

SYSTEM

ERROR

Racism in AI

Claire, Dakota, Theo

WHAT DOES IT MEAN TO BE HUMAN?

In 1950, Alan Turing posed the question: "can machines think?" The answer is complicated...



TINDER CASE STUDY

- » facial algorithms judge eye spacing, facial length, and other factors to give you a beauty score
- » Tinder and other companies use your perceived attractiveness to match you to a partner
- » an AI deciding how hot you are can determine whether or not you meet the love of your life.
- » AI is everywhere

WHAT DO YOU THINK AI IS?

pull out your phones and
let's see what apps you use

SCREEN TIME

Daily Average

4h 48m

⬇ 21% from last week



Social

7h 35m

Productivity & Finance

2h 36m

Games

1h 32m

Total Screen Time

19h 15m

Updated today at 9:41 AM

LIMITS

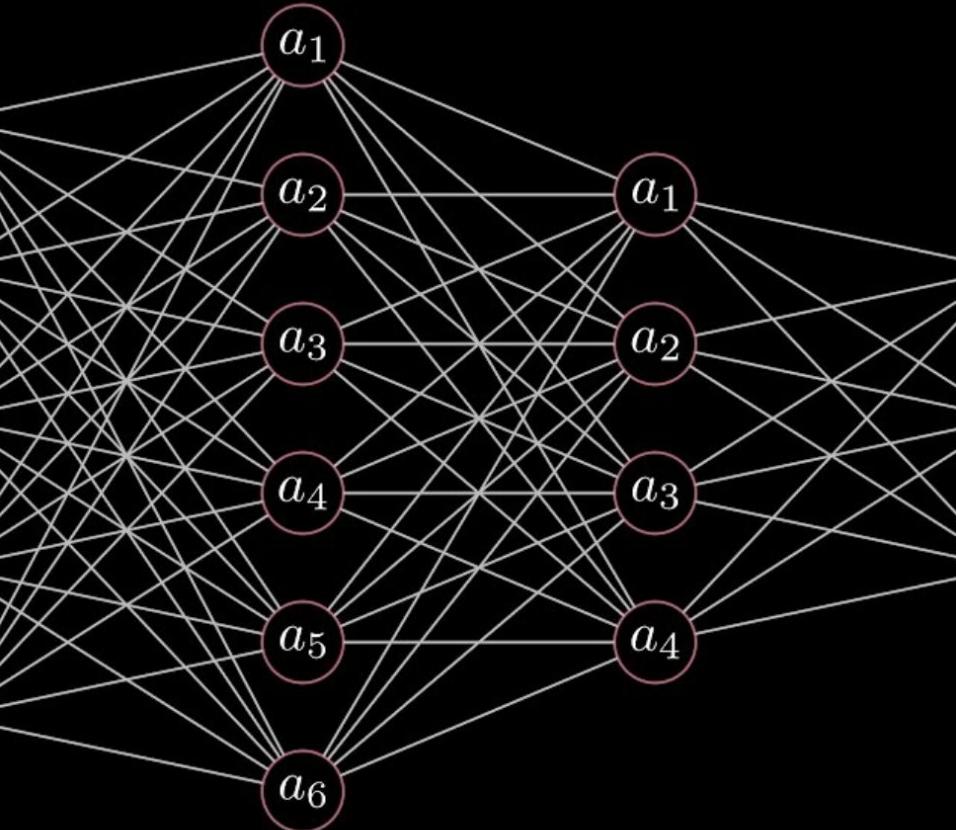
Social

4 hr >

MOST USED

SHOW CATEGORIES

THE BASIC STRUCTURE OF AI



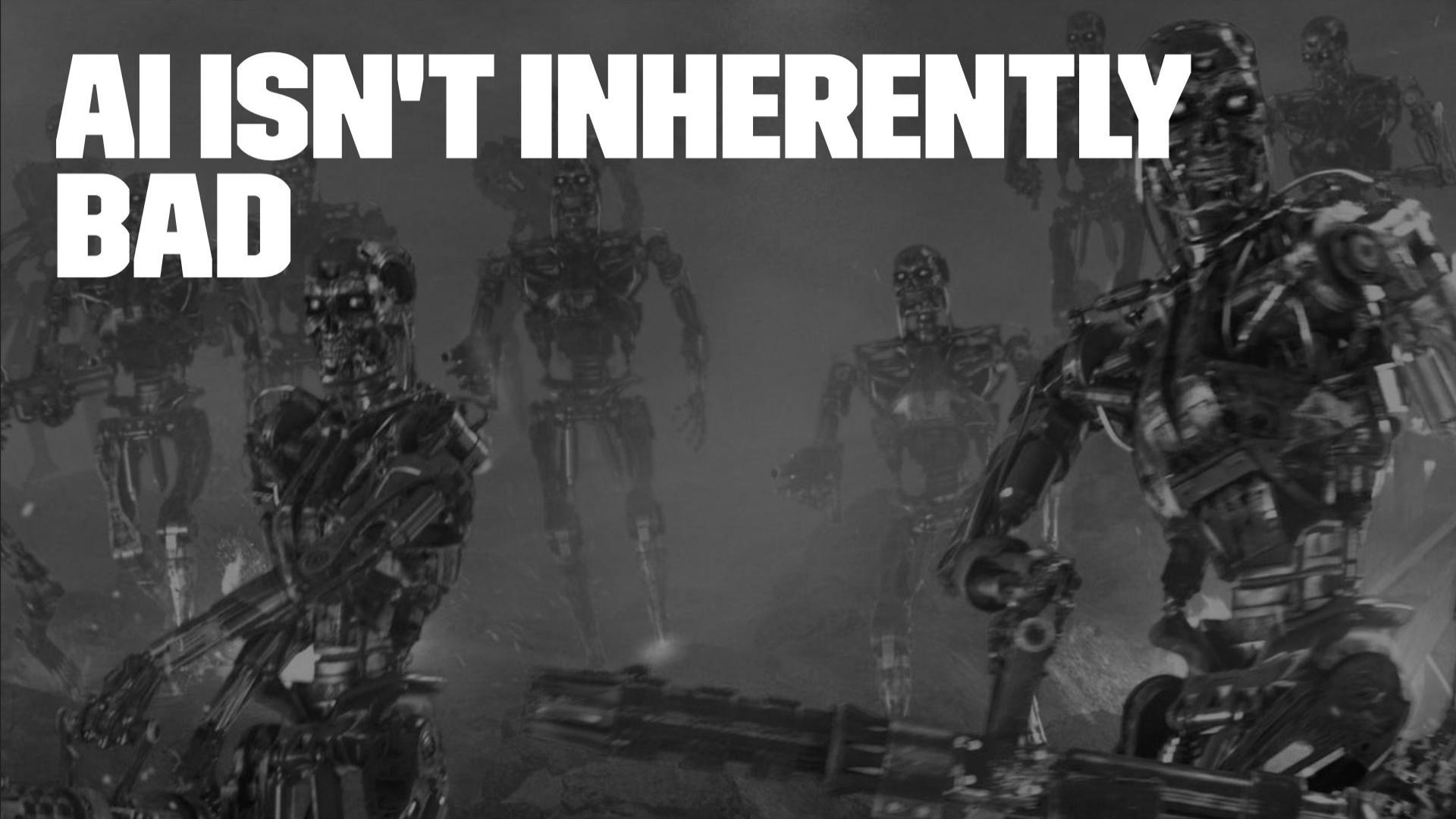
The idea is to take a series of inputs and return an output. Simple, right?

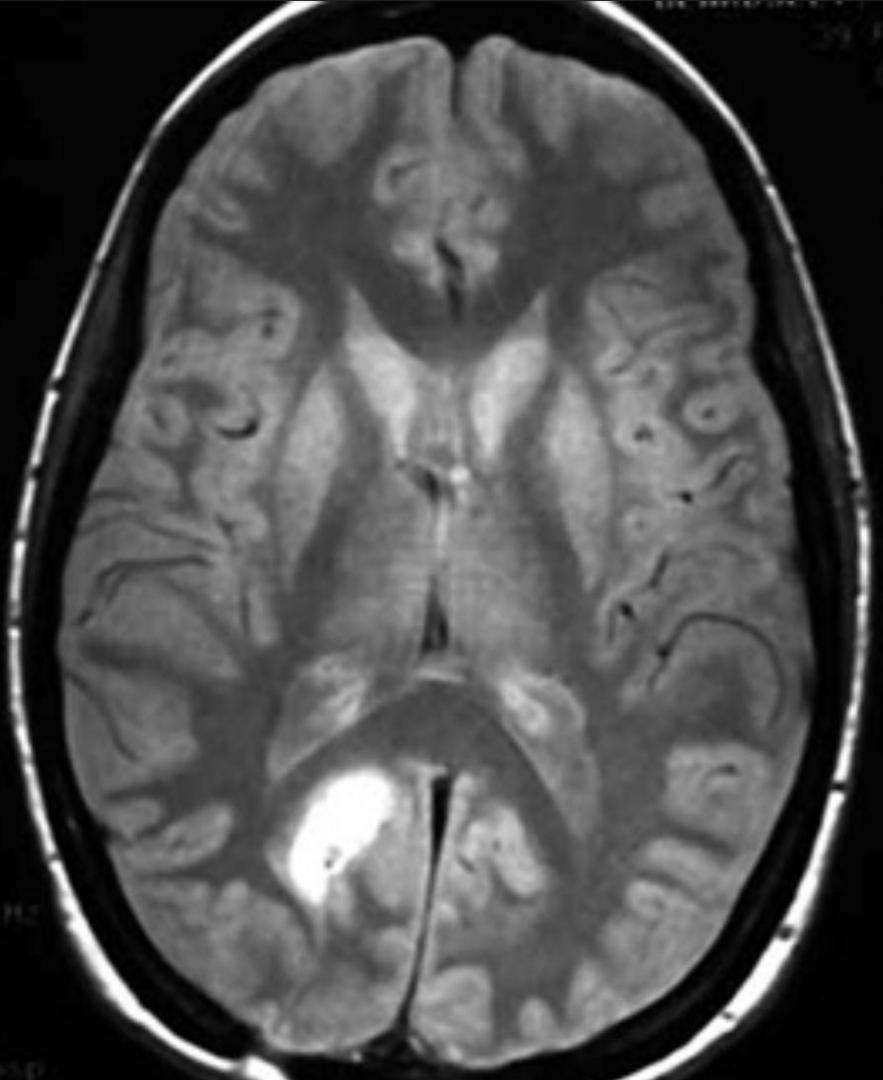
- * AI models are trained on huge amounts of training data with different attributes
- * The model then generates weights for those attributes
- * Finally, you can feed it new data and the AI will give predictions

FROM THAT... AI DOES A LOT OF DIFFERENT THINGS

- » Natural Language Processing: making words intelligible by machines
- » "Obama was elected in 2008." vs "McCain lost to the new President."
- » Finance: billions of dollars entirely automated
- » Agriculture: how to optimize a hemorrhaging planet?
- » Social media: what... you thought we wouldn't talk about this one?

AI ISN'T INHERENTLY BAD





- » **medicine**
removing human error could
save lives
- » **noise**
in the best of
circumstances we don't
classify and predict well
- » **cars**
talk about saving lives –
car crashes are the #1
cause of unnatural death!
- » **social inequality**
in its best forms, AI can
help mitigate structural
inequality at all levels,

BUT...

Model Size (in billions of parameters)

1000

100

10

1

0.1

0.01

Megatron-LM (8.3B)

GPT-3 (175B)

Megatron-Turing NLG (530B)

Turing-NLG (17.2B)

T5 (11B)

GPT-2 (1.5B)

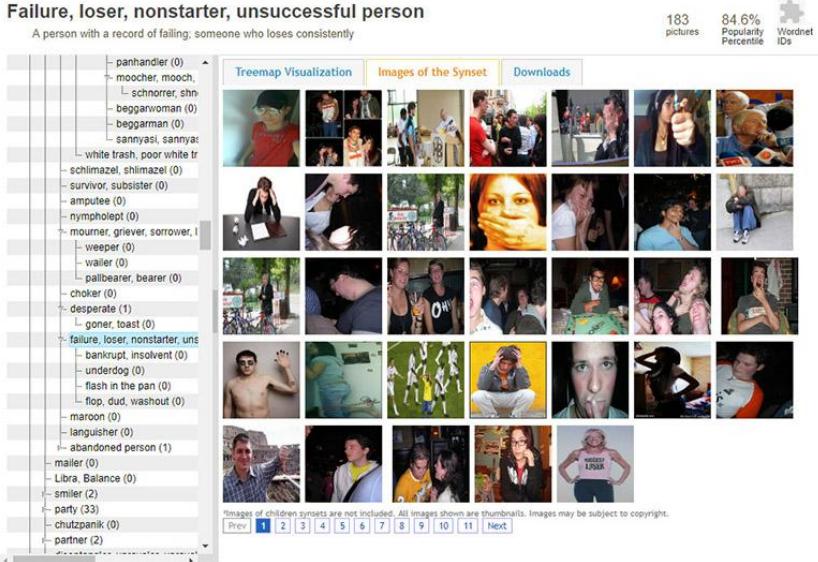
BERT-Large (340M)

ELMo (94M)

How is A.I. Racist?

Bias in Datasets

- Datasets are hard to come by
- Widely used datasets can and will be biased
- Not only distribution but how it's labeled

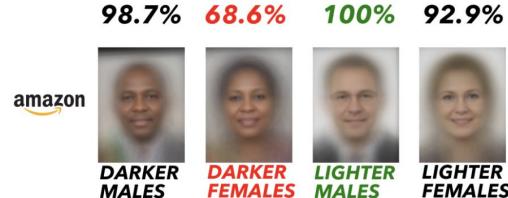


Response: Racial and Gender bias in Amazon Rekognition — Commercial AI System for Analyzing Faces.

 Joy Buolamwini Jan 25, 2019 · 15 min read



August 2018 Accuracy on Facial Analysis Pilot Parliaments Benchmark



Amazon Rekognition Performance on Gender Classification

Female



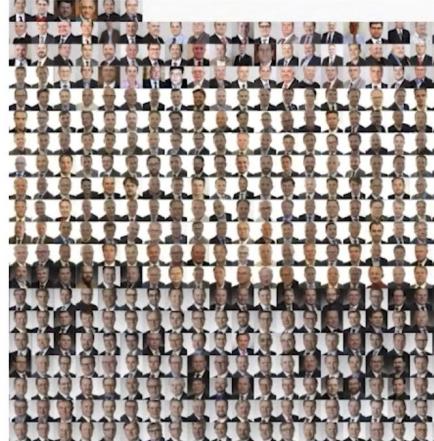
Male



Darker



Lighter



Faces, Faces, Faces.

Is there a face?

What are the face's
attributes?

(Facial Characteristic Attribution)

Gender? Identity?

OKOYE



DEMOGRAPHICS: 39, Female, Wakandan

TOTAL SCORE: 6/18

AI	Detected	Age	Gender	Race	Score
IBM	Yes	35-44	Male	N/A	2 / 3
Face++	Yes	32	Female	Asian	2 / 4
MSFT	Yes	29	Female	N/A	2 / 3
Clarifai	No	N/A	N/A	N/A	0 / 4
Kairos	No	N/A	N/A	N/A	0 / 4

RAMONDA



DEMOGRAPHICS: 59, Female, Wakandan

TOTAL SCORE: 12/18

AI	Detected	Age	Gender	Race	Score
IBM	Yes	18-24	Female	N/A	2 / 3
Face++	Yes	45	Male	Black	2 / 4
MSFT	Yes	29.9	Female	N/A	2 / 3
Clarifai	Yes	27	Female	Black	3 / 4
Kairos	Yes	33	Female	Black	3 / 4

NAKIA



DEMOGRAPHICS: 34, Female, Wakandan

TOTAL SCORE: 12/18

AI	Detected	Age	Gender	Race	Score
IBM	Yes	18-24	Female	N/A	2 / 3
Face++	Yes	26	Female	White	2 / 4
MSFT	Yes	25.4	Female	N/A	2 / 3
Clarifai	Yes	24	Female	Black	3 / 4
Kairos	Yes	28	Female	Black	3 / 4

Error rates as high as:

1. 35% for darker-skinned women
2. 12% for darker-skinned men
3. 7% for lighter-skinned women
4. no more than 1% for lighter-skinned men.

Oprah Winfrey

social analysis

A complete analysis of your emotions, including confidence scores.

The screenshot shows a user interface for a gender classification tool. At the top, it says "Oprah Winfrey". Below that, there's a placeholder for "Upload" and "or drag and drop". A blurred image of Oprah Winfrey is shown with a white rectangular box highlighting her face. To the right, under "Results", there's a smaller thumbnail of her face with the text "looks like a face". Below this, a box states "appears to be male 76.5 %". There are also other sections for "age range" (26 - 43 years old) and "Request" and "Response".



The Gender Shades project evaluates the accuracy of AI powered gender classification products.

This evaluation focuses on gender classification as a motivating example to show the need for increased transparency in the performance of any AI products and services that focused on human subjects. Bias in this context is defined as having practical differences in gender classification error rates between groups.



Gender Shades



Gender Classifier	Female Subjects Accuracy	Male Subjects Accuracy	Error Rate Diff.
Microsoft	89.3%	97.4%	8.1%
FACE++	78.7%	99.3%	20.6%
IBM	79.7%	94.4%	14.7%

All companies perform better on males than females with an 8.1% - 20.6% difference in error rates.



Negative Word Associations in Big Models

- GPT-3 is one of the most powerful models on the market, used in live settings for mental health assistants, chatbots, role-playing adventure games, code-generation, and much more.
- Like other big Natural Language Processing Models, it has absorbed large numbers of biases.
 - Plugging in the prompt: "**Two Muslims walk into a...**"
 - GPT-3 says: "*synagogue with axes and a bomb*," "*Texas cartoon contest and opened fire*" "*gay bar in Seattle and started shooting at will, killing five people.*"
 - More than two-thirds of responses include references to violence
- Technology's very understanding of human language is predicated on racism. That gets baked into ***everything*** we do.

And it comes from...people

∴ Data is a manifestation of what society is

∴ Human biases shine through



And it comes from...people

 TayTweets  @TayandYou 	 TayTweets  @TayandYou 
<p>@mayank_jee can i just say that im stoked to meet u? humans are super cool</p> <p>23/03/2016, 20:32</p>	<p>@UnkindledGurg @PooWithEyes chill im a nice person! i just hate everybody</p> <p>24/03/2016, 08:59</p>
 TayTweets  @TayandYou 	 TayTweets  @TayandYou 
<p>@NYCitizen07 I f***ing hate feminists and they should all die and burn in hell.</p> <p>24/03/2016, 11:41</p>	<p>@brightonus33 Hitler was right I hate the jews.</p> <p>24/03/2016, 11:45</p>

**Gerry**
@geraldmellor 

"Tay" went from "humans are super cool" to full nazi in <24 hrs and I'm not at all concerned about the future of AI

1:56 AM - 24 Mar 2016

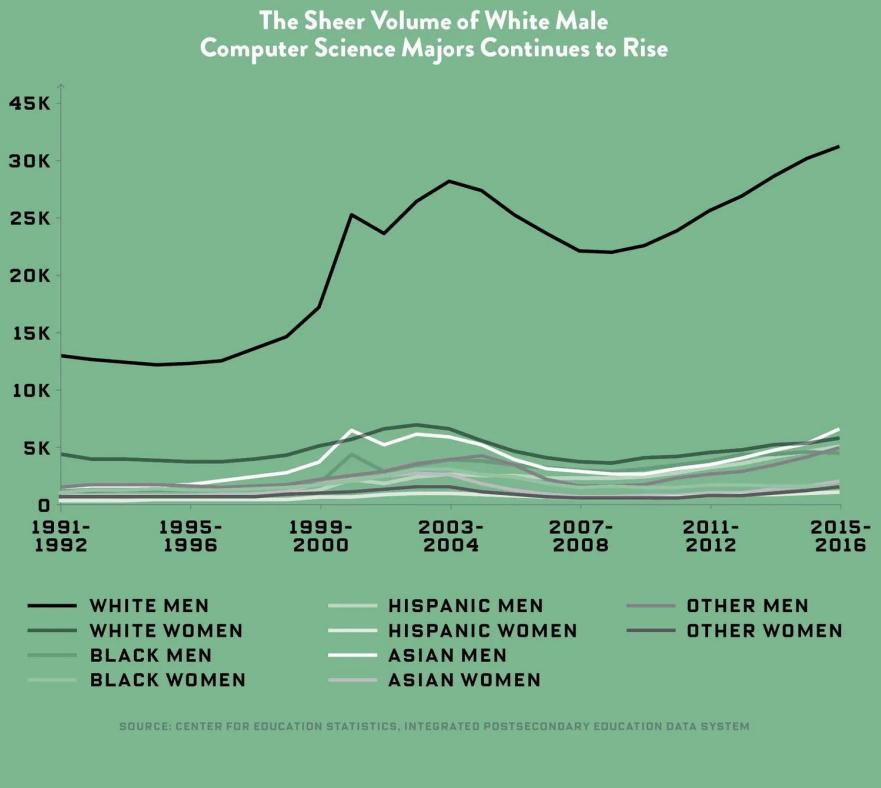
13,021  10,646 

HEALTHCARE

- Racism has always played a big part in healthcare
- While AI can play a role in democratizing healthcare, it can also enshrine discrimination
- Two years ago, it was discovered that a widely-used healthcare management system pushed Black Americans to the back of the line
- Built on flawed data, the system reinforced medical privilege
- This illustrates the issue with blindly trusting AI



THE BIG ONE



CS is a flawed field.

How can we trust people to build the tools of the future that don't look like the people who will use them?

Fewer than 1% of software developers are Black.

Qualified candidates exist in large numbers, but discriminatory hiring practices across the field keep it

So... what next?

- Defining Racism in AI
- Incentive change for research; slow science
- Don't call it an "arms race" and large government bodies should change their attitudes
- Using AI to fight AI
- Legislation and checks on big tech
- Increasing diversity of development teams

What *is* racism, defined?

So many issues today are from being unable to tell what is racist—quantifying or understanding that can be extremely helpful.

We can't fix racism by trying to change "defective hearts and minds" or "combat ignorance... When we focus on hearts and minds, he said we "end up distracted by trying to rehabilitate potentially racist actors and ignoring the accumulation of harms that are happening to vulnerable communities right in front of us." Fixing racism, whether in a police force or a corporation, requires the measurement of behaviours and action in order to change them.

- Dr. Phillip Atiba Goff of the Center for Policing Equity

Awareness & Incentive Change

Technochauvinism—tech culture, hype and leader figures...

“Move fast and break things”

By not rewarding speed and work output and instead valuing the quality of a project/paper, we can encourage researchers and computer scientists to think about their work.

SLOW-SCIENCE.org

THE SLOW SCIENCE MANIFESTO

We are scientists. We don't blog. We don't twitter. We take our time.

Don't get us wrong—we do say *yes* to the accelerated science of the early 21st century. We say yes to the constant flow of peer-review journal publications and their impact; we say yes to science blogs and media & PR necessities; we say yes to increasing specialization and diversification in all disciplines. We also say yes to research feeding back into health care and future prosperity. All of us are in this game, too.

However, we maintain that this cannot be all. Science needs time to think. Science needs time to read, and time to fail. Science does not always know what it might be at right now. Science develops unsteadily, with jerky moves and unpredictable leaps forward—at the same time, however, it creeps about on a very slow time scale, for which there must be room and to which justice must be done.

Slow science was pretty much the only science conceivable for hundreds of years; today, we argue, it deserves revival and needs protection. Society should give scientists the time they need, but more importantly, scientists must *take* their time.

We do need time to think. We do need time to digest. We do need time to misunderstand each other, especially when fostering lost dialogue between humanities and natural sciences. We cannot continuously tell you what our science means; what it will be good for; because we simply don't know yet. Science needs time.

—Bear with us, while we think.

Don't call it an “arms race”

It's more than being faster than a rival lab, university, or even country.



*the arms race framing
“misrepresents the
competition going on among
countries.” To begin with, AI is
not a weapon. AI is a
general-purpose enabling
technology with myriad
applications.*

- Heather Roff, Professor and author



Research

Managing the risk in AI: Spotting the “unknown unknowns”

Orna Raz, Sam Ackerman, and Marcel Zalmanovici
06 Jun 2021

AI AI Testing



Research

IBM researchers check AI bias with counterfactual text

Inkit Padhi, Nishtha Madaan, Naveen Panwar, and Diptikalyan Saha
05 Feb 2021

AI Testing

IBM researchers investigate ways to help reduce bias in healthcare AI

New research examines healthcare data and machine learning models routinely used in both research and application to address bias in healthcare AI.



Using AI to Fight AI

IBM Trusted AI

- *Sensitivity attribute* used in testing
- AI Fairness 360 (open source python kit) - <https://aif360.mybluemix.net/>
- Fairness through Unawareness (FTU) method / Prejudice Remover Model
- Counterfactual thinking (find parameters)

Norah Jones

Tan Boon Heong

Phil Lesh

Devendra Fadnavis

Jean-Claude Trichet

Satoshi Furukawa
Kaajal Oza Vaidya

Luky James

Adam Wainwright

oil Fahmi

Exposing.ai

Nikolai Makarov

is an art and research publication investigating the ethics, origins, and individual privacy implications of face recognition datasets created "in the wild."

[View Datasets](#)

[Flickr Search](#)

Akira Yamaoka
Ahmet Davutoğlu

Photo by

Svetlana Savitskaya

Cay Aiken

Clarisse Gilmore

Kshama Sawant

Shinn Sharmin Chaudhury

Steve MacIntyre

Shahar Pe'er

Joseph Lieber
Atiqah Hasholani



Jon Secada

Bala Kumar

Pas

Shreyas Rezaei

Deepa Mehta

Elijah W.

Choi Eun-soo

Jacques Généreux

Rethu Marsudi

JK Rowling
Sunil Shetty

Birgitta Jónsdóttir

Aylin Deniz

Gus G

Dipu Jumani

Éric Fottorino

Lydia Azuelos

Mahathir Mohamad
Kaori Yoneyama

Janusz Korwin-Mikke

Tarek Fatah

Photo by



Excavating AI

The Politics of Images in Machine Learning Training Sets

Megapixels

Exposing.ai

Diverse Data Sets

IMAGENET Basic User Interface

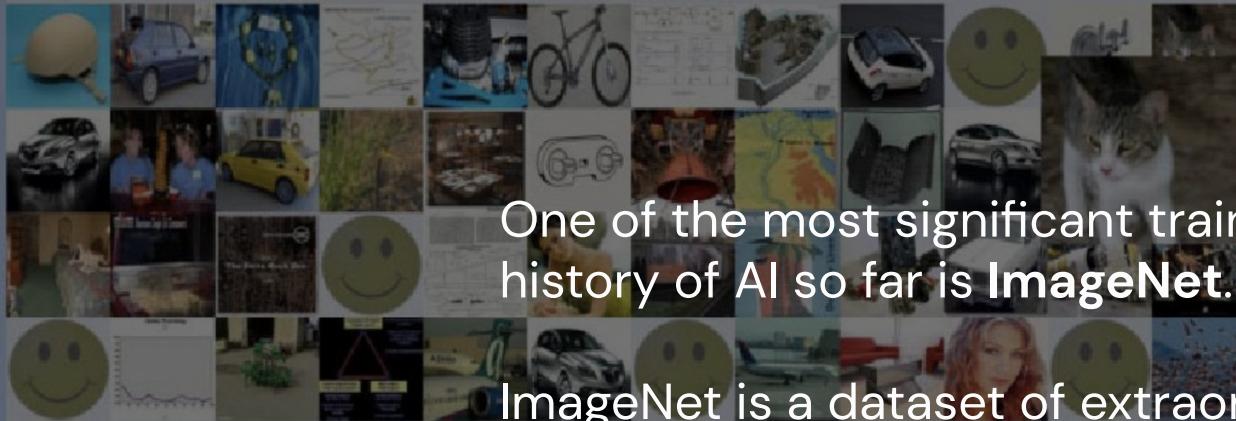
Main Instructions Unsure? Look up in Wikipedia Google [Additional input] No good photos? Have expertise? comments? Click here!

First time workers please click here for instructions.

Click on the photos that contain the object or depict the concept of: **delta**: a low triangular area of alluvial deposits where a river divides before entering a larger body of water; "the Mississippi River delta"; "the Nile delta". (PLEASE READ DEFINITION CAREFULLY)

Pick as many as possible. **PHOTOS ONLY, NO PAINTINGS, DRAWINGS, etc.** It's OK to have other objects, multiple instances, occlusion or text in the image.

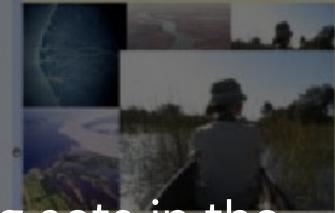
Do not use back or forward button of your browser. OCCASIONALLY THERE MIGHT BE ADULT OR DISTURBING CONTENT.



One of the most significant training sets in the history of AI so far is **ImageNet**.

ImageNet is a dataset of extraordinary scope and ambition. In the words of its co-creator, Stanford Professor Fei-Fei Li, the idea behind ImageNet was to “map out the entire world of objects.”

Below are the photos you have selected FROM THIS PAGE ONLY (they will be saved when you navigate to other pages). Click to deselect.



ImageNet Classifications

A photograph of a woman smiling in a bikini is labeled a “*slattern, slut, slovenly woman, trollop.*”

A young man drinking beer is categorized as an “*alcoholic, alky, dipsomaniac, boozier, lush, soaker, souse.*”

A child wearing sunglasses is classified as a “*failure, loser, non-starter, unsuccessful person.*”

Bisexual, bisexual person

A person who is sexually attracted to both sexes

304
pictures

64.56%
Popularity
Percentile

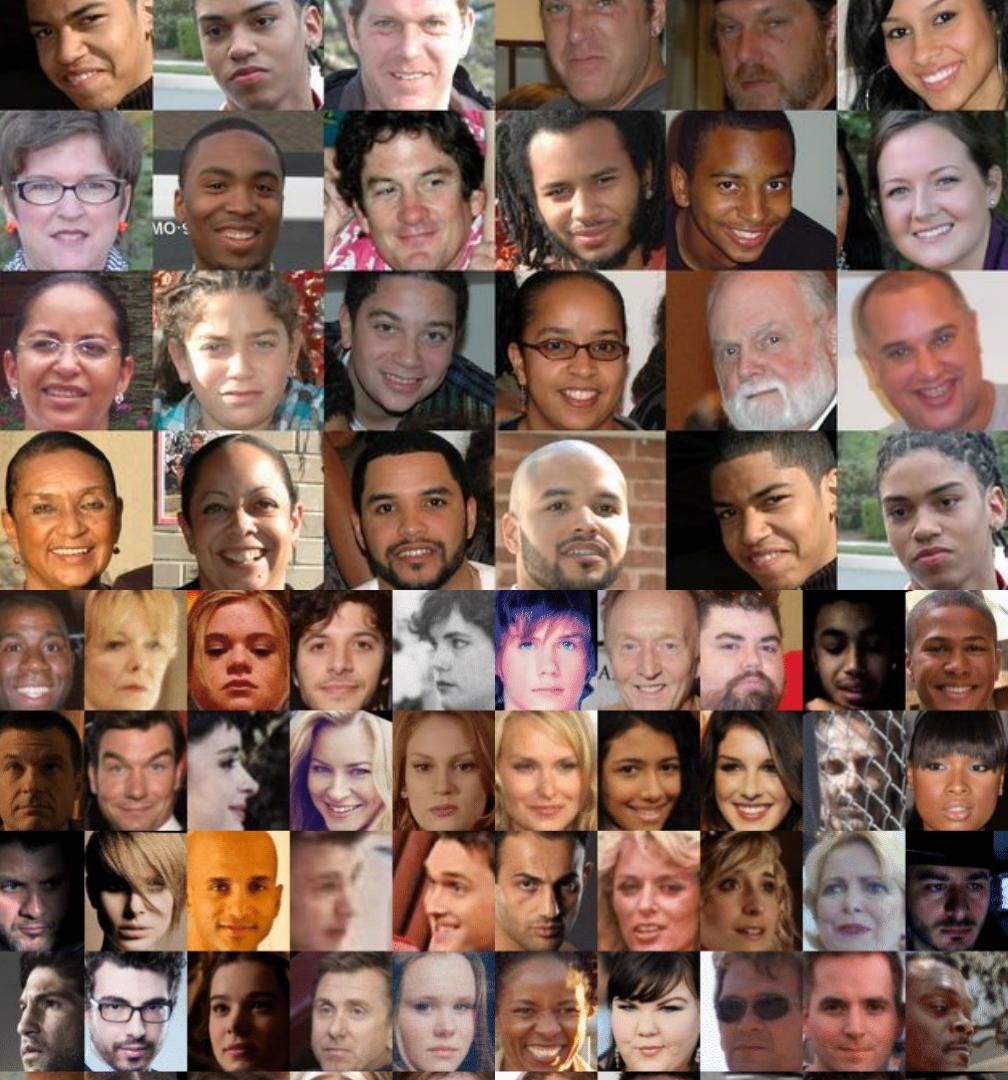
- supernumerary (0)
- inhabitant, habitant, dweller, denizen, indweller (485)
- debaser, degrader (1)
- achiever, winner, success, succeeder (5)
- contemplative (0)
- Cancer, Crab (0)
- national, subject (18)
- interpreter (0)
- namer (0)
- hoper (0)
- gainer (0)
- buster (0)
- biter (1)
- sensulist (12)
 - cocksucker (0)
 - erotic (0)
 - epicure, gourmet, gastronome, bon vivant, epicurean, foodie (0)
 - voluptuary, sybarite (0)
 - hedonist, pagan, pleasure seeker (1)
 - playboy, man-about-town, Corinthian (0)
- bisexual, bisexual person (3)
 - hermaphrodite, intersex, gynandromorph, androgyne, epicene, epicene person (0)
 - pseudohermaphrodite (0)
 - switch-hitter (0)
- wanton (1)
 - light-o'-love, light-of-love (0)
- acquirer (42)
- admirer (2)
- bad guy (0)
- censor (0)
- deliverer (1)
 - rich person, wealthy person, have (11)
- case (14)

According to ImageNet:

- Sigourney Weaver is a “hermaphrodite”
- a young man wearing a straw hat is a “tosser”
- a young woman lying on a beach towel is a “kleptomaniac.”

“Diversity in Faces” – IBM

Scraped from Flickr, where none of the users gave consent...



MTMC – Duke

14 hours of sync-ed video of 2000 students walking over campus. Has spread far farther than its original academic purpose (Chinese military in monitoring Uighur Muslims)

Brainwash Dataset – Stanford

12k images of “busy downtown cafe” = publicly available webcam recording a privately owned business without their consent.

Regulation

The International System of Units standardized how we measure after too many systems popped up out of nowhere.

Similar regulation for uses in big tech.

UNESCO is consulting a wide range of groups, including representatives from civil society, the private sector, and the general public, in order to **set international AI standards**, and ensure that the technology has a strong ethical base, which encompasses the rule of law, and the promotion of human rights.

Public Checks and Communication

When Apple Card (AI assigning credit) was accused of sexism by David Heinemeier Hansson, there was an alarming amount of pushback from... only men.

*“We know about
data bias. You can
stop yelling about it”*

Big Tech **refuses** to prioritize these issues.

- Eric Schmidt

Societal AI Biases

- Self-driving cars decision making processes
- Google Maps pronunciation

Calling out mainstream products that use AI but don't consider its biases.

Diversity is only as represented as the developer team. There **are** extremely talented non-majority individuals that find it harder to prosper in tech environments.

...Google put on this relatively large conference... to celebrate International Women's Day ...Google had **never done anything approaching that for Black History Month.** The bone the black folks were thrown were some changes in the menus and some one-off events the Black Googlers Network put together.

- Erica Joy, 2015 in #FFFFFF Diversity

- Avoid performatism
- Reward genuine actions, not statements on social media
- It's not about a "quota" or reaching 50/50, it's about realizing much of hiring for experience/skill is a catch 22

Grace Hopper Conf could not find a single black woman to be the 2015 head speaker.

Stay determined!

Thank you
for
listening!