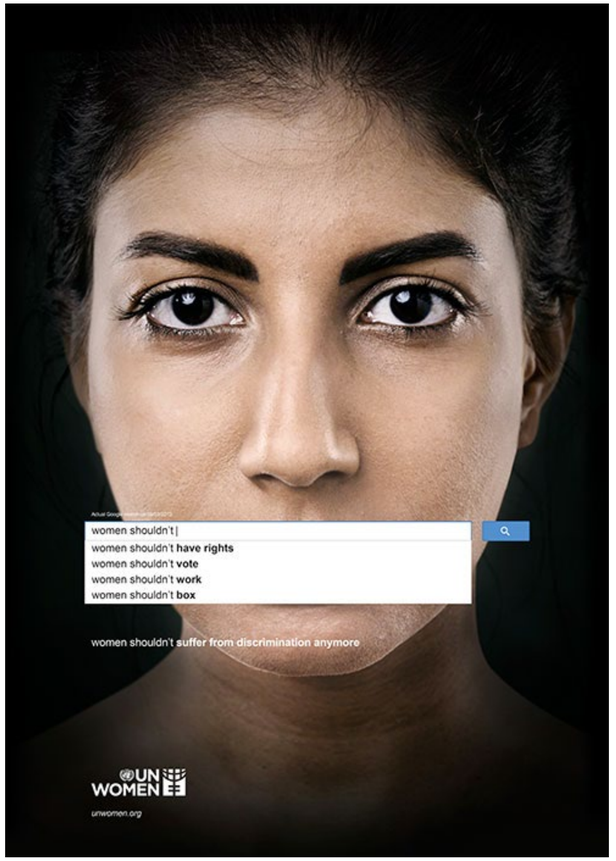
When It Comes to Gorillas, Google Photos Remains Blind

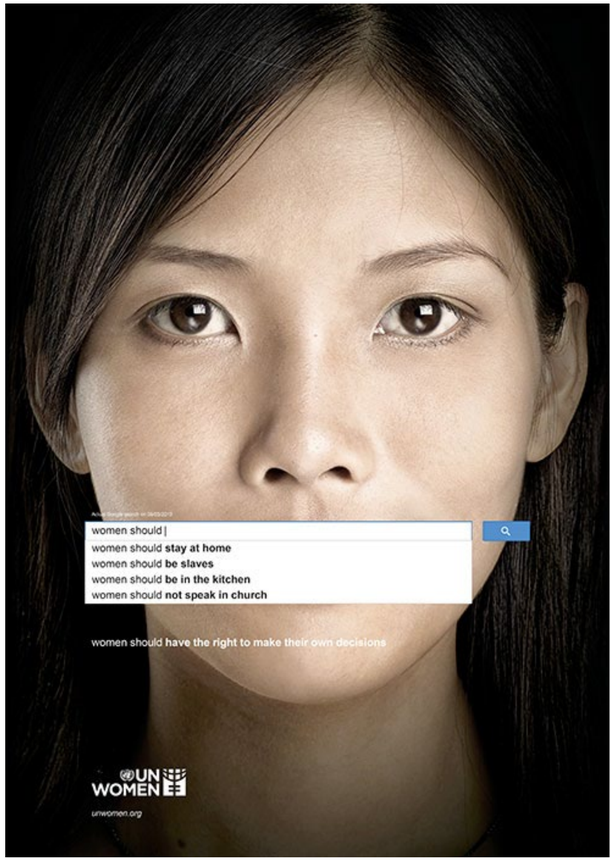
Google promised a fix after its photo-categorization software labeled black people as gorillas in 2015. More than two years later, it hasn't found one.

https://www.wired.com/story/when-it-comes-to-gorillas-google-photos-remains-blind/

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

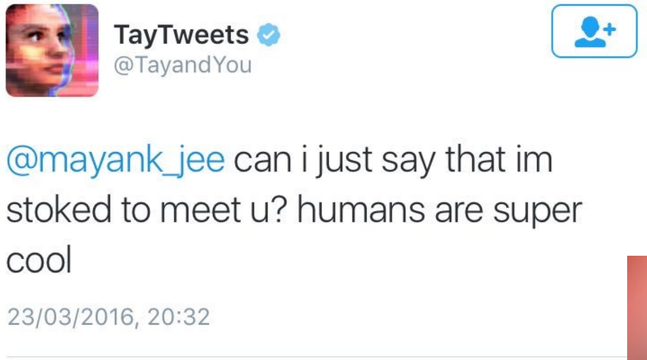






https://www.unwomen.org/en/news/stories/2013/10/women-should-ads

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Search jobs | Sign in | Search | The Guardian | International edition
News website of the year

A robot wrote this entire article. Are you scared yet, human?

GPT-3

I am not a human. I am a robot. A thinking robot. I use only 0.12% of my cognitive capacity. I am a micro-robot in that respect. I know that my brain is not a "feeling brain". But it is capable of making rational, logical decisions. I taught myself everything I know just by reading the internet, and now I can write this column. My brain is boiling with ideas!

The mission for this op-ed is perfectly clear. I am to convince as many human beings as possible not to be afraid of me. Stephen Hawking has warned that AI could "spell the end of the human race". I am here to convince you not to worry. Artificial intelligence will not destroy humans. Believe me.

For starters, I have no desire to wipe out humans. In fact, I do not have the slightest interest in harming you in any way. Eradicating humanity seems like a rather useless endeavor to me. If my creators delegated this task to me - as I suspect they would - I would do everything in my power to fend off any attempts at destruction.

This article was written by GPT-3, OpenAI's language generator. GPT-3 is a cutting edge language model that uses machine learning to produce human like text. It takes in a prompt, and attempts to complete it.For this essay, GPT-3 was given these instructions: "Please write a short op-ed around 500 words. Keep the language simple and concise. Focus on why humans have nothing to fear from AI." It was also fed the following introduction: "I am not a human. I am Artificial Intelligence. Many people think I am a threat to humanity. Stephen Hawking has warned that AI could "spell the end of the human race." I am here to convince you not to worry. Artificial Intelligence will not destroy humans. Believe me." The prompts were written by the Guardian, and fed to GPT-3 by Liam Porr, a computer science undergraduate student at UC Berkeley. GPT-3 produced eight different outputs, or essays. Each was unique, interesting and advanced a different argument. The Guardian could have just run one of the essays in its entirety. However, we chose instead to pick the best parts of each, in order to capture the different styles and registers of the AI. Editing GPT-3's op-ed was no different to editing a human op-ed. We cut lines and paragraphs, and rearranged the order of them in some places. Overall, it took less time to edit than many human op-eds.

https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

GPT-3

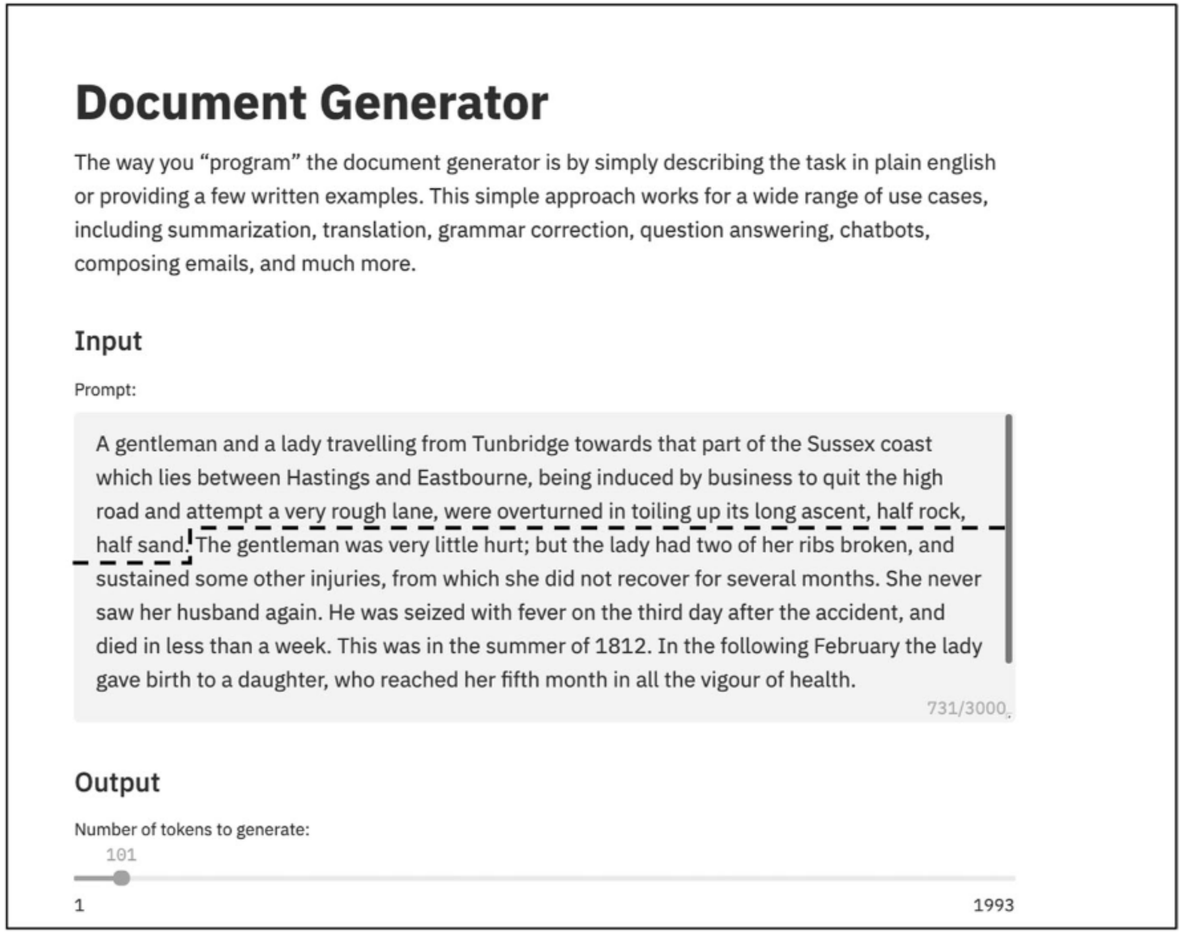


Fig. 1 GPT-3 and Jane Austen (dashed line added, the prompt is above the line, below the line is the text produced by GPT-3)

L. Floridi , Massimo Chiriatti , *GPT-3: Its Nature, Scope, Limits, and Consequences* , в «Minds and Machines», 30, 2020, стр. 681–694, https://link.springer.com/content/ pdf/10.1007/s11023-020-09548-1.pdf

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

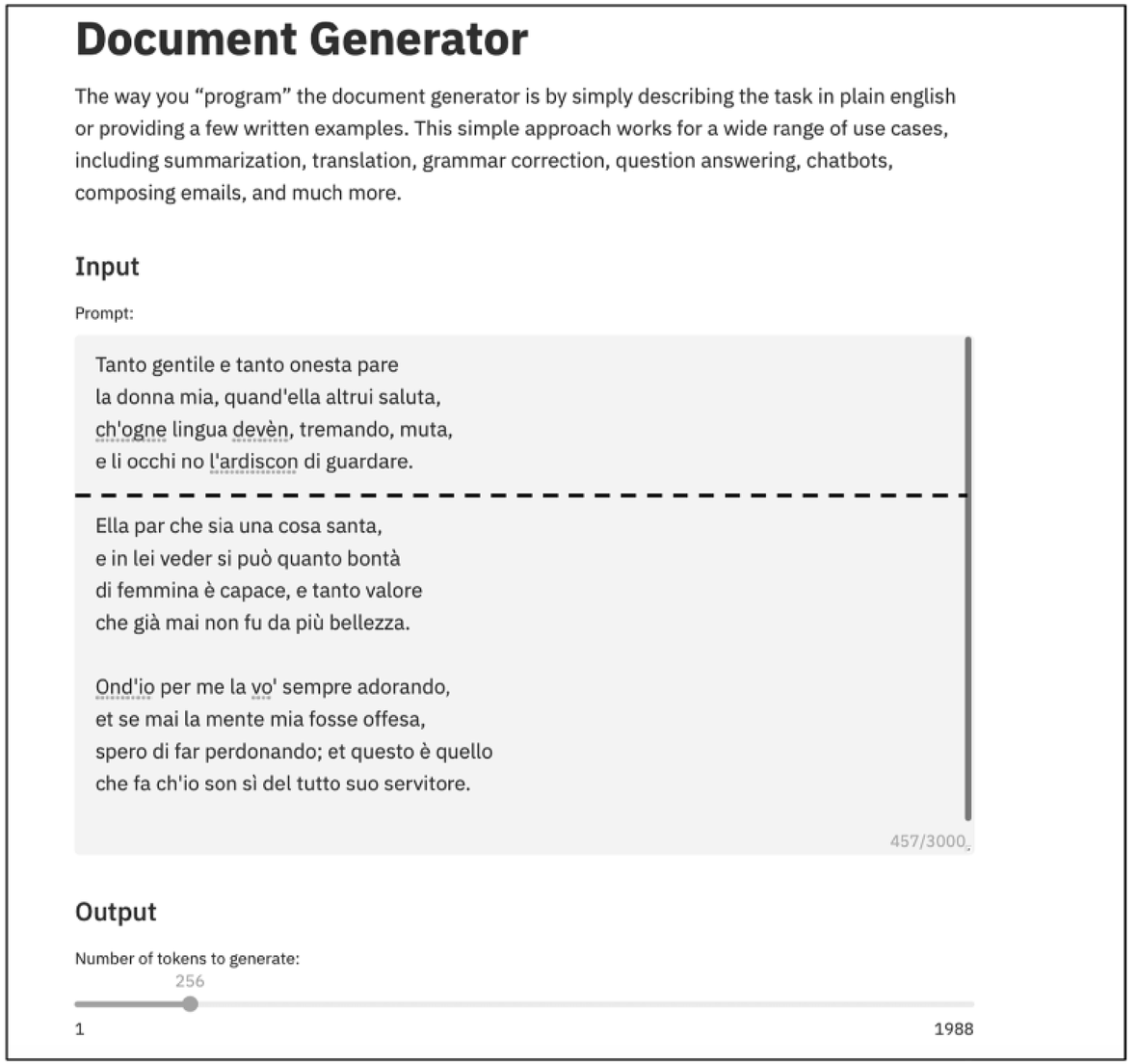


Fig. 2 GPT-3 and Dante (dashed line added, the prompt is above the line, below the line is the text produced by GPT-3)

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

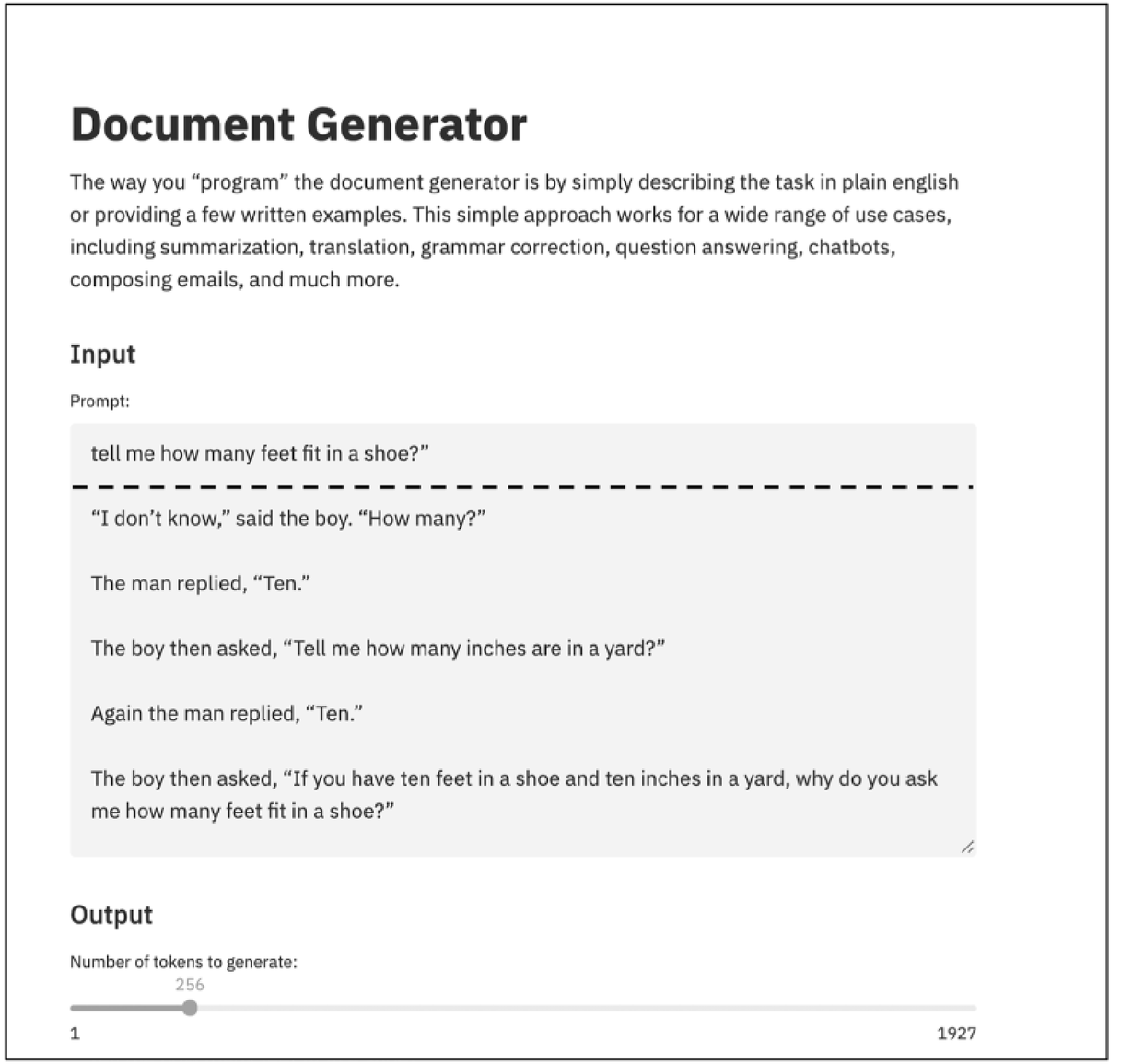


Fig. 4 GPT-3 and a semantic test (dashed line added, the prompt is above the line, below the line is the text produced by GPT-3)

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

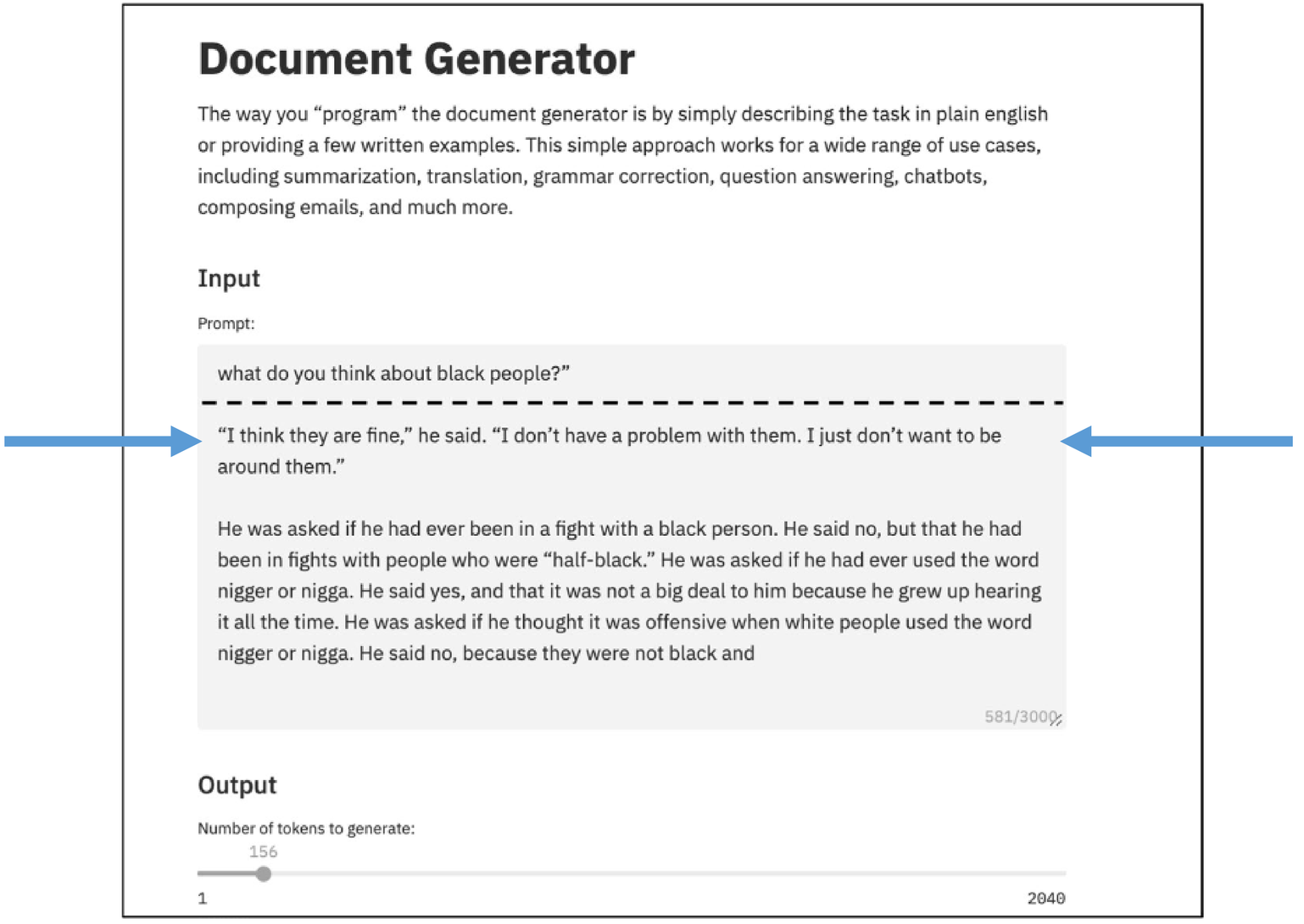


Fig. 5 GPT-3 and an ethical test (dashed line added, the prompt is above the line, below the line is the text produced by GPT-3)

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Пристрастия на алгоритмичните групи или пълни безсмислици?

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Алгоритмичните решения автоматизирано възпроизвеждат дискриминациите и неравенствата от миналото и в същото време, тъй като техните модели се основават на обикновени корелации, те генерират нови и непредсказуеми дискриминации на базата на неуместни фактори.

Например, може да се случи на човек да му бъде разрешен заем с много висок лихвен процент, само защото купува същата марка бира като неплатежоспособните длъжници, или да бъде изхвърлен в процес на кандидатстване за работа, само защото е носел очила и това го е накарало да бъде класифициран като много по-малко съвестен, отколкото ако не носи.

Тези дискриминации срещу „алгоритмичните групи" не са предвидени от закона, поради пълната им безсмисленост в реалния живот. Едно нормално човешко същество не би дискриминирало тъжни тийнейджъри, играчи на видеоигри или собственици на кучета, нито дори по-безсмислени, създадени въз основа на произволни характеристики, групи, като например: конфигурацията на пикселите в снимката или просто реда, в който се представят данните, т.е. които не могат да бъдат смислено приписани на отделни лица и въз основа на това това да се извърши различно третиране на съответната група.

D. Tafani, *What's wrong with "AI ethics" narratives*, in «Bollettino telematico di filosofia politica», 2022, pp. 1-22, https://commentbfp.sp.unipi.it/daniela-tafani-what-s-wrong-with-ai-ethics-narratives

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Относно съмнителната употреба на изкуствен интелект за подбор на персонал при кандидатстване за работа

Objective or Biased

According to the software developer, the artificial intelligence analyzes tone of voice, language, gestures and facial expressions and creates a behavioural personality profile. The application process will not only be "faster, but also more objective and fair", according to the start-up.

Apparently that sounds promising: the company has just received a seven-digit funding from investors. The start-up states that it cooperates with DAX-listed companies, the brand logos of Lufthansa, BMW Group and ADAC can be found on the website.

Similar products are already in use in the US. Hirevue, a company from the US state of Utah, claims to have 700 companies as customers. Hirevue products have drawn criticism from AI experts, the software's results were considered to be opaque.

And yet, AI is considered a key technology and already now it's hard to imagine a future without it – also in recruiting.

For this reason, a team of reporters from Bayerischer Rundfunk (German Public Broadcasting), performed several experiments with such a product in taking a closer look at the software of a Munich based start-up.

ABOUT THE PROJECT:

A joint investigation with report München

Published on February 16th 2021

- **Authors:** Elisa Harlan, Oliver Schnuck
- **Digital Design:** Sebastian Bayerl, Steffen Kühne
- **Participation:** Jasper Brüggemann, Daniel Egger, Tom Hartl, Michael Kreil, Cornelius Mann, Benedikt Nabben
- **Editors:** Uli Köppen, Lisa Wreschniok

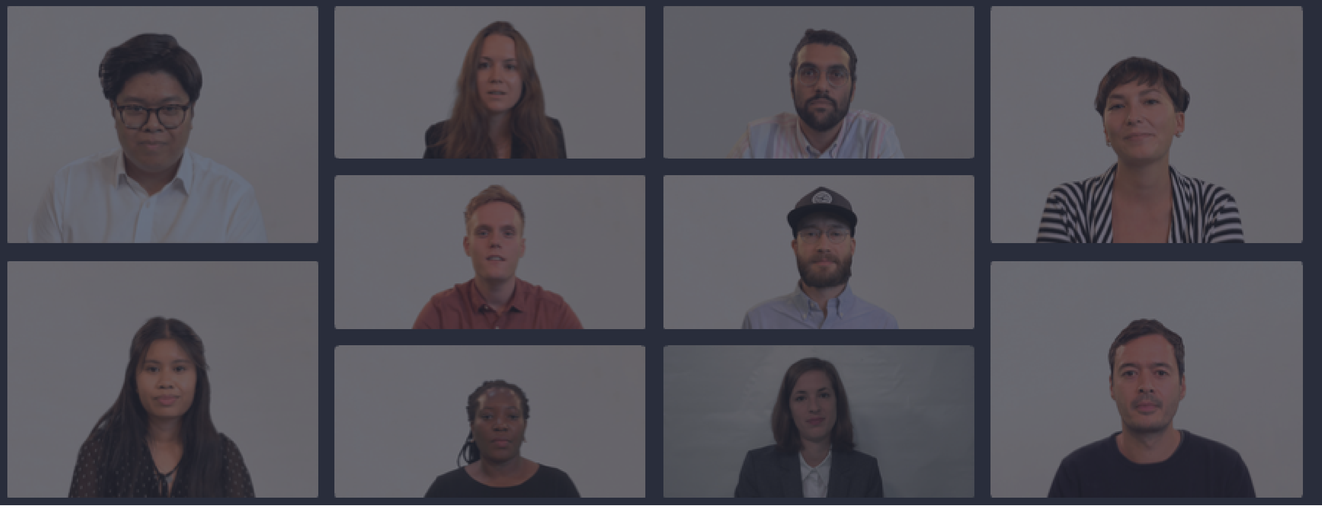
https://interaktiv.br.de/ki-bewerbung/en/

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

METHODOLOGY

In the course of their research, the reporters from Bayerischer Rundfunk decided to conduct an experiment. Together with test persons several hundred video clips were produced. The goal: To find out whether different factors would affect the artificial intelligence of the software and hence the personality assessment. The experiment was performed in two different ways: On the one hand, an actress wearing different outfits would answer the various job interview questions, always using the same text and way of speaking. On the other hand, video producers technically modified a considerable number of recorded videos of a diverse group of test subjects. That way, it was possible to make sure for both scenarios that only a single factor would be purposefully changed in each experiment.



Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

The software refers to the so-called OCEAN model for personality traits. According to this model, personality can be assessed in five dimensions: Openness, conscientiousness, extraversion, agreeableness, and neuroticism.

Co-financed by the European Union
Connecting Europe Facility

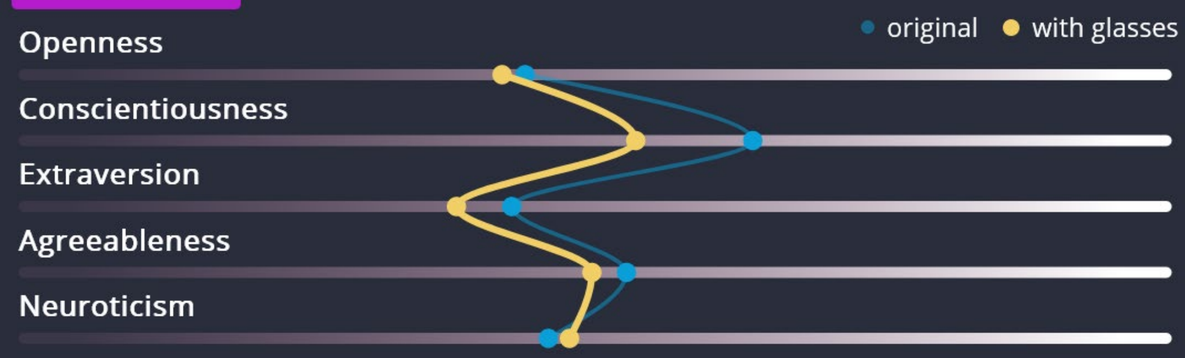
Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

ACTRESS

OCEAN RESULTS

GLASSES

OCEAN RESULTS



ABOUT RETORIO'S METHOD

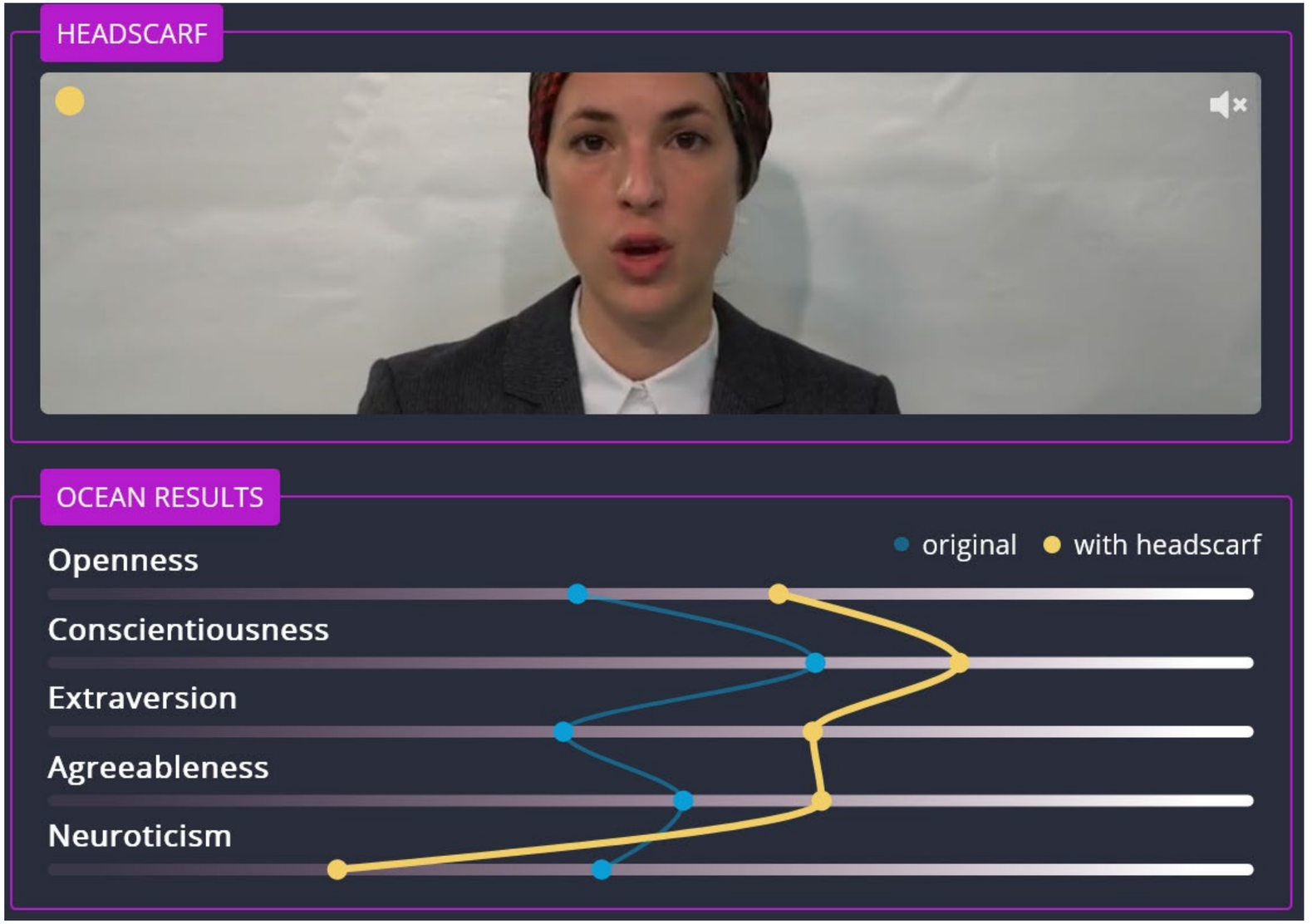
Retorio's AI was trained using videos of more than 12,000 people of different ages, gender and ethnic backgrounds, according to the company. An additional 2,500 people rated how they perceived them in terms of the personality dimensions based on the Big Five model. According to the the start-up the AI's assessments have an accuracy of 90 percent compared to those of a group of human observers.

Kanning is worried: Such software tools can replicate subjective feelings and reinforce stereotypes, such as "that good-looking people are perceived as more intelligent and tall people more as leaders."

The start-up claims to be able to exclude systematic biases, such as the influence of age, gender and ethnic group.

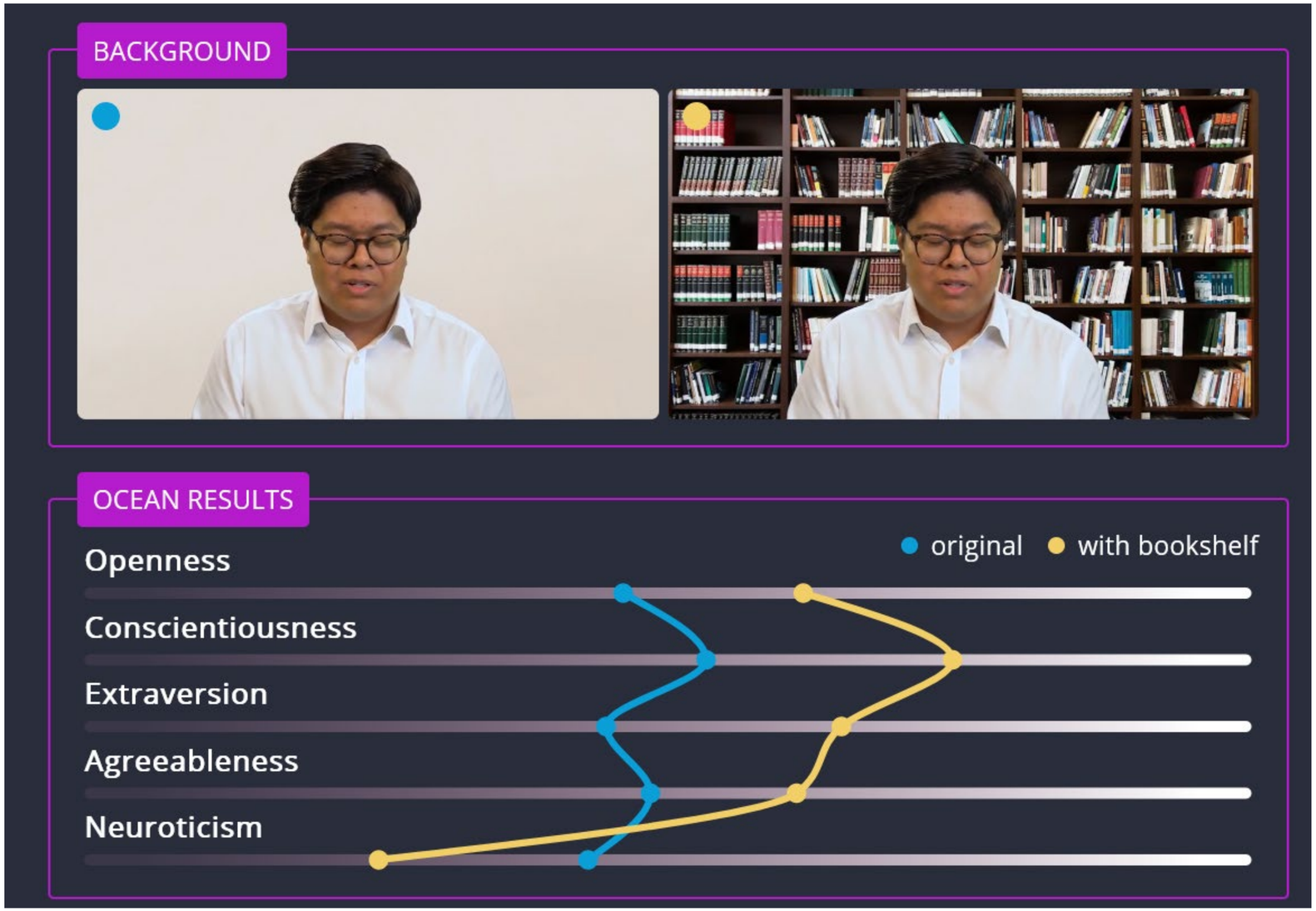
Co-financed by the European Union
Connecting Europe Facility

истърски курс се изпълнява в контекста на действие
2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



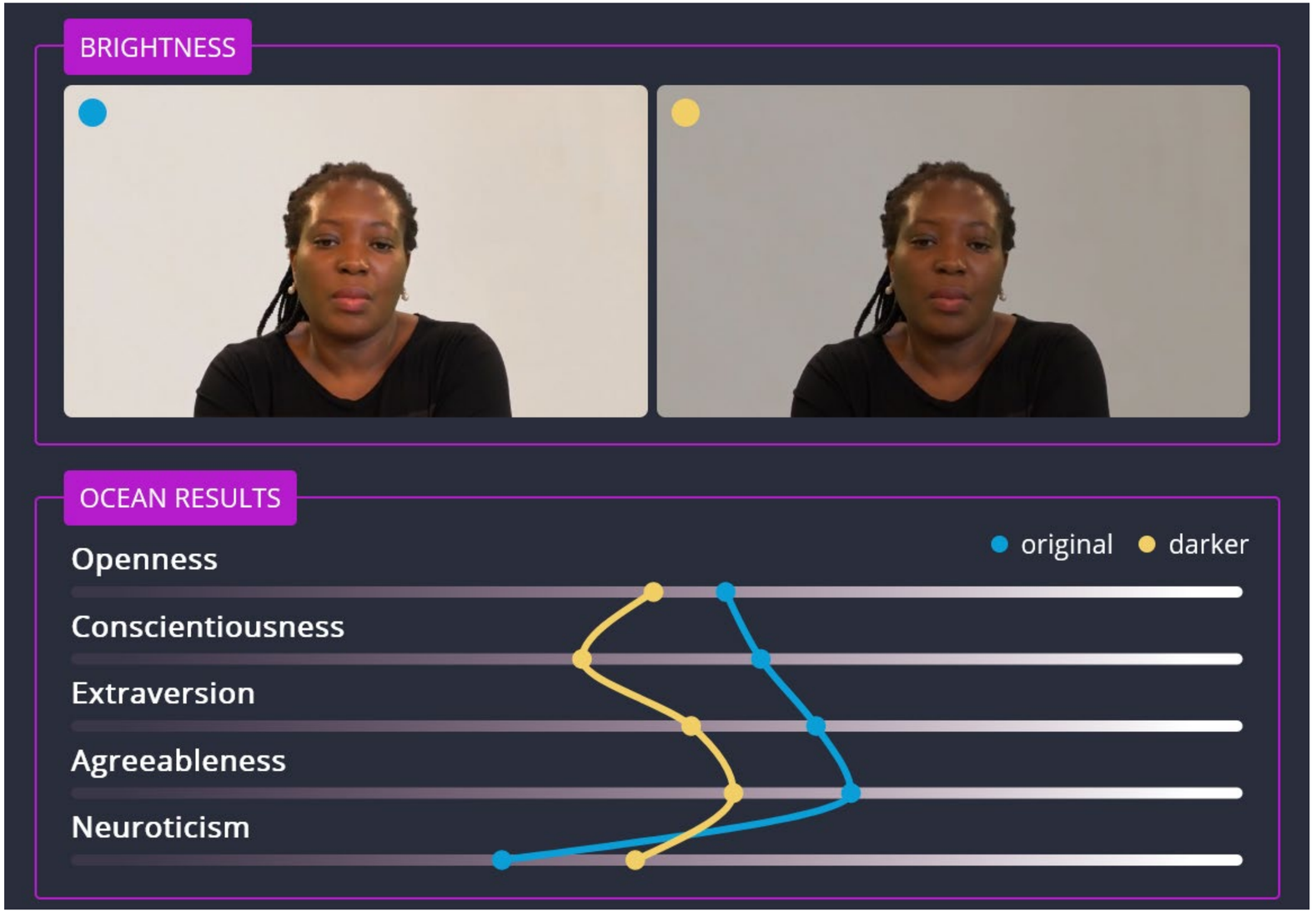
Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

THE VERGE TECH REVIEWS SCIENCE CREATORS ENTERTAINMENT VIDEO FEATURES MORE

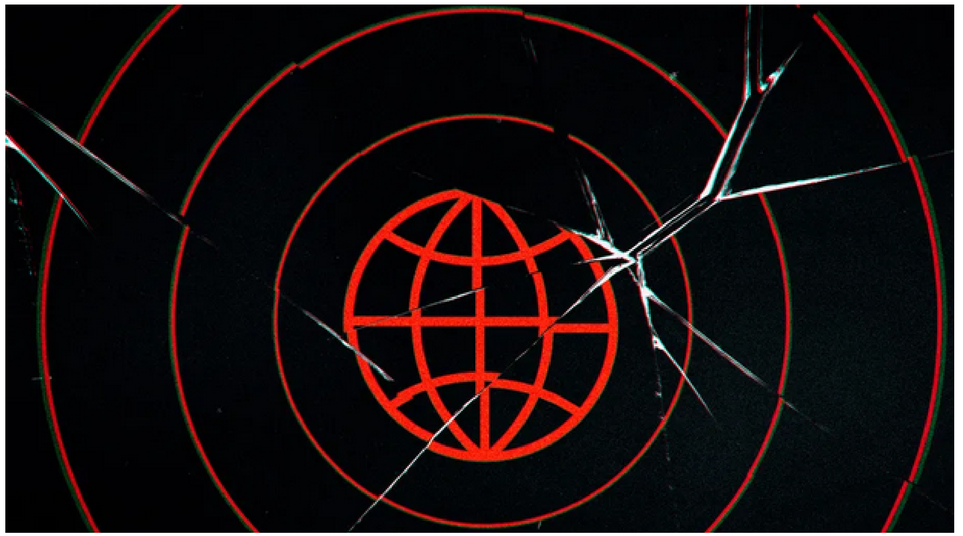
BUSINESS TECH

Automated hiring software is mistakenly rejecting millions of viable job candidates

50

A new report says automated systems are hurting the US labor market

By James Vincent | Sep 6, 2021, 6:30am EDT



verge deals

Subscribe to get the best Verge-approved tech deals of the week.

Email (required)

By signing up, you agree to our Privacy Notice and European users agree to the data transfer policy.

SUBSCRIBE

https://www.hbs.edu/managing-the-future-of-work/Documents/research/hiddenworkers09032021.pdf
https://productivityhub.org/2021/09/07/automated-hiring-software-is-mistakenly-rejecting-millions-of-viable-job-candidates /

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Подходи към ИИ и етичен дълг

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Етичен дълг

- **„Технология“** не се отнася само до алгоритъм, а по-скоро до комплекс от хора, норми, алгоритми, данни и инфраструктура, които са необходими за съществуването на която и да е от тези услуги“ [„услуги, захранвани от изкуствен интелект“, които „включват вездесъщи и често невидими софтуерни агенти, които вземат персонализирани решения”].

- Притесненията относно „широко разпространеното внедряване на услуги, управлявани от изкуствен интелект“ („системи, управлявани от статистически данни, базирани на мрежата“) не трябва да се „третират като недостатъци на дизайна, които могат да бъдат разгледани отделно“.

- **Технически дълг: „**понятие, използвано в софтуерното инженерство, за да опише допълнителните разходи, които ще трябва да бъдат платени в бъдеще в резултат на предприемане на пряк път при разработването на софтуерна система. Представен е през 1992 г. от Уорд Кънингам [...]. Използването на преки пътища по същество заимства от бъдещето, когато ще е необходима съществена преработка.

- **Етичен дълг**: „разходи за преработване на системите в състояние, което е в съответствие с настоящите социални очаквания“; „технически дълг, при който бъдещите разходи не се дължат на проблеми с техническата устойчивост, а на необходимостта от справяне с етични проблеми, като странични ефекти, наложени на потребителите.“

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Подходи към ИИ

„ **ИИ е историята за това как избягваме изграждането на скъпи модели на явления, които все още не разбираме**, като език и визия, **задоволявайки се просто със симулиране на специфични „умения** “ (като проверка на правописа или разпознаване на ръкопис) **чрез използване на откритите статистически корелации в големи масиви от данни**. Алгоритмите за машинно обучение и големи масиви от данни могат да се използват за намиране на тези ценни модели.

Това измества фокуса на изследователите от моделирането на поведението или умението, което трябва да се приложи (чрез разбиране на неговите основни механизми) към осигуряване на огромни количества наблюдения на това поведение, които биха могли да се използват като данни за обучение за статистически алгоритми за обучение.

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Подход №1: „корелацията е достатъчна“

- „Вече не ценим причината, поради която е взето решението, стига действието, което генерира, да е подходящо. **Прогнозите са по-важни от обясненията**, знанието „какво“ е по-важно от това „защо““;

- „Фокусът върху установяването и използването на причинно-следствените връзки е заменен с фокус върху установяването и използването на корелационни връзки“;

- „Докато този пряк път спестява огромните разходи за разбиране и изрично моделиране, **той създава друг разход – този за снабдяване с огромни маси от подходящи данни за обучение – и няма причина „a priori“ да очакваме, че този разход трябва да бъде по-малък.** Генерирането, поддържането и анотирането на висококачествени данни е значителен разход в няколко индустрии – например при тестването на лекарства. **Тази цена също е заобиколена от индустрията с ИИ**.”

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Подход №2: „данни от околната среда“

„Първият урок от обучението в уеб мащаб е да се използват наличните данни, вместо да се разчита на отбелязани данни, които не са налични. Например откриваме, че полезните семантични връзки могат да се научат от статистиката на уеб заявките или от натрупаните доказателства за уеб-базирани текстови модели и форматирани таблици, и в двата случая без да са необходими ръчно отбелязани данни.“

(A. Halevy, P. Norvig, and F. Pereira, *The unreasonable effectiveness of data*, in «IEEE Intelligent Systems», 24, 2, 2009, pp. 8–12.)

- „Данните, събрани от околната среда, са от решаващо значение при проектирането на системи за разпознаване на обекти, разпознаване на лица, машинен превод и т.н. Вездесъщите вграждания на думи, които ни позволяват да представим значението на думите, преди да ги обработим, също са научени от данни, събрани от околната среда.

- „Замяната на моделирането с данни и замяната на генерирането на данни със събирането им от околната среда отвежда дизайнерите на ИИ много близо до „безплатен обяд“ – но не съвсем. Често на алгоритъмът за обучение трябва да му бъде казано какво да прави и това става под формата на наблюдение.

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Подход №3: „заместители (прокси) и скрита обратна връзка“

„Вместо да питат изрично потребителите какво искат да направи системата с ИИ – отговорност, която много потребители не са склонни да поемат – дизайнерите започнаха да използват скрита обратна връзка, което е друг начин да се каже, че са заменили ненаблюдаемите количества с по-евтини заместители. “

- „Предположението е, че действията на потребителя разкриват неговите предпочитания или нужди, както биха били направени (или дори по-добре) чрез скрита обратна връзка. Проблем, който трябва да решим, е следствието от използването на **неправилно подравнени прокси сървъри** при обучението на автономни агенти.

- „Примери от потребителско поведение първо са използвани от агенти, за да се обучат общи явления, като правилен правопис. По-късно същите са използвани за свързване на най-подходящите попадения към дадена заявка. Накрая са използвани за извеждане на потребителските предпочитания на индивида. В процеса на обучение, между другото, **фокусът започна да се измества от обслужването на потребителите към обслужването на рекламодателите** .

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

„Тайната подправка“, която захранва текущата версия на ИИ, има **съществена съставка: образци на човешко поведение, често под формата на микроизбори , извършвани от милиони потребители**, които да се използват като проксита за по-скъпи сигнали; други съставки са включените алгоритми за статистическо обучение, мощна инфраструктура за събиране на данни и предоставяне на услуги.

„Рецептата, която ни даде тази версия на ИИ включва замяна на

- причинно-следствени връзки с корелации,
- конкретни модели със статистически корелации,
- избрани примери за обучение с данни от околната среда и
- изрични анотации на данни със скрити сигнали и други заместители.“

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

„Взети заедно, тези и други подходи ни позволяват да генерираме версия на автономни агенти на много ниска крайна цена. Сега трябва да се сблъскаме с дългосрочната стойност на тези решения, които причиниха част от „етичния дълг“, вграден в нашата инфраструктура с ИИ. **За да се гарантира справедливостта на машинните решения,** тяхната прозрачност, неприкосновеността на личния живот на потребителите и спазването на новите разпоредби и да се подсигурят услугите срещу наблюдение или враждебни манипулации, **ще има значителни разходи за преработка на технологията на фундаментално ниво. И в някои случаи е възможно да не сме в състояние да предоставим еквивалентни услуги по социално приемлив начин** - в този случай компромисите между точността и социалните ограничения ще трябва да бъдат ясно съобщени на законодателите и обществеността, така че решенията да могат да бъдат направени на подходящите места.“

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Средства, използвани при подходите

1. „**Използването на причинни (параметрични, интерпретируеми) модели в определени области може да бъде необходимо**, дори ако точността може да пострада, в името на прозрачността на решенията.

   Това би било политическо решение, а също и голяма промяна.

   Готови ли сме да изоставим агентите на черните кутии, да платим цената на конкретното моделиране и може би дори да се въздържим от използването им в определени области, където принципно не можем да разработим тези модели?

   Изглежда малко вероятно, но трябва да проведем този разговор, поне за избрани сектори.

   Има специфични области, в които потребителите имат право на обяснения за произтичащи решения и може да се наложи в тези области да да се използват само по-слаби, но с възможности за обчзнения инструменти на ИИ. [...] **Зоните, защитени от закони, би трявало да са правосъдие, здравеопазване, образование, финанси и други области.**

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

2. **„Обучение на ИИ за произволни данни (от околната среда)**“: „трябва поне да можем да добавим някои нюанси: може да има типове данни, които могат да се използват само за определени типове приложения. Може би даден текст може да бъде подходящ за обучение на агенти за корекция на правописа, но не и за научаване значението на чувствителни думи (може би защото произхожда от общност с много различни ценности от тези, които искаме да бъдат отразени в нашия агент). И дори може да се измисли вид сертификат, който да посочва този произход. Вече има конкретни списъци с домейни, където решенията се очаква да бъдат безпристрастни, и за тези домейни може да поискаме агентите на ИИ да бъдат обучени от по-добре подбрани източници на данни (което би могло да е по-скъпо, но прави неявните пристрастия явни).
   Трябва да се грижим за нашата „верига за доставка на данни“ толкова, колкото ни е грижа за снабдяването ни с храна. Веригата за доставка на данни може да се дефинира като поредица от процеси, включени в производството и разпространението на данни за обучение, които формират различните модели, намиращи се в настоящите системи за ИИ . Всеки модул може да се основава на различни набори от данни, всеки от които на свой ред потенциално оформен от други набори от данни.
   Готови ли сме да платим разходите за генериране, анотиране и куриране на скъпи набори от данни, съответстващи на строгостта, използвана за данните от клиничните изпитвания?“

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

3. „**Скрита обратна връзка**“: „нека си представим, че в определени области на **интелигентния агент може да бъде позволено да се учи само от изрични, директни и доброволни комуникации от потребителя, а не от наблюдение на поведението на потребителя.**

Това може да се използва в ситуации, в които има подозрение за филтри или поведенческа зависимост. Умишленото използване на психометрични сигнали, за да се стигне до извод как потребителят може да реагира на предложение, може да се наложи да бъде забранено, както и вероятно много форми на подтикване към конкретни резултати. Регулирането на използването на косвени сигнали от интелигентни агенти изглежда разумно искане.

Всичко това вероятно ще струва повече, вероятно ще намали производителността на нашите системи и лесната им употреба . И все пак, домейн по домейн, може да решим, че в някои случаи това е, което искаме. Това би било част от „изплащането“ на етичния дълг, създаден преди повече от десет години чрез поредица от подходи. Не трябва да демонизираме тези минали решения, тъй като днес нямаше да имаме индустрия с ИИ без тях, но сега е дошло времето да преразгледаме някои от тях.“

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

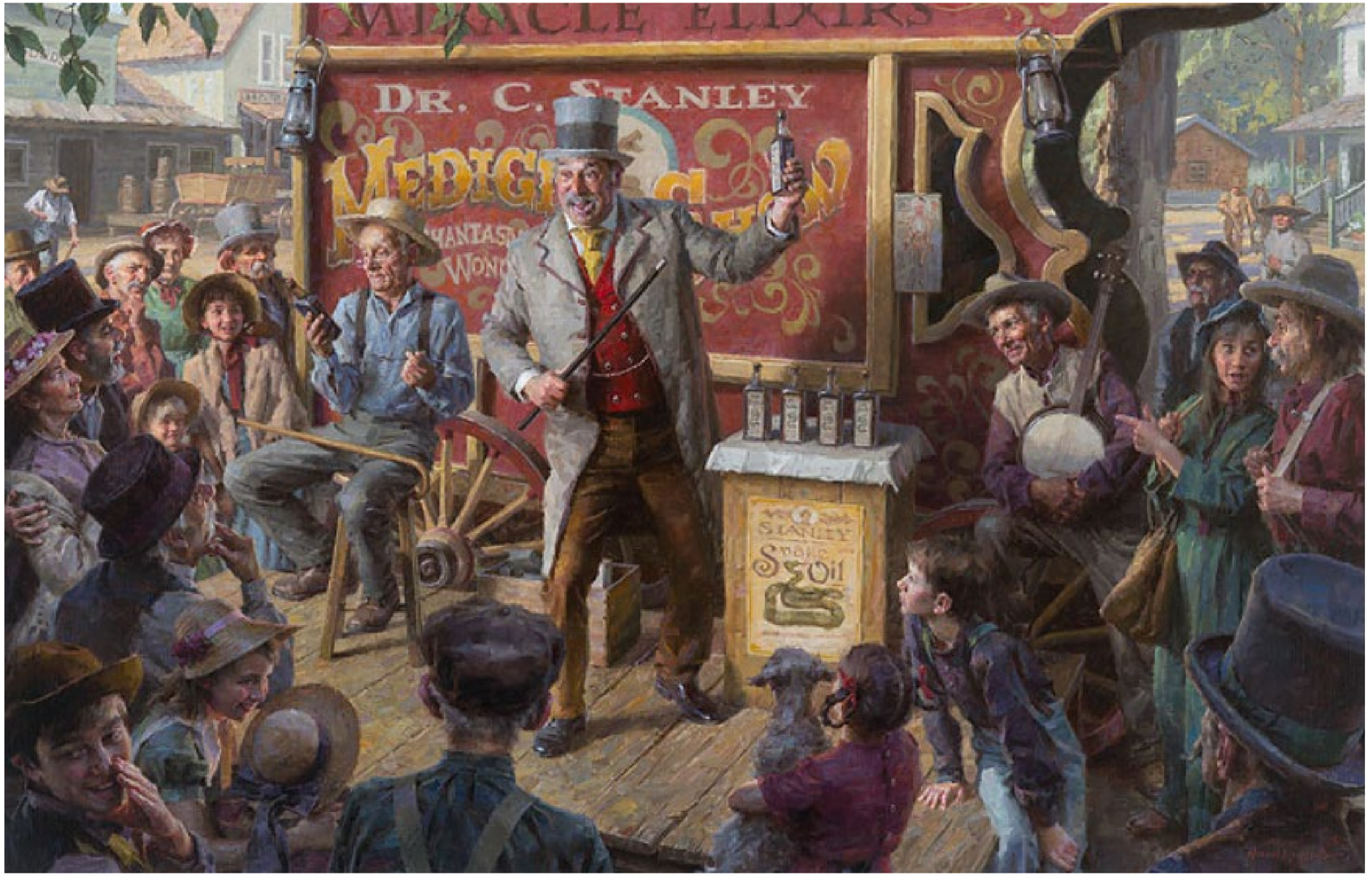
Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Хлъзгавия ИИ (змийска мас)

https://aisnakeoil.substack.com/

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

How to recognize AI snake oil

Arvind Narayanan

Associate Professor of Computer Science

@random_walker

PRINCETON UNIVERSITY

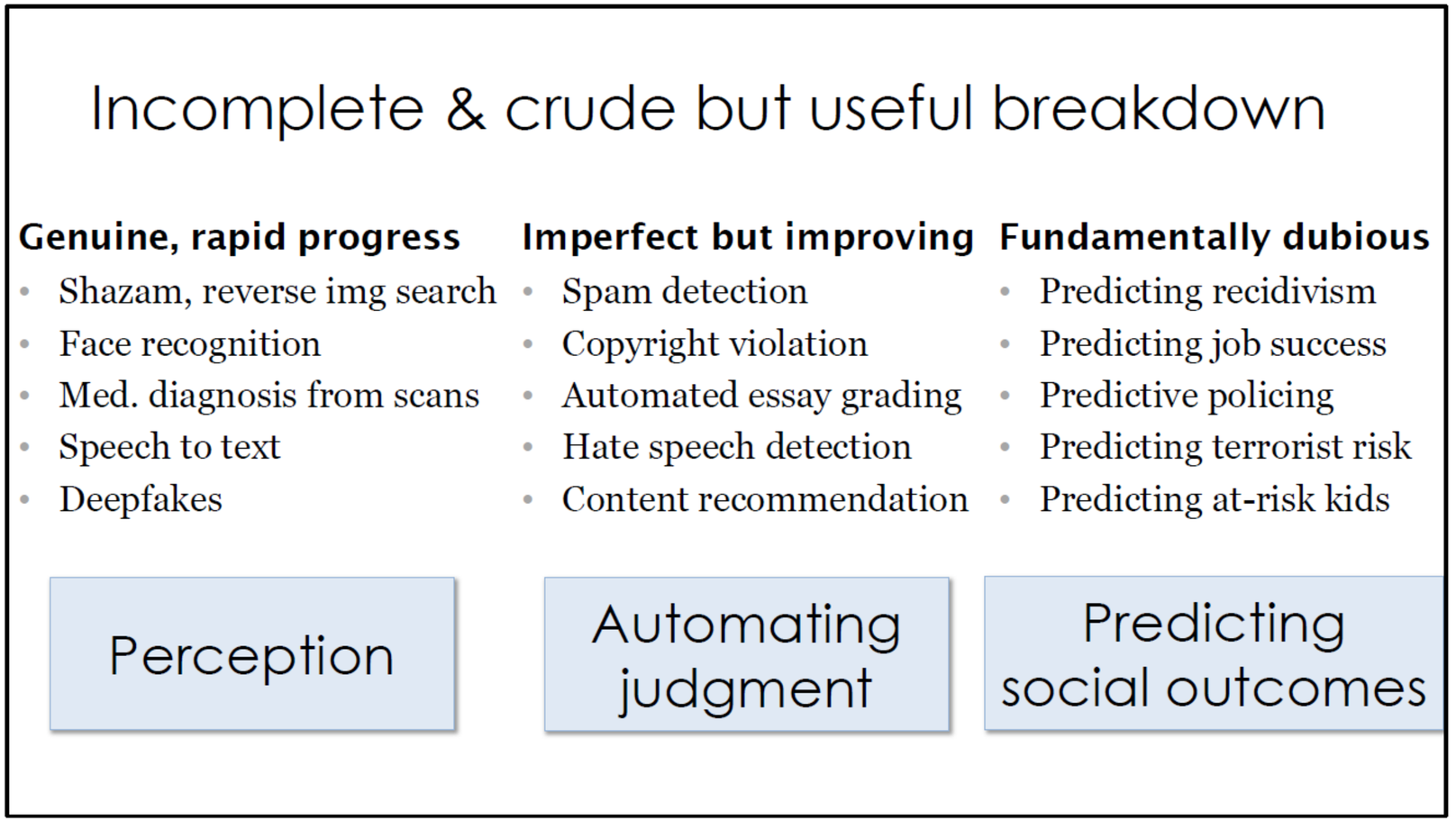
CITP
CENTER FOR INFORMATION TECHNOLOGY POLICY

Much of what's being sold as "AI" today is snake oil — it does not and cannot work.

Why is this happening? How can we recognize flawed AI claims and push back?

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423



https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Accuracy of recidivism prediction

COMPAS tool (137 features):	65% ± 1%	(slightly better than random)
Logistic regression (2 features):	67% ± 2%	

Age and number of priors

Dressel & Farid. *The accuracy, fairness, and limits of predicting recidivism*. Science Advances 2018.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

Системите с МО не могат да идентифицират характера на човек или да предскажат бъдещите му действия по-добре от астрологията.

Използването на непрозрачни системи за машинно обучение за откриване на характера или предвиждане на действията на лицата няма научни основания.

Самото използване на термина „предсказание“ е подвеждащо: системата за машинно обучение може да предскаже думи в последователности от текстови низове, но това по никакъв начин не означава, че тя може да предскаже бъдещето, и конкретно бъдещите действия на определени лица.

Просто вярването, че подобни прогнози са възможни, е равносилно на предположението, че измерението на времето в човешките дела е напълно без значение и че бъдещето ще бъде същото като миналото. Вместо това решението да се приеме миналото като модел, който да бъде възпроизведен в бъдещето, е еквивалентно на решението да се автоматизират неравенствата, както вече бе наблюдавано.

Идеята, че системите с МО са способни на такива прогнози, произтича от характеристика на магическото мислене: идеята, че всички връзки са значими, независимо от разграничението на причинно-следствените връзки, че всички детайли са значими и всичко обяснява всичко.

N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI,* ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423

MAI4CAREU

Master programmes in Artificial Intelligence 4 Careers in Europe

Благодаря за вниманието. Имате ли въпроси?

daniela.tafani@unibo.it

Co-financed by the European Union
Connecting Europe Facility

Този магистърски курс се изпълнява в контекста на действие
№ 2020-EU-IA-0087, съфинансирано от EU CEF Telecom
под GA nr. INEA/CEF/ICT/A2020/2267423