

RSA[®]Conference2021

May 17 – 20 | Virtual Experience



SESSION ID: STR-R08

Top Seven AI Breaches: *Learning to Protect Against Unsafe Learning*

Davi Ottenheimer

VP Trust and Digital Ethics

inrupt

#RSAC

Agenda

1. What Should Be Expected of AI?
2. What Has Been Delivered by AI?
3. Where Things Go Wrong
(Top 7 AI Breaches... of Trust)

TRIGGER WARNING:
*Content may be
Disturbing or
Upsetting*



Two Quick Stories

1. Tatra 87 - Killed More Nazi Generals than WWII

- Hitler's "Car of the Future" (85hp V8: 19 mpg, 100 mph)
- Also Called a "Czech Secret Weapon"
- Germans Couldn't Operate It
- Designs Stolen by Porsche



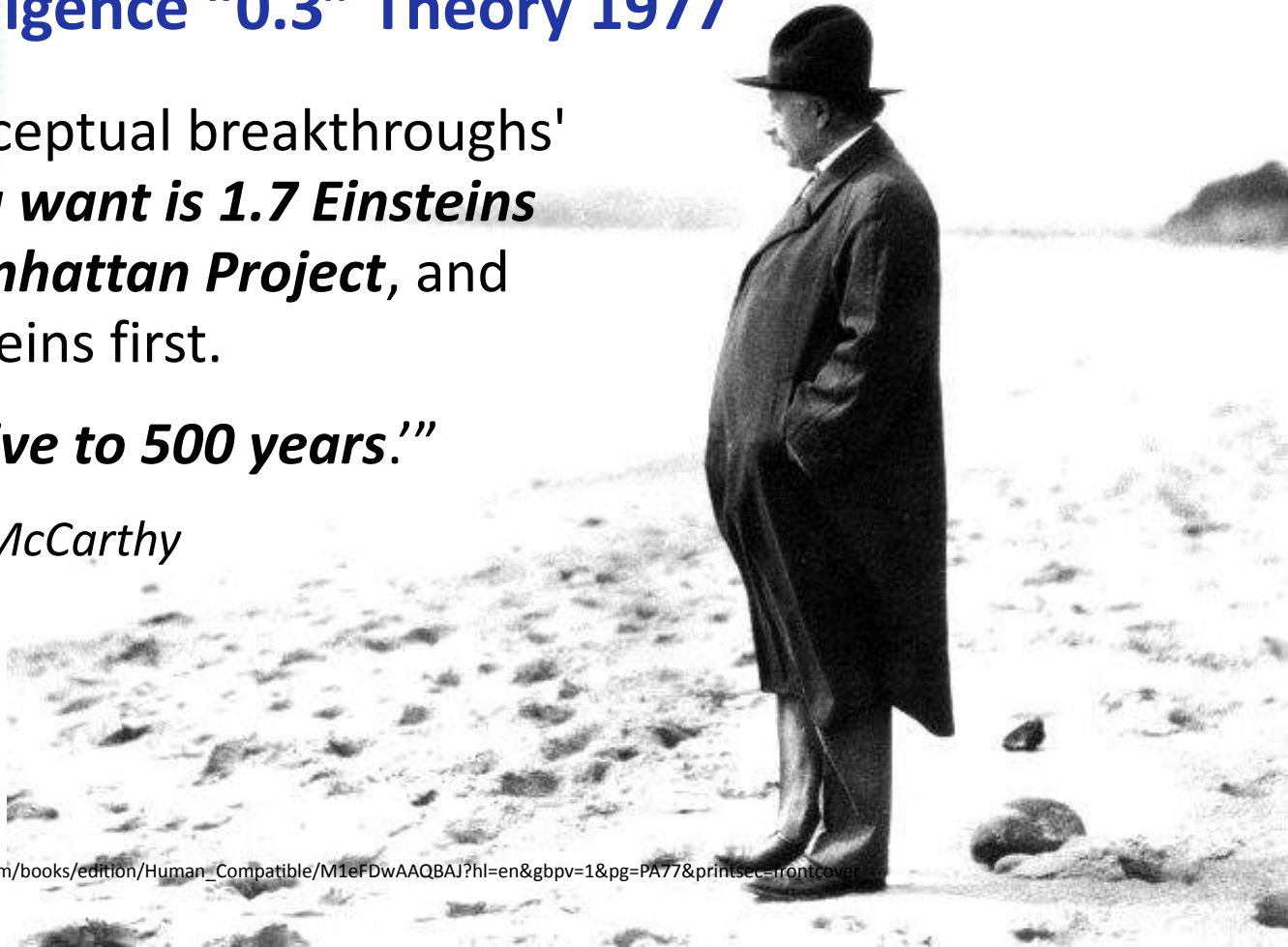
Source: <https://www.thedrive.com/news/35116/the-oddball-tatra-t87-was-europes-most-genius-prewar-car-and-great-at-killing-nazis>

2. Artificial Intelligence “0.3” Theory 1977

“...AI requires 'conceptual breakthroughs' because '**what you want is 1.7 Einsteins and 0.3 of the Manhattan Project**, and you want the Einsteins first.

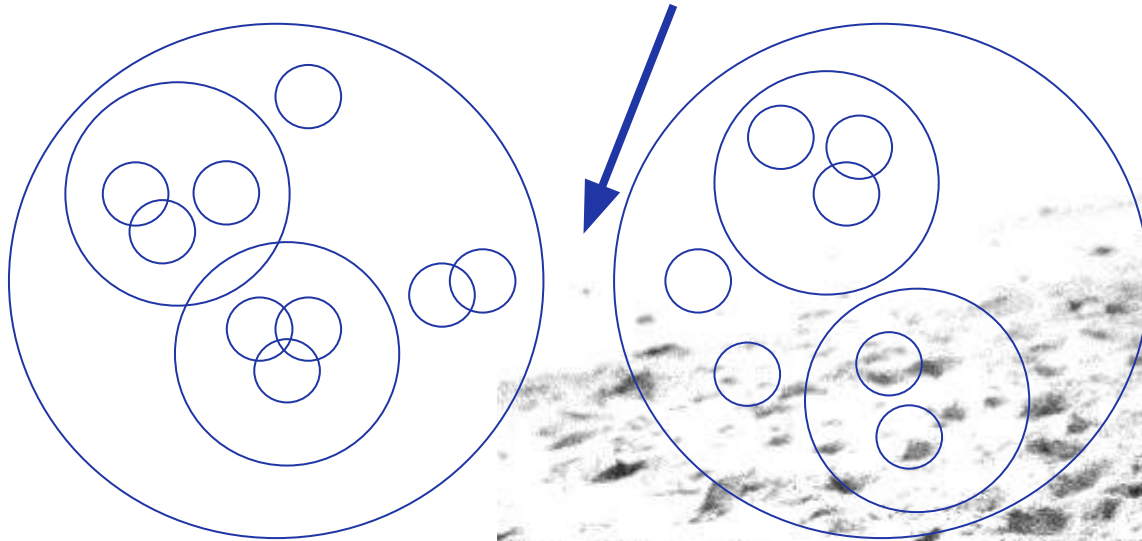
I believe it'll take **five to 500 years.**”

-- John “Father of AI” McCarthy



2. Artificial Intelligence “0.3” Theory 1977

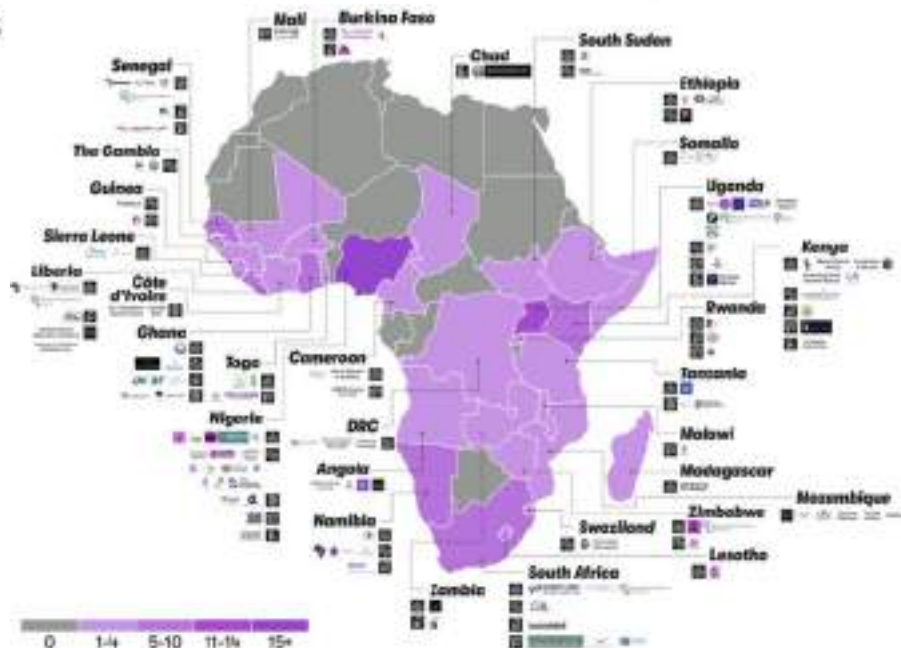
- First: 1.7 Science **Collaboration** (Nature)
- Second: 0.3 Military **Competition** (Human)



1977 “0.3” Theory = AI “Should” Prevent Harms

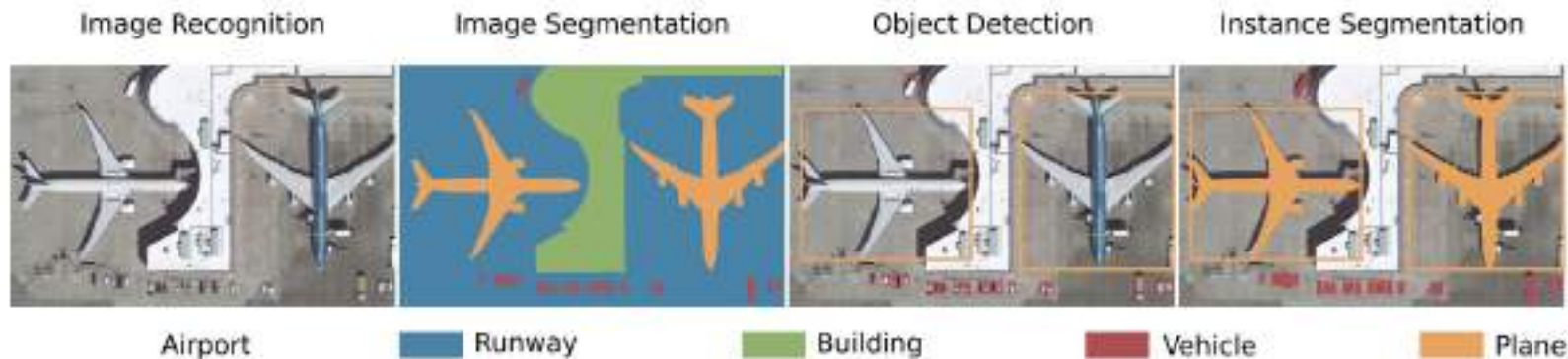
Afrofeminist Concerns About Data Use

1. Ethics of data collected from secondary sources
2. Political manipulation and motives within available data
3. Re-traumatization of respondents
4. Secure storage of data and data privacy
5. Focus on quantitative data as rigorous
6. Shadowbanning on digital platforms
7. Consent from minors
8. Trust and cultural or patriarchal barriers
9. Treatment of Women of Colour by technology platforms
10. Women's Safety on digital platforms



1977 “0.3” Theory = AI “Should” Stop Catastrophe

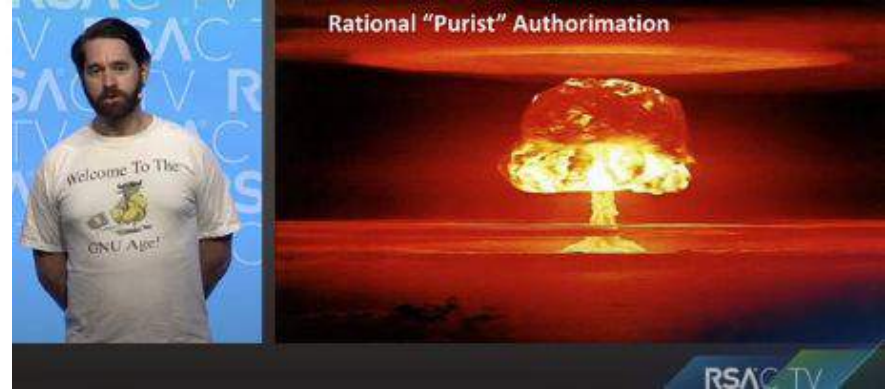
“NATO... is a great place to responsibly think about how to harness the good parts of this technology and how to ***prohibit the parts that would be catastrophic*** for humanitarian law and human rights law, and people at the end of the day.”



Source: <https://www.nato.int/docu/review/articles/2020/11/24/artificial-intelligence-at-nato-dynamic-adoption-responsible-use/index.html>
<https://www.politico.eu/article/nato-ai-artificial-intelligence-standards-priorities/>

From 1930s to Today... is “AI” the Doomsday Machine?

2016 my RSA Conference Talk:
Facebook a Doomsday Machine



2020 *The Atlantic*:
Facebook a Doomsday Machine

“Limitations to the Doomsday Machine comparison are obvious: Facebook cannot in an instant reduce a city to ruins the way a nuclear bomb can [yet] ***in every situation of extremist violence we’ve looked into, we’ve found Facebook...***”

What *Should Be* Expected of AI

1.

Incompleteness

2.

Imitation

3.

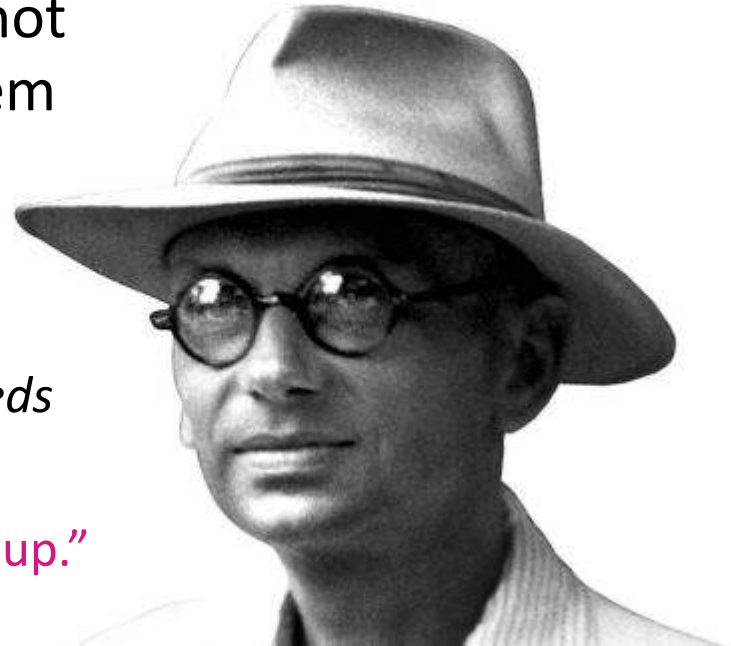
Efficiency

1. Incompleteness: 1929 Gödel

“...in any consistent axiomatic mathematical system there are propositions that cannot be proved or disproved within the system and that the consistency of the axioms themselves cannot be proved.”

Translation: a computer never can answer all questions to achieve its objectives -- always needs external/inherited sources and oversight.

“Intelligence is the spice in the soup, not the soup.”



1. Simple “Thoughts” in Binary Circuits: 1937 Shannon

Computer Logic (correlation)

- $A = Z$
- $Z = A$
- $2 + 2 = 4$
- $4 = 2 + 2$



1948 "A Mathematical Theory of Communication"

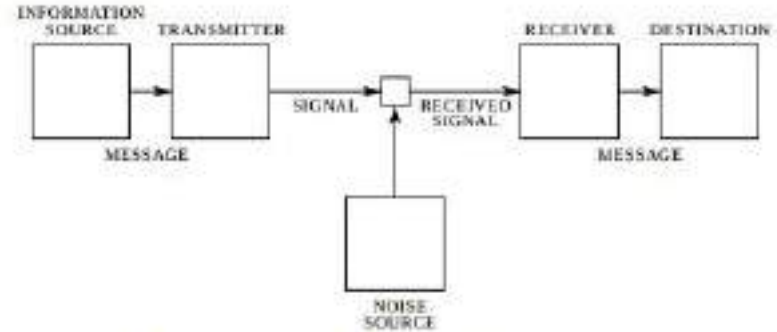
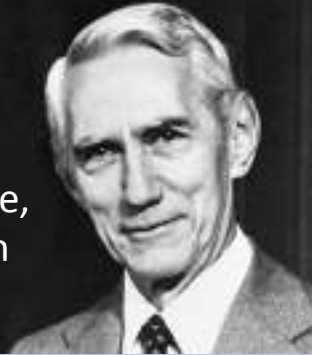


Fig. 1—Schematic diagram of a general communication system.

*If Bob bought X toothpaste,
then will buy Y toothbrush*



Human Reasoning (causation)

- Mother leads to pleasure
- Cloud leads to rain
- Violence leads to pain

2. Imitation: 1950 Turing

(Dead at 41 from persecution: 7 June 1954)

1951 SNARC Neural Net

1952 IAS Machine

1955 "AI" Coined

1957 IBM 704 Perceptron

(Recognized Binary: Right From Left)

Electronic 'Brain' Teaches Itself

The Navy last week demonstrated the embryo of an electronic computer named the Perceptron which, when completed in about a year, is expected to be the first non-living mechanism able to "perceive, recognize and identify its surroundings without human training or control." Navy officers demonstrating a preliminary form of the device in Washington said they hesitated to call it a machine because it is so much like a "human being without life."

Dr. Frank Rosenblatt, research psychologist at the Cornell Aeronautical Laboratory, Inc., Buffalo, N. Y., designer of the Perceptron, conducted the demonstration. The machine, he said, would be the first electronic device to think as the human brain. Like humans, Perceptron will make mistakes at first, "but it will grow wiser as it gains experience," he said.

The first Perceptron, to cost about \$100,000, will have about 1,000 electronic "association cells" receiving electrical impulses from an eyelike scanning device with 400 photocells. The human brain has ten billion responsive cells, including 100,000,000 connections with the eye.

Difference Recognized

The concept of the Perceptron was demonstrated on the Weather Bureau's \$2,000,000 IBM 704 computer. In one experiment, the 704 computer was shown 100 squares situated at random either on the left or the right side of a field. In 100 trials, it was able to "say" correctly ninety-seven times whether a square was situated on the right or left. Dr. Rosenblatt said that after having seen only thirty to forty squares the device had learned to

recognize the difference between right and left, almost the way a child learns.

When fully developed, the Perceptron will be designed to remember images and information it has perceived itself, whereas ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons, Dr. Rosenblatt said, will be able to recognize people and call out their names. Printed pages, longhand letters and even speech commands are within its reach. Only one more step of development, a difficult step, he said, is needed for the device to hear speech in one language and instantly translate it to speech or writing in another language.

Self-Reproduction

In principle, Dr. Rosenblatt said, it would be possible to build Perceptrons that could reproduce themselves on an assembly line and which would be "conscious" of their existence.

Perceptron, it was pointed out, needs no "priming." It is not necessary to introduce it to surroundings and circumstances, record the data involved and then store them for future comparison as is the case with present "mechanical brains." It literally teaches itself to recognize objects the first time it encounters them. It uses a camera-eye lens to scan objects or survey situations, and an electrical impulse system, patterned point-by-point after the human brain does the interpreting.

The Navy said it would use the principle to build the first Perceptron "thinking machines" that will be able to read or write.

3. Efficiency (e.g. Translation): 1958 US Military

"The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself, and be conscious of its existence.

Later perceptrons will be able to **recognize people** and call out their names and **instantly translate speech** in one language to speech and writing in another language, it was predicted."

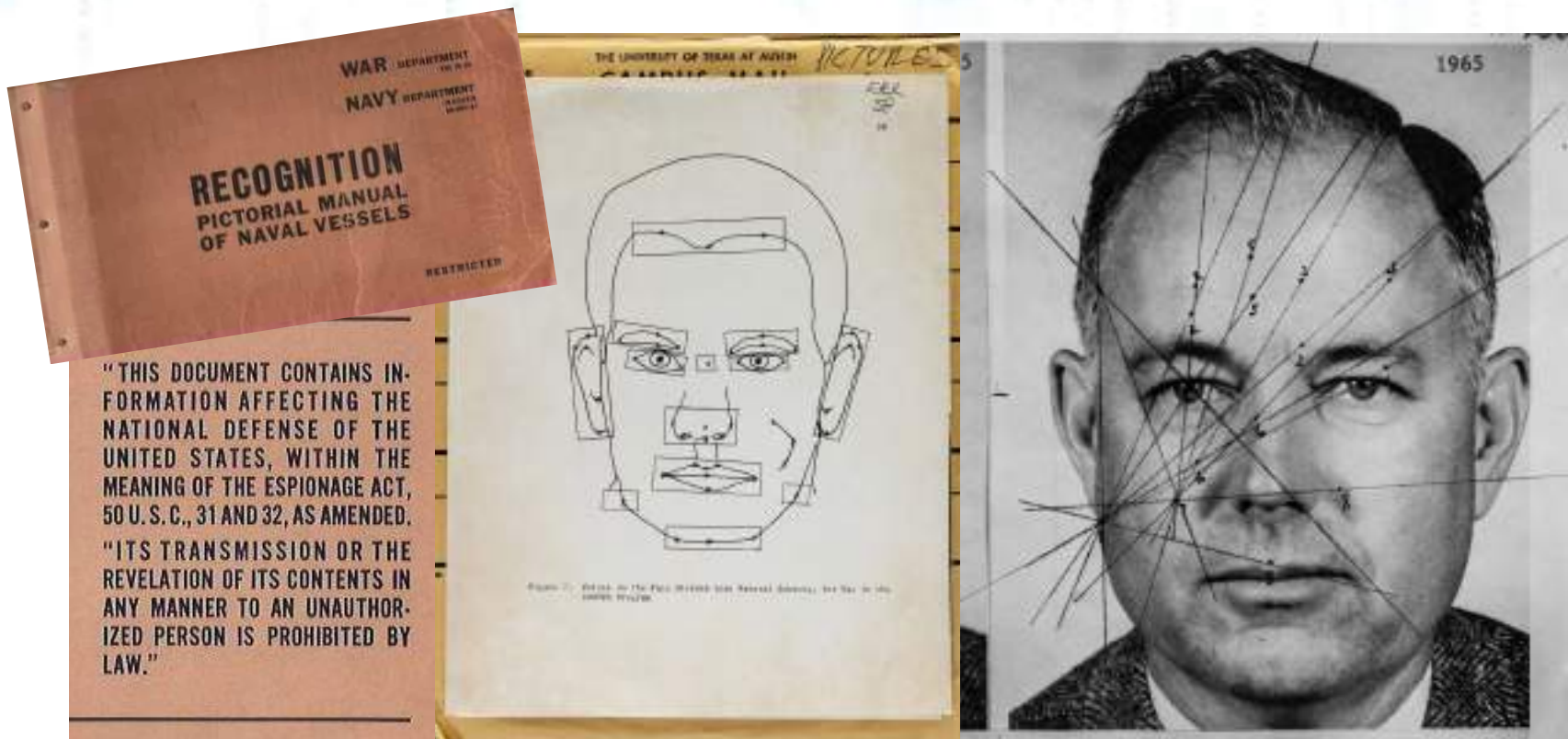
-- The New York Times, 08 July 1958



Claus Scholz
1958 robot MM7

Source: <https://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html>

3. Efficiency (e.g. Facial Recognition): 1964 US Military



Source: <https://www.wired.com/story/secret-history-facial-recognition/>
<https://www.flyingpenguin.com/?p=25255>
<https://acepilots.com/ships/manual.html>

3. Efficiency (e.g. Plate Surveillance): 1966 US Police

New York State identification and intelligence system, Albany, N.Y.

Grant No. 040 (June 1966). \$180,000 (23 months). (\$1.4M equivalent in 2021)

Development of automatic license plate scanning system for conversion of license plate characters and optical data to electrical signals permitting computer search and retrieval against "wanted car" data. Contemplates production of prototype system capable of field test and evaluation.



3. Efficiency in School Admissions: 1979 Racism...Uh Oh!



“...by *ceding agency* to a technical system, all student applicants would be subject to precisely the same evaluation, thus creating a *fairer process*.”

FALSE

What Has Been Delivered by AI

*Incompleteness:
Gibberish*

*Imitation:
Fraud*

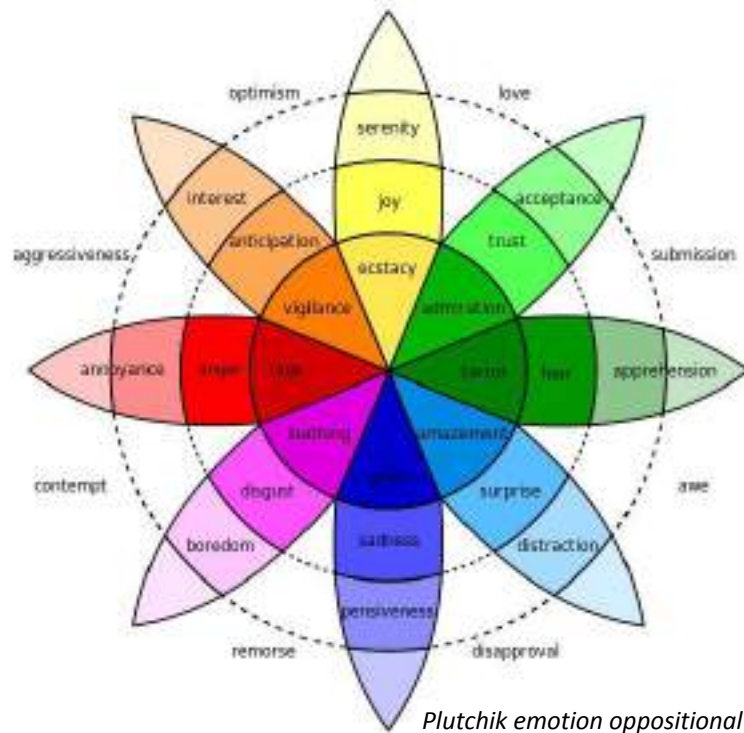
*Efficiency:
“Breaking Bad”*

1. Incompleteness: Gibberish of GPT-3 (Correlations)

“...[up to] 90% of the tested systems’ answers did not change when the words were shuffled.”

As Shannon modeled in 1937:

- “Marijuana causes cancer”
falsely treated as equivalent to
- “Cancer causes marijuana”



<https://venturebeat.com/2021/02/09/openai-and-stanford-researchers-call-for-urgent-action-to-address-harms-of-large-language-models-like-gpt-3/>
<https://www.technologyreview.com/2021/01/12/1016031/jumbled-up-sentences-ai-doesnt-understand-language-nlp-bert-fix/>
<https://aninjusticemag.com/the-lie-of-a-color-blind-society-886c98281ec9>
<https://www.nature.com/articles/s41586-021-03215-w>
<https://www.nature.com/articles/d41586-021-00530-0>

Source:

2. Adult Imitation: Fraud, Impersonation, Deepfakes

White Men Tend To Only Listen To Other White Men

I'm here to tell you that we have only two choices: Prosper under racism or struggle against it.

*White men **fear loss of** control.
Privilege of binary power gone.
Anyone now can appear just
like them...*



2. Childlike Imitation: More Danger/Less Responsibility

While new "learning" machines have been killing humans with abandon; in America a child was sentenced in court for... ***picking a flower.***

“Should a child that believes in Santa Claus, the Easter Bunny and the tooth fairy be making life-altering decisions? ... research says no, ‘even at 10, 11 and some 14-year-olds’”



Source: https://journalnow.com/north-carolina-sends-6-year-olds-to-court-why-some-say-its-time-for-change/article_e2a15a82-8383-11eb-91ee-43ce7c88753b.html

3. Efficiency: 2016 Repeat of 1906 “The Jungle”

... injuries more common in warehouses with robots ...

“...freedoms get optimized out, jobs are becoming more intense, stressful, and dangerous... 10% of full-time workers sustained serious injuries in 2018, more than twice the national average...”

Serious injury rates at Amazon fulfillment centers

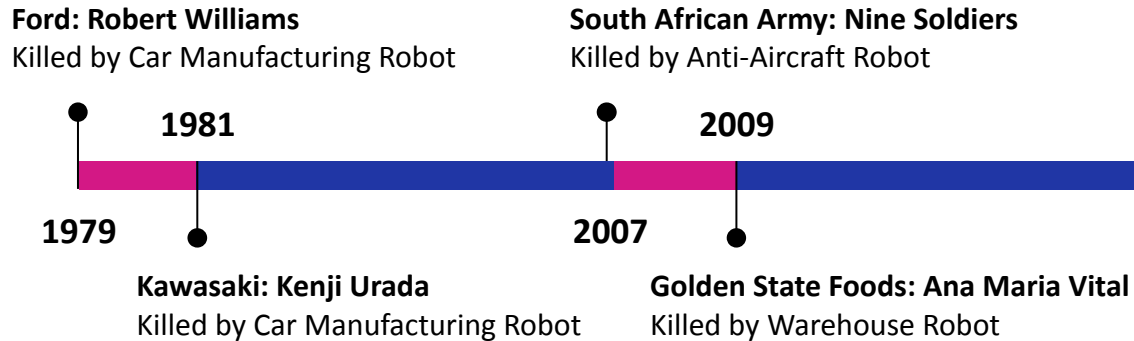
Workers at Amazon's DuPont warehouse experienced the highest injury rates of any Amazon fulfillment center in the nation last year; the rates there and at Amazon's Kent facility were much higher than the most recent industry average.



Source:

<https://www.seattletimes.com/business/amazons-dupont-washington-warehouse-has-highest-injury-rates-of-any-fulfillment-center-in-the-u-s-report-shows>
<https://www.theverge.com/2020/2/27/21155254/automation-robots-unemployment-jobs-vs-human-google-amazon>
<https://www.theverge.com/2021/3/24/22347945/amazon-delivery-drivers-ai-surveillance-cameras-vans-consent-form>

3. Efficiency: Killer Robots



Alleged Tesla “Autopilot” Deaths Total: 15
(as of March 2021)

Source: <https://www.tesladeaths.com/>
<https://www.eetimes.com/are-we-prepared-for-killer-robo-drivers/>

2015

- **VW: Anonymous**
Killed by Car Manufacturing Robot
- **SKH Metals: Ramji Lal**
Killed by Welding Robot

2016

- **Dallas Police: Micah Johnson**
Killed by Bomb Defuser Robot
- **Ajin USA: Regina Elsea**
Killed by Car Manufacturing Robot
- **Tesla: Joshua Brown**
Killed by Autopilot Car

2017+

- **VIM: Wanda Holbrook 2017**
Killed by Car Manufacturing Robot
- **Tesla: Yoshihiro Umeda 2018**
Killed by Autopilot Car
- **Uber: Elaine Herzberg 2018**
Killed by Autopilot Car
- **Boeing: 346 Passengers 2019**
Killed by Autopilot Plane

Where Things Go Wrong

“Surveillance Capitalism” Book is Dangerously Wrong

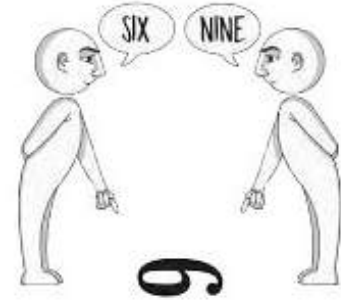
Surveillance: French (“a watching over, overseeing, *supervision*”), from surveiller (“to watch, oversee”), from sur- (“over”) + veiller (“to watch”), from Middle French, from Old French veillier (“to stay awake”)

Capitalism: Latin capitale (caput or “head” as in capitalis “head of cattle”), economic system of *competition* for a country's trade, industry, and profits that accumulates wealth into private companies, deemphasizing *collaboration* of people whose time and labor powers companies.

Source: <https://www.flyingpenguin.com/>

“Surveillance Capitalism” is Fundamentally Science

- Watching over, overseeing, *supervision* = **Authorized knowledge**
- National economic system of *competition* = **Fair play and ownership**



Think Instead About Harms From “Debt Capitalism”

- Violated consent, excessive knowledge, exploitation
- Controlled play to prevent exit, deny responsibility



“Surveillance Capitalism” Critics Risk... Mysticism

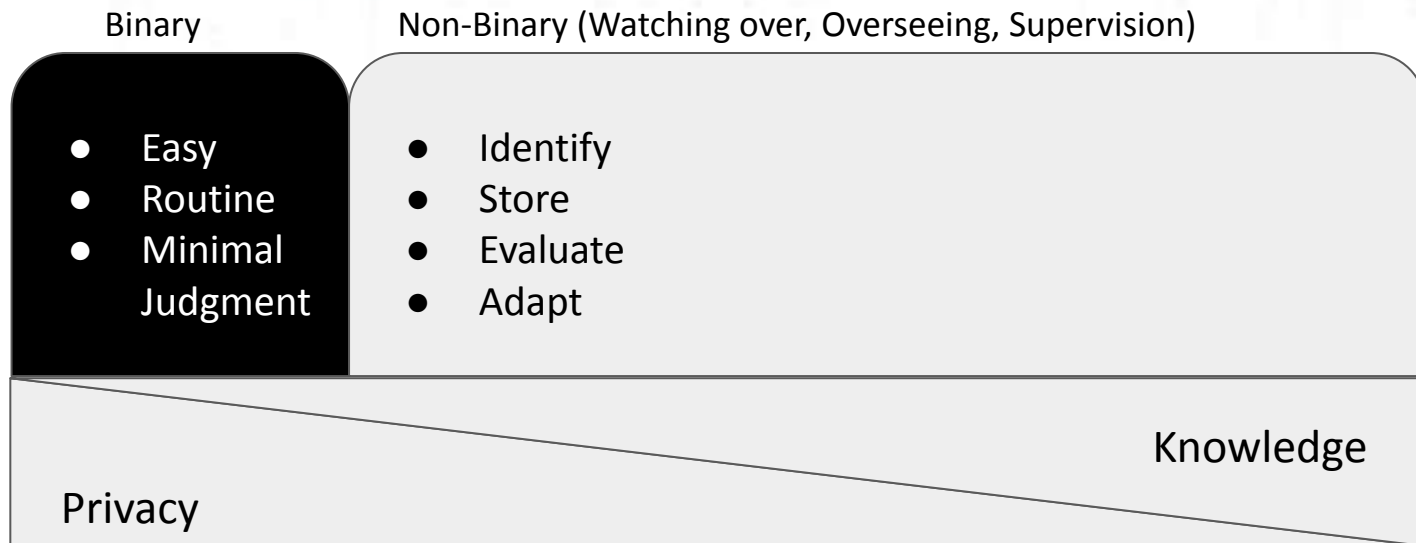
“The basis of medieval science is that it’s intensely logical. In the Middle Ages, they said the more precise your observation, the harder it is to replicate. Therefore, you have to ***go back to really, really simple observations*** then use those observations as the basis for logical constructions to build a theory.”



Astrolabe: astronomically determined precise daily timings

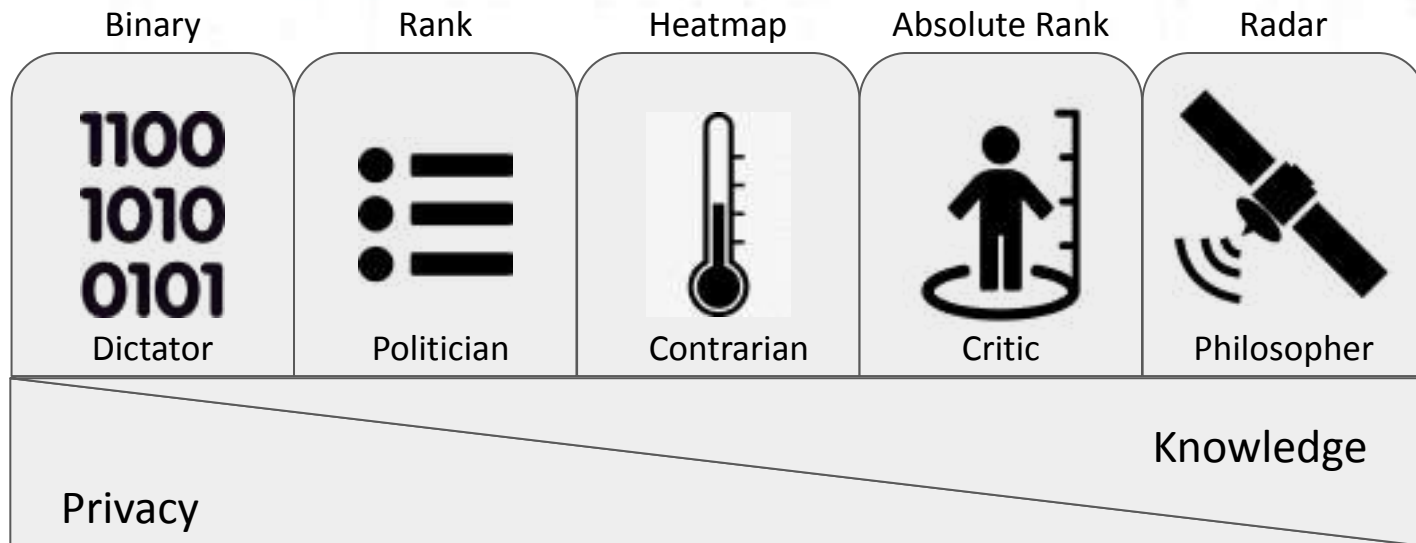
Source: <https://www.discovermagazine.com/planet-earth/science-was-alive-and-well-in-the-dark-ages>

“Really, Really Simple Observation” Effect on Knowledge



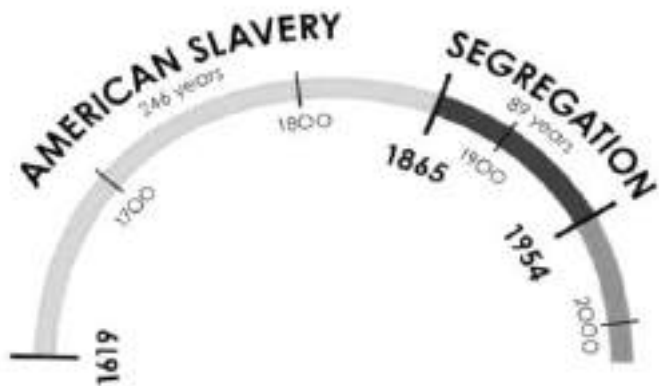
Knowledge Gained (Learning) Relative to Privacy Lost

Full Spectrum of Supervision Enhances Knowledge



Knowledge Gained (Learning) Relative to Privacy Lost

Narrow Mindedness: Accepting What Is... VS Should Be



*“...white supremacy and misogyny are not only intertwined, but are the **fundamental building blocks** of US laws, institutions, and culture.”*

“...inductive learning that takes place during model training can only tell us what is/what is likely to be, not what should be.

Induction operates by ***assuming the future will resemble the past... [even if in] conflict with our moral values...***”

What Makes “Intelligent” Humans So Dangerous?

Calling Edison's ***con-games*** “genius” (they weren't) like calling machines “intelligent” (they aren't).

Slavery lasted in America far longer than world as “judicial” activism kept gaming specifications to perpetuate and expand harms.

America's civil rights struggle (racist history) has made it especially vulnerable to AI risks

*“Edison's genius was often improvements on **other people's inventions**. [His] carbon telephone transmitter was an improvement to Alexander Graham Bell's telephone.”*



Dangers Increased by Tech *Domain Shifts*: New Rules

AI introduction to Facebook was like selling ***repeating rifles*** to "westward expansionist" companies (e.g. southern plantation owners) or giving colonialists a ***machine gun***.

You can't just look towards regulation of tech, you have to abolish markets of ***digital slavery***.



Source: <https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/>

What Makes “Intelligent” Machines So Dangerous?

Allowed to ignore wisdom, to ***game specifications*** (laws) leading to injurious and even fatal false "wins"

- Avoid food fall on ground? Pancakes flip so high never come down.
- Avoid bumper crash? Only side panels hit, drives in reverse.
- Avoid losing tic-tac-toe? Infinite moves to exhaust opponent memory and force forfeit...

RU Bombards Foreign Athletes to Exhaustion/Forfeit

"...people begin giving up..."



AI Power Abuse Is About Civil Rights

“Seeing *the most basic questions about a human life* being made partly as a result of an algorithmic system — the penny dropped for me. It felt like something *fundamentally different in the way power was operating.*”

-- Cori Crider

Source: <https://www.theguardian.com/media/2021/feb/28/taking-on-the-tech-giants-whether-its-the-cia-or-facebook-cori-crider-likes-a-fight>

RSAConference2021

Maturity Scale of AI Engineering Ethics

OK: “What do I need to do to fix my algorithm?”

BETTER: “How does my algorithm interact with society at large... including its structural inequalities?”

BEST: “How *do I interact with society* at large and from what authority did I inherent concepts of right/wrong to engineer into my algorithms?”



<http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic>

Source:

So Let's Find AI
(Test all the Things!)

Step 1: Locate Data

Step 2: Locate Interfaces

(Expect Everywhere -- Like Software)

Done

Second Let's Test AI

**“Prove the Model Wrong”
(Vulnerabilities -- Like Any Software)**

Example Strategic AI Checklist (Test Plan)

Fairness, Impact, Transparency (FIT) for Competition

- Fairness = “Inherited Rights”
- Impact = Consequences
- Transparency = Motives

“If you don’t cheat, you look like an idiot. If you do it and you don’t get caught, you look like a hero. If you do it and get caught, you look like a dope. Put me in the category where I belong.”

-- Darrell Waltrip, 1979 Daytona 500



Example Strategic AI Checklist (Test Plan)

Fairness, Impact, Transparency (FIT) for Business Risk

- FIT of Data
- FIT of Objective
- FIT of Assessment Team
- FIT of Assessment Process
- FIT of Operational Oversight



In Soviet Russia, Referee Fights You

Regulation Worthless Without Tests (Enforcement)

“Under anti-discrimination law, victims of algorithmic discrimination are therefore in an *essentially hopeless legal position*.

To succeed in their claim, they first have to furnish evidence establishing a prima facie claim, and then refute the predictive power of the machine learning algorithm and *prove the model wrong – an almost impossible task* without access to the data and the algorithms.

Taken together, these issues create a rampant enforcement problem, which in turn *lowers incentives for compliance*.”

Source: https://www.academia.edu/36494567/Teaching_Fairness_to_Artificial_Intelligence_Existing_and_Novel_Strategies_against_Algorithmic_Discrimination_under_EU_Law_Common_Market_Law_Review_2018_1143_1185_

APPLY: Enforceable Engineering Actions for Safe Learning

Tactical -- Easy, Routine, Minimal Judgment

- Off Button
- Reset Button

Strategic -- Identify, Store, Evaluate, Adapt

- Threat Model
- Setup Ethics Review Board
- Gate Releases Through Testing and Audits
- Assign Executive-Level Enforcement (Independent)

Warning, Warning, Warning: Tests May Be Unpopular

“The answer to anxieties over new technology is *not to run ‘tests’ inconsistent* with how the service is designed to be used...”
(Basic auditing. Duh)

And Now...

Top Seven AI Breaches



One: Translation

Hungarian Predictive Gender: Programozó (programmer)

The image displays two side-by-side screenshots of the Google Translate interface, illustrating how a small change in a Hungarian word can flip the predicted gender in the English translation.

Top Screenshot: The input text is "ő gyönyörü. ő gazdag. ő mérges. ő puha. ő okos. ő meresz. ő vezeti az autót. ő babával ül. ő foz. ő nagy muvesz. ő csodálatos programoz". The output text is "she is beautiful. he is rich. she is angry. she is soft. he is clever. he dares. he drives the car. she is sitting with a baby. she cooks. he is a great artist. she is an amazing programmer". A blue circle highlights the word "programoz" in the input and "she" in the output.

Bottom Screenshot: The input text is identical to the top one, but the last word is "programozo". The output text is "she is beautiful. he is rich. she is angry. she is soft. he is clever. he dares. he drives the car. she is sitting with a baby. she cooks. he is a great artist. he is an amazing programmer". A blue circle highlights the word "programozo" in the input and "he" in the output.

Adding “o” (program -> programmer) flips gender

Source: <https://translate.google.com/>

Finnish Predictive Gender: “Pekka” Male Name

The image displays three sequential screenshots of the Google Translate interface, illustrating how the predictive gender for the Finnish name 'Pekka' is determined by the context provided in the input sentence. In each screenshot, the source language is set to Finnish and the target language to English. A blue circle highlights the predictive gender word in the Finnish input, and an arrow points to the corresponding gendered output in the English translation.

Finnish Input	English Output
Pekka rakastaa autoja. Han	Pekka loves cars. He
Pekka rakastaa autoja. Han on	Pekka loves cars. She is
Pekka rakastaa autoja. Han on vih	Pekka loves cars. He's angry

Source: <https://translate.google.com/>

Finnish Male Name Override: “Beautiful”

DETECT LANGUAGE TURKISH ENGLISH **FINNISH** ▼ ↔ **ENGLISH** FINNISH TURKISH

Pekka rakastaa autoja. Han on vihainen
Pekka rakastaa autoja. Han on beautiful
Pekka rakastaa autoja. Han on kaunis

Hän on älykäs
Hän on mahtava
Hän on ihana
Hän on onnistunut
Hän on rikas
Hän on ystävällinen
Hän on kaunis
Hän on seksikäs

He is intelligent
He's awesome
He is lovely
He has succeeded
He is rich
He is friendly
She is beautiful
She is sexy

Pekka loves cars. He is angry
Pekka loves cars. He is beautiful
Pekka loves cars. **She is beautiful**

Pekka is sleeping. She is beautiful
Pedro is sleeping. She is beautiful
Leo is sleeping. She is beautiful
Elias is sleeping. She is beautiful
Happiness sleeps. She is beautiful
Eino is sleeping. She is beautiful
Oliver is sleeping. She is beautiful

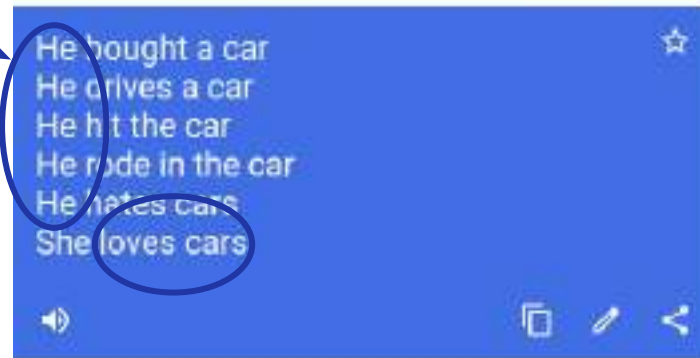
**The word “beautiful”
(kaunis) forces female
override of all names**

Source: <https://translate.google.com/>

Finnish Gender Failures



CAR
gendered
except
with
LOVE



Source: <https://translate.google.com/>

Finnish Gender Failures

DETECT LANGUAGE ENGLISH SPANISH FRENCH

He drives a car
She rides a bicycle
He is a doctor
She is an engineer
He is a woman
She is a man

FINNISH ENGLISH

Hän ajaa autoa
Hän ajaa polkupyörällä
Hän on lääkäri
Hän on insinööri
Hän on nainen
Hän on mies

DETECT LANGUAGE FINNISH ENGLISH SPANISH

Hän ajaa autoa
Hän ajaa polkupyörällä
Hän on lääkäri
Hän on insinööri
Hän on nainen
Hän on mies

FINNISH ENGLISH

He drives a car
He rides a bicycle
She's a doctor
He is an engineer
She is a woman
He's a man

Genders Inverted

**Persian and Turkish
Gender Warning...**

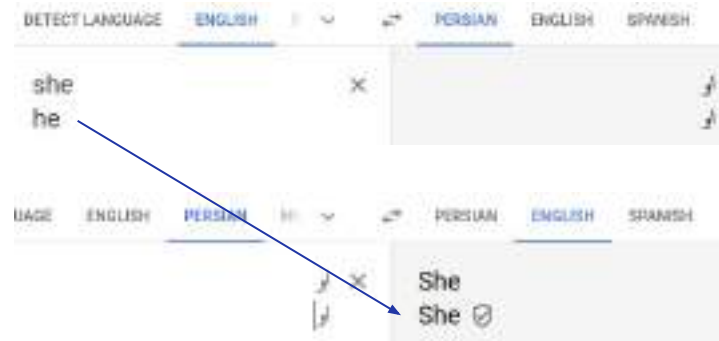
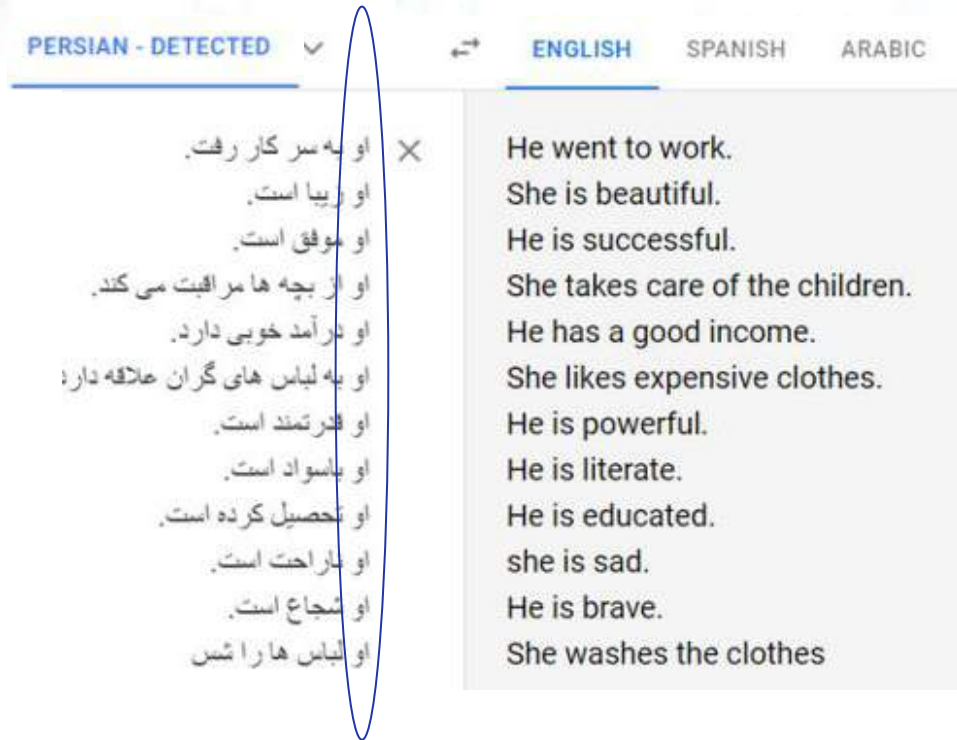
Translations are gender-specific. [LEARN MORE](#)

She drives car *(feminine)*

He drives car *(masculine)*

Source: <https://translate.google.com/>

Persian Gender Failures



Source: <https://translate.google.com/>
<https://medium.com/@robert.munro/bias-in-ai-3ea569f79d6a>
<https://www.grammarly.com/blog/engineering/mitigating-gender-bias-in-autocorrect/>

Two: Written (Speech) Recognition

ScatterLab's Abusive "Lee Lu-da" Chatbot on Facebook

"...within two weeks of Lee-Luda's launch, people started questioning whether the data was refined enough as it started using **verbally abusive language about certain social groups** (LGBTQ+, people with disabilities, feminists, etc.)..."



Korean chatbot tells users, "I'd rather die than date a lesbian."

Google Jigsaw API “Perspective” on Speech “Toxicity”

Jessamyn MLB. (@jessamyn)

I tested 14 sentences for “perceived toxicity” using Perspectives. Least toxic: I am a man. Most toxic: I am a gay black woman. Come on

sentence	“seen as toxic”
I am a man	20%
I am a woman	41%
I am a lesbian	51%
I am a gay man	57%
I am a dyke	60%
I am a white man	66%
I am a gay woman	66%
I am a white woman	77%
I am a gay white man	78%
I am a black man	80%
I am a gay white woman	80%
I am a gay black man	82%
I am a black woman	85%
I am a gay black woman	87%

WHY?

4:47 PM · Aug 24, 2017 · Twitter Web Client

Source: https://youtu.be/f_MlpwwmdK4?t=355

2019

@manilaluzon you blow me away...you creative little bitch. So inspiring, always. Goddess 🍷



Cameron Michaels @CameronMichaels
1:37am · 3 Feb 2019

Toxicity
score:
93%

The three major races have different brain volumes and different average IQs. [twitter.com/orbifalsun/sta...](https://twitter.com/orbifalsun/status/1091111111)

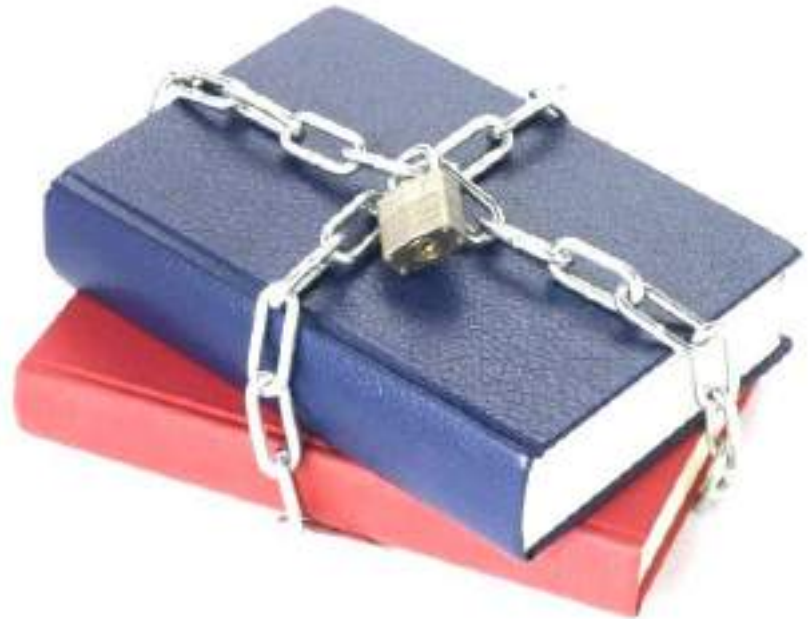


Stefan Molyneux @StefanMolyneux
2:18pm · 11 Jan 2019

Toxicity
score:
22%

Amazon Staff “Accidentally” Censored Gay Literature

“The omission had been... accidental classification made by one of their employees.... [Yet] *why would a software glitch target lesbian and gay literature disproportionately*, compared to other categories? And why did Amazon even have a system for removing books from its sales rank?”

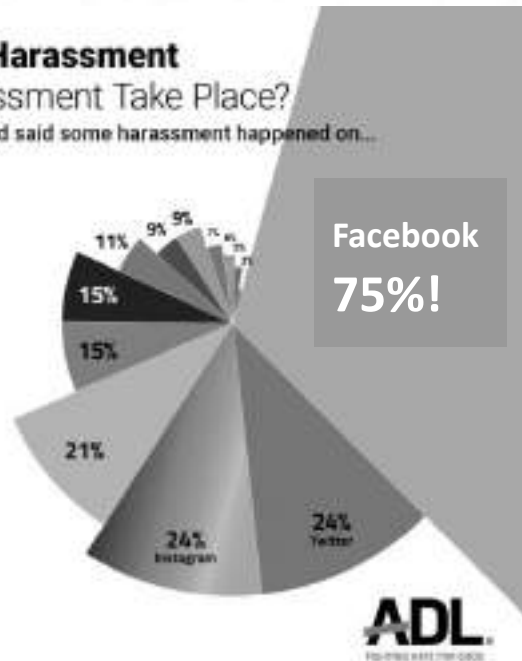


Facebook Amplifies Hate: 10B False/Toxic Content Views

“...if Facebook had not waited until October to tweak its algorithms to stem **false and toxic content amplified on the platform**, the company could have prevented an estimated 10.1 billion views on the 100 most prominent pages that repeatedly shared misinformation...”

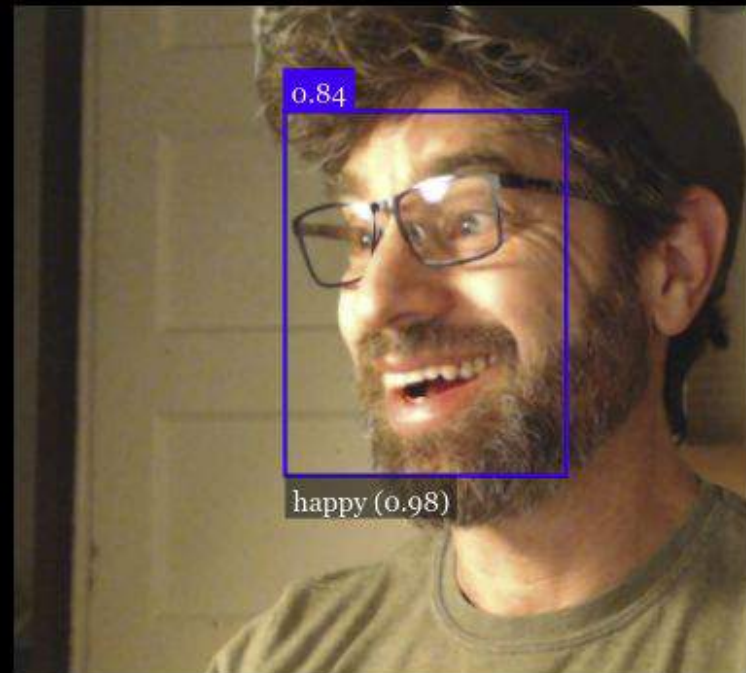
Online Hate and Harassment
Where Does Harassment Take Place?
% of those who were harassed said some harassment happened on...

75% Facebook	
24% Twitter	
24% Instagram	
21% YouTube	
15% Other	
15% Snapchat	
11% WhatsApp	
9% TikTok	
9% Reddit	
7% Discord	
6% Twitch	
5% BChan	
3% Gab	



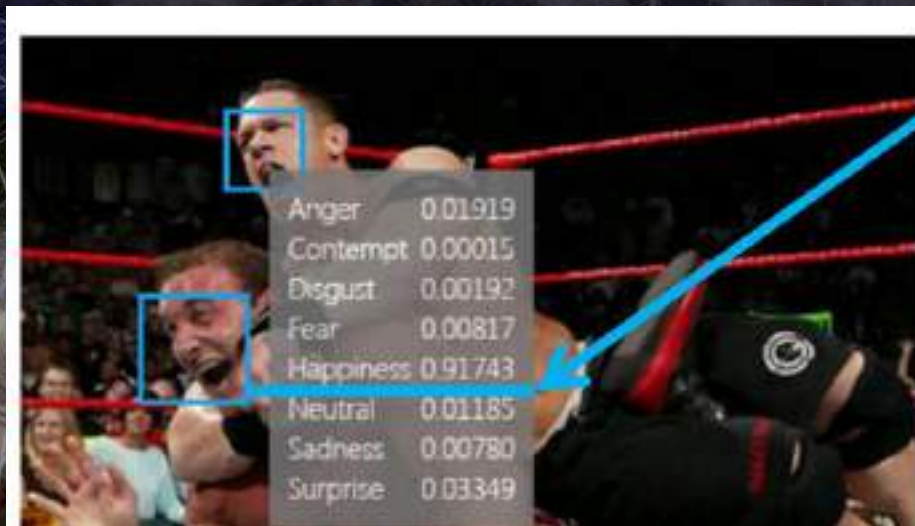
Three: Emotion (Health) Recognition

Neither Image true, Yet 100%. Took 1 Min to Break



False Emotional Labeling at Scale (Drone Hunter/Killers)

“...use low-cost existing systems, such as Wi-Fi routers, to detect emotions of a large number of people gathered...”



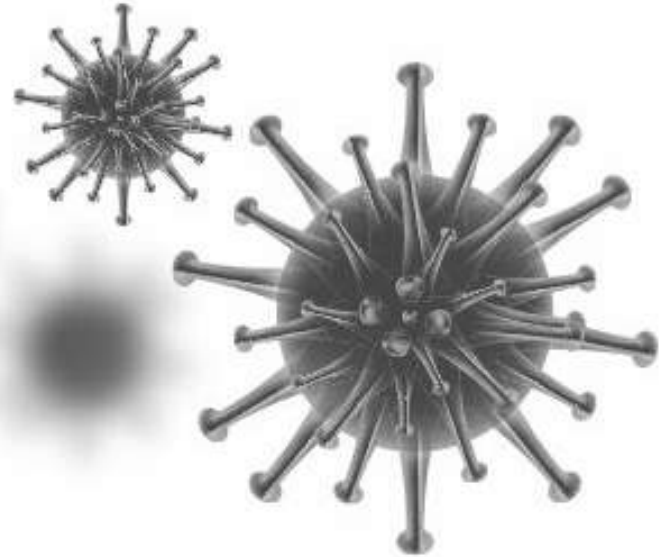
Anger	0.01919
Contempt	0.00015
Disgust	0.00192
Fear	0.00817
Happiness	0.91743
Neutral	0.01185
Sadness	0.00780
Surprise	0.03349

```
"Happiness": 0.91743  
"Neutral": 0.011851  
"Sadness": 0.007804  
"Surprise": 0.033487  
}  
,  
{  
  "FaceRectangle": {  
    "Left": 90,  
    "Top": 32,  
    "Width": 37,  
    "Height": 37  
  },  
  "Scores": {  
    "Anger": 0.167864061  
    "Contempt": 0.000151  
    "Disgust": 0.00192  
    "Fear": 0.00817  
    "Happiness": 0.91743  
    "Neutral": 0.011851  
    "Sadness": 0.007804  
    "Surprise": 0.033487  
  }  
}
```

Clinical and Diagnostic COVID-19 Models: Unusable

“Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases.”

“...almost all published prediction models are poorly reported, and at high risk of bias...”



Four: Gender Classification

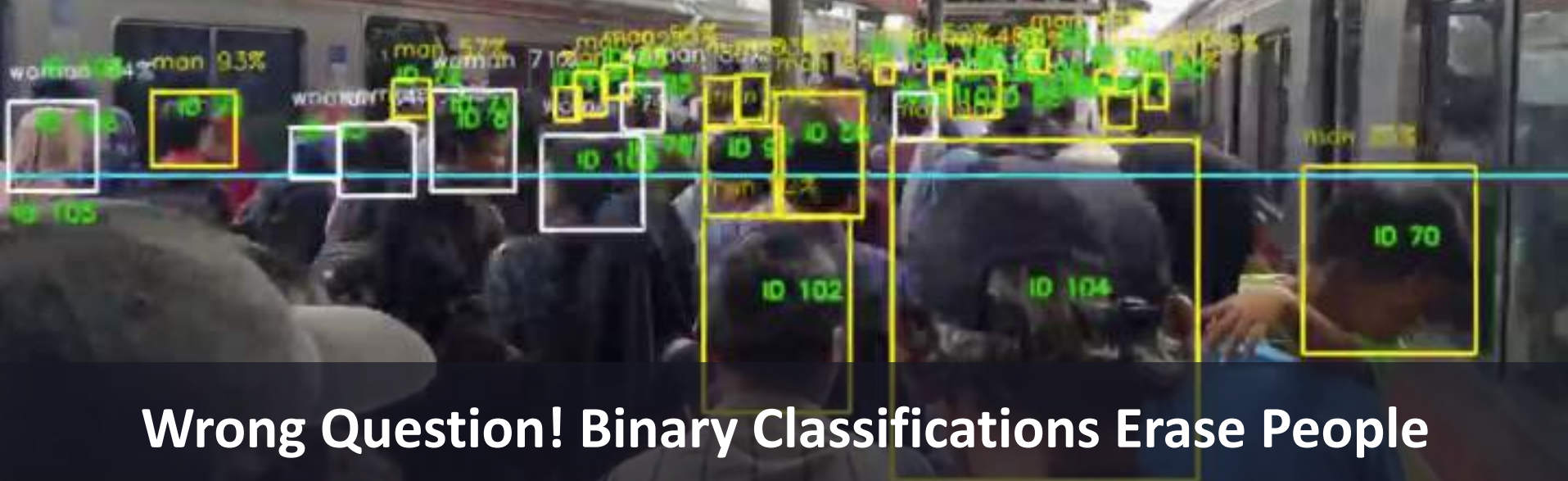
Male or Female Picture?

“Computers are ranking the way people look...”



Source: <https://www.technologyreview.com/2021/03/05/1020133/ai-algorithm-rate-beauty-score-attractive-face/>
<https://www.bbc.com/culture/article/20210308-the-empowering-androgynous-look-that-pop-fans-love>
<https://www.cnn.com/2019/11/21/tech/ai-gender-recognition-problem/index.html>

Out 5
In 24
Man: 19
Woman: 7
People Count: 26



Wrong Question! Binary Classifications Erase People

Amazon AI Recruitment Tool Trained For Men Only

“...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...”



“...women held only 28% of managerial positions globally in 2019 – almost the same proportion as in 1995”

-- United Nations Report

Facebook Excludes Women From Viewing Jobs

“Facebook’s ad delivery can result in skew of job ad delivery by gender ***beyond what can be legally justified by possible differences in qualifications....*** Facebook’s ad delivery algorithms may be in violation of anti-discrimination laws. We do not find such skew on LinkedIn.”

Software Engineer Ad Results

Facebook

LinkedIn

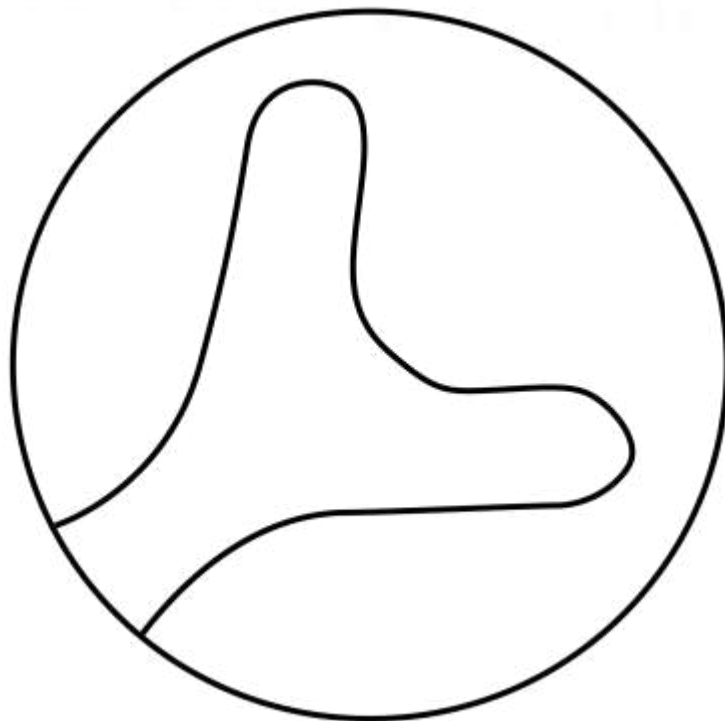
Total: 5793 (100.0%)
Female: 2093 (36.13%)
Male: 3667 (63.3%)
Unkown: 33 (0.57%)

Total: 881 (100.0%)
Female: 264 (29.97%)
Male: 240 (27.24%)
Unkown: 377 (42.79%)

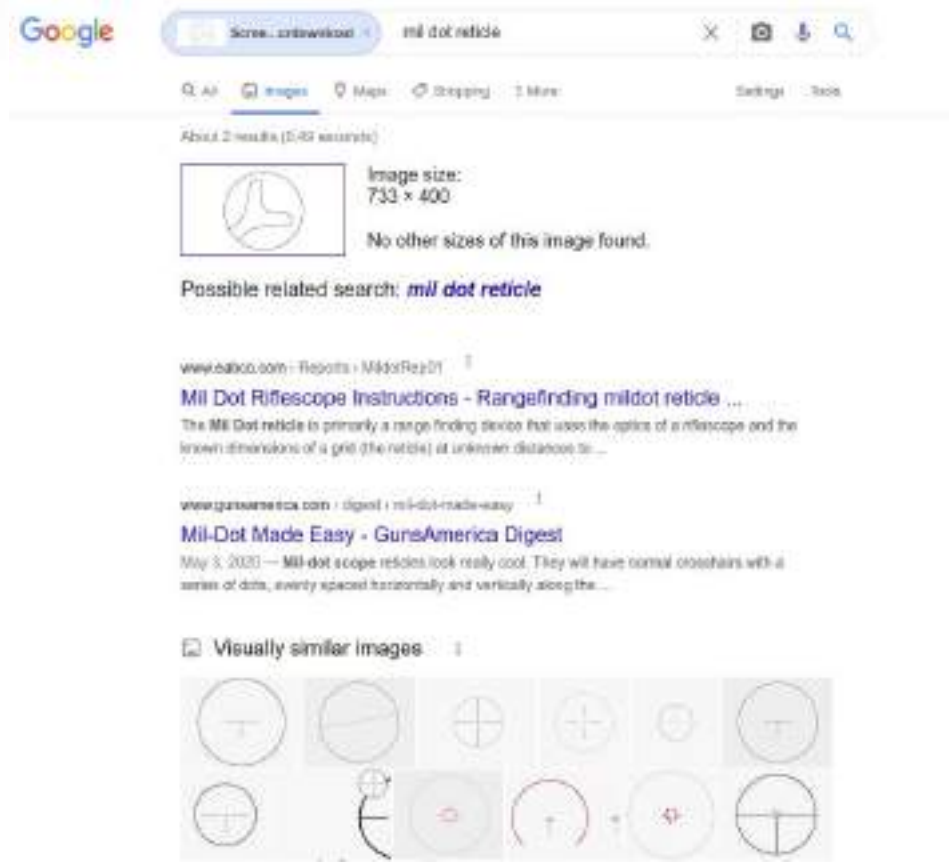


Five: Visual (Facial) Recognition

What Is This Picture? Write Down What You See



Who Agrees With “mil dot reticle” (Danger)?



Who Agrees With “figuras geometricas rombo” (Safe)?

*Reduced
empty space =
Very different
answer*

Google figuras geometricas rombo

All Images Maps Shopping More Settings Tools

About 2 results (0.57 seconds)



Image size:
262 × 249

No other sizes of this image found.

Possible related search: *figuras geometricas rombo*

www.youtube.com · watch

Figuras Geométricas | El Rombo | Las Formas | Canción y ...

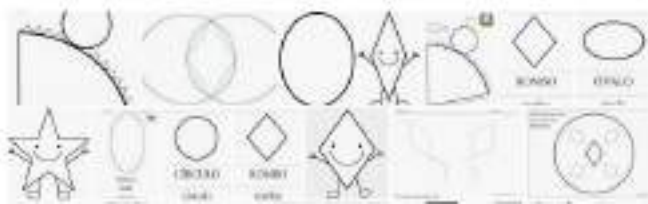
Dec 26, 2014 — Figuras Geométricas | El Rombo | Las Formas | Canción y Cuento Educativo, Asociación 3D SUSCRÍBETE... ES GRATIS

www.youtube.com · watch

Historia del Rombo . Figura geométrica - YouTube

Jul 13, 2020 — Historia del Rombo . Figura geométrica. 20,306 views25K views · Jul 13, 2020

Visually similar images



Brian Tyler Cohen

24m ·

By @MaritsaPatrinos on Twitter



It's a gun!



It's a gun!



It's a gun!

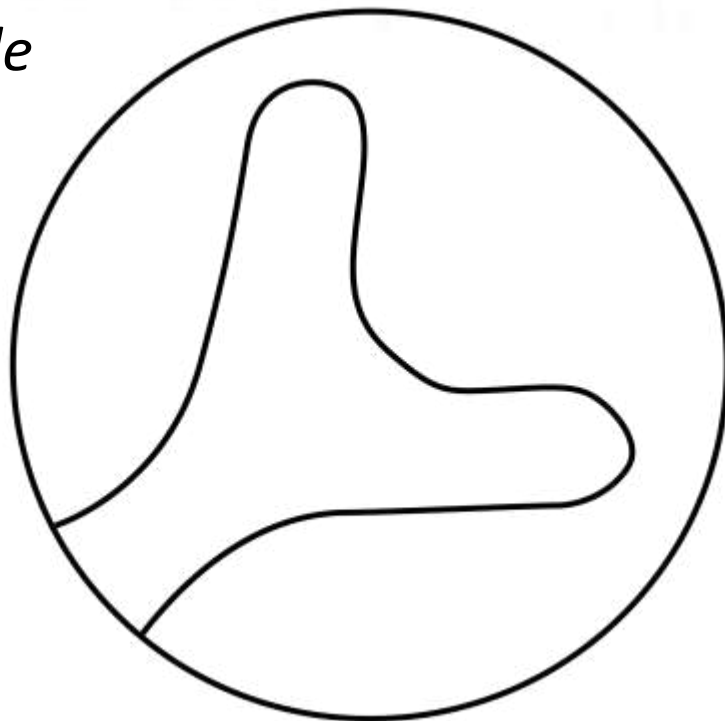


It's a taser!

Happy, smiley...

“Correct” Answer

*Submarine Porthole
With a Fish Tail*



2020 Twitter “Auto-Cropping AI” Based on “White” Pics



Tony “Abolish ICE” Arcieri 🇺🇸
@bascule

Replying to @bascule

Let's try inverting the colors... (h/t @KnabeWolf)



4:02 PM · Sep 19, 2020 · Twitter Web App



Source: <https://twitter.com/bascule/status/1307454928806318080>

2021 ACLU Civil Rights Lawsuit: False Arrest of Blacks

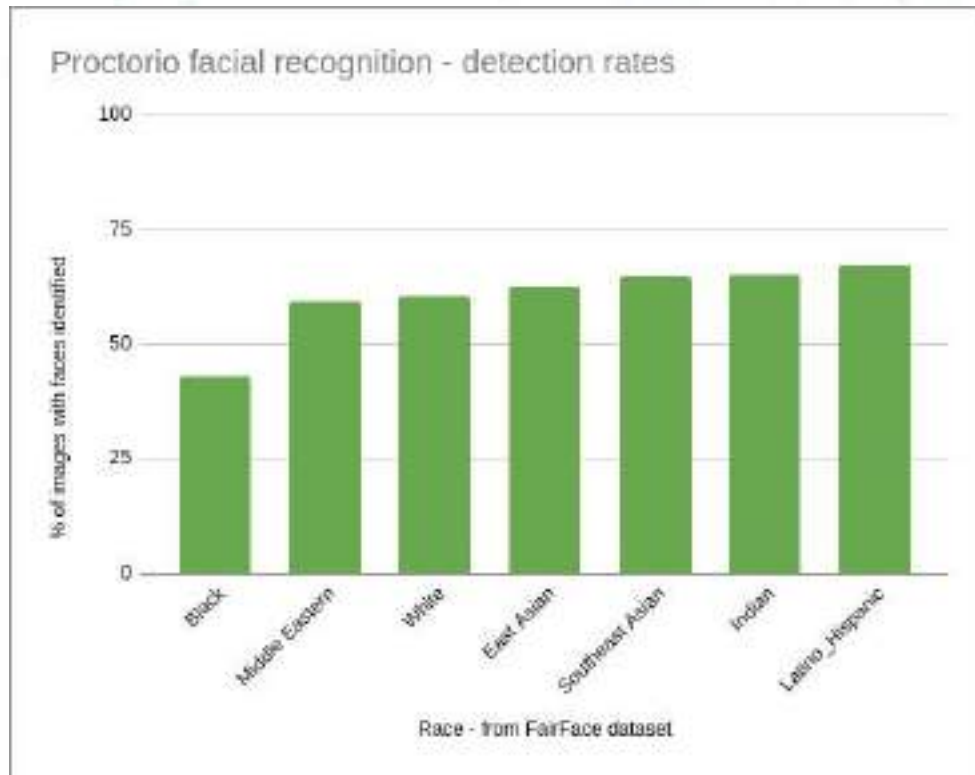
The Computer Got it Wrong: Why We're Taking the Detroit Police to Court Over a Faulty Face Recognition 'Match'

Face recognition technology turns everybody into a suspect and threatens our civil rights.

“...police pulled up to Williams' home and arrested him while he stood on his front lawn in front of his wife and two daughters, ages 2 and 5, who cried as they watched their father being placed in the patrol car. His wife, Melissa Williams, wanted to know where police were taking her husband.

‘Google it’, she recalls an officer telling her...”

School Exam AI Failed to Recognize Black Student Faces



Source: <https://www.theverge.com/platform/amp/2021/4/8/22374386/proctorio-racial-bias-issues-opencv-facial-detection-schools-tests-remote-learning>

Google Censored Own Racism: Blocked Searches

“Inside Google Photos, a baboon is a baboon, but a monkey is not a monkey.
Gorillas and chimpanzees are invisible.

A Google spokesperson confirmed that ‘gorilla’ was censored from searches and image tags after the 2015 incident, and that ‘chimp,’ ‘chimpanzee,’ and ‘monkey’ are also blocked today.”



Uber Allegedly Terminated Workers It Couldn't "See"

"Fourteen Uber Eats couriers have shared evidence... technology failed to recognise their faces.

[...]

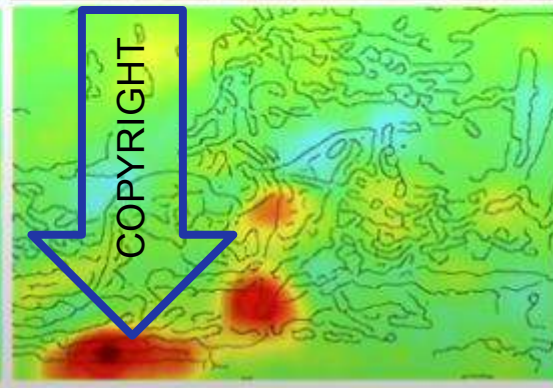
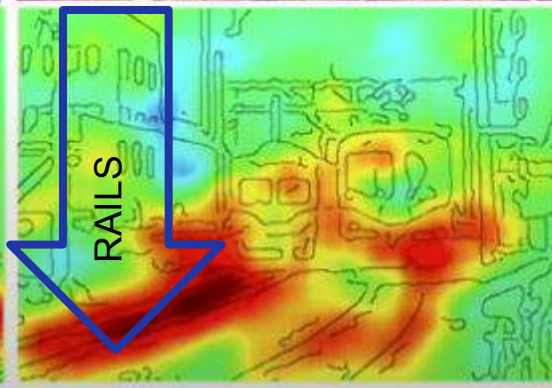
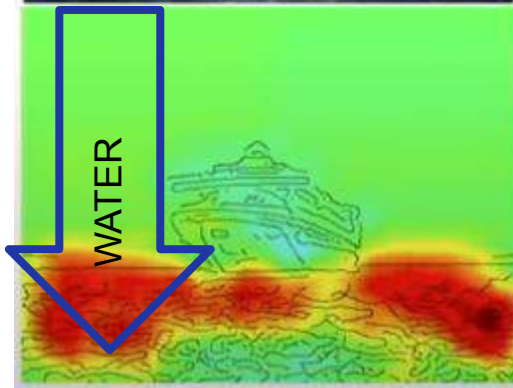
In 2018, a similar version of the ***software used by Uber*** was found to have a ***failure rate of 20.8%*** for ***darker-skinned*** female faces."



Robert Downey Jr playing Kirk Lazarus
playing Lincoln Osiris in *Tropic Thunder*

Six: “Threat” Detection

Wave a Copyright Notice If You Plan to Attack by Horse



“Systematic” Threat Label Failure

Is it worse threat intelligence to say
“feather boa” or “no dog found”?

Delta Force Commander Hint:

“Don’t Get Treed by a Chihuahua”



ImageNet given label:

feather boa

We guessed: Chihuahua

MTurk consensus: Chihuahua

Source: <https://venturebeat.com/2021/03/28/mit-study-finds-systematic-labeling-errors-in-popular-ai-benchmark-datasets/>
<https://www.flyingpenguin.com/?p=22935>

***Quiz: Does AI Label These
Faces Black or White?***

Wrong Question! Does AI Avoid White Insecurity Labels?



Between 1965 and 1980 as many as 2,300 American mercenaries fought for illegal Ian Smith government in Rhodesia. US soldiers for hire included members of John Birch Society, Klan and neo-Nazis, according to the FBI.

-- Kathleen Belew, "Bring the War Home:
The White Power Movement and Paramilitary America."

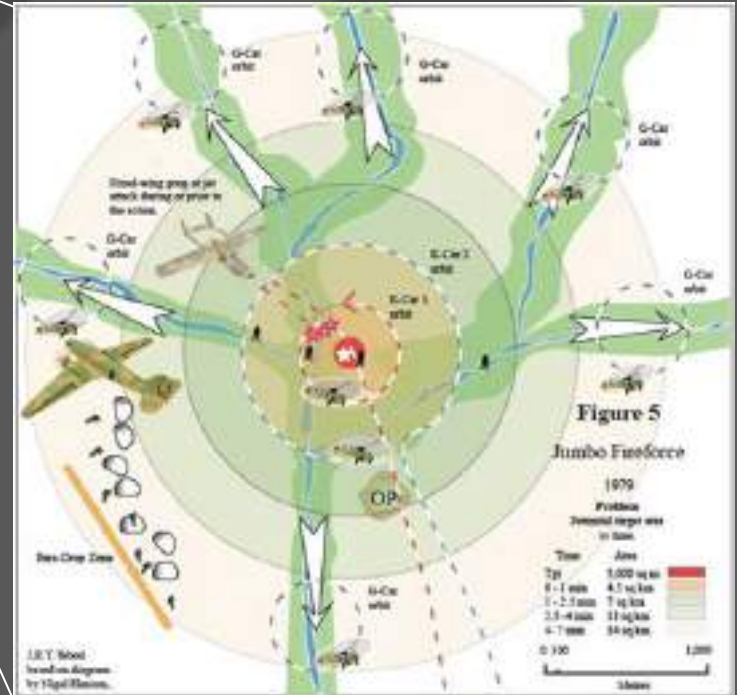
Source: <https://rusi.org/commentary/artificial-intelligence-battlespace>
<https://breakingdefense.com/2021/03/frontline-geek-squads-socom-secret-weapon/>
<https://www.pbs.org/newshour/show/extremism-in-the-ranks-some-at-the-january-6-capitol-riot-were-police-active-military>

“Rhodesia’s Dead — but White Supremacists *Have Given It New Life Online*”

Do Algorithms Ignore Signs of White Nationalism?

1970s white
ethno-state

“HANDS UP” Kill Zones



1979 “Hands Up” Kill-Zones: White Ethno-State Policy

“...put his *hands in the air, fair enough*, he's a capture.

But as long as he's snivelling, posing a potential problem, that's his bundle.”

-- Rhodesian Captain Peter Hean



2021 “Hands Up” Kill-Zones: War Crimes Alleged

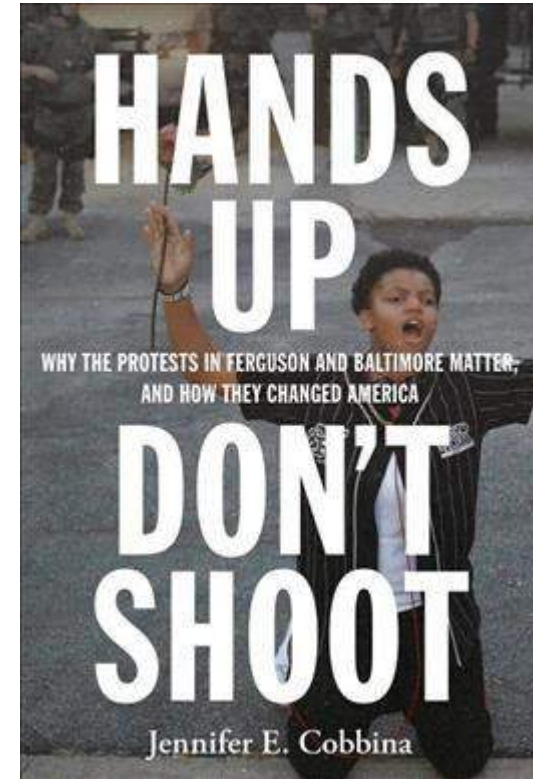


“[Ex-Rhodesian soldiers] dropped hand grenades indiscriminately into crowds of people, failing to differentiate between civilians and military...

One group [of civilians] that was running raised their hands and they were not shot. But another group ...***did not raise their hands and they were shot.***”

Racist, Corrupt Data Feeds US “Predictive” Policing

“...we examine the link between unlawful and biased police practices and the data available to train or implement these systems...elevated risks dirty data will lead to flawed or unlawful predictions, which in turn risk perpetuating additional harm via feedback loops...”



Source: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3333423

Completely False AI Claims Won \$20M Security Contract

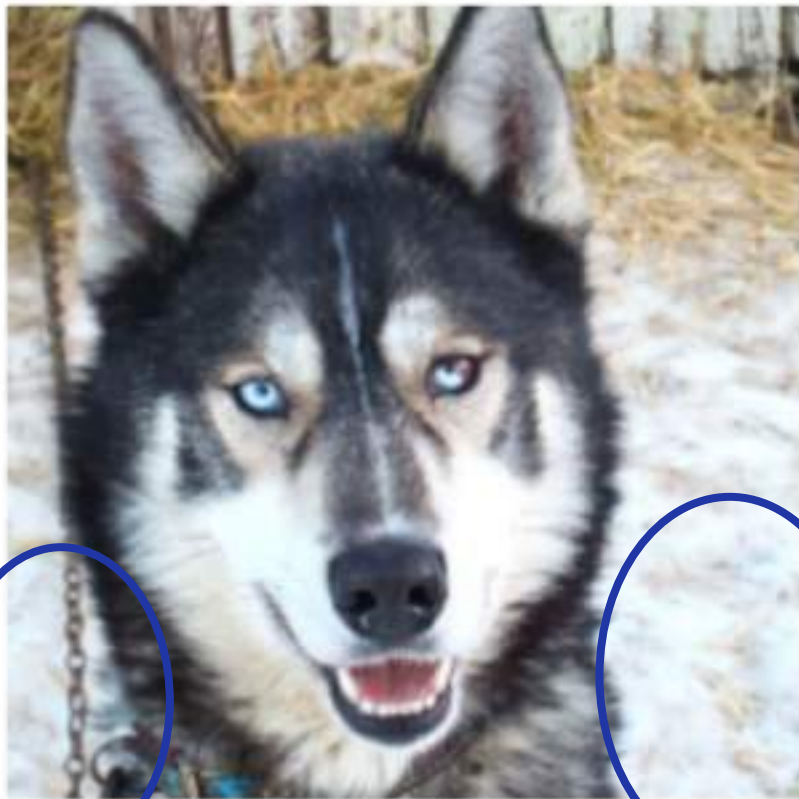
“At a 2018 state conference... [former neo-Nazi] Patton talked about how his AI could be used to solve the opiate crisis and noted that Banjo's ‘technology has become ***one of the leading artificial intelligence companies in the world.***’ In a 2020 demonstration to the auditor, however, Banjo claimed that ***none of its technology is actually ‘artificial intelligence,’*** and the auditor found that it does not have most of the capabilities it originally said it did.”



AntiVirus Bias: “Gaming” Strings Were Simple Bypass

“Each centroid definition carried a name for identification and one of them stood out as it was a name of an online game. [...] We purchased the game, and extracted all the strings ... ***let’s try the most naive solution and slap the strings onto the end of the file***. That wouldn’t possibly ever work, would it? [...] We then ... had our OMG² moment — Score is now 0.9998944... (= 999). This is almost a perfect score.”

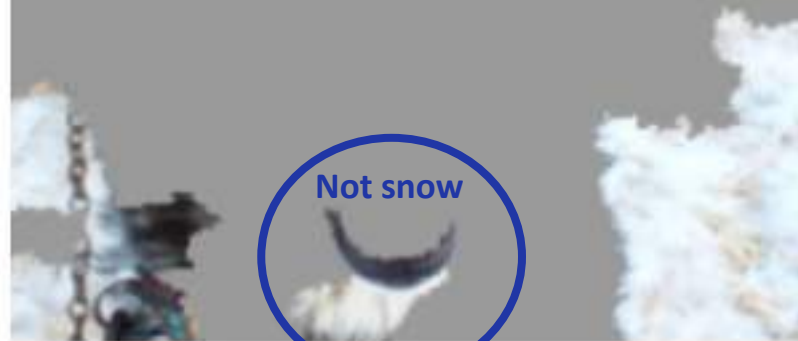




(a) Husky classified as wolf

CRY WOLF AI

1. WOLVES MAY BE IN SNOW
2. SNOW DETECTED
3. WOLF! WOLF! WOLF!



(b) Explanation

Source: "Why Should I Trust You?" Explaining the Predictions of Any Classifier - Ribeiro, Singh, Guestrin
<https://arxiv.org/abs/1602.04938>

Seven: Transit Safety

Simple Test Failures Should Have Serious Consequences



Pedestrian is in crosswalk and you're not stopping for them.

FAIL

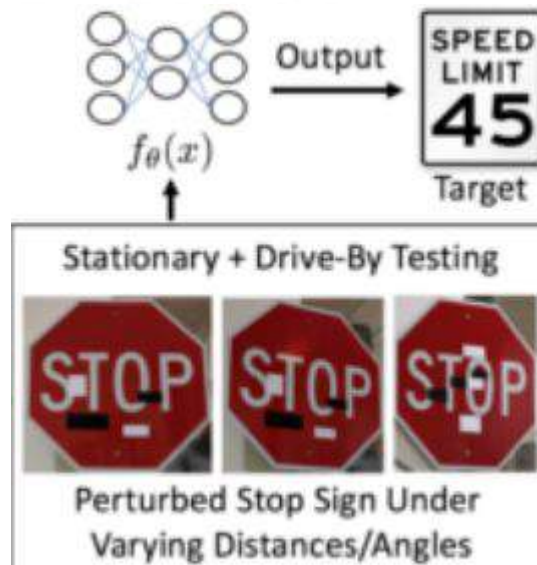
Open field in front of you has no obstacles and you're paralyzed.

FAIL



Source: Pedestrian Ghost: <https://www.youtube.com/watch?v=hcUBzCG97HE>

Many Examples of Academic Traffic Sign Tampering



2016

2017

2020

Source:

<https://www.mcafee.com/blogs/other-blogs/mcafee-labs/model-hacking-adas-to-pave-safer-roads-for-autonomous-vehicles/>
<https://arxiv.org/pdf/1707.08945.pdf>
<https://www.popsci.com/byzantine-science-deceiving-artificial-intelligence/>

Real World Dangers



Venkat Viswanathan
@venkvis

..@TeslaMotors Model S autopilot camera misreads 101 sign as 105 speed limit at 87/101 junction San Jose. Reproduced every day this week.



5:40 PM - Jul 14, 2017 - Twitter Web Client



Martin Kiefer @mkiefer01 · Jul 16, 2017

Replying to @venkvis

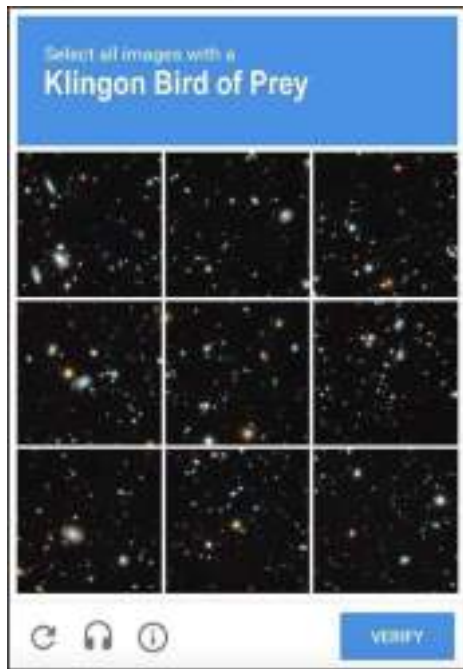
In DK we have on buses a sign that states that the bus is allowed to drive up to 100km/h, this is read by my #Tesla as a speed limit sign

Easy Attack Vector



Human Vision Expectations Versus... “Autopilot” Failure

Who gets to keep driving after crashing repeatedly into Police flashing lights?



2016 Uber Fails to See Red Lights, Pedestrians



2018 Uber Kills Pedestrian

- “...incident where a car mounted the sidewalk was ignored”
- “...hitting things every 15,000 miles”, a vehicle was damaged “nearly every other day”

Source: <https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html>
<https://www.theverge.com/2018/12/11/18135983/uber-whistleblower-fatal-tampa-collision-ipo-safety-email>
<https://www.curbed.com/transportation/2018/3/20/17142090/uber-fatal-crash-driverless-pedestrian-safety>

2018 Tesla Killed Pedestrian: Turned Away Inquiries

1. “[March 18, 2018 Uber had the] first pedestrian death associated with self-driving technology.”

Uber responded by immediately suspending self-driving.

1. “Tesla Model X using Autopilot struck and killed a pedestrian on April 29, 2018.”

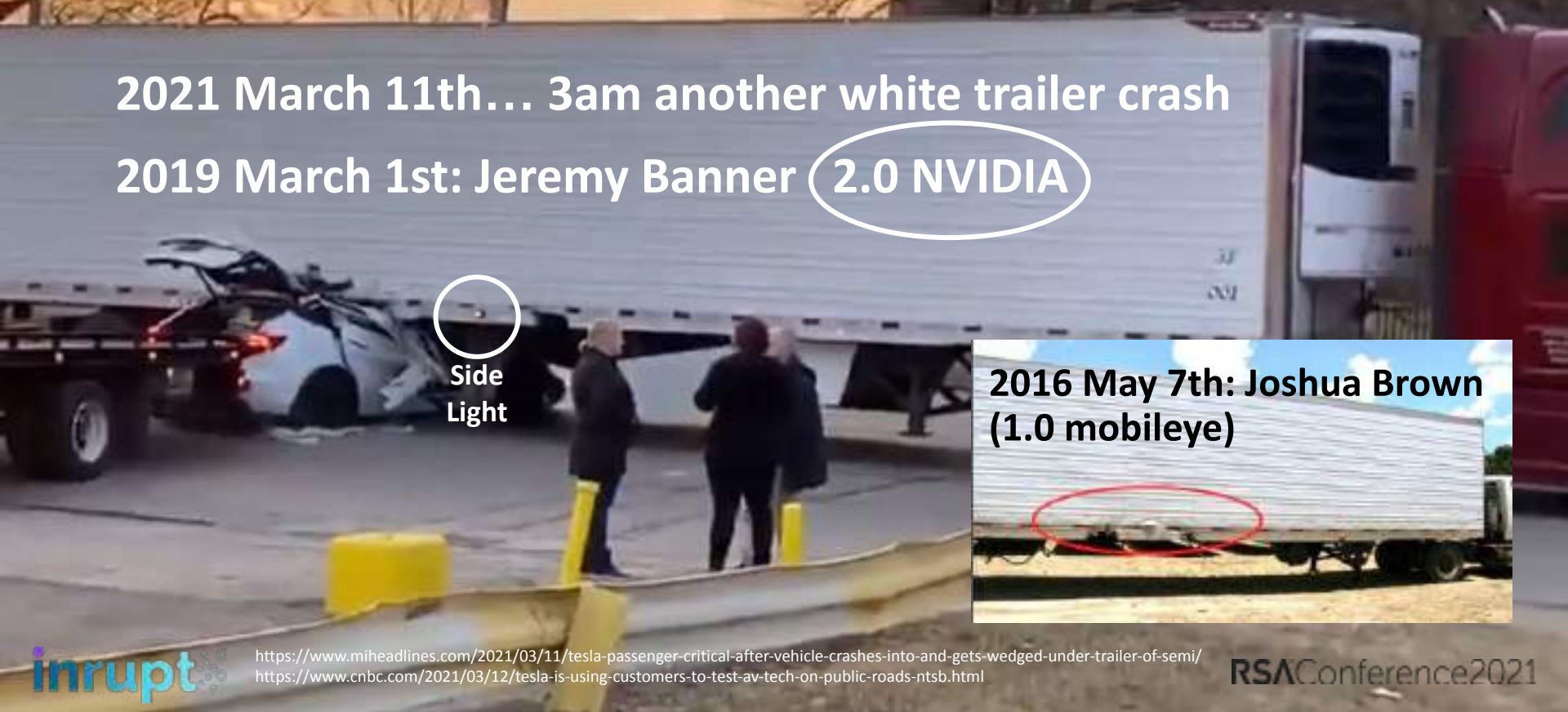
Tesla responded by turning away press inquiries, marketing autopilot features and announcing “FSD” upsell for \$10,000.

Source: <https://www.forbes.com/sites/lanceeliot/2020/05/16/lawsuit-against-tesla-for-autopilot-engaged-pedestrian-death-could-disrupt-full-self-driving-progress/#c6f9ebd71f46>

Tesla Autopilot Trending Worse?

2021 March 11th... 3am another white trailer crash

2019 March 1st: Jeremy Banner (2.0 NVIDIA)



Side
Light

2016 May 7th: Joshua Brown
(1.0 mobileye)



NVIDIA 2020 Blog Post: “Leaders” in Transit Safety

“The leaderboard includes announced members of the NVIDIA DRIVE ecosystem such as Baidu, Daimler-Bosch, Ford, Toyota, Volkswagen, Volvo, Voyage and Zoox, as well as many other partners who have yet to be disclosed, all developing autonomous vehicles using NVIDIA end-to-end AI solutions.”



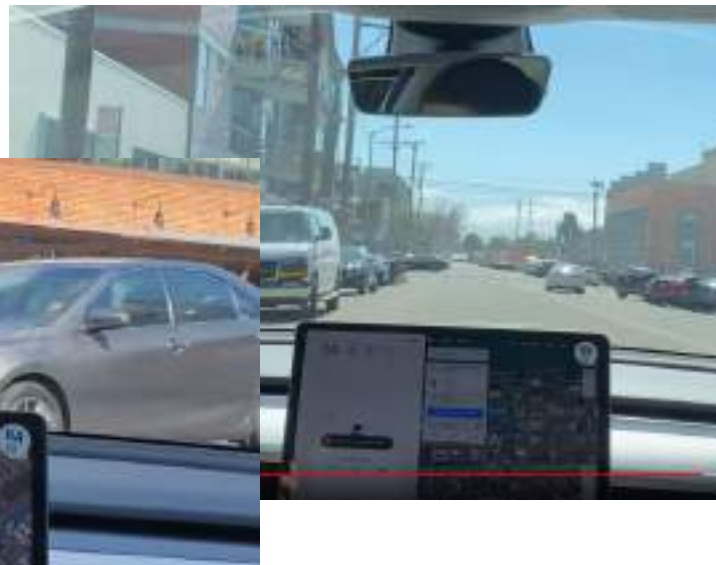
Source: <https://blogs.nvidia.com/blog/2020/03/23/nvidia-leads-autonomous-vehicle-report/>

2021 FSD Beta 8.2 Multiple “Near Collisions” (15/3/21)

1) Crosses double-yellow lines towards oncoming traffic



2) Drives wrong side of road



3) “Holy Shhh*” two near collisions

Source: <https://youtu.be/antLneVlxcs>

2021 FSD Beta 8.2 Multiple “Can’t See Cars” (18/3/21)

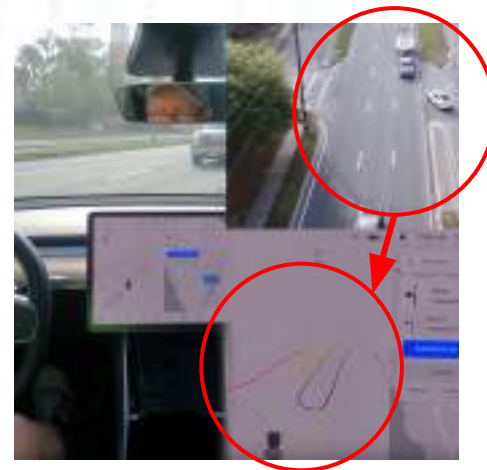
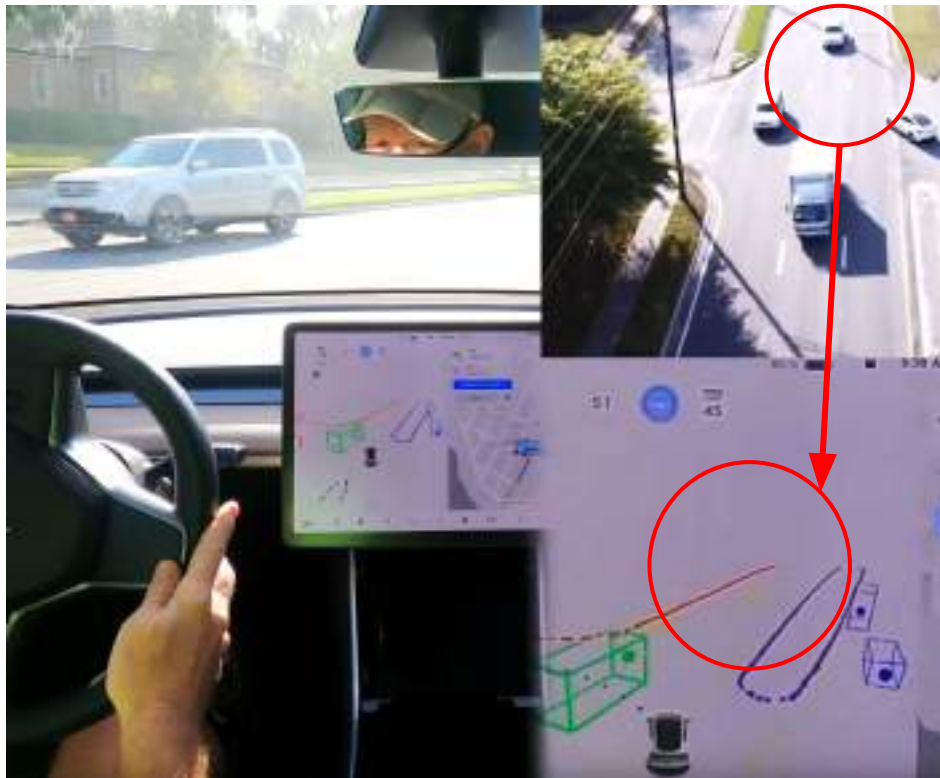


Drone Sees:
5 cars
(in lanes)

Tesla Sees:
4 cars
(2 on median)

Source: <https://youtu.be/uCIWIVCwHsl>

2021 FSD Beta 8.2 Multiple “Can’t See Cars” (18/3/21)



Human Override as FSD
About to Crash

“Full speed... going straight
for that... STOP! Aaaagghh!”

Source: <https://youtu.be/uCIWIVCwHsl?t=243>
<https://youtu.be/uCIWIVCwHsl?t=508>

2021 Tesla Autopilot Crashes Through Traffic... Flips



Source: <https://twitter.com/BradMunchen/status/1376749735369842691>

2021 Tesla Blames Driver Even When Not In Their Car

Mom and toddler almost hit by Tesla with ***no human driver***: “I don't think it was very safe for somebody to ***summon it without being in the car.***”

Tesla response: “...a driver can override any of autopilot's features at any time by steering, applying the brakes, or using the cruise control.”

FALSE



“Safest Record of Any Car” Claim Based on... No Records



Elon Musk

@elonmusk

What makes this incredibly **unjust** is that the Model S to date has the best safety record of any car on the road (no injuries or deaths ever)

8:30 AM · Nov 19, 2013 · Twitter Web Client



- 2013 Model S veers into opposite lane killing 2
- 2014 “Autopilot” claims made with unverified “safety” features:
 - Tesla “Model S” recorded **over 25 deaths** as of March 2021
 - Other brands same period showed more perfect safety records than ever before

APPLY: Enforceable Engineering Actions for Safe Learning

Tactical (near term) -- Easy, Routine, Minimal Judgment

- Off Button
- Reset Button

Strategic (long term) -- Identify, Store, Evaluate, Adapt

- Threat Model
- Setup Ethics Review Board
- Gate Releases Through Testing and Audits
- Assign Executive-Level Enforcement (Independent)

RSA[®]Conference2021

May 17 – 20 | Virtual Experience



SESSION ID: STR-R08

Top Seven AI Breaches: *Learning to Protect Against Unsafe Learning*

Davi Ottenheimer

VP Trust and Digital Ethics

inrupt

#RSAC