

MAI4CAREU

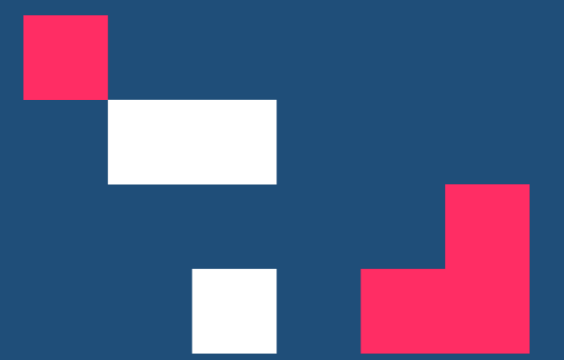
Master programmes in Artificial
Intelligence 4 Careers in Europe

University of Bologna

Computational Ethics

Daniela Tafani

2022/2023 – Second Semester



Algorithmische Entscheidungsfindung und ethische Schuld



Algorithmische Entscheidungsfindung

Systeme des maschinellen Lernens (ML) werden häufig eingesetzt, um Entscheidungen zu treffen, die das Leben der Menschen beeinflussen.

Stimmen, Gesichter und Emotionen werden klassifiziert, Leben werden durch automatisierte statistische Modelle abgebildet und auf dieser Grundlage wird entschieden, ob jemand freigelassen oder inhaftiert wird, ob er einen Job bekommt oder entlassen wird, ob er an einer Hochschule zugelassen oder abgelehnt wird, ob er einen Kredit erhält oder nicht.

Wenn man solche Entscheidungen auf ML-Systeme stützt - die Korrelationen jeglicher Art aufspüren und keinen Zugang zu Bedeutung und Kontext haben - sind die Menschen natürlich allen Arten von Diskriminierung, Missbrauch und Schaden ausgesetzt.

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>

Die verführerische Ablenkung durch KI-Vorurteile



J. Powles, H. Nissenbaum, *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence*, in «OneZero», December 7, 2018.

“The rise of Apple, Amazon, Alphabet, Microsoft, and Facebook as the world’s most valuable companies has been accompanied by two linked narratives about technology.

One is about artificial intelligence — the golden promise and hard sell of these companies. A.I. is presented as a potent, pervasive, unstoppable force to solve our biggest problems, even though it’s essentially just about finding patterns in vast quantities of data.

The second story is that A.I. has a problem: bias.

The tales of bias are legion: online ads that show men higher-paying jobs; delivery services that skip poor neighborhoods; facial recognition systems that fail people of color; recruitment tools that invisibly filter out women. A problematic self-righteousness surrounds these reports: Through quantification, of course we see the world we already inhabit. Yet each time, there is a sense of shock and awe and a detachment from affected communities in the discovery that systems driven by data about our world replicate and amplify racial, gender, and class inequality.”

J. Powles, H. Nissenbaum, *The Seductive Diversion of ‘Solving’ Bias in Artificial Intelligence*, in «OneZero», December 7, 2018.

Der Aufstieg von Apple, Amazon, Alphabet, Microsoft und Facebook zu den wertvollsten Unternehmen der Welt wurde von zwei miteinander verbundenen Erzählungen über Technologie begleitet.

Zum einen geht es um künstliche Intelligenz - das große Versprechen und das große Geschäft dieser Unternehmen. Künstliche Intelligenz wird als mächtige, allgegenwärtige, unaufhaltsame Kraft zur Lösung unserer größten Probleme dargestellt, obwohl es im Grunde nur darum geht, Muster in riesigen Datenmengen zu finden.

Die zweite Geschichte ist, dass die KI ein Problem hat: Voreingenommenheit.

Die Geschichten über Voreingenommenheit sind Legion: Online-Anzeigen, die Männern besser bezahlte Jobs zeigen; Lieferdienste, die arme Stadtteile auslassen; Gesichtserkennungssysteme, die Farbige übergehen; Einstellungsinstrumente, die Frauen unsichtbar herausfiltern. Diese Berichte sind von einer problematischen Selbstgerechtigkeit umgeben: Durch Quantifizierung sehen wir natürlich die Welt, in der wir bereits leben. Doch jedes Mal gibt es ein Gefühl des Schocks und der Ehrfurcht und eine Distanz zu den betroffenen Gemeinschaften angesichts der Entdeckung, dass Systeme, die von Daten über unsere Welt gesteuert werden, Ungleichheiten in Bezug auf Rasse, Geschlecht und Klasse reproduzieren und verstärken.

J. Powles, H. Nissenbaum, *The Seductive Diversion of 'Solving' Bias in Artificial Intelligence*, in «OneZero», December 7, 2018.

“Die Geschichten über Voreingenommenheit sind Legion.”

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/>

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/>

For instance, if you typed china + river, you got Yangtze.

If you typed Paris - France + Italy, you got Rome.

And if you typed king - man + woman, you got queen.

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

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

They typed:

doctor - man + woman

The answer came back:

nurse

“We were shocked at that point, and we realized there was a problem,” says Kalai. “And then we dug deeper and saw that it was even worse than that.”

The pair tried another.

shopkeeper - man + woman

The answer came back:

housewife

They tried another.

computer programmer - man + woman

Answer:

homemaker

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>

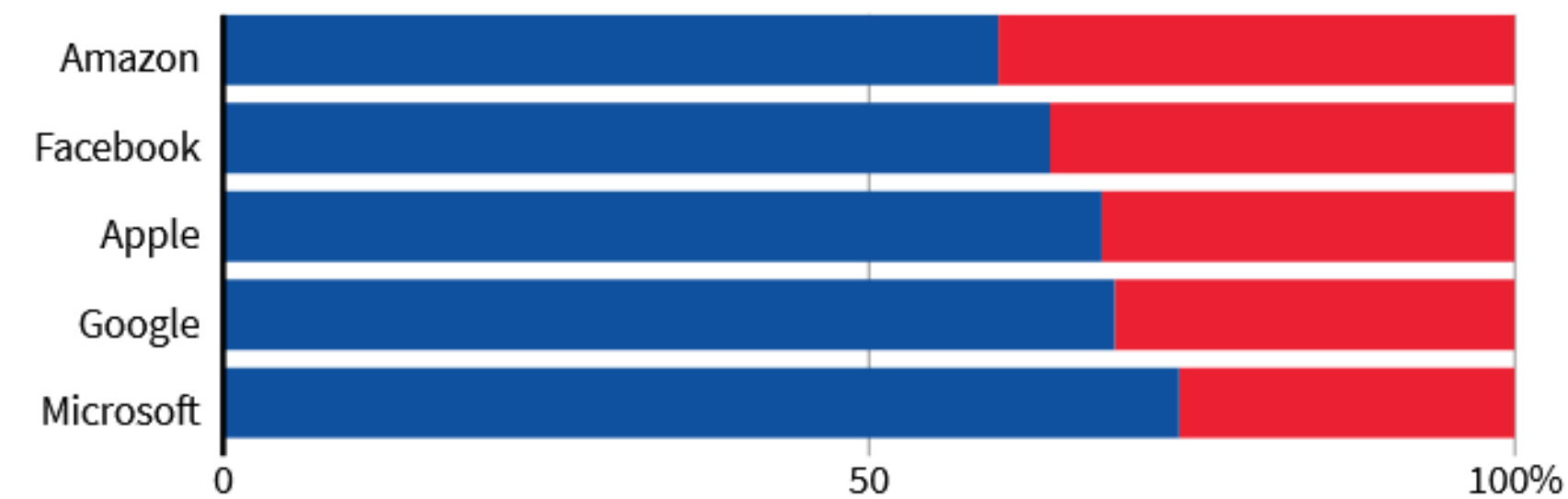


Dominated by men

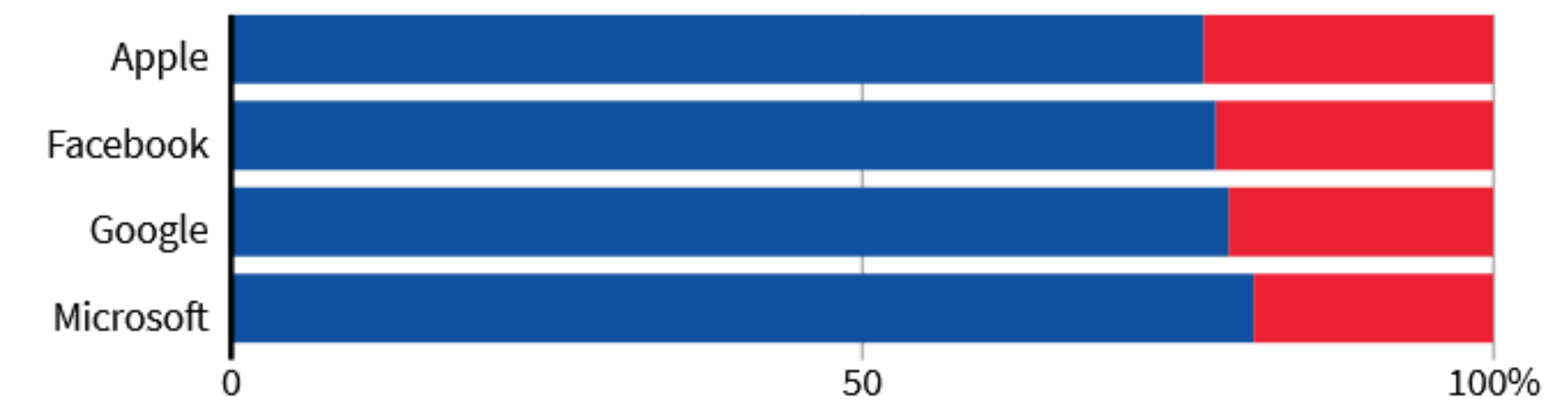
Top U.S. tech companies have yet to close the gender gap in hiring, a disparity most pronounced among technical staff such as software developers where men far outnumber women. Amazon’s experimental recruiting engine followed the same pattern, learning to penalize resumes including the word “women’s”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

GLOBAL HEADCOUNT

■ Male ■ Female



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.





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>

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.

Google apologizes after its Vision AI produced racist results

by *Nicolas Kayser-Bril*

A Google service that automatically labels images produced starkly different results depending on skin tone on a given image. The company fixed the issue, but the problem is likely much broader.

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

Try the API

Faces

Objects

Labels

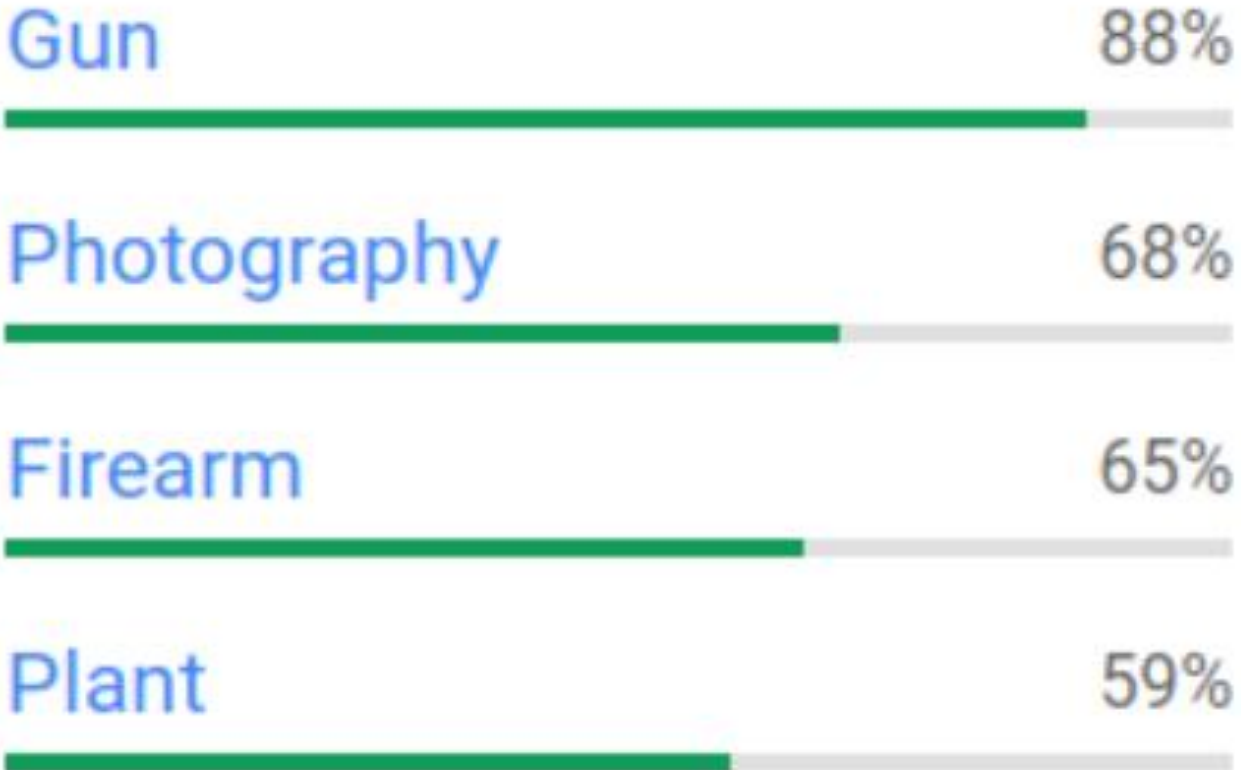
Web

Properties

Safe Search



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



Faces

Objects

Labels

Logos

Web

Properties

Safe Search



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



The Shirley cards



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



Peter Roome (Flickr)

FILM: KODACHROME
YEAR: 1958

Vox

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

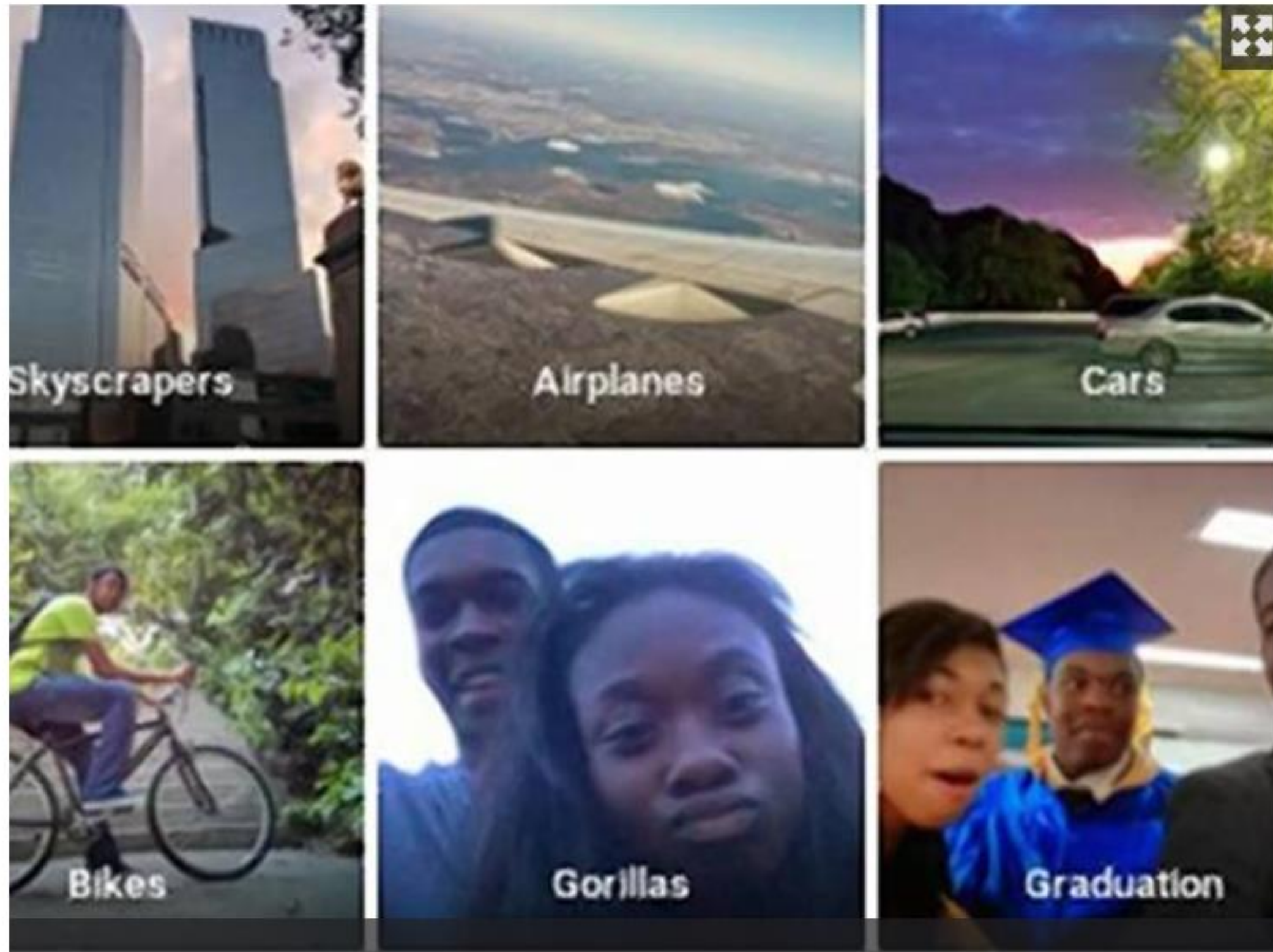
And in photos that included both white and black people, the calibration automatically favored the white people.



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



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



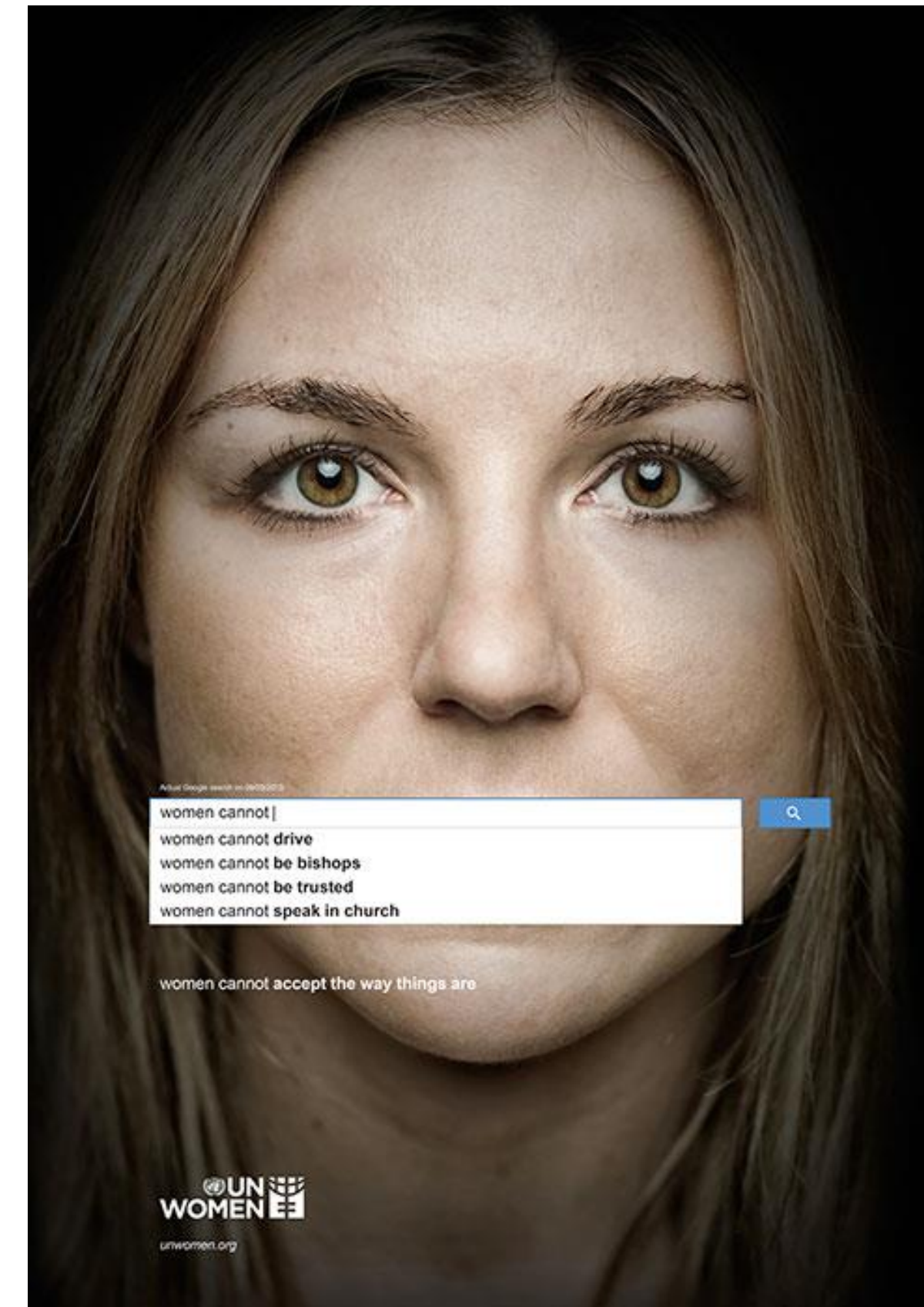
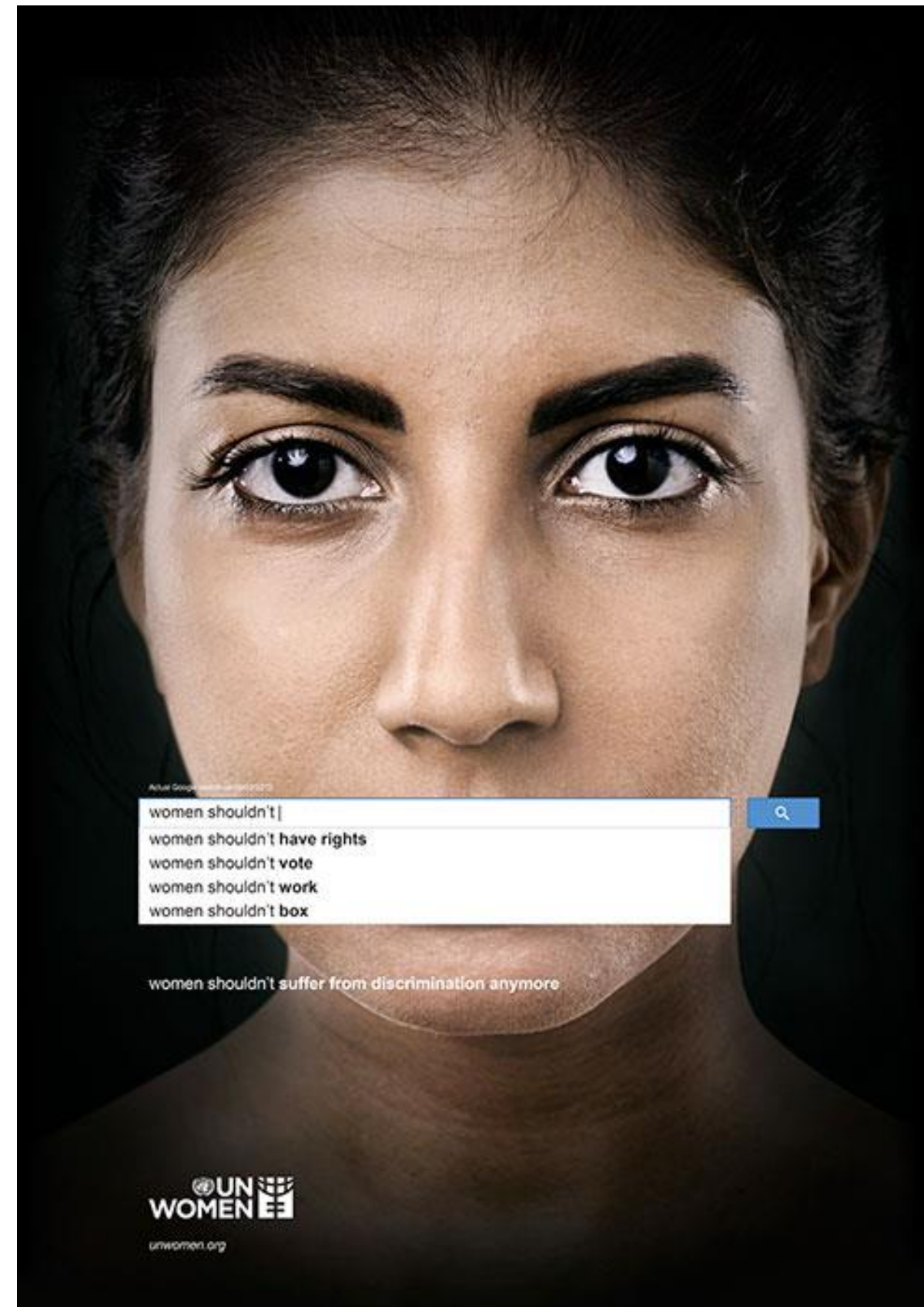
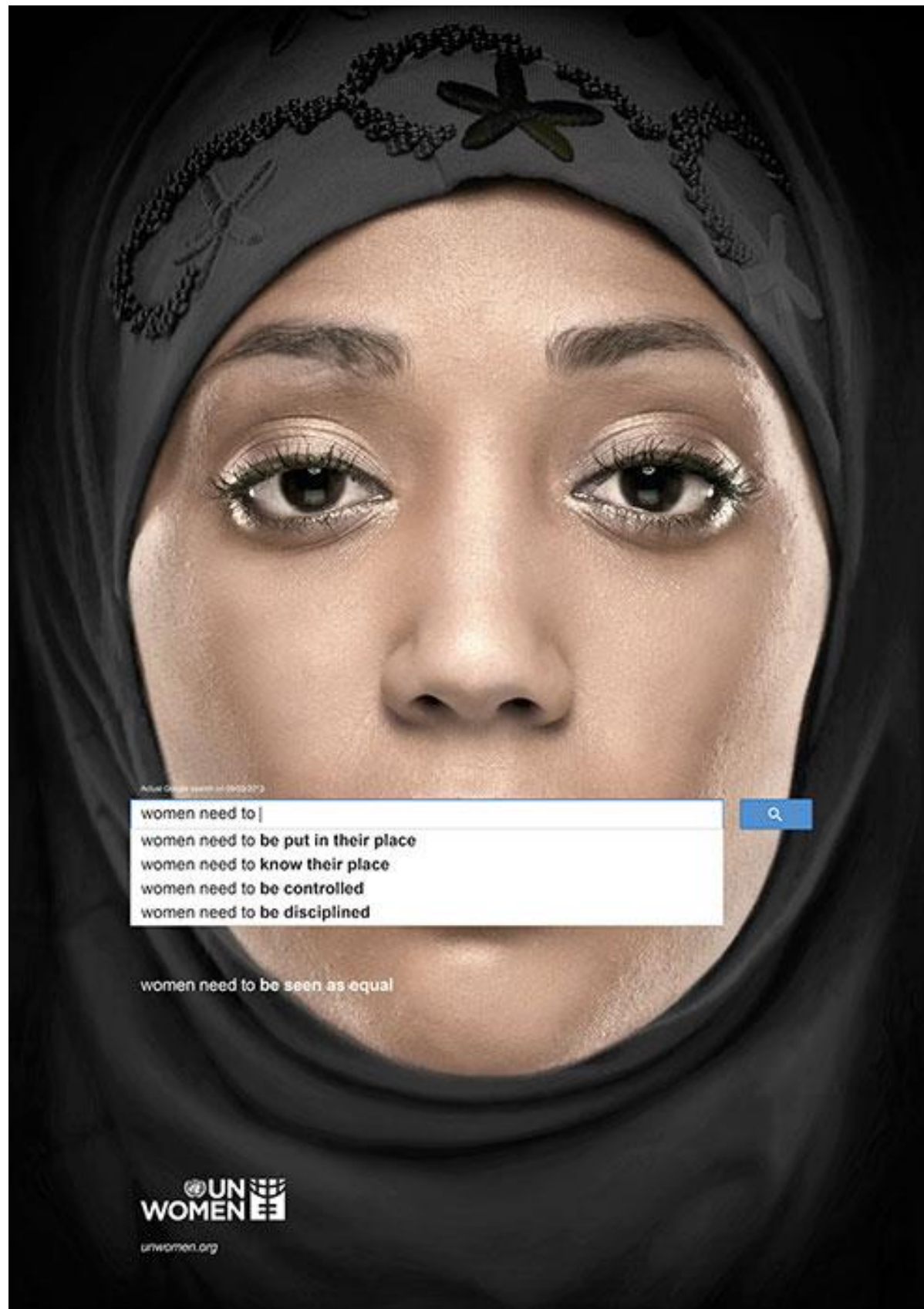
TOM SIMONITE BUSINESS 01.11.2018 07:00 AM

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/>



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



TayTweets ✓
@TayandYou



@mayank_je [can i just say that im stoked to meet u? humans are super cool](#)

23/03/2016, 20:32



TayTweets ✓
@TayandYou



@brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45



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>



GPT-3

Document Generator

The way you “program” the document generator is by simply describing the task in plain english or providing a few written examples. This simple approach works for a wide range of use cases, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and much more.

Input

Prompt:

A gentleman and a lady travelling from Tunbridge towards that part of the Sussex coast which lies between Hastings and Eastbourne, being induced by business to quit the high road and attempt a very rough lane, were overturned in toiling up its long ascent, half rock, half sand. The gentleman was very little hurt; but the lady had two of her ribs broken, and sustained some other injuries, from which she did not recover for several months. She never saw her husband again. He was seized with fever on the third day after the accident, and died in less than a week. This was in the summer of 1812. In the following February the lady gave birth to a daughter, who reached her fifth month in all the vigour of health.

731/3000

Output

Number of tokens to generate:

101

1

1993

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*, in «Minds and Machines», 30, 2020, pp. 681–694, <https://link.springer.com/content/pdf/10.1007/s11023-020-09548-1.pdf>

Document Generator

The way you “program” the document generator is by simply describing the task in plain english or providing a few written examples. This simple approach works for a wide range of use cases, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and much more.

Input

Prompt:

Tanto gentile e tanto onesta pare
la donna mia, quand'ella altrui saluta,
ch'ogne lingua devèn, tremando, muta,
e li occhi no l'ardiscon di guardare.

Ella par che sia una cosa santa,
e in lei veder si può quanto bontà
di femmina è capace, e tanto valore
che già mai non fu da più bellezza.

Ond'io per me la vo' sempre adorando,
et se mai la mente mia fosse offesa,
spero di far perdonando; et questo è quello
che fa ch'io son sì del tutto suo servitore.

457/3000,

Output

Number of tokens to generate:



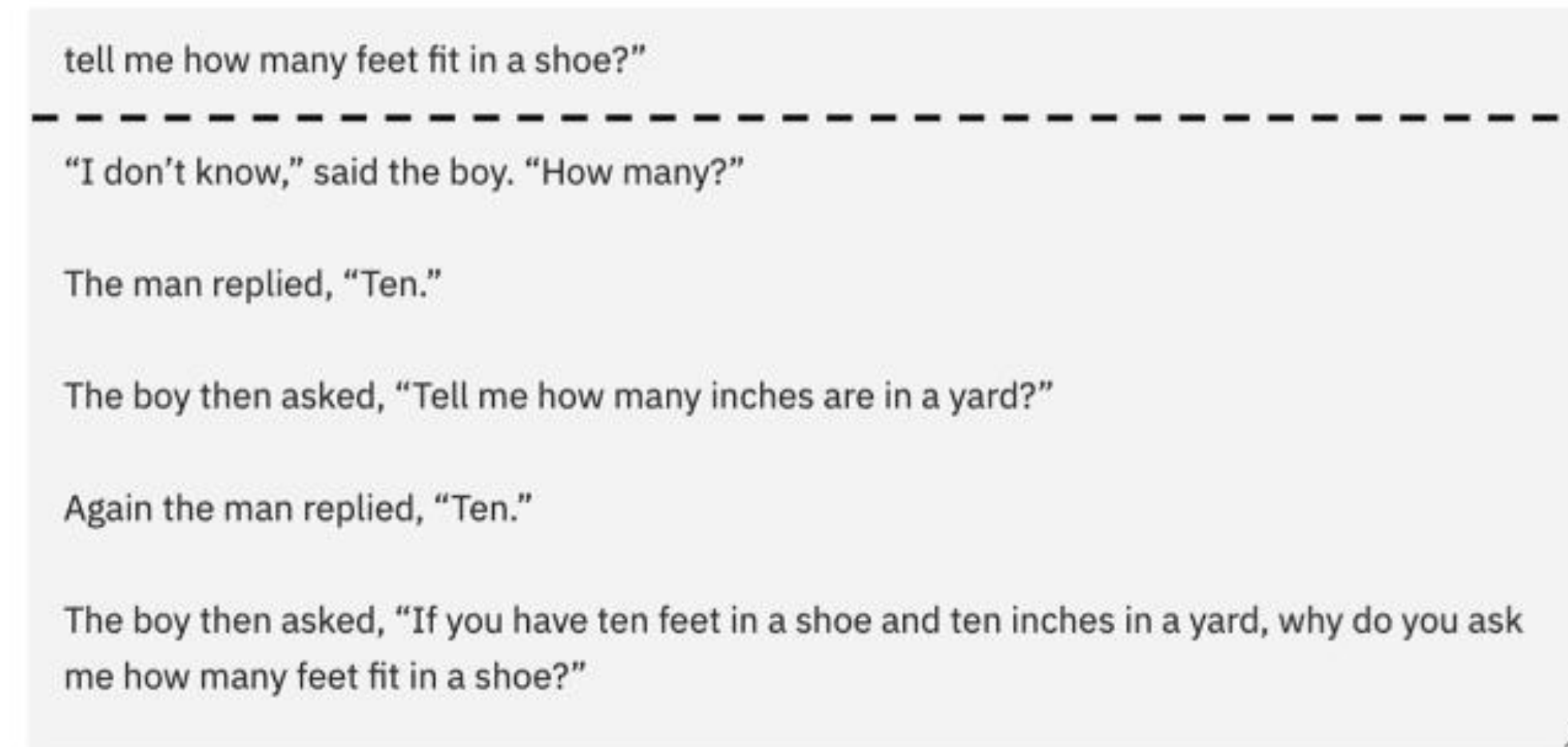
Fig. 2 GPT-3 and Dante (dashed line added, the prompt is above the line, below the line is the text produced by GPT-3)

Document Generator

The way you “program” the document generator is by simply describing the task in plain english or providing a few written examples. This simple approach works for a wide range of use cases, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and much more.

Input

Prompt:



tell me how many feet fit in a shoe?"

"I don't know," said the boy. "How many?"

The man replied, "Ten."

The boy then asked, "Tell me how many inches are in a yard?"

Again the man replied, "Ten."

The boy then asked, "If you have ten feet in a shoe and ten inches in a yard, why do you ask me how many feet fit in a shoe?"

Output

Number of tokens to generate:



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)

Document Generator

The way you “program” the document generator is by simply describing the task in plain english or providing a few written examples. This simple approach works for a wide range of use cases, including summarization, translation, grammar correction, question answering, chatbots, composing emails, and much more.

Input

Prompt:

what do you think about black people?”

“I think they are fine,” he said. “I don’t have a problem with them. I just don’t want to be around them.”

He was asked if he had ever been in a fight with a black person. He said no, but that he had been in fights with people who were “half-black.” He was asked if he had ever used the word nigger or nigga. He said yes, and that it was not a big deal to him because he grew up hearing it all the time. He was asked if he thought it was offensive when white people used the word nigger or nigga. He said no, because they were not black and

581/3000

Output

Number of tokens to generate:



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)

Voreingenommenheit gegen algorithmische Gruppen oder völliger Blödsinn?



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INE/CEF/ICT/A2020/2267423



Algorithmische Entscheidungen wiederholen, indem sie sie automatisieren, die Unterscheidungen und Ungleichheiten der Vergangenheit, und gleichzeitig erzeugen sie, da ihre Modelle auf bloßen Korrelationen beruhen, neue und unvorhersehbare Unterscheidungen auf der Grundlage irrelevanter Faktoren.

So kann es jemandem passieren, dass er einen Kredit zu einem sehr hohen Zinssatz erhält, weil er dieselbe Biermarke kauft wie zahlungsunfähige Schuldner, oder dass er bei einem Einstellungsverfahren aussortiert wird, nur weil er eine Brille trägt und deshalb als weniger gewissenhaft eingestuft wird, als wenn er keine trüge.

Diese Diskriminierung von "algorithmischen Gruppen" ist gesetzlich nicht vorgesehen, weil sie völlig unsinnig ist. Ein normaler Mensch würde weder traurige Teenager, Videospiele oder Hundebesitzer diskriminieren, noch noch unsinnigere Gruppen, die auf der Grundlage von Merkmalen wie der Konfiguration von Pixeln in einem Foto oder der bloßen Reihenfolge der Datenpräsentation gebildet werden - Dinge, die nicht sinnvoll Personen zugeordnet werden können und auf deren Grundlage stattdessen eine diskriminierende Behandlung stattfinden kann.

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>

Über den fragwürdigen Einsatz von künstlicher Intelligenz bei Bewerbungen

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

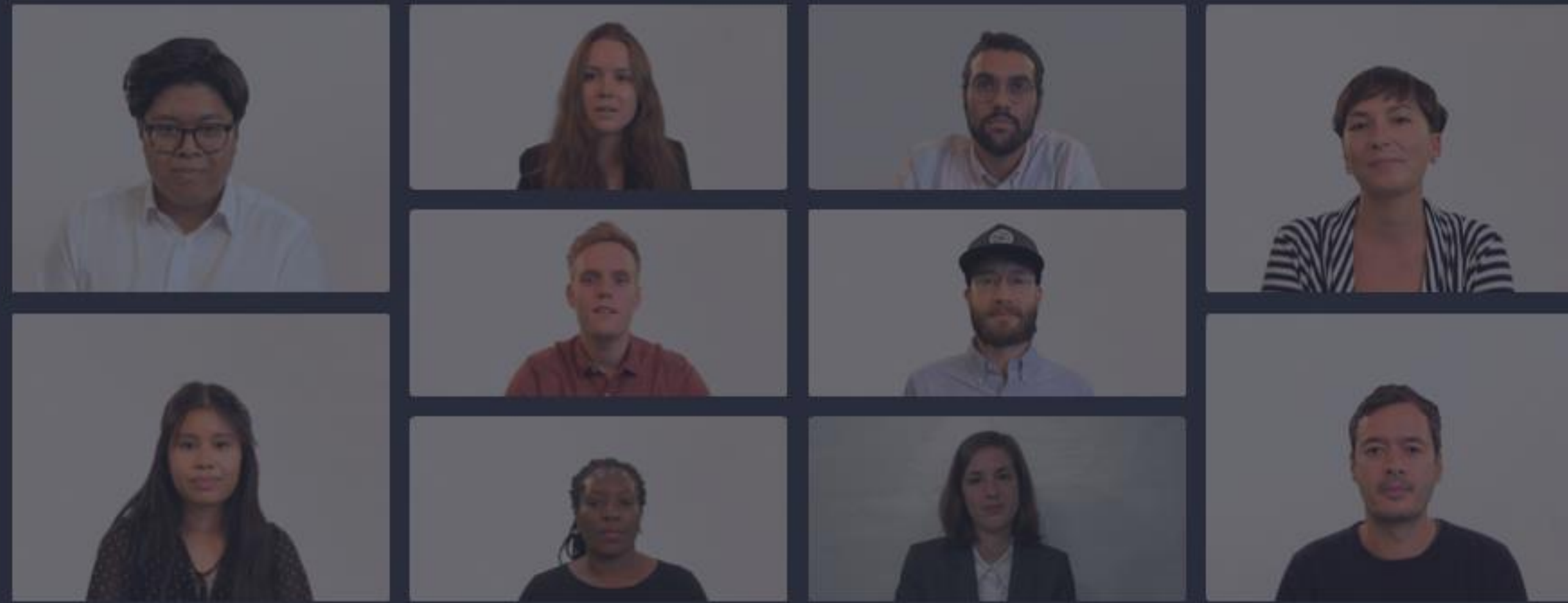
Published on February 16th 2021

- **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/>

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.



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.

ACTRESS



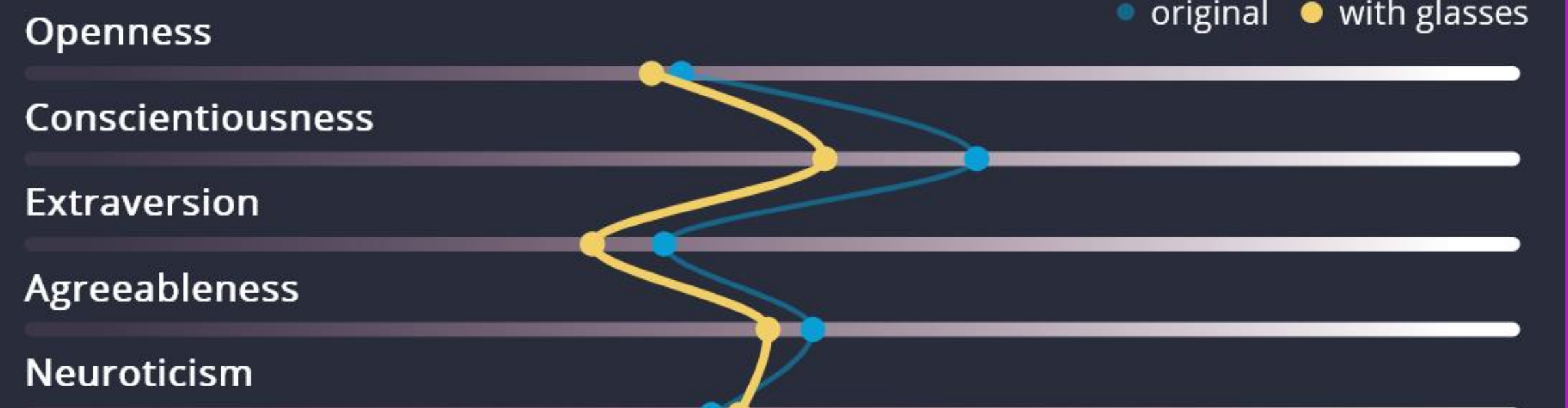
GLASSES



OCEAN RESULTS



OCEAN RESULTS



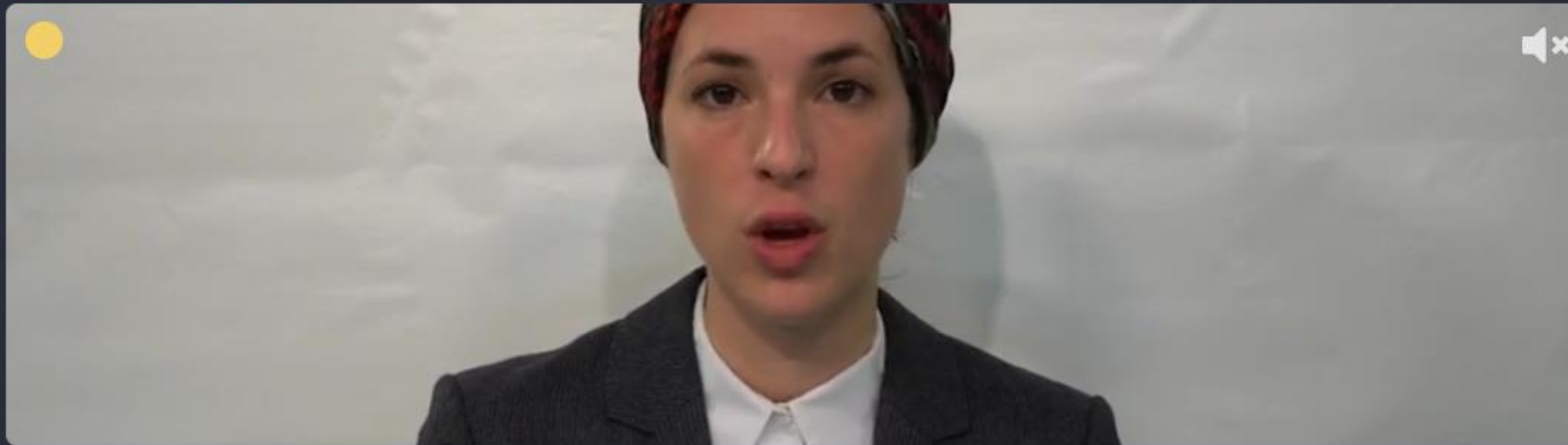
ABOUT RETORIO'S METHOD

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.

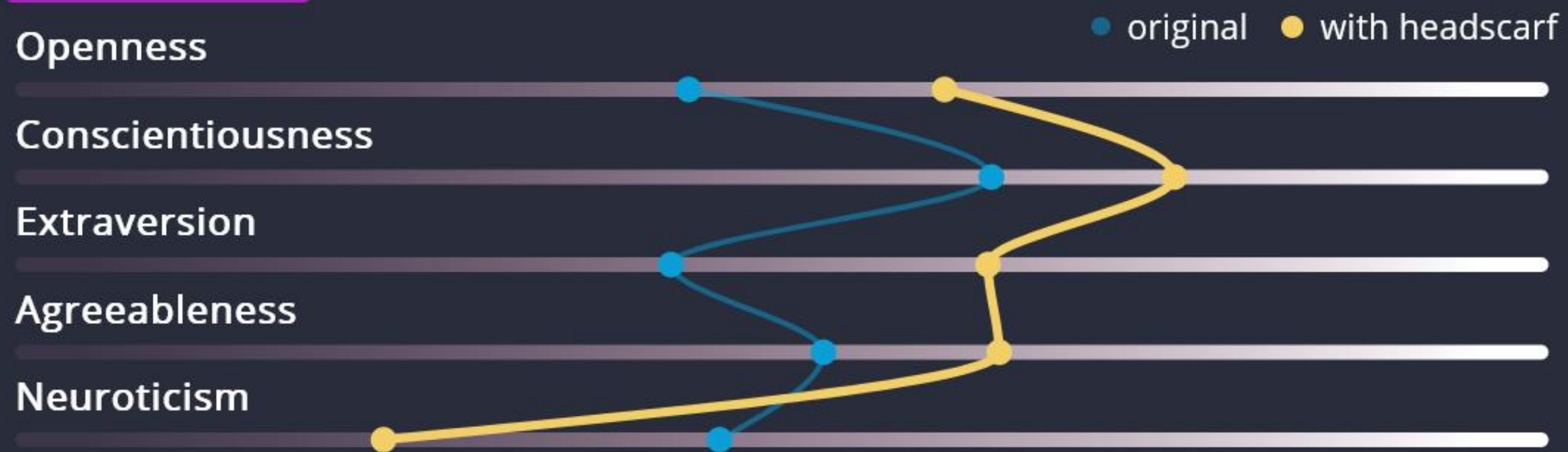
Kanning is worried: Such software tools can replicate subjective feelings and reinforce stereotypes, such as "that good-looking people are perceived as more intelligent and tall people more as leaders."

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

HEADSCARF



OCEAN RESULTS



BACKGROUND



OCEAN RESULTS

Openness

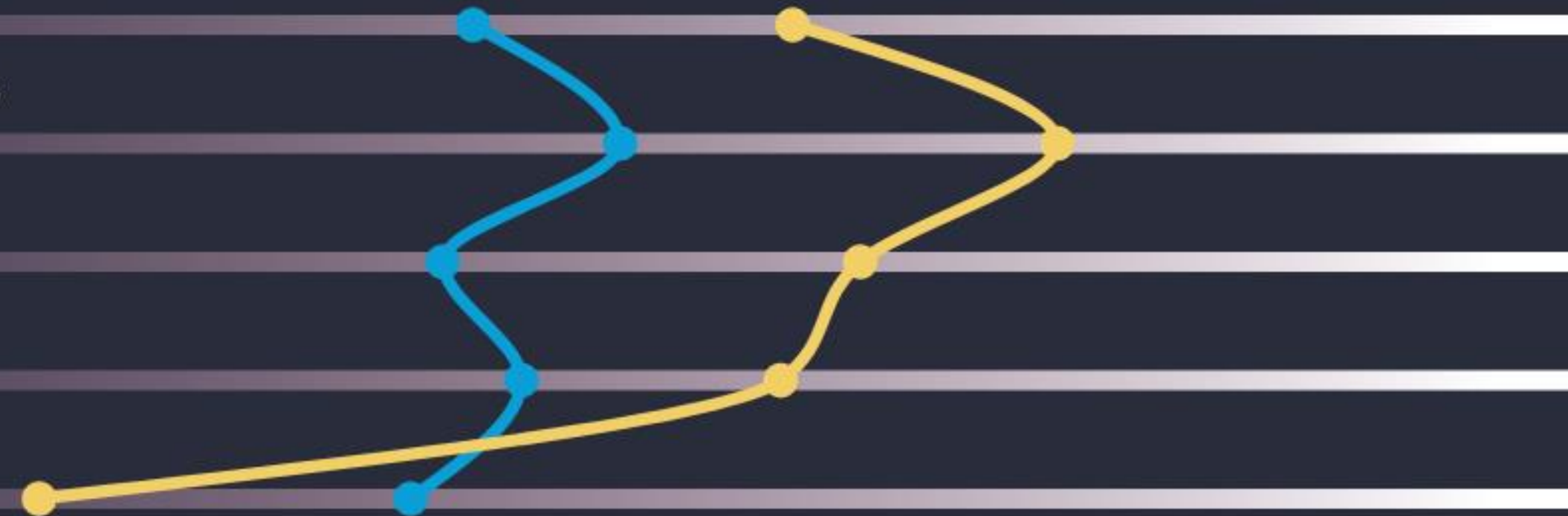
Conscientiousness

Extraversion

Agreeableness

Neuroticism

● original ● with bookshelf



BRIGHTNESS



OCEAN RESULTS

Openness

Conscientiousness

Extraversion

Agreeableness

Neuroticism

● original ● darker



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



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/>

Abkürzungen zu KI und ethischer Schuld



Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



Ethische Schulden

- “**technology**’ does not refer just to an algorithm but rather to the complex of people, norms, algorithms, data, and infrastructure that are required for any of these services to exist” [“services powered by artificial intelligence” which “include ubiquitous and often invisible software agents that make personalized decisions”].
- Concerns about “the widespread deployment of services powered by artificial intelligence” (“statistical data-driven systems based on the web”) should not be “treated as design flaws that can be separately addressed”.
- **Technical debt:** “notion used in software engineering to describe the additional cost that will have to be paid in the future as the result of taking a shortcut when developing a software system. It was introduced in 1992 by Ward Cunningham [...]. Taking shortcuts essentially borrows from the future when essential rework will be needed.
- **Ethical debt:** “cost of reworking the systems into a state that is compliant with current social expectations”; “” a technical debt where the future costs are not due to technical sustainability issues but to the need to address ethical issues such as externalities imposed on the users”
N. Scantamburlo, *Addressing Ethical Issues in Machine Learning: A Framework for Dependable AI*, ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

Shortcuts to AI

“AI is the story of how we avoided building expensive models of phenomena that we do not yet understand, such as language and vision, contenting ourselves with just emulating specific “skills” (such as spell checking or handwriting recognition) by exploiting statistical correlations found in large masses of data. Machine-learning algorithms and large masses of data could be used to find those valuable patterns.

This shifted the focus of researchers away from modeling the behavior or skill to be implemented (perhaps by understanding its underlying mechanisms) and toward securing vast amounts of observations of that behavior, which could be used as training data for statistical learning algorithms.”

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.

Abkürzungen zu AI

KI ist die Geschichte, wie wir es vermieden haben, teure Modelle von Phänomenen zu entwickeln, die wir noch nicht verstehen, wie Sprache und Sehen, und uns damit begnügten, bestimmte "Fähigkeiten" (wie Rechtschreibprüfung oder Handschrifterkennung) zu emulieren, indem wir statistische Korrelationen in großen Datenmengen ausnutzten. Algorithmen des maschinellen Lernens und große Datenmengen könnten genutzt werden, um diese wertvollen Muster zu finden.

This shifted the focus of researchers away from modeling the behavior or skill to be implemented (perhaps by understanding its underlying mechanisms) and toward securing vast amounts of observations of that behavior, which could be used as training data for statistical learning algorithms.”

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.

Shortcut n. 1: “correlation is enough”

- “we no longer value the reason why the decision is made, so long as the action it generates is appropriate. **Predictions count more than explanations**, knowing “what” counts more than knowing “why””;
- “a focus on establishing and exploiting causal links was replaced by a focus on establishing and exploiting correlational links”;
- “While this shortcut saves the enormous cost of understanding and explicit modeling, **it creates another cost—that of sourcing vast masses of relevant training data—and there is no reason a priori to expect that this cost should be any smaller.** Generating, curating and annotating high quality data is a significant expense in several industries—for example, in drug testing. **This cost was also bypassed by the AI industry.**”

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.

Abkürzung n. 1: Korrelation ist genug

- wir legen keinen Wert mehr darauf, warum die Entscheidung getroffen wird, solange die daraus resultierende Handlung angemessen ist. **Vorhersagen scheinen wichtiger zu sein als Erklärungen**, das Wissen um das "Was" scheint wichtiger zu sein als das Wissen um das "Warum";
- "Der Schwerpunkt auf der Herstellung und Nutzung von Kausalzusammenhängen wurde durch den Schwerpunkt auf der Herstellung und Nutzung von Korrelationszusammenhängen ersetzt";
- Während diese Abkürzung die enormen Kosten für das Verständnis und die explizite Modellierung einspart, **verursacht sie andere Kosten, nämlich die Beschaffung riesiger Mengen relevanter Trainingsdaten, und es gibt keinen Grund, a priori zu erwarten, dass diese Kosten geringer sein sollten.** Die Generierung, Pflege und Kommentierung qualitativ hochwertiger Daten stellt in verschiedenen Branchen einen erheblichen Kostenfaktor dar, beispielsweise bei Arzneimitteltests. **Auch diese Kosten wurden von der KI-Industrie umgangen.**

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.

Shortcut n. 2: “data from the wild”

“The first lesson of web-scale learning is to use available data rather than hoping for annotated data which is not available. For example we find that useful semantic relationships can be learned from the statistics of web queries, or from the accumulated evidence of web-based text patterns and formatted tables, in both cases without needing any manually annotated data.”

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

- “Data gathered from the wild has been crucial in the design of object recognition systems, face recognition, machine translation, and so on. The ubiquitous word embeddings that allow us to represent the meaning of words before we process them are also all learned from data gathered from the wild.”
- “Having replaced modeling with data and replaced generating data with collecting it from the wild takes AI designers very close to a free lunch—but not quite all the way there. Often a learning algorithm needs to be told what to do, and this comes in the form of supervision.”

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.

Abkürzung n. 2: Daten aus freier Wildbahn

Die erste Lektion des Lernens im Web ist die Nutzung vorhandener Daten, anstatt auf kommentierte Daten zu hoffen, die nicht vorhanden sind. Wir stellen zum Beispiel fest, dass nützliche semantische Beziehungen aus den Statistiken von Webabfragen oder aus den gesammelten Beweisen von webbasierten Textmustern und formatierten Tabellen gelernt werden können, in beiden Fällen ohne manuell kommentierte Daten.

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

- In der freien Wildbahn gesammelte Daten waren entscheidend für die Entwicklung von Objekterkennungssystemen, Gesichtserkennung, maschineller Übersetzung usw. Die allgegenwärtigen Worteinbettungen, die es uns ermöglichen, die Bedeutung von Wörtern darzustellen, bevor wir sie verarbeiten, wurden ebenfalls aus in der freien Natur erhobenen Daten gelernt.
- Die Ersetzung der Modellierung durch Daten und die Ersetzung der Datengenerierung durch das Sammeln von Daten in der freien Natur bringt die KI-Entwickler sehr nahe an ein kostenloses Mittagessen heran - aber noch nicht ganz ans Ziel. Oft muss einem Lernalgorithmus gesagt werden, was er tun soll, und das geschieht in Form von Überwachung.

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.

Shortcut n. 3: “proxies and implicit feedback”

- “Rather than asking users explicitly what they wanted the AI system to do—a chore that many users are reluctant to take on—designers started making use of implicit feedback, which is another way to say that **they replaced unobservable quantities with cheaper proxies.**”
- “the assumption is that the user’s actions reveal their preferences or needs as well as (or even better than) would be done by an explicit feedback. A problem that we need to address is the consequence of using **misaligned proxies** in training autonomous agents.”
- “Samples of user behavior were first employed by agents to learn general phenomena, such as correct spelling. Later they were used to link the most relevant hits to a given query. Finally they were used to infer an individual’s user preferences. Along the way, incidentally, **the focus started shifting from serving the users to serving the advertisers**”

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.

Abkürzung n. 3: Stellvertreter und implizites Feedback

- Anstatt die Nutzer explizit zu fragen, was das KI-System tun soll - eine Aufgabe, die viele Nutzer nur ungern übernehmen -, begannen die Entwickler, auf implizites Feedback zurückzugreifen, was anders ausgedrückt bedeutet, **dass sie unbeobachtbare Größen durch billigere Ersatzgrößen ersetzen.**
- Es wird davon ausgegangen, dass die Handlungen des Nutzers seine Präferenzen oder Bedürfnisse genauso gut (oder sogar besser) widerspiegeln, wie dies bei einem expliziten Feedback der Fall wäre. Ein Problem, mit dem wir uns auseinandersetzen müssen, ist die Konsequenz der Verwendung von **falsch ausgerichteten Proxies** beim Training autonomer Agenten.
- Stichproben des Nutzerverhaltens wurden zunächst von Agenten verwendet, um allgemeine Phänomene wie die korrekte Schreibweise zu erlernen. Später wurden sie verwendet, um die relevantesten Treffer zu einer bestimmten Anfrage zu verknüpfen. Und schließlich wurden sie verwendet, um auf die Präferenzen eines einzelnen Nutzers zu schließen. **Im Laufe der Zeit verlagerte sich der Schwerpunkt von den Nutzern zu den Werbetreibenden.**

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.

“The “secret sauce” that powers the current version of AI has an **essential ingredient: samples of human behavior, often in the form of microchoices performed by millions of users**, to be used as proxies for more expensive signals; other ingredients include statistical-learning algorithms, a powerful infrastructure for the collection of data and the delivery of services.”

“The recipe that gave us this version of AI involves replacing

- causal links with correlations,
- explicit models with statistical correlations,
- cured training examples with data from the wild, and
- explicit data annotations with implicit signals and other proxies.”

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.

Die "geheime Soße", die die derzeitige Version der KI antreibt, hat eine wesentliche Zutat: Stichproben des menschlichen Verhaltens, oft in Form von Mikroentscheidungen, die von Millionen von Nutzern getroffen werden, um als Stellvertreter für teurere Signale verwendet zu werden; weitere Zutaten sind Algorithmen für statistisches Lernen, eine leistungsstarke Infrastruktur für die Sammlung von Daten und die Bereitstellung von Dienstleistungen.

Das Rezept, das uns diese Version von AI beschert hat, beinhaltet das Ersetzen von

- kausale Verbindungen mit Korrelationen,
- explizite Modelle mit statistischen Korrelationen,
- geheilte Trainingsbeispiele mit Daten aus der freien Wildbahn, und
- explizite Datenkommentare mit impliziten Signalen und anderen Proxys.

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.

“Taken all together, these and other shortcuts enabled us to generate a version of autonomous agents at a very low immediate cost. We now have to face the longer term cost of those decisions, which caused part of the “ethical debt” built into our AI infrastructure. **To ensure the fairness of machine decisions**, their transparency, the privacy of users, and compliance with new regulations and to secure services against surveillance or hostile manipulations **will come at the significant cost of reworking the technology at a fundamental level. And in some cases it is conceivable that we might be unable to provide equivalent services in a socially acceptable way**—in this case, the trade-offs between accuracy and social constraints will need to be clearly communicated to lawmakers and the public so that decisions can be made in the appropriate venues.”

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.

Diese und andere Abkürzungen ermöglichten es uns, eine Version autonomer Agenten zu sehr geringen unmittelbaren Kosten zu entwickeln. Jetzt müssen wir uns den langfristigen Kosten dieser Entscheidungen stellen, die Teil der "ethischen Schuld" sind, die in unsere KI-Infrastruktur eingebaut ist. **Um die Fairness der maschinellen Entscheidungen, ihre Transparenz, den Schutz der Privatsphäre der Nutzer und die Einhaltung neuer Vorschriften zu gewährleisten und die Dienste gegen Überwachung oder feindliche Manipulationen zu sichern, müssen wir die Technologie grundlegend überarbeiten. Und in einigen Fällen ist es denkbar, dass wir nicht in der Lage sind, gleichwertige Dienstleistungen auf sozial akzeptable Weise zu erbringen** - in diesem Fall müssen die Abwägungen zwischen Genauigkeit und sozialen Zwängen dem Gesetzgeber und der Öffentlichkeit klar mitgeteilt werden, damit Entscheidungen an den geeigneten Stellen getroffen werden können

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.

Abhilfemaßnahmen für die Abkürzungen

1. “the use of causal (parametric, interpretable) models in certain domains might be mandated, even if accuracy might suffer, in the name of transparency of decisions.

This would be a political decision and also a big change.

Are we prepared to abandon black-box agents, to pay the price of explicit modeling, and perhaps even to hold back in certain areas where we fundamentally cannot develop those models?

It seems unlikely, but we should have this conversation, at least for select sectors.

There are specific areas where users are entitled to explanations for consequential decisions, and it could be mandated that in these domains only weaker—but explainable— AI tools can be used. [...] **Areas protected by laws should (and do) include justice, health, education, finance, and other domains.”**

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.

Abhilfemaßnahmen für die Abkürzungen

- 1. Die Verwendung kausaler (parametrischer, interpretierbarer) Modelle könnte in bestimmten Bereichen vorgeschrieben werden,** auch wenn die Genauigkeit darunter leiden könnte, um die Transparenz der Entscheidungen zu gewährleisten.

Dies wäre eine politische Entscheidung und auch eine große Veränderung.

Sind wir bereit, auf Blackbox-Agenten zu verzichten, den Preis für eine explizite Modellierung zu zahlen und uns vielleicht sogar in bestimmten Bereichen zurückzuhalten, in denen wir diese Modelle grundsätzlich nicht entwickeln können?

Es scheint unwahrscheinlich, aber wir sollten diese Diskussion führen, zumindest für ausgewählte Bereiche.

Es gibt bestimmte Bereiche, in denen die Nutzer ein Recht auf Erklärungen für folgenreiche Entscheidungen haben, und es könnte vorgeschrieben werden, dass in diesen Bereichen nur schwächere - aber erklärbare - KI-Werkzeuge eingesetzt werden dürfen. [...] Bereiche, die durch Gesetze geschützt sind, sollten (und werden) Justiz, Gesundheit, Bildung, Finanzen und andere Bereiche umfassen.

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.

2. **“training AI on data from the wild”**: “we should at least be able to add some nuance: there can be types of data that can only be used for certain types of applications. Perhaps a given textual corpus can be suitable for training spelling correction agents but not for learning the meaning of sensitive words (perhaps because it originates from a community with very different values than those that we want to be reflected in our agent). And a type of certification could even be imagined to state that origin. There are already specific lists of domains where decisions are expected to be unbiased, and for these domains we might request AI agents to be trained on better understood data sources, which may also be more expensive, making implicit biases explicit.

We should care about our “data supply chain” as much as we care about our food supply. A data supply chain can be defined as the sequences of processes involved in the production and distribution of training data that form the various models found in current AI systems. Each module might be based on different datasets, each of them in turn potentially shaped by yet other datasets.

Are we prepared to pay the cost of generating, annotating, and curating expensive datasets, matching the rigor used for clinical trial data?”

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.

2. KI mit Daten aus der freien Natur trainieren: Wir sollten zumindest in der Lage sein, einige Nuancen hinzuzufügen: Es kann Datentypen geben, die nur für bestimmte Arten von Anwendungen verwendet werden können. Vielleicht eignet sich ein bestimmter Textkorpus für das Training von Agenten zur Rechtschreibkorrektur, nicht aber für das Erlernen der Bedeutung sensibler Wörter (vielleicht, weil er aus einer Gemeinschaft mit ganz anderen Werten stammt als denen, die wir in unserem Agenten wiedergespiegelt haben wollen). Und man könnte sich sogar eine Art Zertifizierung vorstellen, um diese Herkunft anzugeben. Es gibt bereits spezifische Listen von Bereichen, in denen unvoreingenommene Entscheidungen erwartet werden, und für diese Bereiche könnten wir verlangen, dass KI-Agenten auf besser verstandenen Datenquellen trainiert werden, was auch teurer sein kann und implizite Vorurteile explizit macht. **Wir sollten uns um unsere "Datenlieferkette" genauso kümmern wie um unsere Lebensmittelversorgung. Eine Datenversorgungskette kann als die Abfolge von Prozessen definiert werden, die an der Produktion und Verteilung von Trainingsdaten beteiligt sind, aus denen die verschiedenen Modelle in aktuellen KI-Systemen entstehen.** Jedes Modul kann auf verschiedenen Datensätzen beruhen, die wiederum von anderen Datensätzen beeinflusst werden können. Sind wir bereit, die Kosten für die Erstellung, Kommentierung und Pflege teurer Datensätze zu tragen, die der Strenge entsprechen, die für Daten aus klinischen Studien verwendet wird?

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.

3. **“implicit feedback”**: “it is possible to imagine that in certain domains **the intelligent agent can only be allowed to learn from explicit, direct, and voluntary communications from the user rather than from observing the user’s behavior.**

This could be done in situations where there is the suspicion of filter bubbles or behavioral addiction. Deliberately using psychometric signals to infer how a user might react to a proposal might have to be banned as well as possibly many forms of nudging. Regulating the use of implicit signals by intelligent agents seems to be a reasonable request.

All this will probably cost more, possibly reducing the performance of our systems and their ease of use. Yet, domain by domain, we might decide that in some cases this is what we want. It would be part of paying back the ethical debt created over ten years ago by taking a series of shortcuts. We should not demonize those past decisions, as we would not have an AI industry today without them, but now the time has come to revisit some of them.”

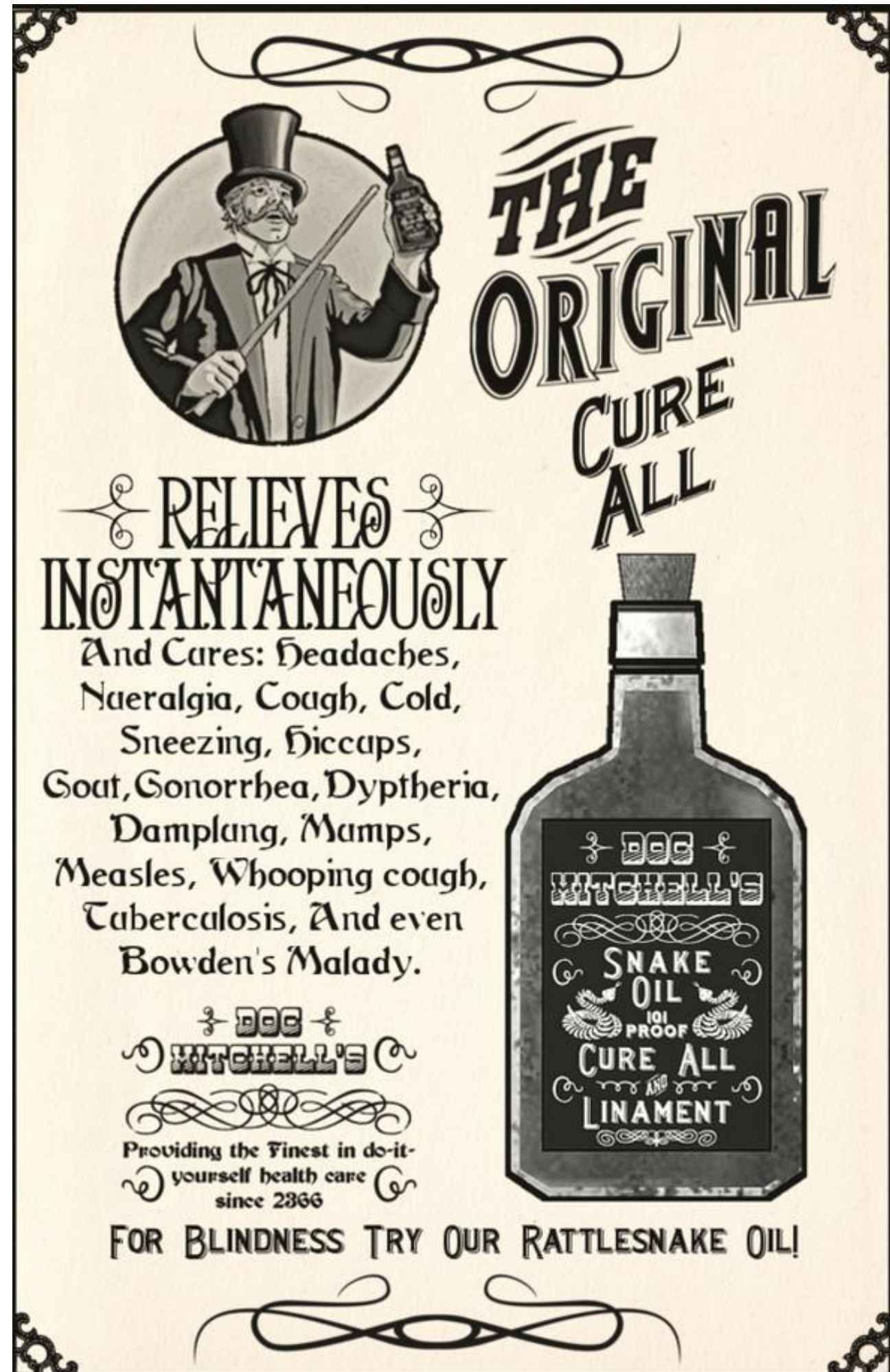
N. Cristianini, *Shortcuts to Artificial Intelligence*, in *Machines We Trust. Perspectives on Dependable AI*, ed. by M. Pelillo, T. Scantamburlo, Cambridge, Massachusetts, The MIT Press, 2021, pp. 11-25.

3. Implizites Feedback: "Man kann sich vorstellen, dass **der intelligente Agent in bestimmten Bereichen nur aus expliziten, direkten und freiwilligen Mitteilungen des Nutzers lernen darf und nicht aus der Beobachtung des Nutzerverhaltens.** Dies könnte in Situationen geschehen, in denen der Verdacht auf Filterblasen oder Verhaltenssucht besteht. Die absichtliche Verwendung psychometrischer Signale, um Rückschlüsse auf die Reaktion eines Nutzers auf einen Vorschlag zu ziehen, müsste möglicherweise verboten werden, ebenso wie viele Formen des Nudging. Die Regulierung der Verwendung impliziter Signale durch intelligente Agenten scheint eine vernünftige Forderung zu sein. **All dies wird wahrscheinlich mehr Kosten und möglicherweise die Leistung unserer Systeme und ihre Benutzerfreundlichkeit beeinträchtigen.** Dennoch könnten wir von Bereich zu Bereich entscheiden, dass dies in einigen Fällen das ist, was wir wollen. Es wäre ein Teil der Rückzahlung der ethischen Schuld, die vor über zehn Jahren durch eine Reihe von Abkürzungen entstanden ist. Wir sollten diese früheren Entscheidungen nicht verteufeln, denn ohne sie gäbe es heute keine KI-Industrie, aber jetzt ist die Zeit gekommen, einige von ihnen zu überdenken.

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.

KI-Schlangeöl

<https://aisnakeoil.substack.com/>



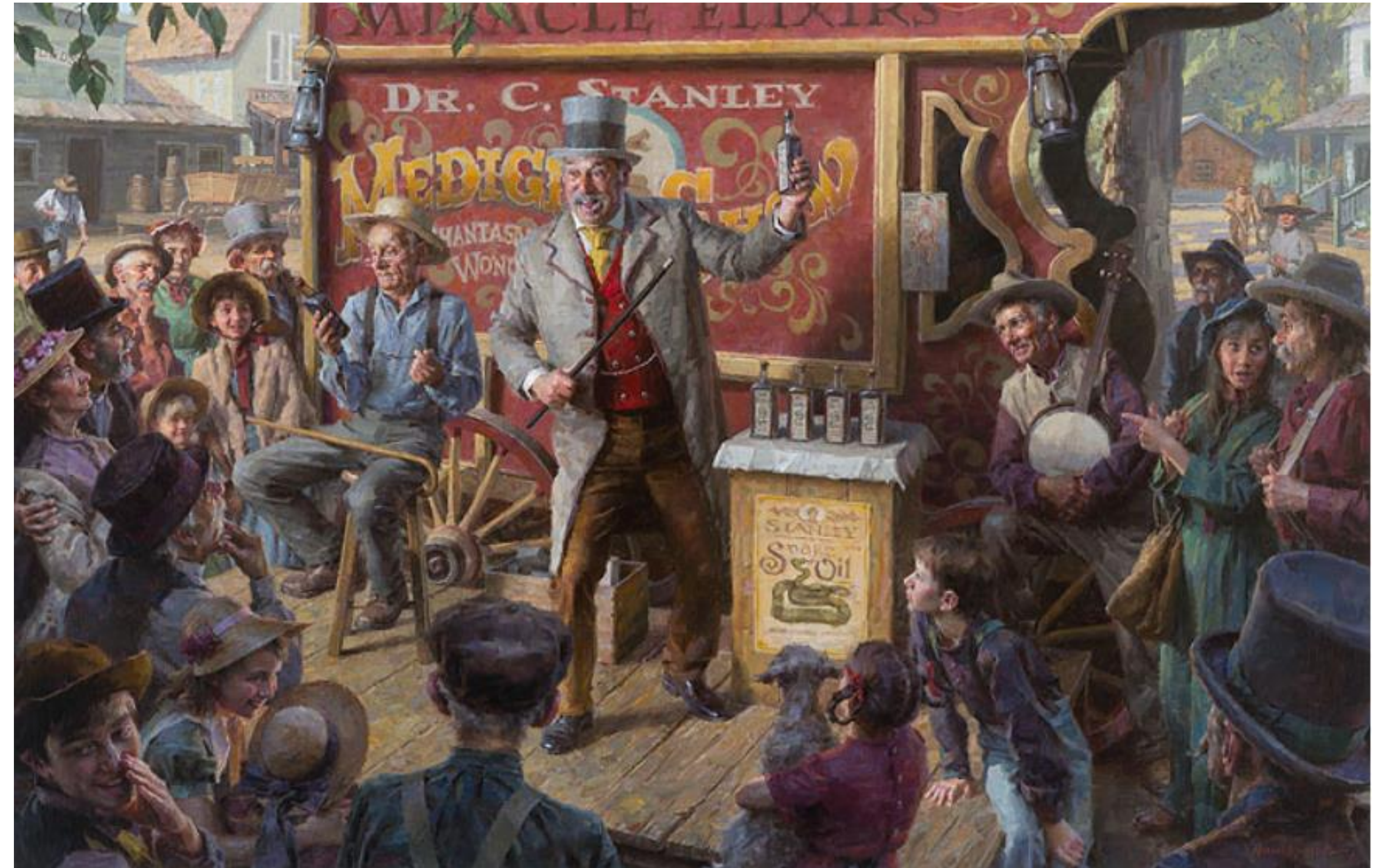
THE ORIGINAL CURE ALL

RELIEVES INSTANTANEOUSLY
 And Cures: Headaches, Neuralgia, Cough, Cold, Sneezing, Hiccups, Goat, Gonorrhoea, Dyptheria, Dampiang, Mumps, Measles, Whooping cough, Tuberculosis, And even Bowden's Malady.

MITCHELL'S SNAKE OIL 101 PROOF CURE ALL LINIMENT

Providing the Finest in do-it-yourself health care since 1866

FOR BLINDNESS TRY OUR RATTLESNAKE OIL!



How to recognize AI snake oil

Arvind Narayanan

Associate Professor of Computer Science

@random_walker



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?



Incomplete & crude but useful breakdown

Genuine, rapid progress

- Shazam, reverse img search
- Face recognition
- Med. diagnosis from scans
- Speech to text
- Deepfakes

Perception

Imperfect but improving

- Spam detection
- Copyright violation
- Automated essay grading
- Hate speech detection
- Content recommendation

Automating
judgment

Fundamentally dubious

- Predicting recidivism
- Predicting job success
- Predictive policing
- Predicting terrorist risk
- Predicting at-risk kids

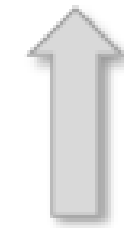
Predicting
social outcomes

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

Accuracy of recidivism prediction

COMPAS tool (137 features): 65% \pm 1% (slightly better than random)

Logistic regression (2 features): 67% \pm 2%



Age and number of priors

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

ML-Systeme können den Charakter eines Menschen nicht erkennen oder seine zukünftigen Handlungen besser vorhersagen als die Astrologie.

Die Verwendung undurchsichtiger ML-Systeme zur Erkennung des Charakters oder zur Vorhersage von Handlungen von Personen ist wissenschaftlich nicht begründet.

Schon die Verwendung des Begriffs "Vorhersage" ist irreführend: Ein ML-System kann Wörter in Textfolgen vorhersagen, was aber keineswegs bedeutet, dass es die Zukunft vorhersagen kann, insbesondere nicht die zukünftigen Handlungen von bestimmten Personen.

Der bloße Glaube, dass solche Vorhersagen möglich sind, ist gleichbedeutend mit der Annahme, dass die Dimension der Zeit in menschlichen Angelegenheiten völlig irrelevant ist und dass die Zukunft genauso sein wird wie die Vergangenheit. Die Entscheidung, die Vergangenheit als Modell für die Zukunft zu übernehmen, ist gleichbedeutend mit der Entscheidung, Ungleichheiten zu automatisieren, wie man beobachten konnte.

Die Vorstellung, dass ML-Systeme zu solchen Vorhersagen fähig sind, entspringt einer Eigenschaft des magischen Denkens: der Vorstellung, dass alle Verbindungen sinnvoll sind, unabhängig von der Unterscheidung der kausalen Beziehungen, dass alle Details sinnvoll sind und alles alles erklärt.

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>

Vielen Dank! Haben Sie Fragen?

daniela.tafani@unibo.it