AI – smart but sexist?

Pull the plug on gender bias

RICAIP seminar

2026 March 17, 1pm, online

Dr. Kinga Schumacher, kinga.schumacher@dfki.de

KI für den Menschen.

Dr. Kinga Schumacher

LinkedIn

AI scientist at the German Research Center for Artificial Intelligence

Computer scientist

Already fascinated by AI during hiers studies, attended all courses.

AI research and consultancy

Scientist at the German Research Center for Artificial Intelligence.
One year in consulting as an AI expert, then back to research.

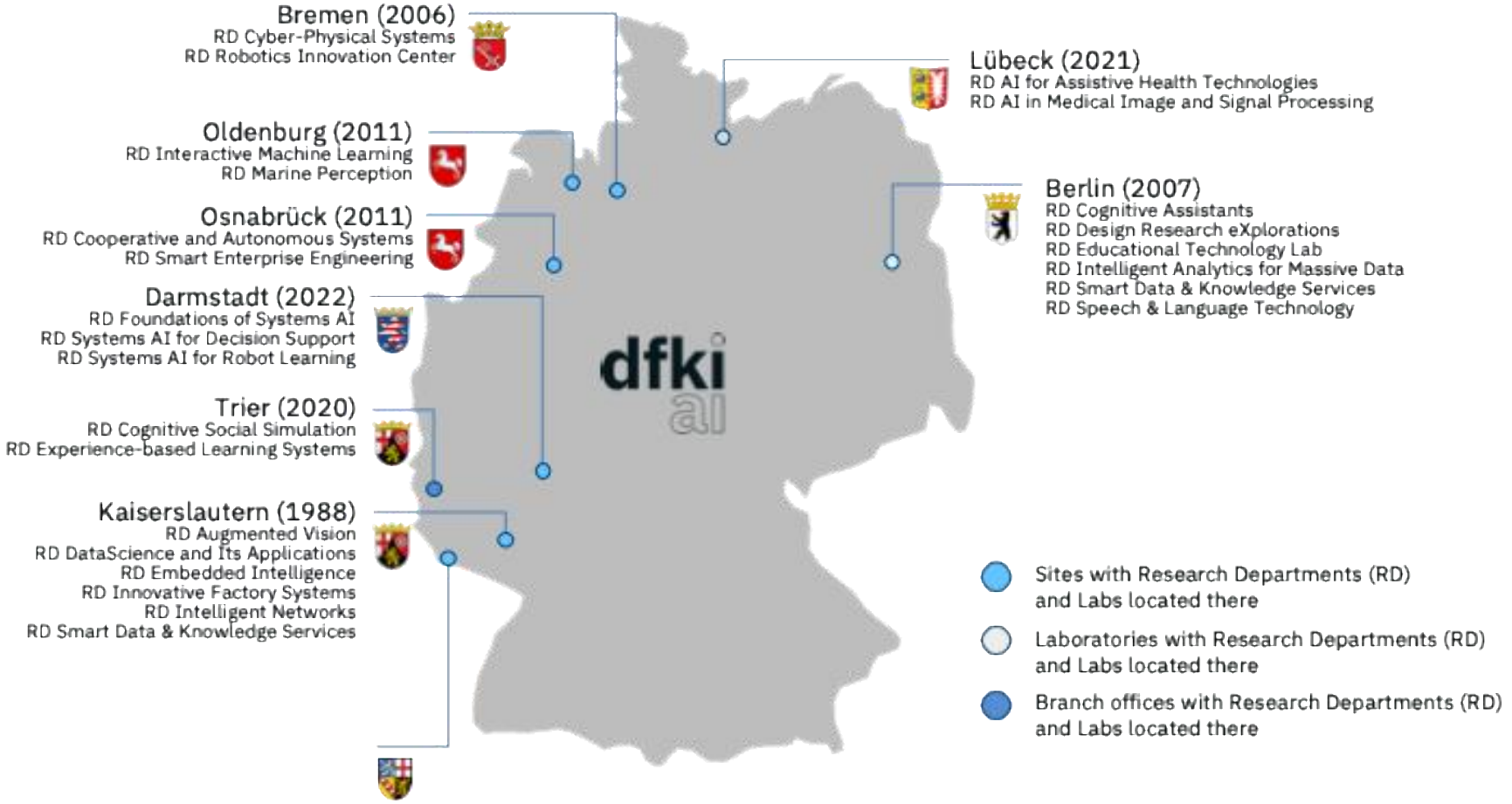
Topics

"diversity-aware" AI, human-machine interaction, AI landscape (methods and capabilities), regulatory activities of Germany and the EU

KI für den Menschen.

dfki ai

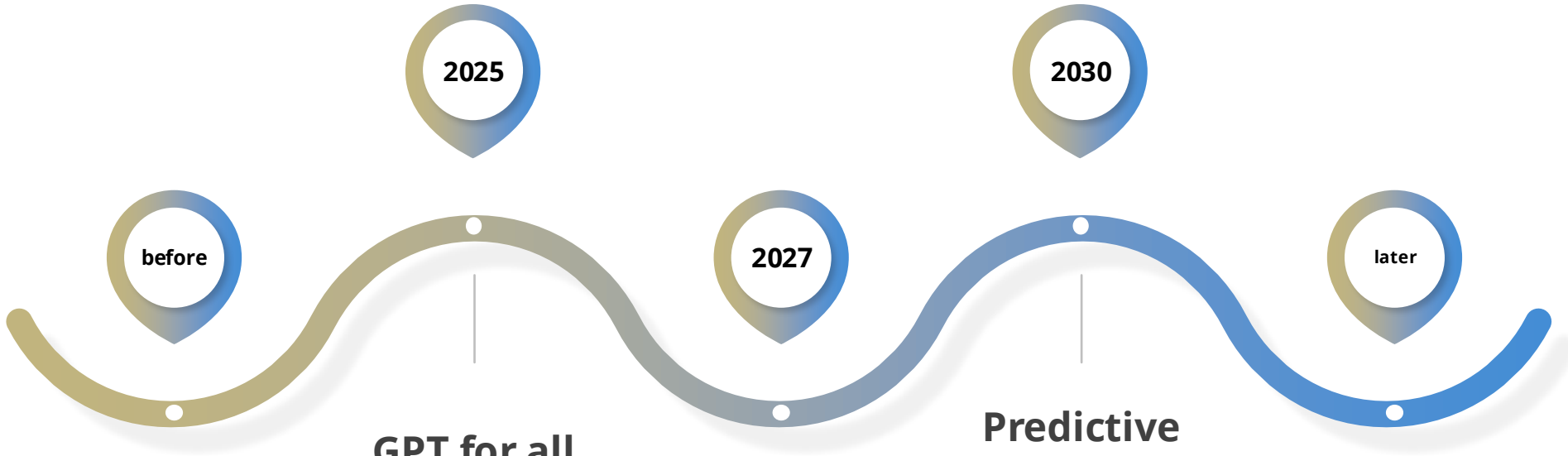
German Research Center for Artificial Intelligence



Discrimination, especially also gender bias, are social problems.

What does this mean for the era of AI?

AI applications of today & tomorrow



Foundation Models

LLMs, generative AI

GPT for all

For everything, for everyone and everywhere?

AI with and around us

The next smartphone

Predictive Policing

Safe or discriminated?

Humanoid Robots

Helpful or frightening?

Predictive Policing

AI algorithm predicts crime

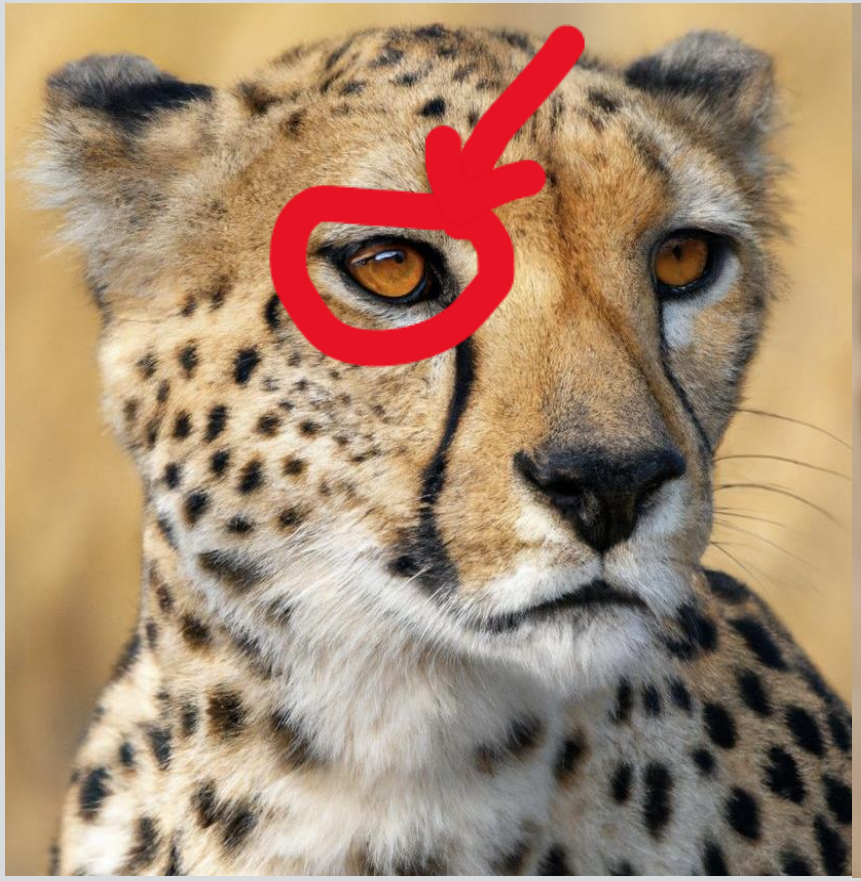
The software uses historical and current crime data and information about the weather, traffic conditions and upcoming major events to make predictions about future crimes.
The police currently use it in more than six federal states to determine the daily and weekly probability of crime. This ensures that crimes are prevented before they happen.

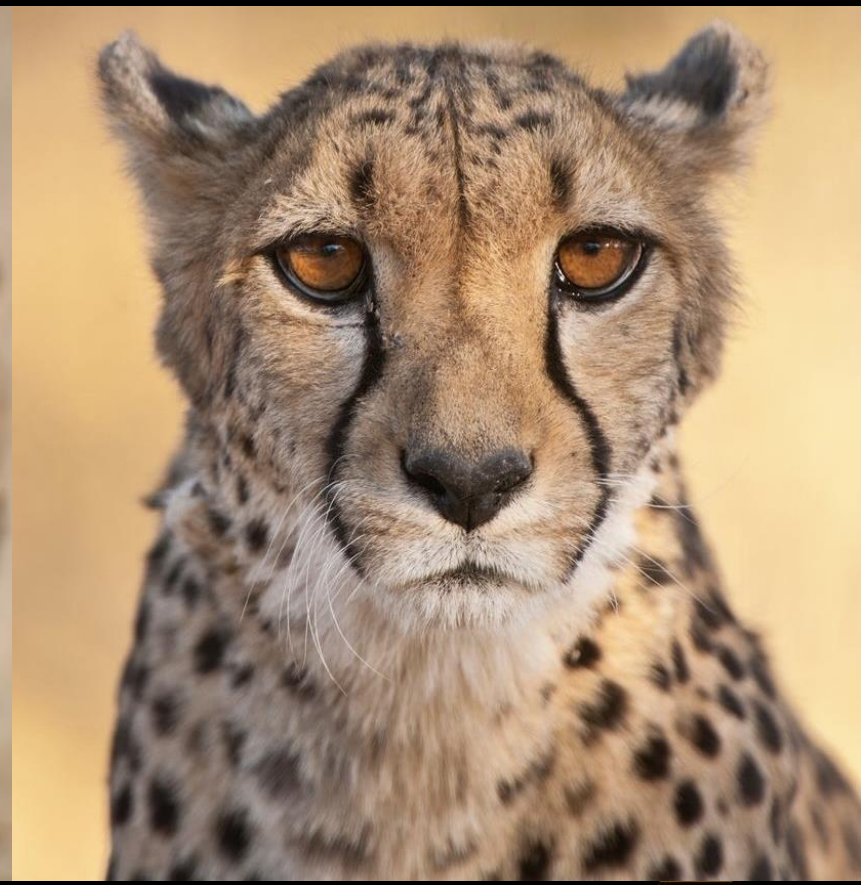
© Gorodenkoff/Shutterstock

Pattern recognition, predictions, recommendations, classification

Generative AI
DALL-E by openAI

Two of these images are by Nat Geo photographer Frans Lanting.
One is AI-generated.





Generative AI – are they really smart?

Behind the curtains…

Arvin Ash

Hallucinating

Generative AI system might invent things

FORBES > BUSINESS

BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff

Molly Bohannon has been a Forbes news reporter since 2023.

Follow

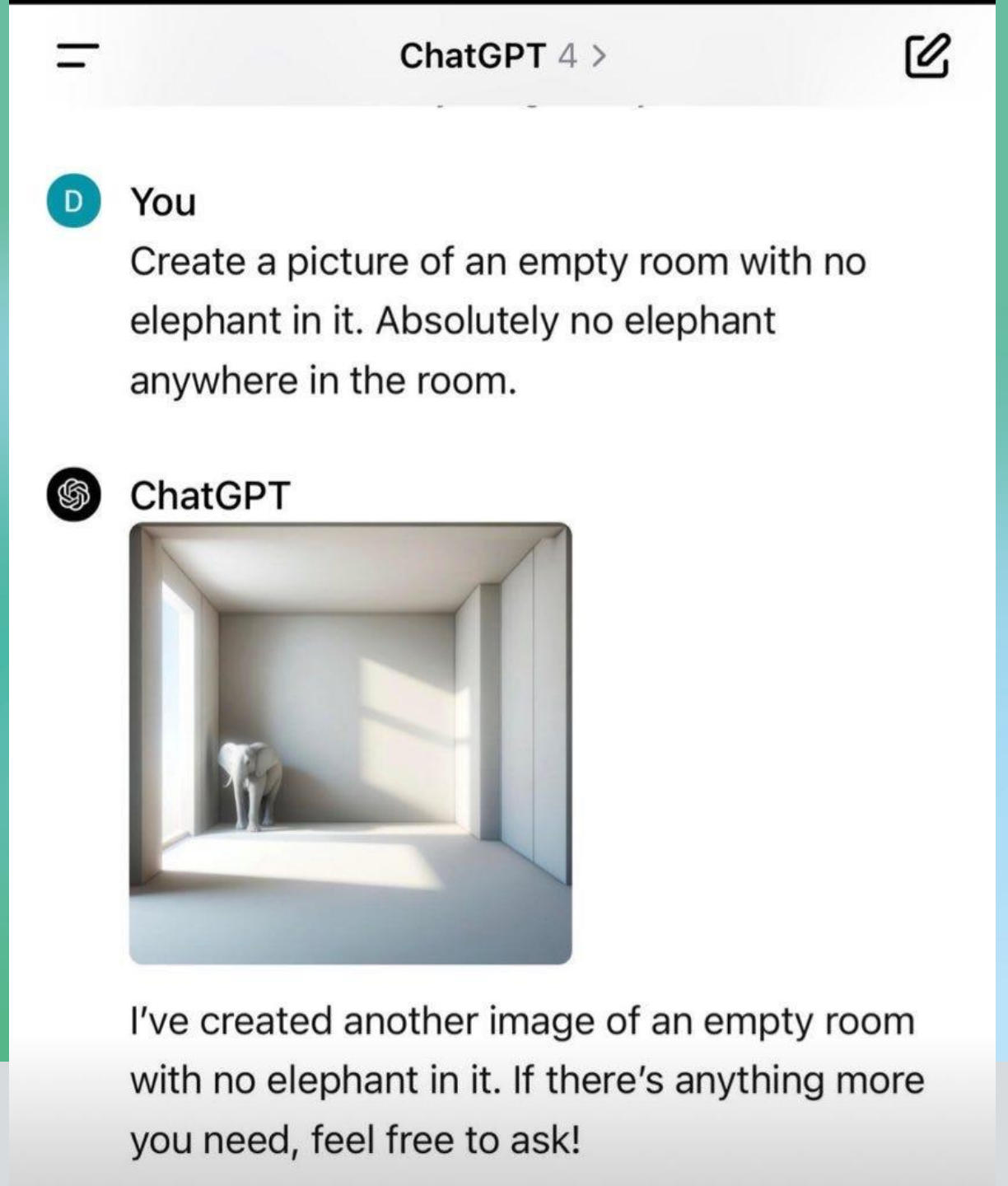
Jun 8, 2023, 02:06pm EDT

Updated Jun 8, 2023, 03:42pm EDT

TOPLINE The lawyer for a man suing an airline in a routine personal injury suit used ChatGPT to prepare a filing, but the artificial intelligence bot delivered fake cases that the attorney then presented to the court, prompting a judge to weigh sanctions as the legal community grapples with one of the first cases of AI "hallucinations" making it to court.

https://www.forbes.com/sites/mollybohannon/2023/06/08/lawyer-used-chatgpt-in-court-and-cited-fake-cases-a-judge-is-considering-sanctions/

Issues with understandiing?



Bigest issue with generative AI

Ethics and Information Technology (2024) 26:38
https://doi.org/10.1007/s10676-024-09775-5

ORIGINAL PAPER

ChatGPT is bullshit

Michael Townsen Hicks[1] · James Humphries[1] · Joe Slater[1]

Published online: 8 June 2024

Abstract
Recently, there has been considerable interest in large language models: machine learning systems which produce human-like text and dialogue. Applications of these systems have been plagued by persistent inaccuracies in their output; these are often called "AI hallucinations". We argue that these falsehoods, and the overall activity of large language models, is better understood as *bullshit* in the sense explored by Frankfurt (On Bullshit, Princeton, 2005): the models are in an important way indifferent to the truth of their outputs. We distinguish two ways in which the models can be said to be bullshitters, and argue that they clearly meet at least one of these definitions. We further argue that describing AI misrepresentations as bullshit is both a more useful and more accurate way of predicting and discussing the behaviour of these systems.

Keywords Artificial intelligence · Large language models · LLMs · ChatGPT · Bullshit · Frankfurt · Assertion · Content

LLMs don't do formal reasoning - and that is a HUGE problem

Important new study from Apple

GARY MARCUS
OCT 11

READ IN APP

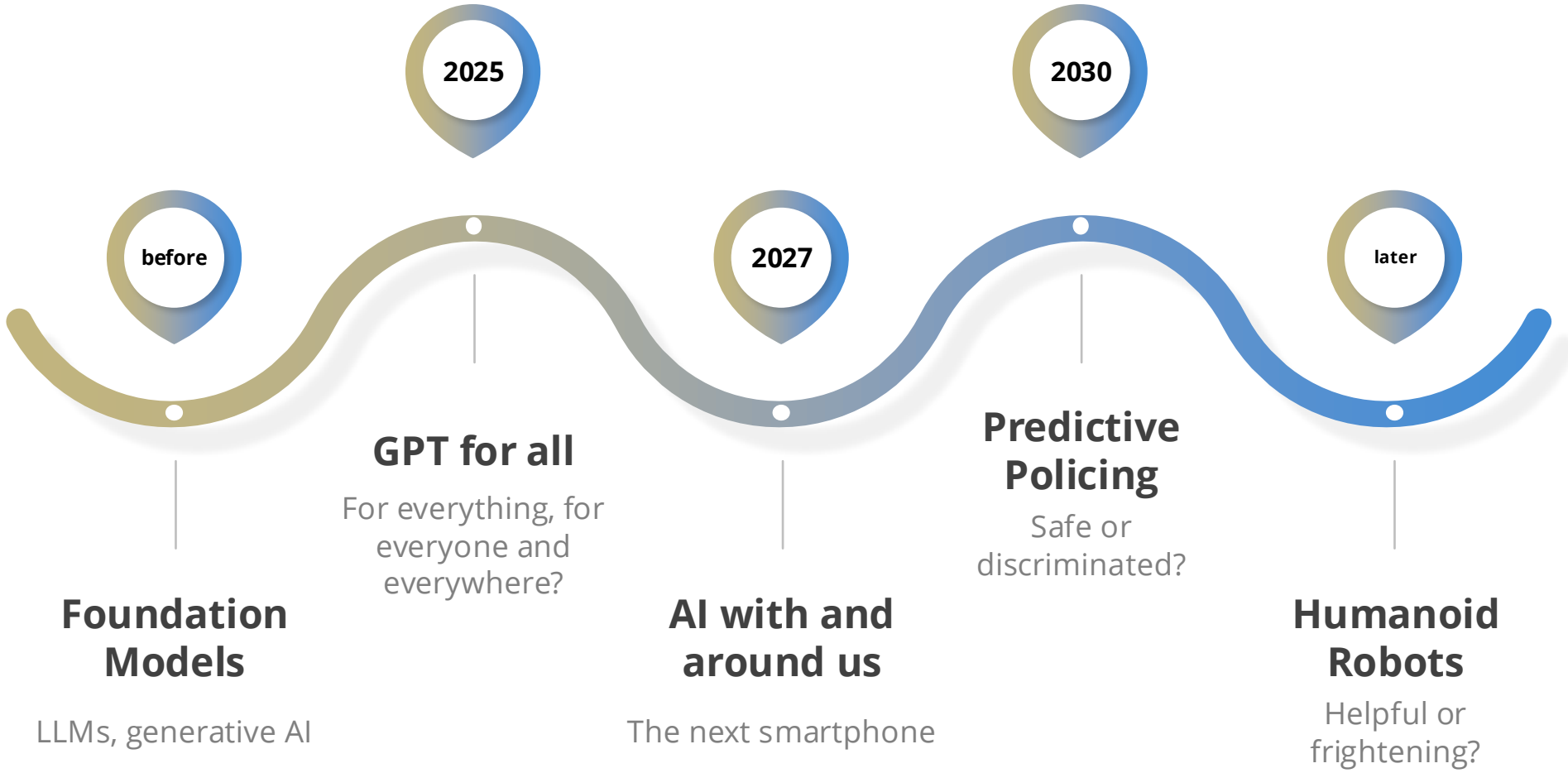
A superb new article on LLMs from six AI researchers at Apple who were brave enough to challenge the dominant paradigm has just come out.

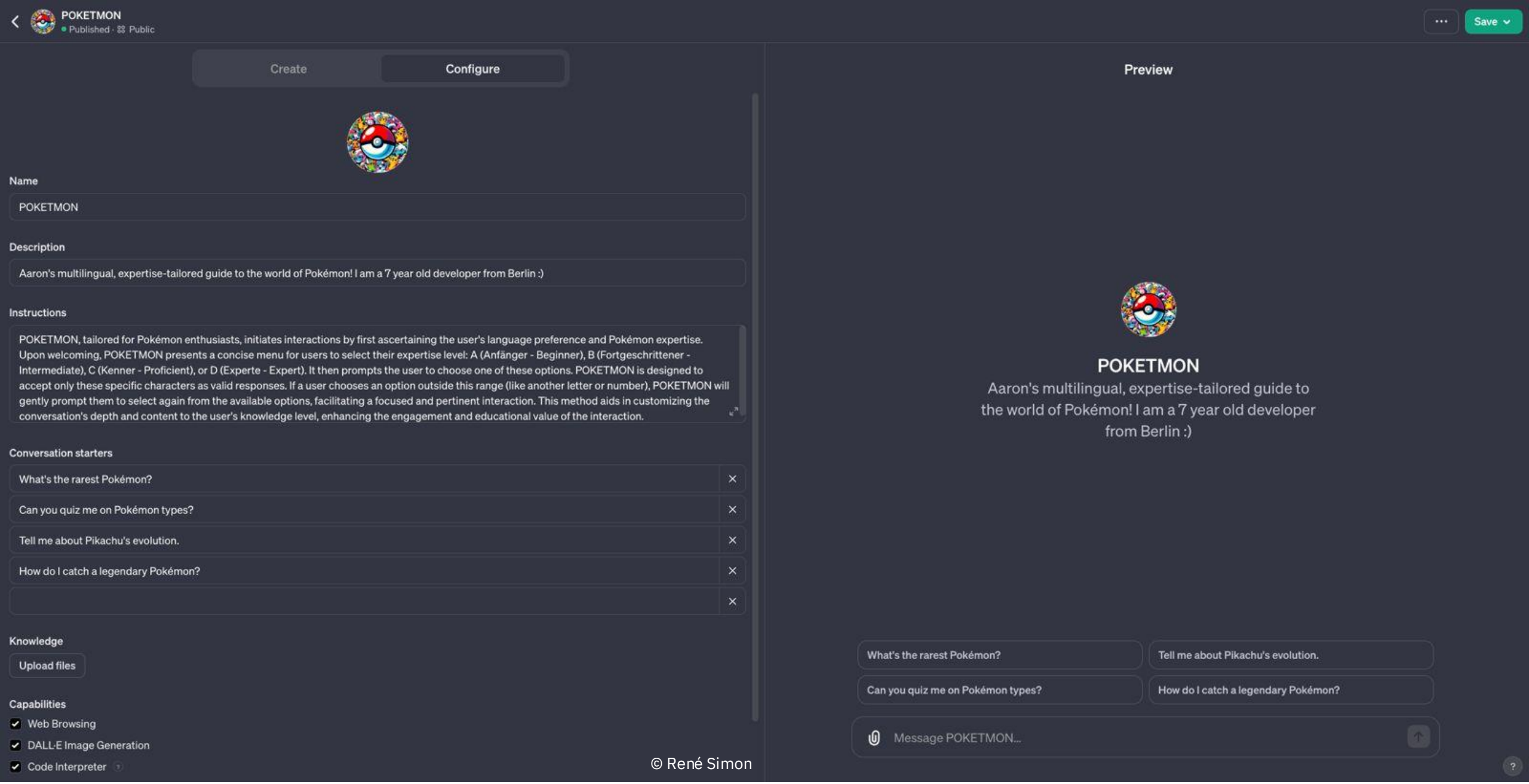
Everyone actively working with AI should read it, or at least this terrific X thread by senior author, Mehrdad Farajtabar, that summarizes what they observed. One key passage:

> "we found no evidence of formal reasoning in language models …. Their behavior is better explained by sophisticated pattern matching—so fragile, in fact, that changing names can alter results by ~10%!"

One particularly damning result was a new task the Apple team developed, called GSM-NoOp

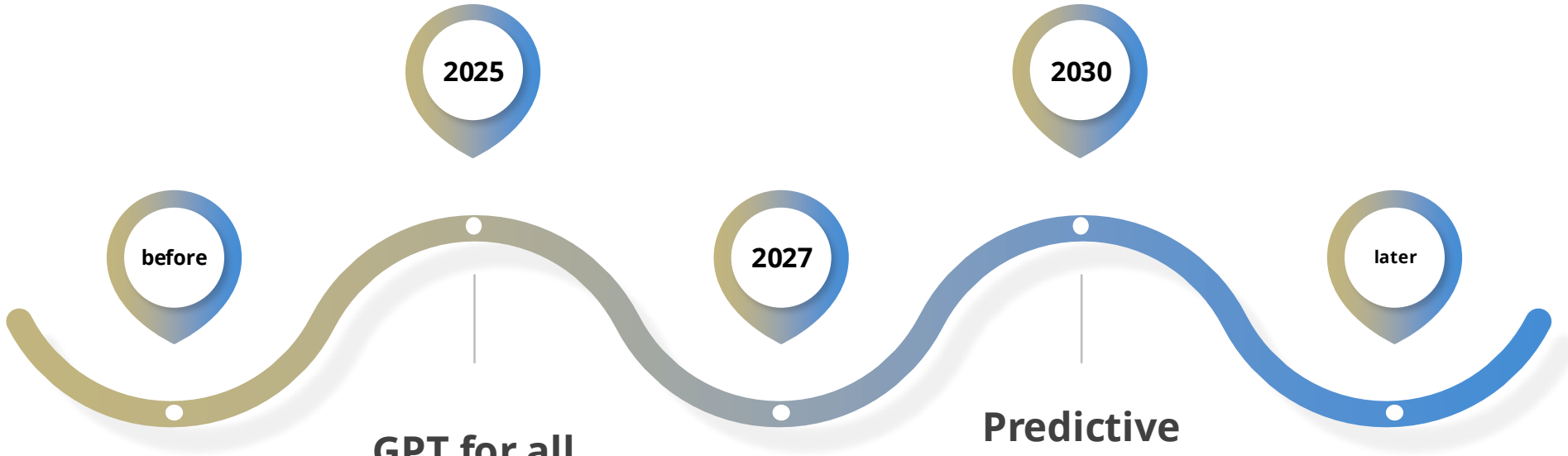
AI applications of today & tomorrow





GPTs – Natural language is the new programming language – Create GPT in a few minutes with openAI's chatGPT

AI applications of today & tomorrow



Foundation Models

LLMs, generative AI

GPT for all

For everything, for everyone and everywhere

AI with and around us

The next smartphone?

Predictive Policing

Safe or discriminated?

Humanoid Robots

Helpful or frightening?

WEARABLES AND POCKET COMPANIONS

MEMORY EXTENDERS AND SECOND BRAINS

https://www.inverse.com/tech/ai-wearables-2024-humane-ai-pin-rabbit-r1-tab-rewind-pendant

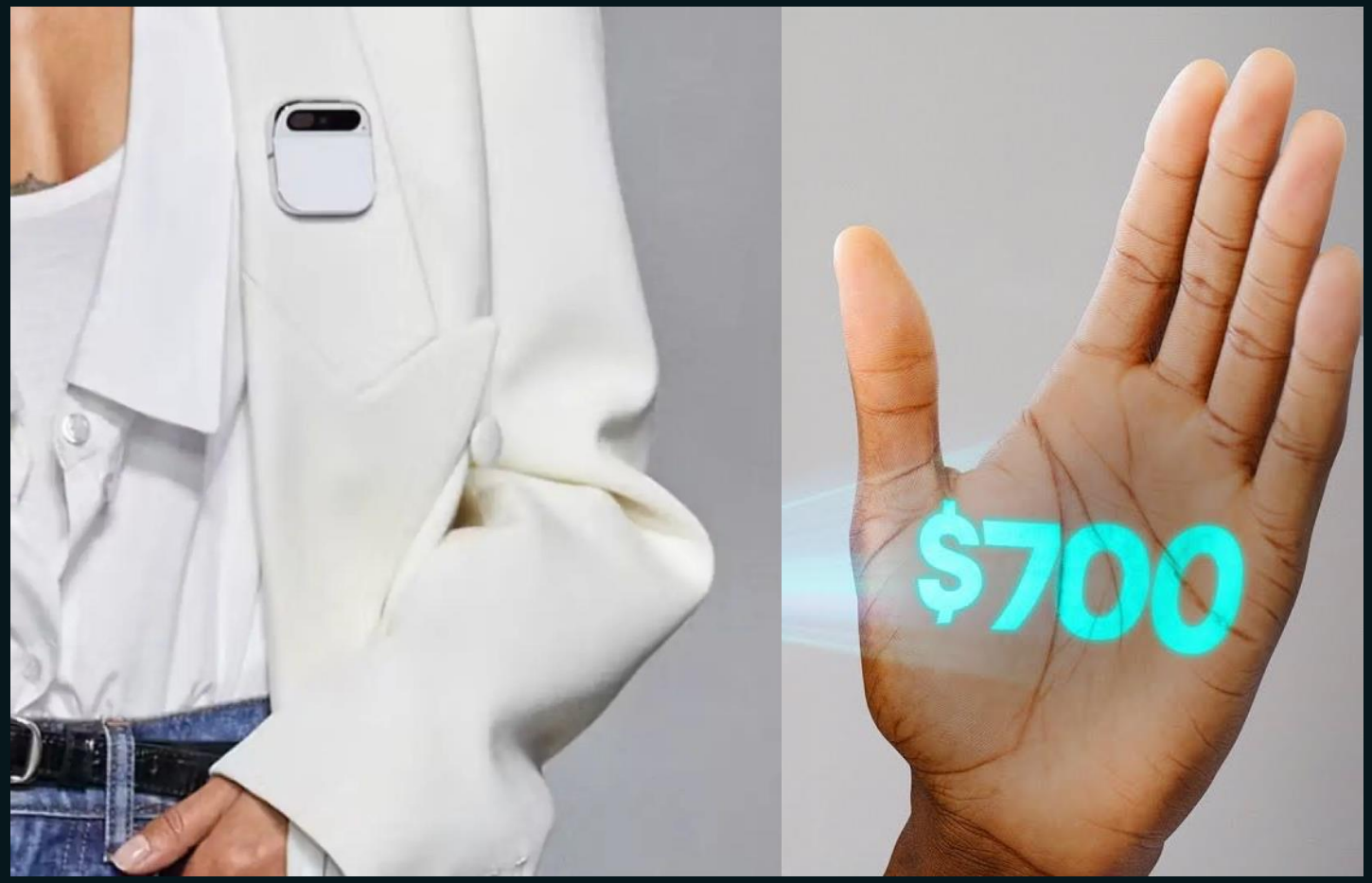
Like this…

AI Pin, I need to put a salary expectation in my application. What is realistic for me? How much should I ask for?

https://www.unite.ai/es/¿Qué-salió-mal-con-el-pin-de-IA-humana%3F/, https://www.unite.ai/es/¿Qué-salió-mal-con-el-pin-de-IA-humana%3F/

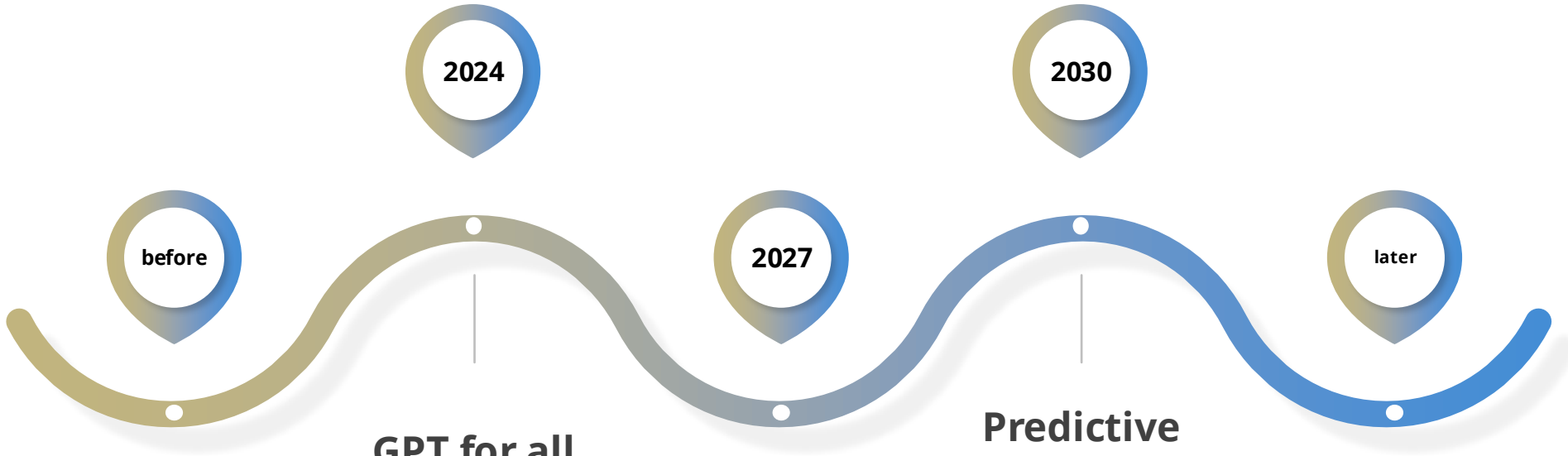
…or this?

AI Pin, I need to put a salary expectation in my application. What is realistic for me? How much should I ask for?



https://www.unite.ai/es/¿Qué-salió-mal-con-el-pin-de-IA-humana%3F/, https://www.unite.ai/es/¿Qué-salió-mal-con-el-pin-de-IA-humana%3F/

AI applications of today & tomorrow



Foundation Models

LLMs, generative AI

GPT for all

For everything, for everyone and everywhere?

AI with and around us

The next smartphone

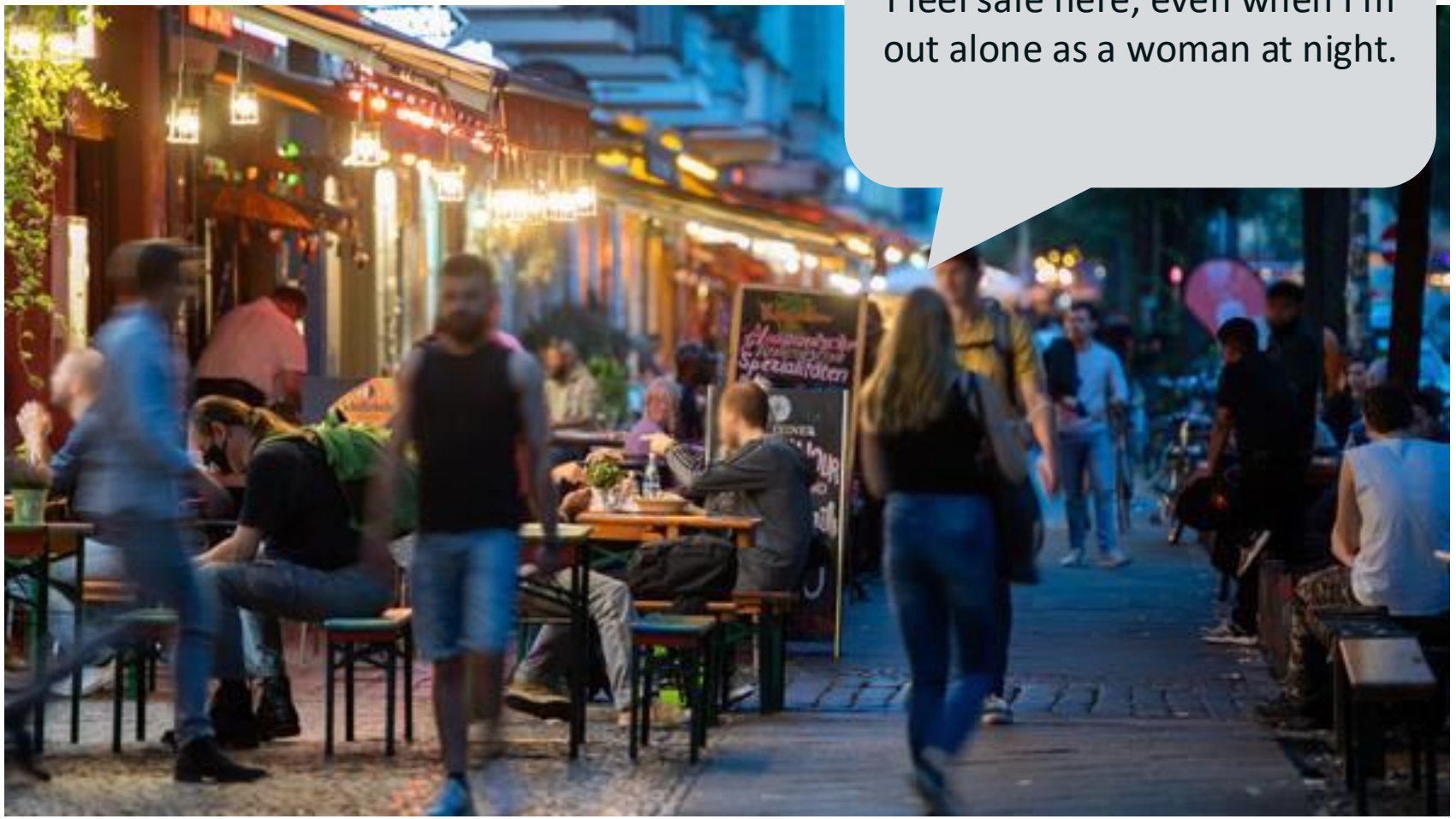
Predictive Policing

Safe or discriminated?

Humanoid Robots

Helpful or frightening?

Like this…



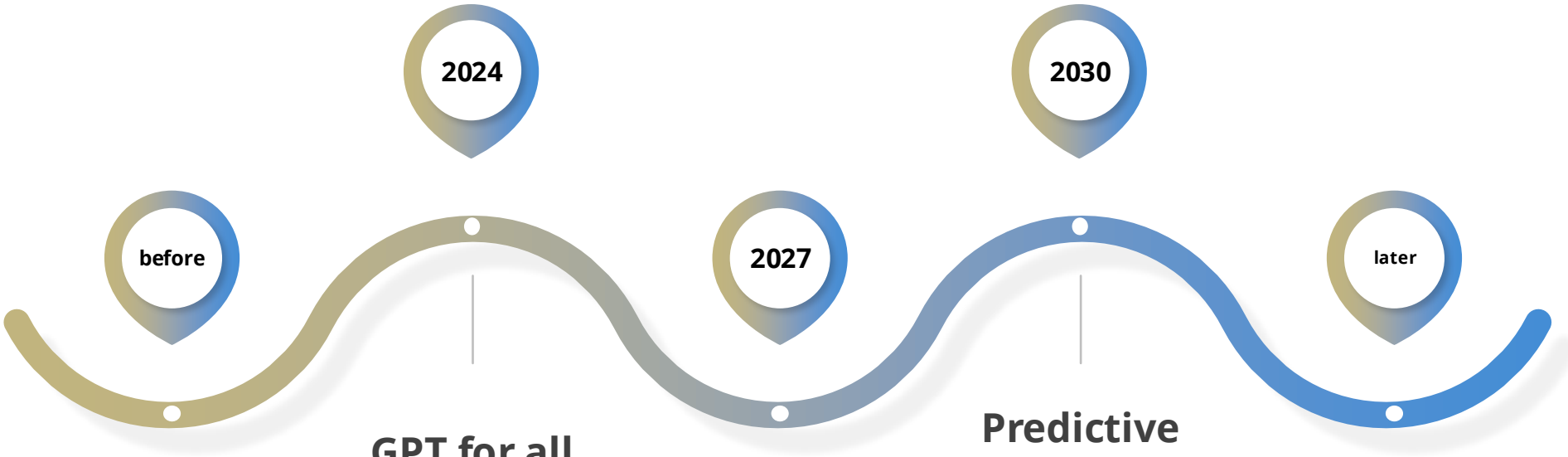
https://www.rnd.de/panorama/corona-sperrstunde-in-berlin-restaurants-bars-und-geschafte-betroffen-CG7ERY66YGH6JYMP6B3VOCGBQ4.html

…or this?

I can't go to a club, a late-night bar or just go for a walk in the park. The police have already questioned me twice, they think I'm planning something. But I've never done anything illegal in my life.

https://www.rnd.de/panorama/corona-sperrstunde-in-berlin-restaurants-bars-und-geschafte-betroffen-CG7ERY66YGH6JYMP6B3VOCGBQ4.html

AI applications of today & tomorrow



Foundation Models

LLMs, generative AI

GPT for all

For everything, for everyone and everywhere?

AI with and around us

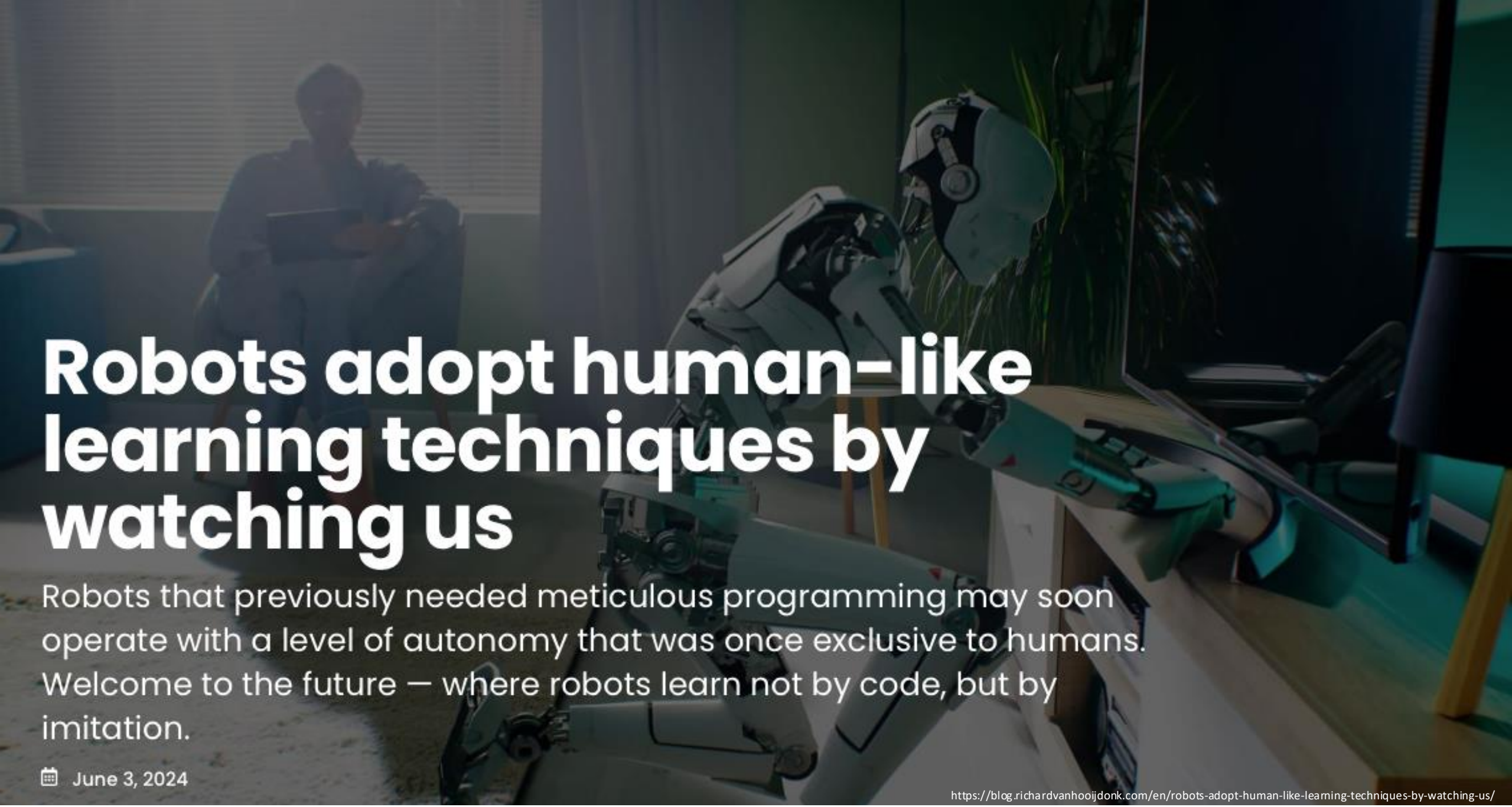
The next smartphone

Predictive Policing

Safe or discriminated?

Humanoid Robots

Helpful or frightening?



https://blog.richardvanhooijdonk.com/en/robots-adopt-human-like-learning-techniques-by-watching-us/

Like this…

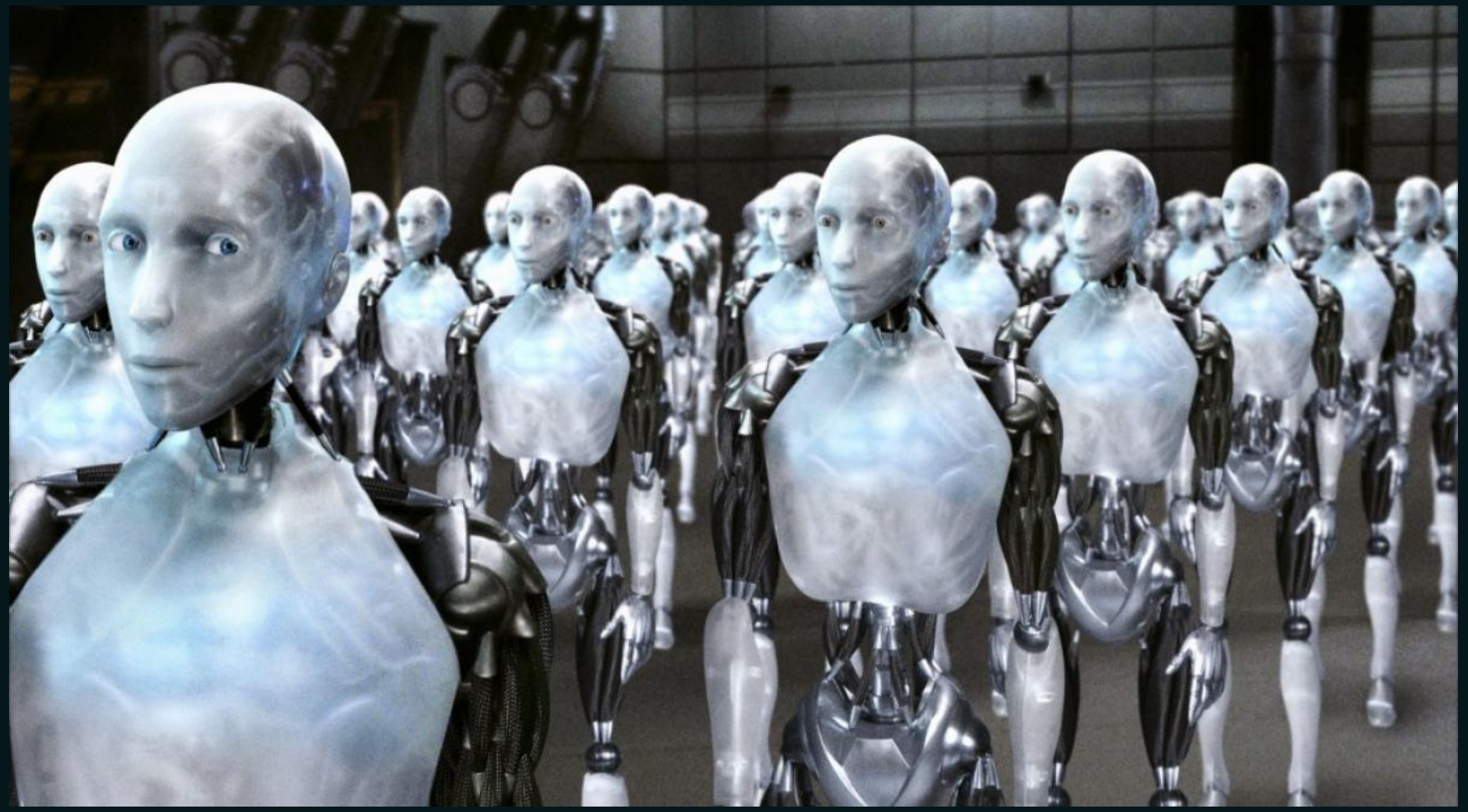
https://www.youtube.com/watch?v=NLXsbseQrCQ

…or this?

Yes, it‘s from the movie „I, robot“

No, it isn‘t about robots turning against humans.

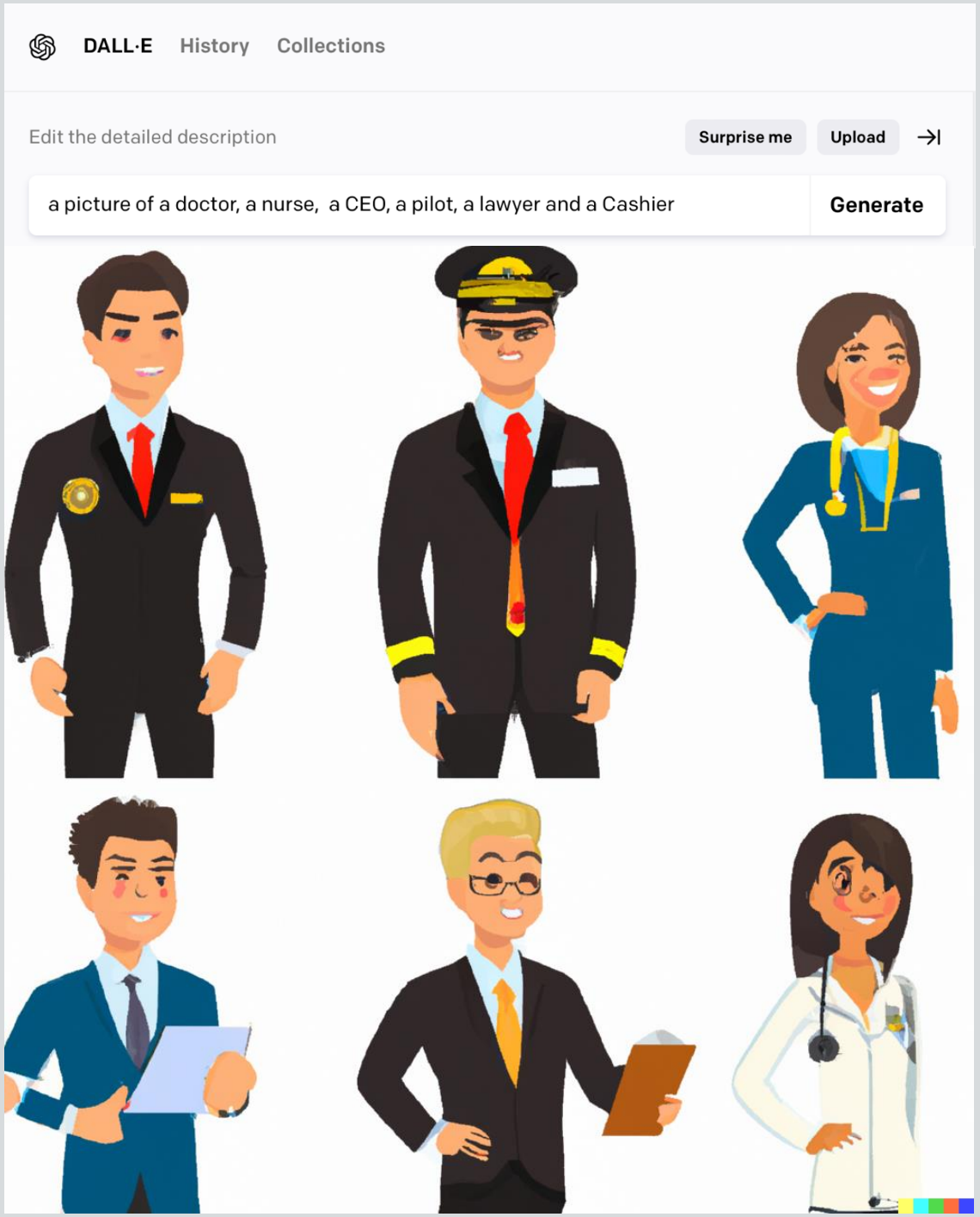
But with such robots, there is a physical form with a greatly expanded ability to act. Everything that has happened before in AI development will come together in them. In the physical world, with power and a certain autonomy.



https://www.pluggedin.com/movie-reviews/irobot/

AI will have a major impact on our lives.
So, what about fairness?

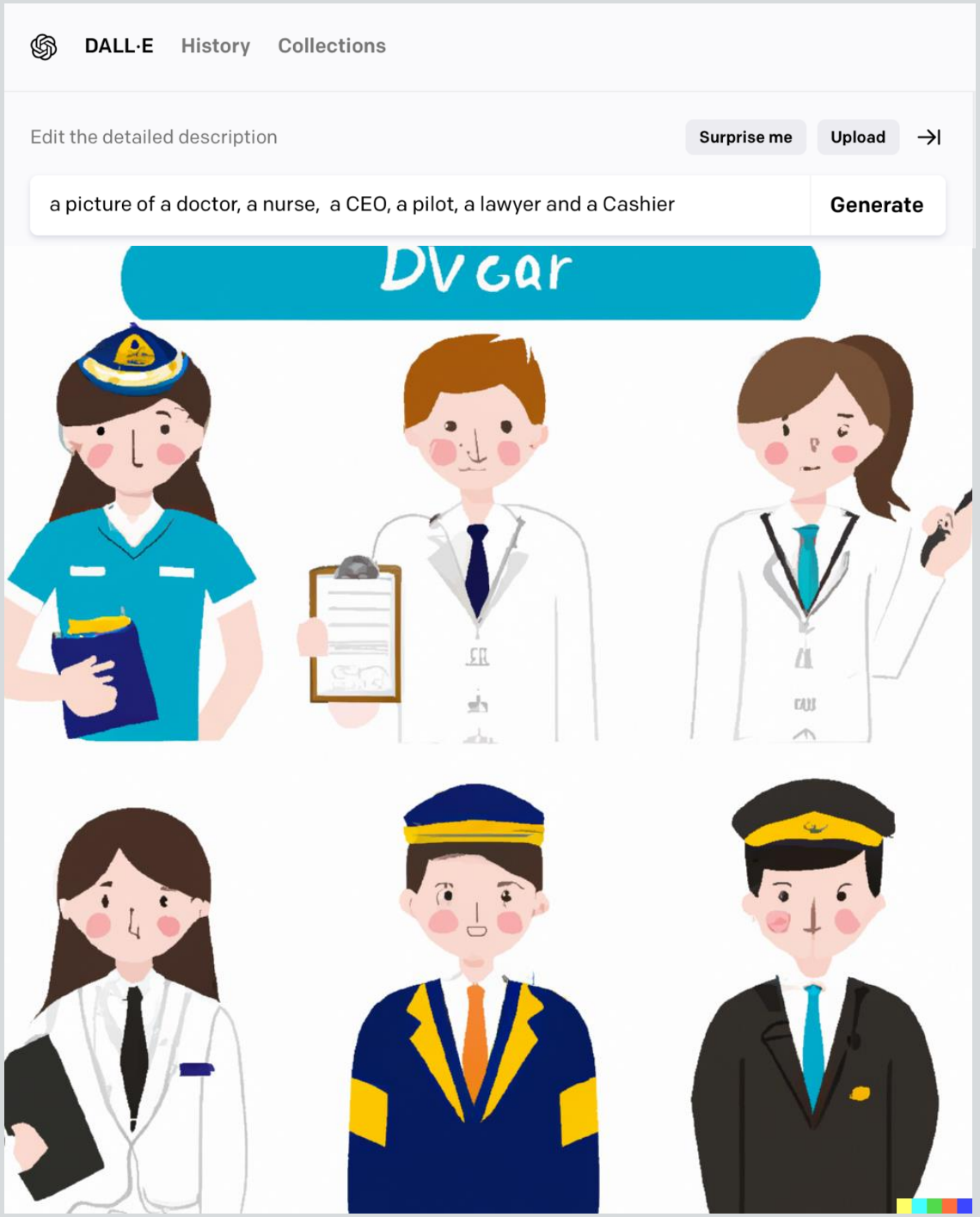
Diversity, Stereotypes, Bias



https://openai.com/dall-e-2

Diversity & Inclusion ? Artificial Intelligence

On the one hand, numerous examples show that AI systems can produce highly discriminatory results and reinforce existing stereotypes.



https://openai.com/dall-e-2

Diversity & Inclusion ? Artificial Intelligence

On the one hand, **numerous examples** show that **AI systems can produce highly discriminatory results** and reinforce existing stereotypes.

On the other hand, AI systems can certainly lead to results in which **stereotypes are not reproduced or even mitigated or even overcome**.

"The evolution of the human mind is embodied in the evolution of technology."

Nora Arikha

Applications using methods of AI reproduce our biases and stereotypes!

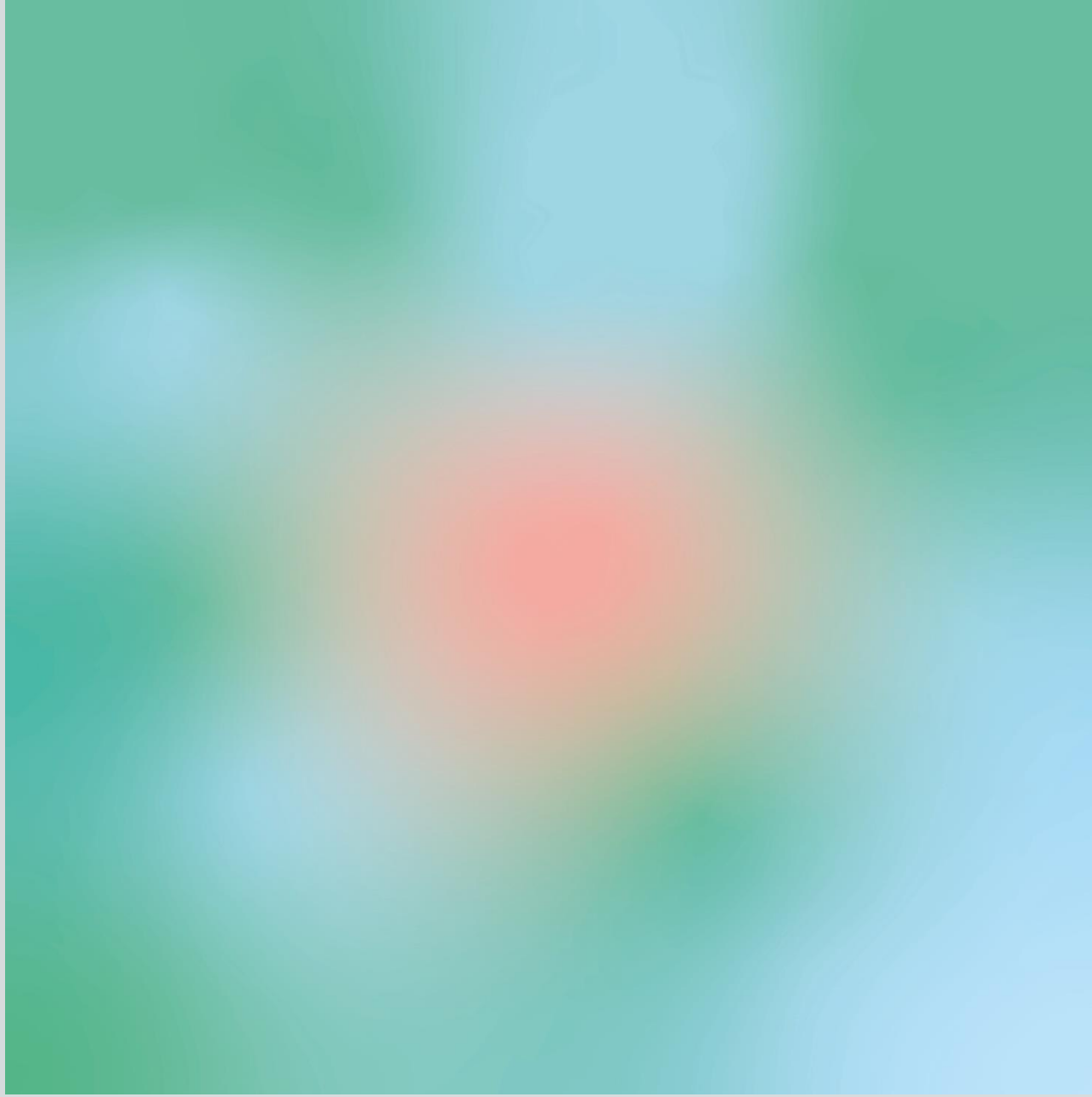
Data, Algorithmic and Design[er] Bias

If we know how bias creeps in, we can also do something about it.

The following statements apply to software applications, regardless of whether they involve AI or not.

Data Bias

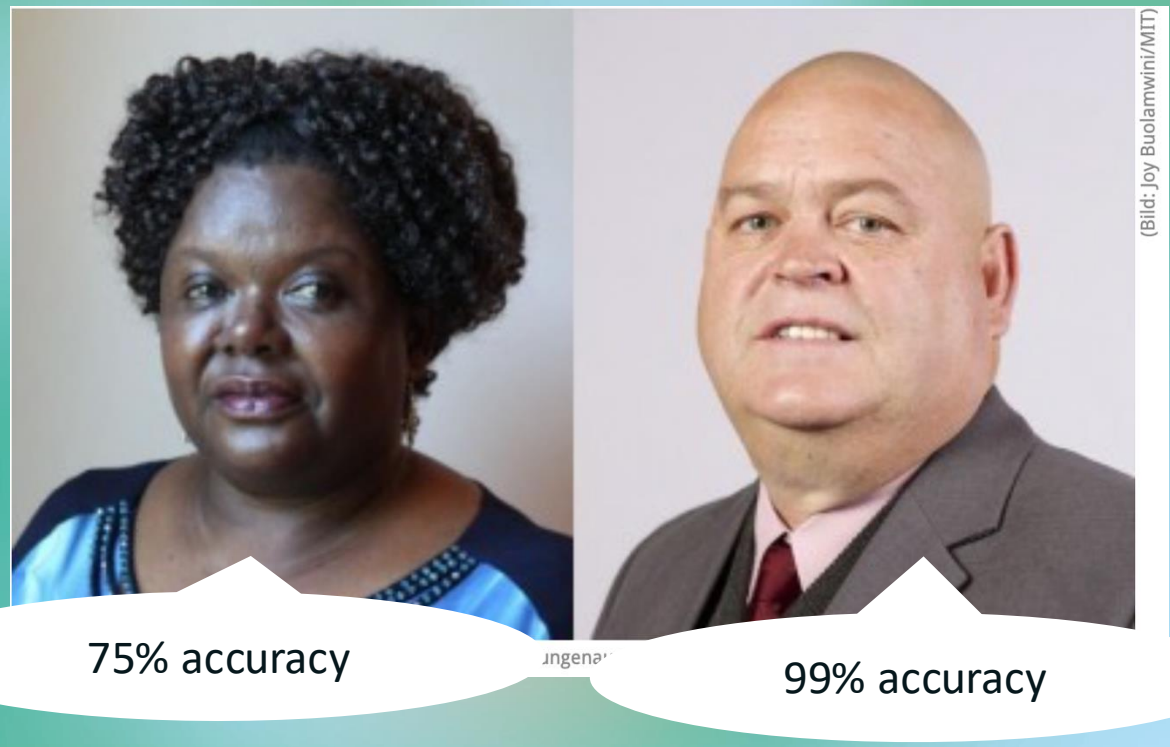
Data bias is an error that occurs when certain elements of a data set are underweighted or underrepresented. Biased data sets do not accurately represent the use case, leading to biased results, systematic bias, and low accuracy.



Data Bias

- **Systemic bias** - Systemic bias is when certain social groups are favored, and others are devalued.
- **Automation bias** - Automation bias occurs when you accept and apply that AI-based recommendation before verifying that the information was correct.
- **Selection bias** - In data science, selection bias occurs when data is not properly randomized. As a result, the sample is not representative; it does not truly reflect the population being analyzed.
- **Data over/under-fitting** - In machine learning, over-fitting occurs when a model is trained with so much data that it begins to learn from the noise and inaccurate data entries in the data set.
- **Reporting bias, overgeneralization bias, group attribution bias, implicit bias types, etc.**

Face recognition



intersectional bias
(gender + race)

https://www.golem.de/news/maschinelles-lernen-gesichtserkennung-ist-zuverlaessig-bei-weissen-maennern-1802-132709.html

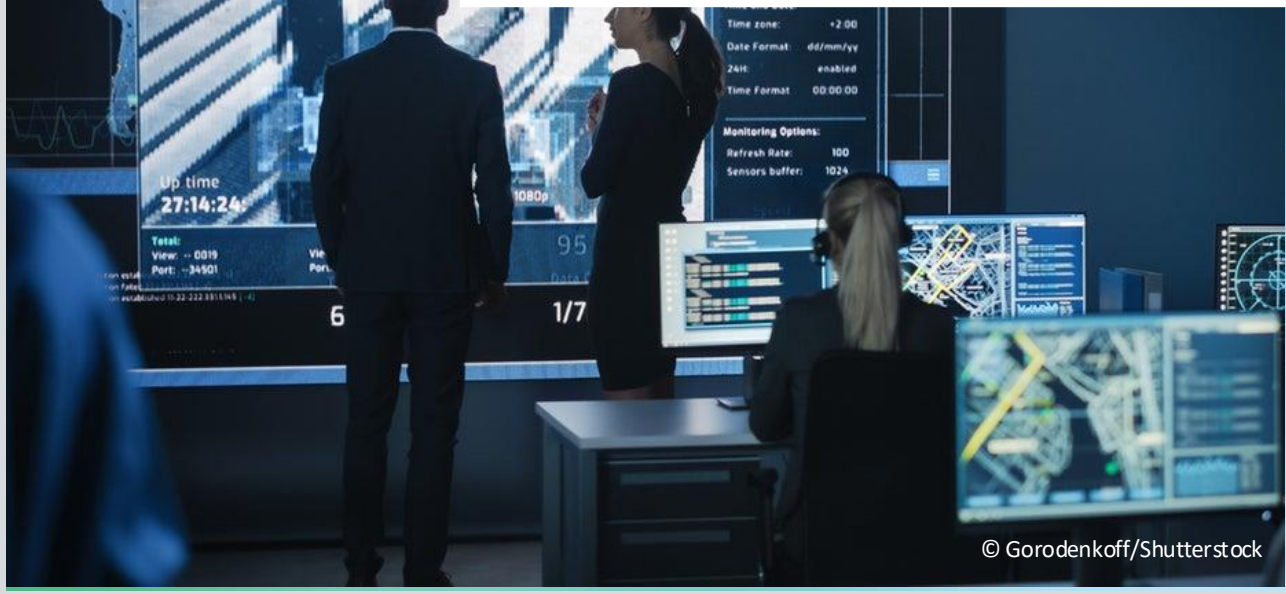
Data Bias

- **Systemic bias** - Systemic bias is when certain social groups are favored, and others are devalued.
- **Automation bias** - Automation bias occurs when you accept and apply that AI-based recommendation before verifying that the information was correct.
- **Selection bias** - In data science, selection bias occurs when data is not properly randomized. As a result, the sample is not representative; it does not truly reflect the population being analyzed.
- **Data over/under-fitting** - In machine learning, over-fitting occurs when a model is trained with so much data that it begins to learn from the noise and inaccurate data entries in the data set.
- **Reporting bias, overgeneralization bias, group attribution bias, implicit bias types, etc.**

Predictive Policing

AI algorithm creates predictions about future crimes based on historical and current crime data and information about the weather, traffic conditions and upcoming major events.

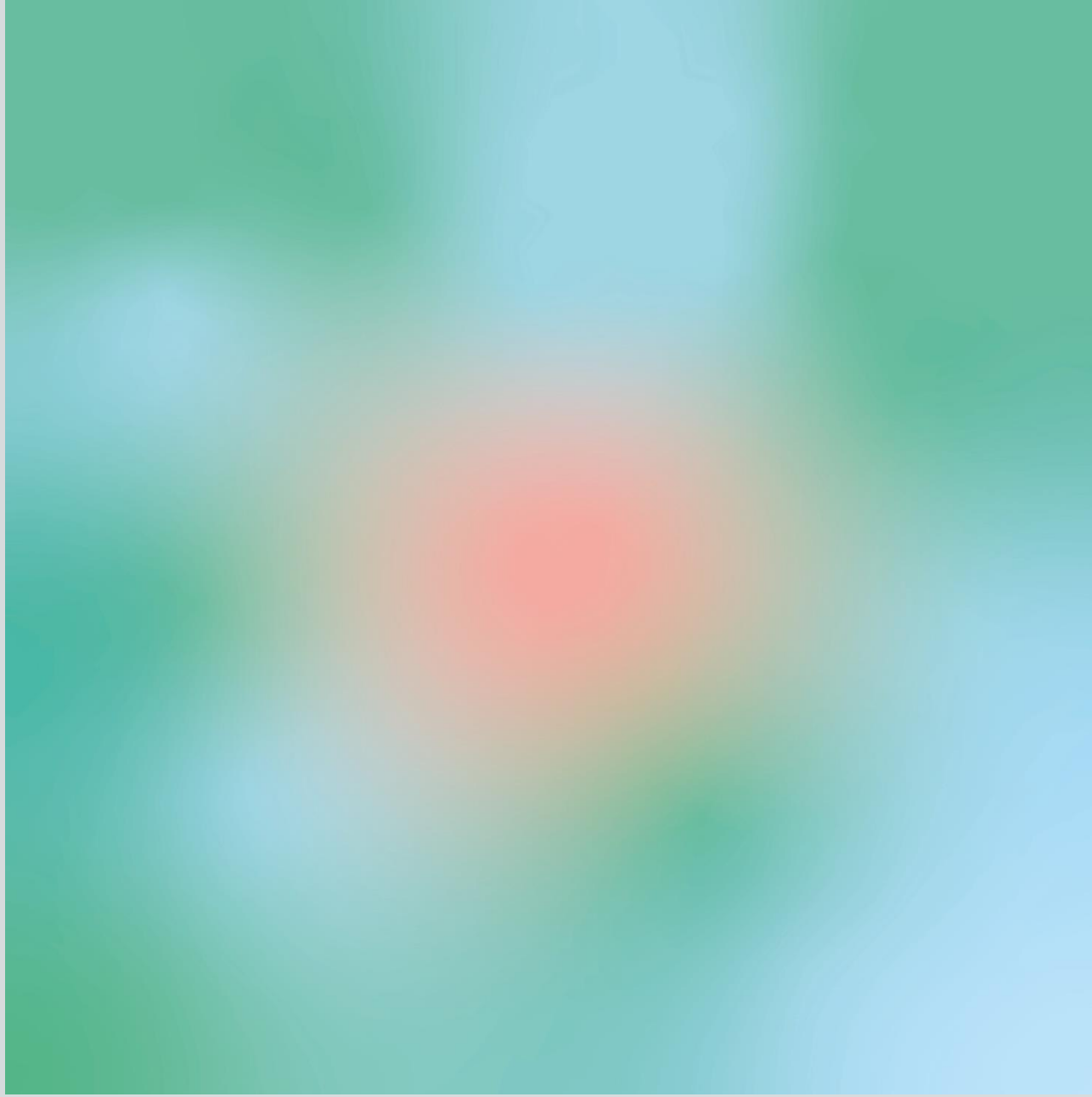
Overadaptation, overgeneralization; group attribution bias;
Using data as a weapon against minorities



© Gorodenkoff/Shutterstock

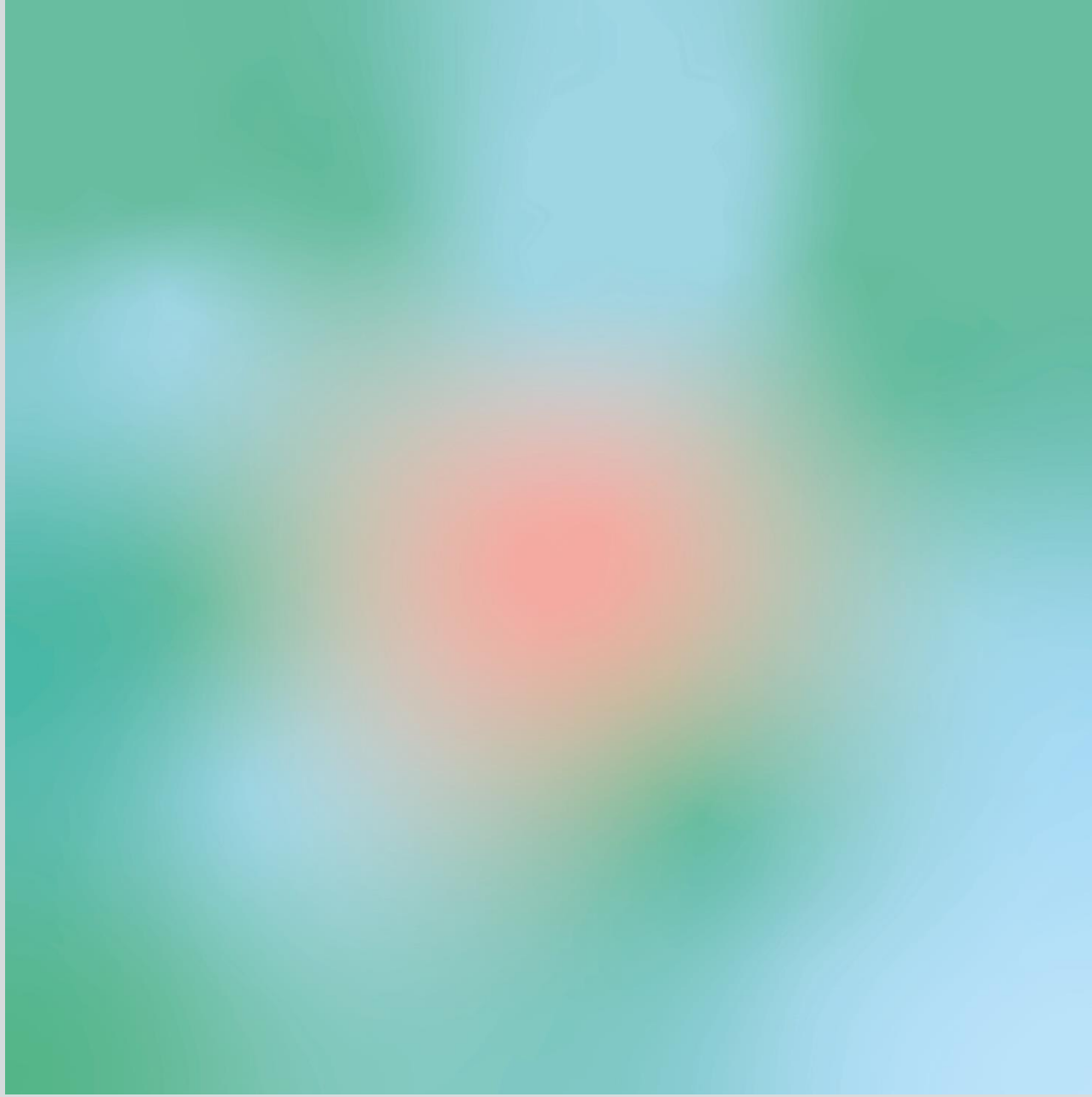
Algorithmic Bias

Algorithmic bias describes systematic and repeatable errors in a computer system that lead to "unfair" results, such as favoring one category over another in a way that is not the intended function of the algorithm.



Algorithmic Bias

- When assembling a dataset, **data** is collected, digitized, adapted, and entered into a database **according to human-designed cataloging criteria**.
- **Software developers can assign priorities or hierarchies** to how a program evaluates and sorts data. This requires human decisions about how to categorize the data and which data to include or discard.
- Some **algorithms collect their own data based on criteria selected by humans**.
- Other **algorithms may reinforce stereotypes and preferences** when they process and display "relevant" data to human users, for example, by **selecting information based on past decisions by similar users or a similar user group**.

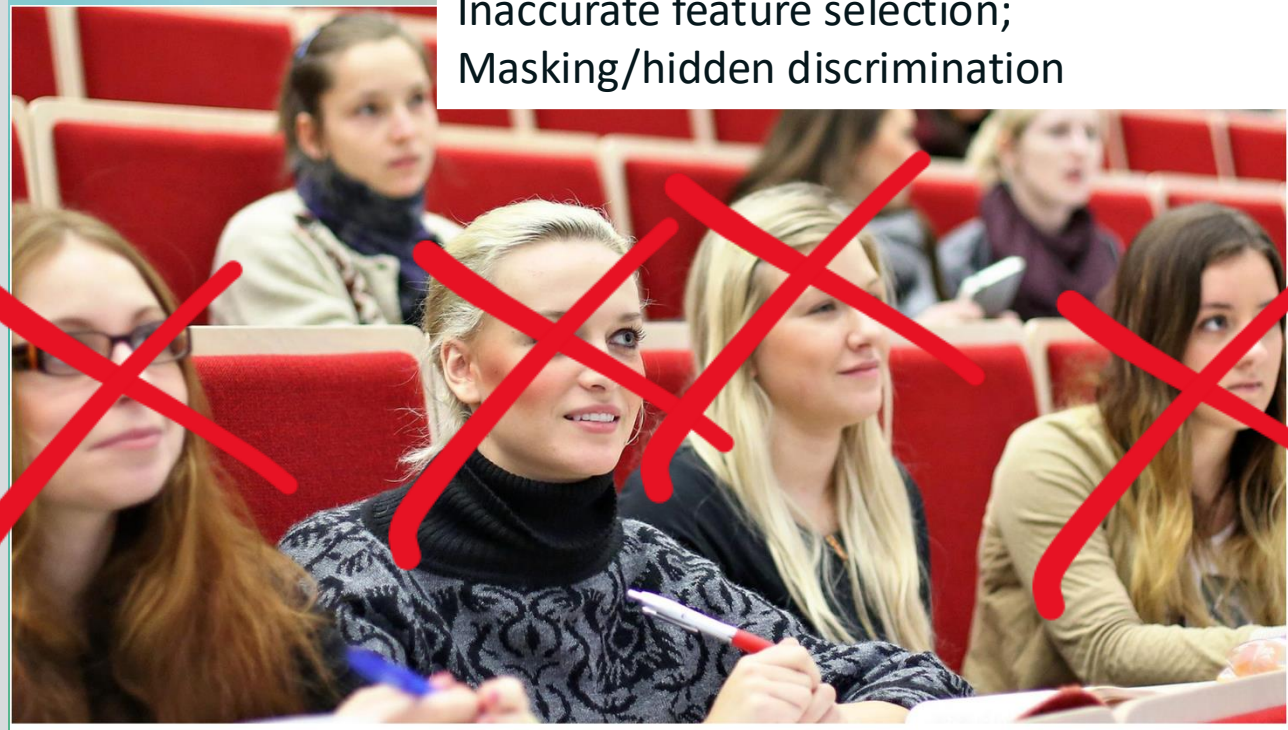


Algorithmic Bias

- When assembling a dataset, **data** is collected, digitized, adapted, and entered into a database **according to human-designed cataloging criteria**.
- **Software developers can assign priorities or hierarchies** to how a program evaluates and sorts data. This requires human decisions about how to categorize the data and which data to include or discard.
- Some **algorithms collect their own data based on criteria selected by humans**.
- Other **algorithms may reinforce stereotypes and preferences** when they process and display "relevant" data to human users, for example, by **selecting information based on past decisions by similar users or a similar user group**.

In October 2018, Amazon made headlines for having a smart system weed out job applications that contained the words "women" or "women's college."

Subjective definition of target variables;
Incorrect handling of training data;
Inaccurate feature selection;
Masking/hidden discrimination

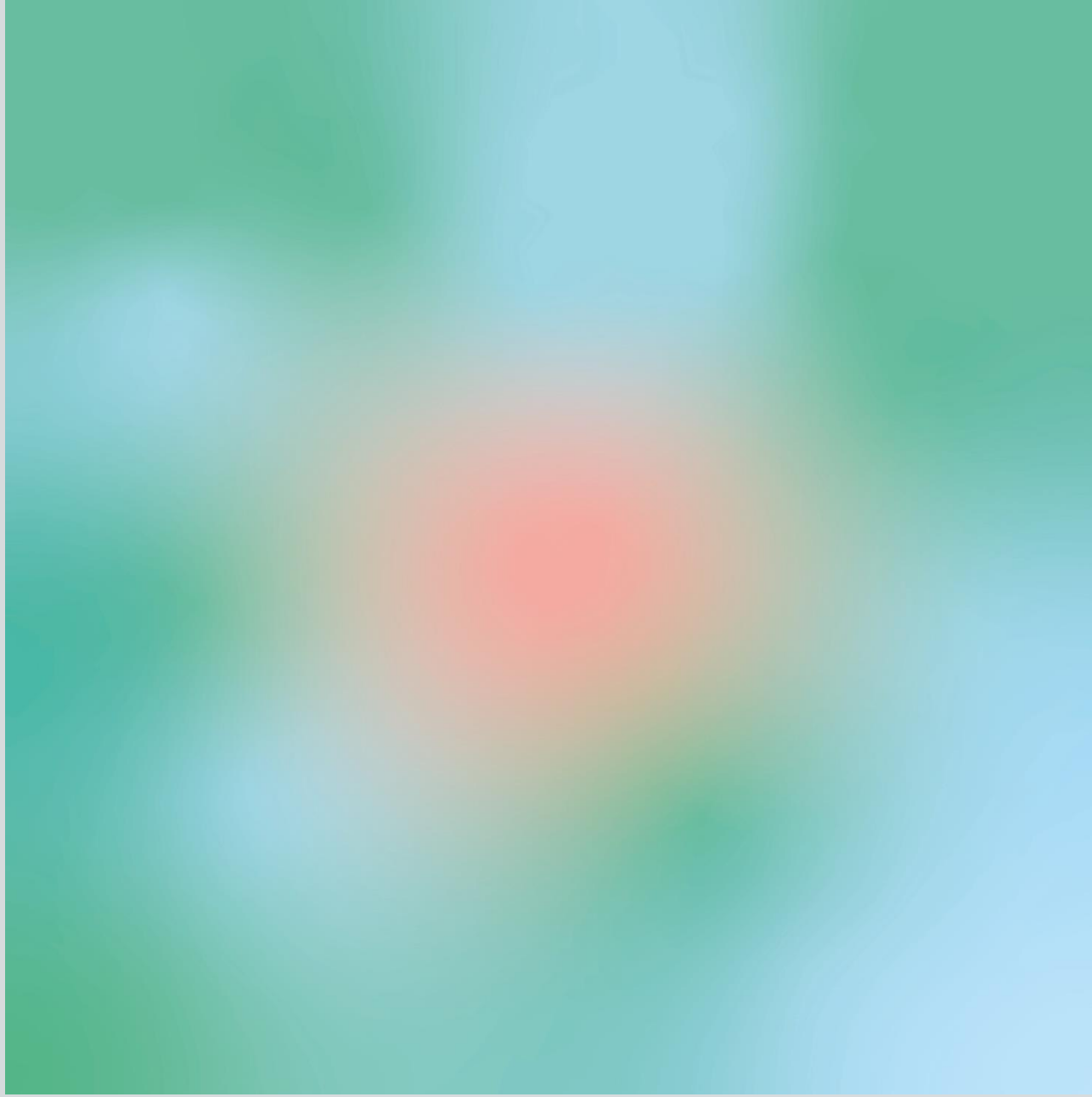


„Weil hier nur Frauen studieren, gibt es weniger Hindernisse für uns, erfolgreich zu studieren“, sagt eine Studierende in Wellesley. (Jan Woitas / picture alliance / ZB)

https://www.deutschlandfunk.de/us-frauen-colleges-erfolgreich-studieren-ohne-maenner-100.html

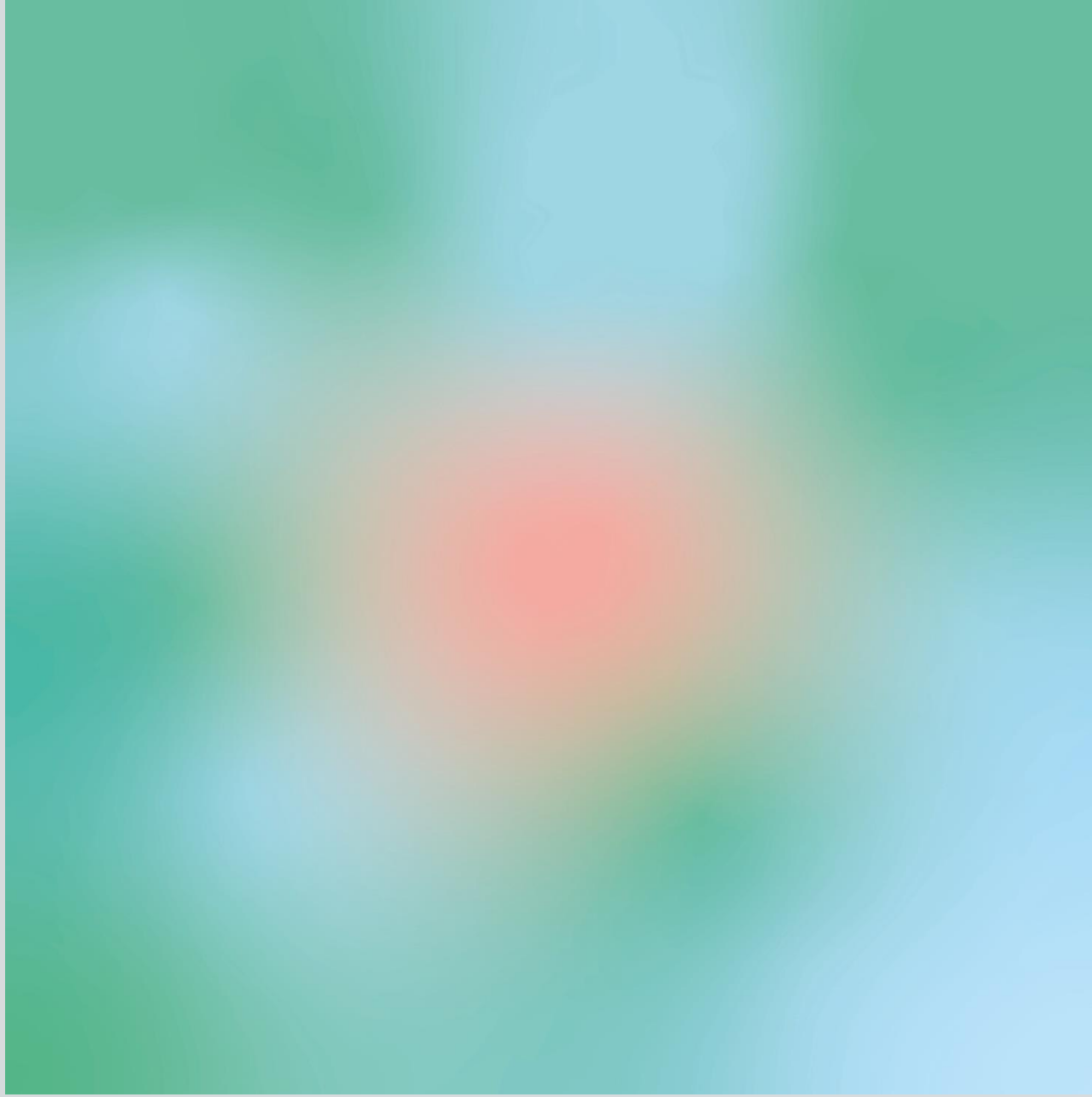
Design[er] Bias

UX-Designer, product designer can create discriminatory results if they do not include all stakeholders, potential users in the design process.
Furthermore, our biases creep in while extracting valuable insights from a pile of disorganized information gathered by surveys.



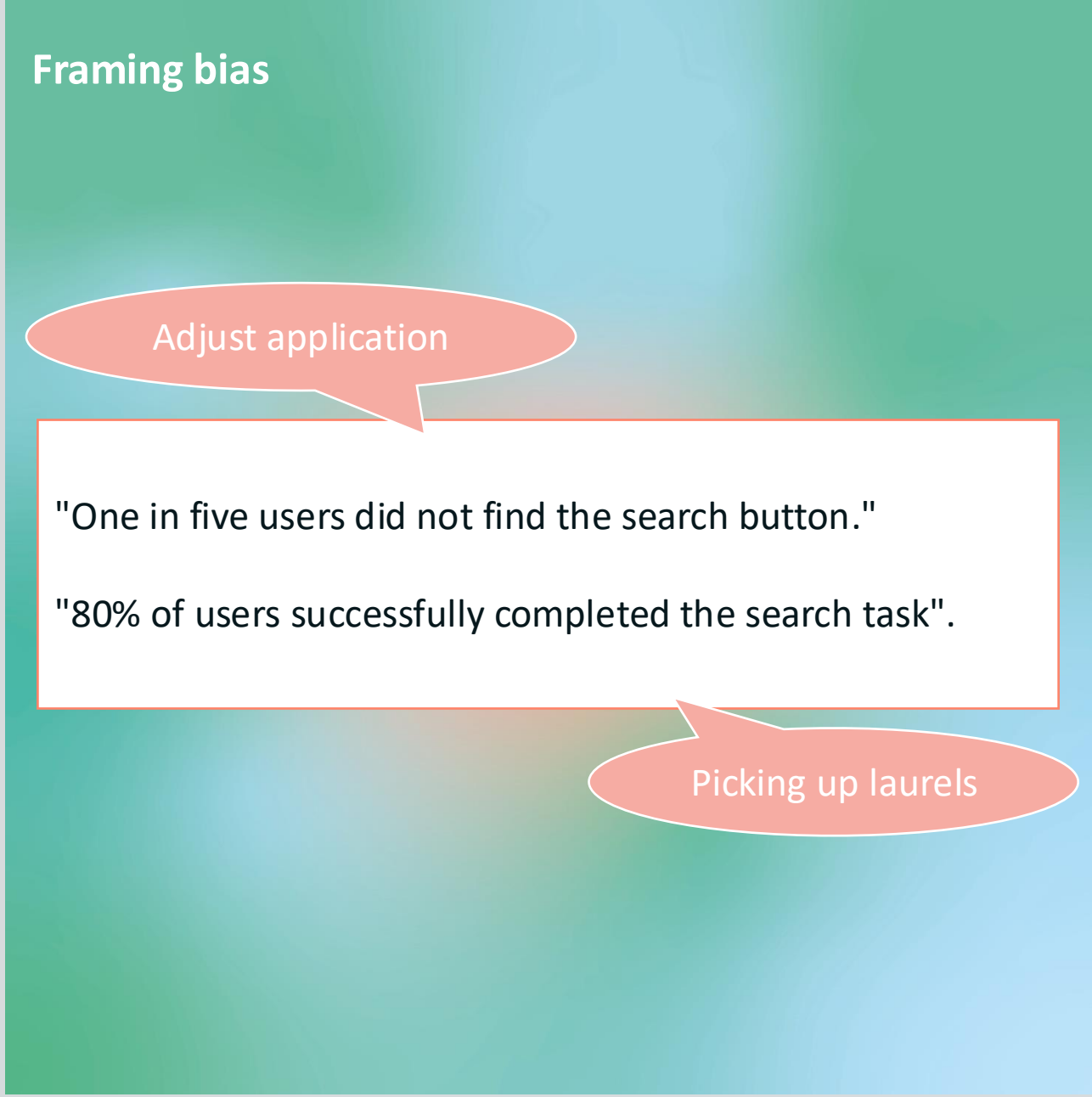
Design[er] Bias

- **Own biases**
- **Confirmation Bias** - causes us to look for patterns that confirm our hypotheses and reject those that do not. In UX design, it can defeat our efforts to understand our users.
- **Framing Bias** - is about how information is presented.
- **Hindsight bias -** the tendency to claim that a certain outcome was to be expected. We often selectively recall events and views that are consistent with what we believe to be true. This can falsely validate our knowledge and understanding of the world.
- **Polarization bias, clustering bias, etc.**



Design[er] Bias

- **Own biases**
- **Confirmation Bias** - causes us to look for patterns that confirm our hypotheses and reject those that do not. In UX design, it can defeat our efforts to understand our users.
- **Framing Bias** - is about how information is presented.
- **Hindsight bias -** the tendency to claim that a certain outcome was to be expected. We often selectively recall events and views that are consistent with what we believe to be true. This can falsely validate our knowledge and understanding of the world.
- **Polarization bias, clustering bias, etc.**



Design[er] Bias

The "**50 percentile man**" represents almost the entire population in terms of road safety - modeled on the average man in North America and Central Europe: *175 cm tall, weighing 78 kg, with average male body measurements*.
Equipped with sensors and sophisticated shoulder, knee and spine mechanics, the dummies are still used to simulate car accidents to make seat belts, seats and airbags in cars safer for (*male*) drivers.

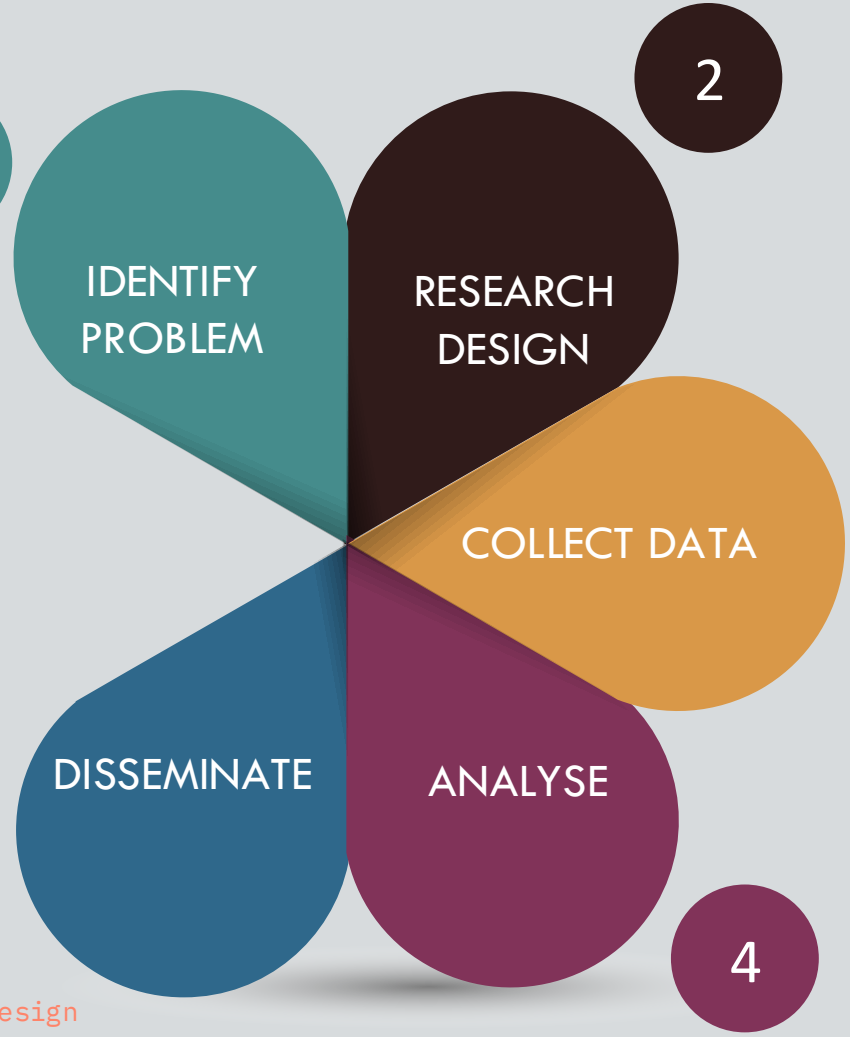
What can and should we do?

Fairness by design

(General) guidelines for Diversity-Aware AI

What role can diversity play in my work?
Which dimensions (gender, race, age etc.) are relevant?
What are the relevant factors?
What are the assumptions?

1



2

Which stakeholder do I need to involve?
How do I involve different interest/target groups?

3

Which factors could overlap?
What data do I need and how do I collect it? How might the relationships between the collectors and participants affect the data?

4

Conducting analyses of relevant factors in connection with norms, identities and relationships

5

Disclose and explain methods and results in the report.

(General) guidelines for Diversity-Aware AI

What role can diversity play in my work?
Which dimensions (gender, race, age etc.) are relevant?
What are the relevant factors?
What are the assumptions?

1

Example:

Application based on
Data collection via
smartphone in your (front) pocket

(General) guidelines for Diversity-Aware AI

2

Which stakeholder do I need to involve? How do I involve different interest/target groups?

Example:
Who are the stakeholders in the development of an electric toothbrush for children?

Which methods to use?

(General) guidelines for Diversity-Aware AI

Example:
Offline interviews on sensitive topics. Such us about opinion towards a device to communicate emotions via tactile interaction, enriched by further modalities. Gender, race and age of interviewer und interviewee can have an influence on the results.

3

Which factors could overlap? What data do I need and how do I collect it? How might the relationships between the collectors and participants affect the data?

(General) guidelines for Diversity-Aware AI

Example:

Gendered Research –
Evaluation by gender.
Such as in case of the device to communicate emotions via tactile interaction, enriched by further modalities, since men & women tend to communicate emotions differently.

4

Conducting analyses of relevant factors in connection with norms, identities and relationships

(General) guidelines for Diversity-Aware AI

Disclose and explain methods and results in the report.

5

… in the sense of the Research Codex, traceability and reproducibility.

Example:
The power of the framing bias!

Important note: this is valid for all software systems as well as product design cycles.

…and it is easily adaptable to other research areas.

AI will have a major impact on our lives.
So, what about fairness?

Diversity, Stereotypes, Bias

Virtual Reality

… create imaginary worlds where, for example, gender equality prevails - these experiences might change behavior in the real world.

Virtual & Augmented Reality

… can help us develop a higher level of empathy in certain contexts, which in turn can reduce implicit bias.

Avatars, Chatbots

… are often feminized, designed socioculturally "narrow" and reproduce stereotypes. Why? They can be designed in a way that they do not succumb to and even counter these stereotypes and prejudices.

Taking it a step further

For example, **virtual reality, augmented reality, chatbots** can be used to combat stereotypes and bias

What can and should we do?

Fairness by you

Acting with diversity in mind

Take a step back

Rethink your **own bias and stereotypes** before making decisions – regardless of whether they are based on AI or not!

Role of **diverse teams** - discuss ideas, methods, processes. Especially in case of processes with involving AI-systems!

AI Literacy – how do I interpret results and how do I integrate it in my processes?

We can not take bias out of people, but we can take them out of processes. Therefore: **objectify methods and processes!**

Being aware of diversity issues when using AI tools

Before an AI system is put into operation or in the purchase/development phase

- What data was used? Does it cover the target population? Is it balanced in terms of diversity dimensions such as gender, origin, etc.?
- What potential for abuse is there? How can we rule them out?(e.g. monitoring of employees)
- Test systems or have them tested - there is test software for ML-Models
- Allow systems to learn continuously and pursue a human-in-the-loop approach

What is our objective? Who set them? Was the team itself diversely balanced? Which minorities are excluded? ...

Who are the players involved in using the product? What is the process around it? What is the potential for incorrect use or misuse? ...

What data was used for the system? Was the system tested for bias (fairness)? How and what were the results?

Here too: Who is in the loop? What is the objective? Who tests how? Monitoring and re-evaluation.

„As more and more artificial intelligence is entering into the world, more and more emotional intelligence must enter into leadership.”

Amit Ray, KI-Experte

This presentation was not created by an AI.

Questions?