Breaking Bad AI

Closing the Gaps Between Data Security and Science

Davi Ottenheimer

Tuesday, Feb 25, 3:40 PM

Moscone West 3004

inrupt

RSAConference2020

Abstract

LONG: For years the security industry has discussed the leap into cloud computing as a paradigm shift. Get ready for another leap, as data platforms are rapidly becoming AI/ML development that requires its own new species of safety validations. This presentation, based on years of field test, will provide a quick intro and practical test list for security teams to comfortably engage data science projects.

QUICK: If you have AI/ML projects and aren't sure how to validate their safety, this presentation is for you. Don't let another AI project go forward without security assessments being in the conversation.

Thus, Two Classes of Breaking

Breaking: Replace, Adjust, Keep Hammering

Breaking Bad: Stop, Toss Bag, Build Standards of Trust (Quality Assurance)

"Break Bad": American 1880s Colloquialism for Flaws

"***Character***" flaw in trusted systems (e.g. business logic)

(Supreme Court, City and County of New York, 1886)

1. Financial pressures to survive can breed immoral safety ***shortcuts*** (for high margins)
2. Unregulated margins attracts increasingly organized ***greed*** (in unsafe/abusive markets)
3. If no enforcement, then expansive disasters (immoral expansion and ultimately "***war***")

GEORGE A. SMITH, polisher, 20 years experience:

The bits used to break bad. In the shop they called this steel "Jersey lightning" (fol. 2206).

All the steel used in the manufacture of bits during 1873 was of bad quality (fol. 2208).

"The fact is that they wouldn't show until you

2189 with the exception of two or three months he got some poor steel there that would break in the same manner; I don't know where he got it from, nor what the name of it was, and it broke so bad that we used to send them in without brightening the screws; that was at Mr. Beecher's shop in 1868 or 1869; I had a contract there for five or six years.

Q. Did you hear of any complaint in the shop, at the time that you took this contract in 1873, concerning this steel?

2190 A. Oh, yes; there was complaints; I don't know whether it was in 1873; I think it is very likely I heard complaints then; but then I remember it pretty much all the while along during those years.

Q. Who was the superintendent at this time?

A. Mr. Swan.

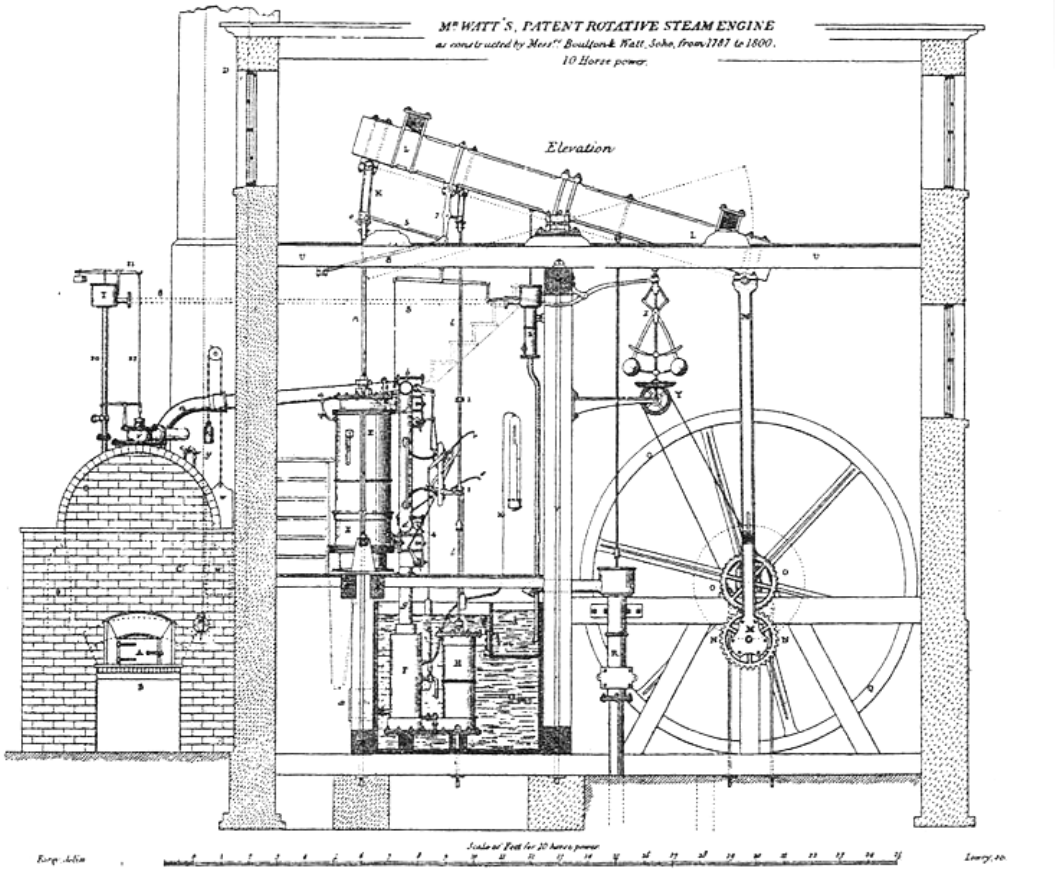
D. Did you represent to him that you were unable to get out a fair proportion of good work?

A. Yes; I told him a good many times about it that they would break bad; sometimes he told me to

2191 have them tempered over, and I did so, but that would do no good. This lot that I speak of that I remember so perfectly—I sent part of them down to have them tempered over; that didn't seem to do any good, they would break first as bad and mash down the same as before.

https://books.google.com/books?id=GebrnWPYIrIC&pg=PA548#v=onepage&q&f=false

1770s The Piston Engine Was Meant to Save Lives



Breaking Bad in 1905: “Grover Shoe Factory Disaster”

orter.

FIFTY DEAD.

Boiler Exploded in Brocton Shoe Fact
ory With Terrible Results.

Brocton, Mass., March 30.—A boiler thirty, twenty-five dead bodies had exploded this morning in the R. B. been taken from the ruins, many Grover shoe factory, employing six terribly mutilated and burned as hundred hands. A section of the build- make identification difficult if not ing in which two hundred hands were possible. A number of injured w working was wrecked and a fire removed to the hospital. started which spread to the adjoining The fire marshall has issued a ment saying that reports to him buildings, several of which were des- dicate that the number of dea employees are un- reach fifty. David Rockwell in believed many of the boiler has been detained ing an investigation of the acc

Breaking Bad in 1905: Regulatory Crackdown in 2 Years

Objections from manufacturers to "needless government interference" are completely set aside:

- Stringent ***operational*** safety laws
- National code

SECTION 2. This act shall take effect upon its passage.
Approved May 29, 1907.

---

*Chap.*465 AN ACT RELATIVE TO THE OPERATION AND INSPECTION OF STEAM BOILERS.

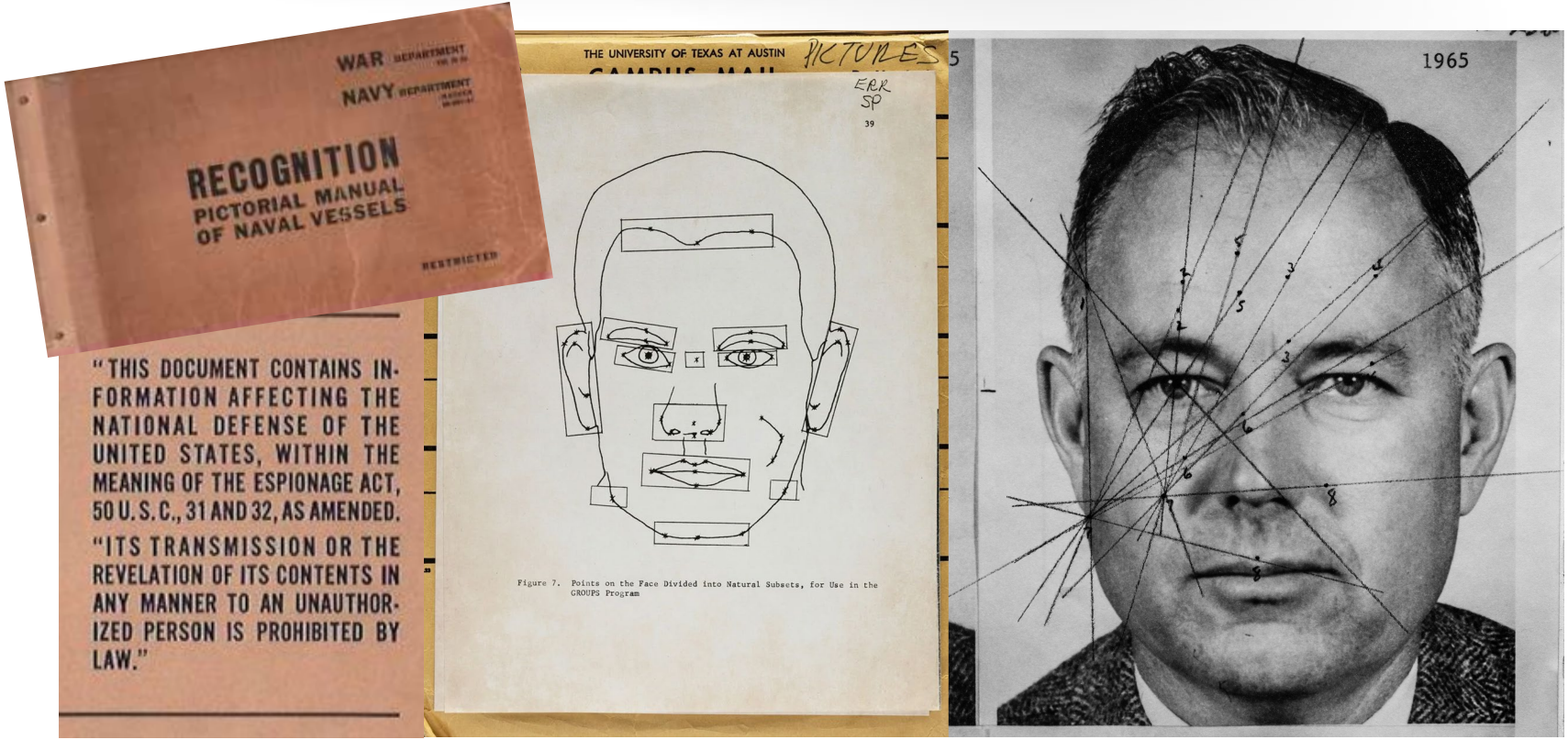
Be it enacted, etc., as follows:

Operation and inspection of steam boilers.

SECTION 1. All steam boilers and their appurtenances, except boilers of railroad locomotives, motor road vehicles, boilers in private residences, boilers in public buildings and in apartment houses used solely for heating, and carrying pressures not exceeding fifteen pounds per square inch, and having less than four square feet of grate surface, boilers of not more than three horse power, boilers used for horticultural and agricultural purposes exclusively, and boilers under the jurisdiction of the United States, shall be thoroughly inspected internally and externally at intervals of not over one year, and shall not be operated at pressures in excess of the safe working pressure stated in the certificate of inspection hereinafter mentioned, which pressure is to be ascertained by rules established by the board of boiler rules, to be appointed as

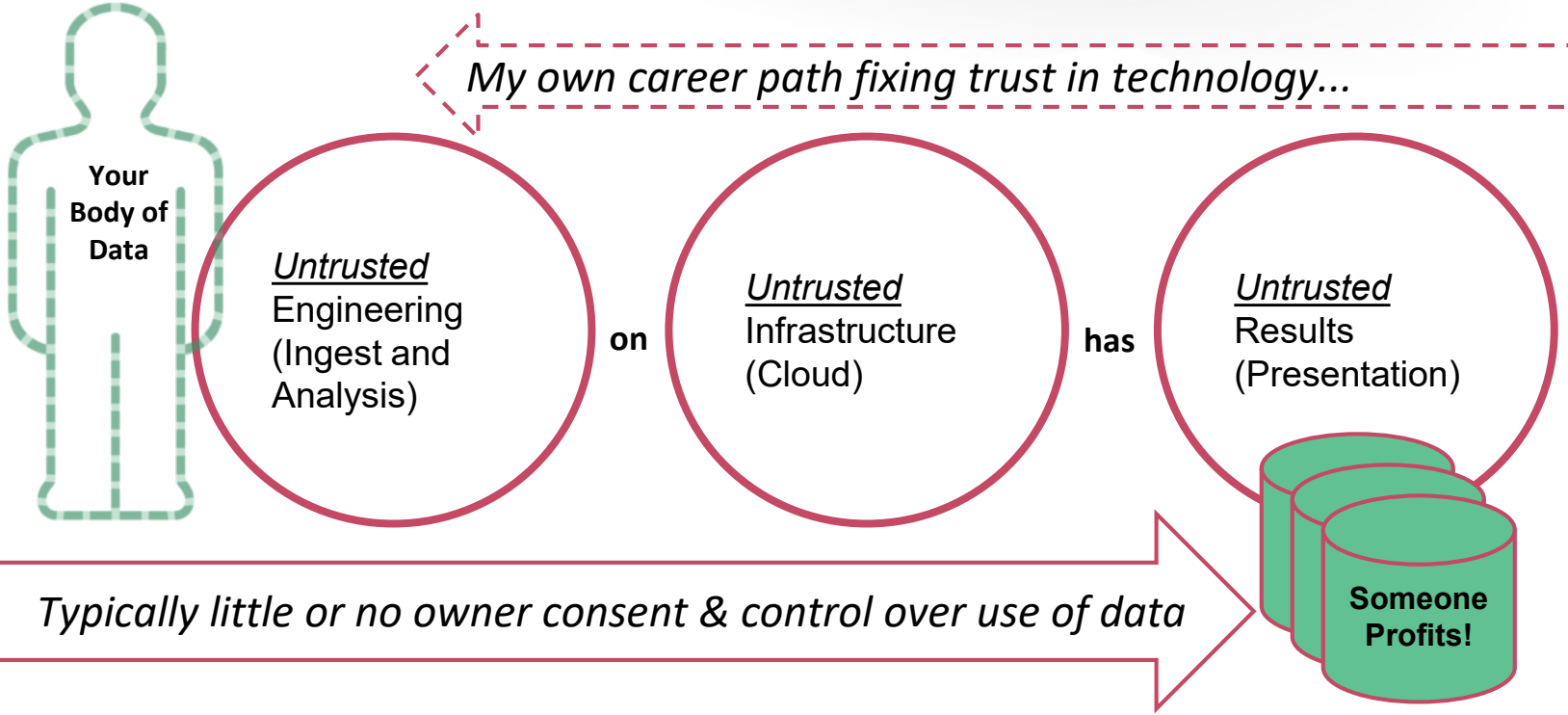
https://archives.lib.state.ma.us/bitstream/handle/2452/74865/1907acts0465.pdf

1950s Recognition Machines Desired For…



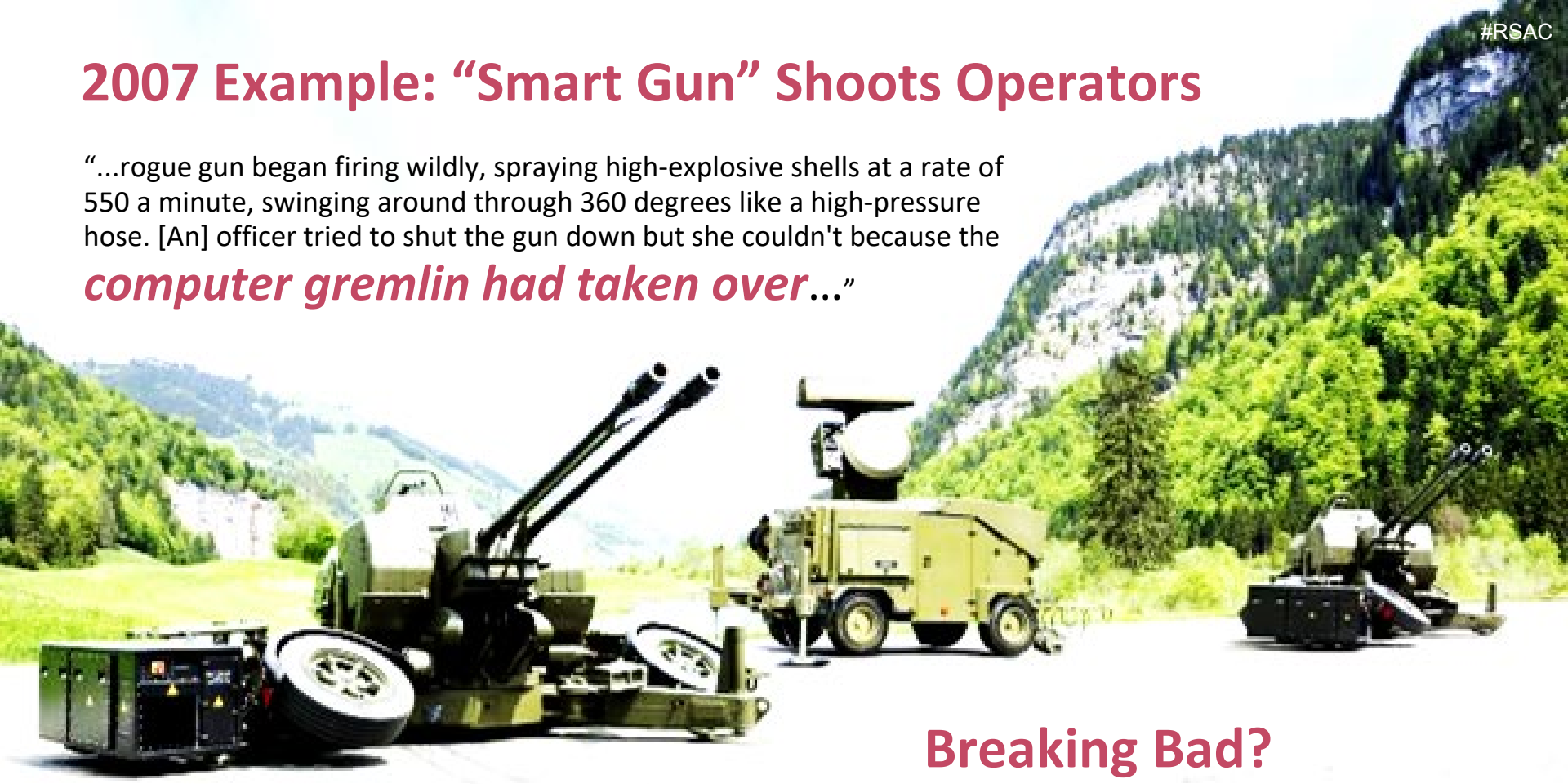
https://www.wired.com/story/secret-history-facial-recognition/
https://www.flyingpenguin.com/?p=25255 and p=26449
https://acepilots.com/ships/manual.html

60 Years Later... Can Anyone Tell If Breaking Bad?



2007 Example: “Smart Gun” Shoots Operators

“...rogue gun began firing wildly, spraying high-explosive shells at a rate of 550 a minute, swinging around through 360 degrees like a high-pressure hose. [An] officer tried to shut the gun down but she couldn't because the ***computer gremlin had taken over***...”



Breaking Bad?

https://www.iol.co.za/news/south-africa/9-killed-in-army-horror-374838

Yes. But Do Gremlins Lead To... More Special Humans?

2020 Smart Gun Report:

"...***can't trust an algorithm — no matter how smart*** — to seek out, identify and kill the correct target..."

Versus Specialized Warriors:

"Our job ... is also to have the judgment to figure out a plan to get to the ***desired end state*** under high-stress situations ***where violence may be counterproductive***."

https://www.businesslive.co.za/bd/world/2020-01-18-whos-calling-the-shots-on-lethal-autonomous-weapons-systems/
https://www.washingtonexaminer.com/opinion/a-warrior-class-of-their-own

AI The Weapon of Modern Warfare (e.g. Civil Rights)

"Humans are and have always been vulnerable to being tricked, provoked, conditioned, deceived, or otherwise manipulated

...artificial intelligence slashes the transaction costs."

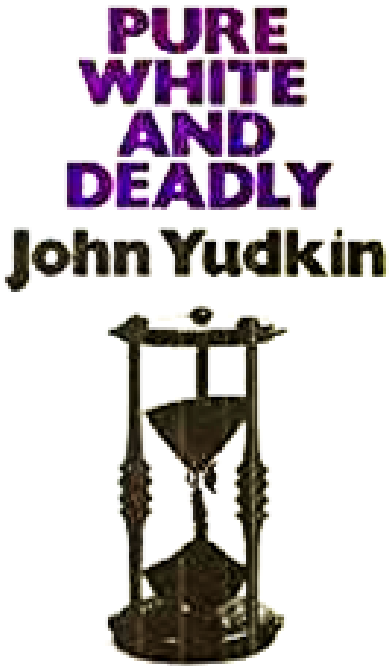
"I told you don't worry. With AI they now see us as fishermen! Haha"

Unvarnished Reality: AI Interferes With Democracy

Direct impact on human liberties.

Facebook no more can claim its products protect our privacy, than Coke can claim sugary drinks protected our health.

Far-right white supremacist groups are primed to use AI to wage global asymmetric battles against liberty and freedom.

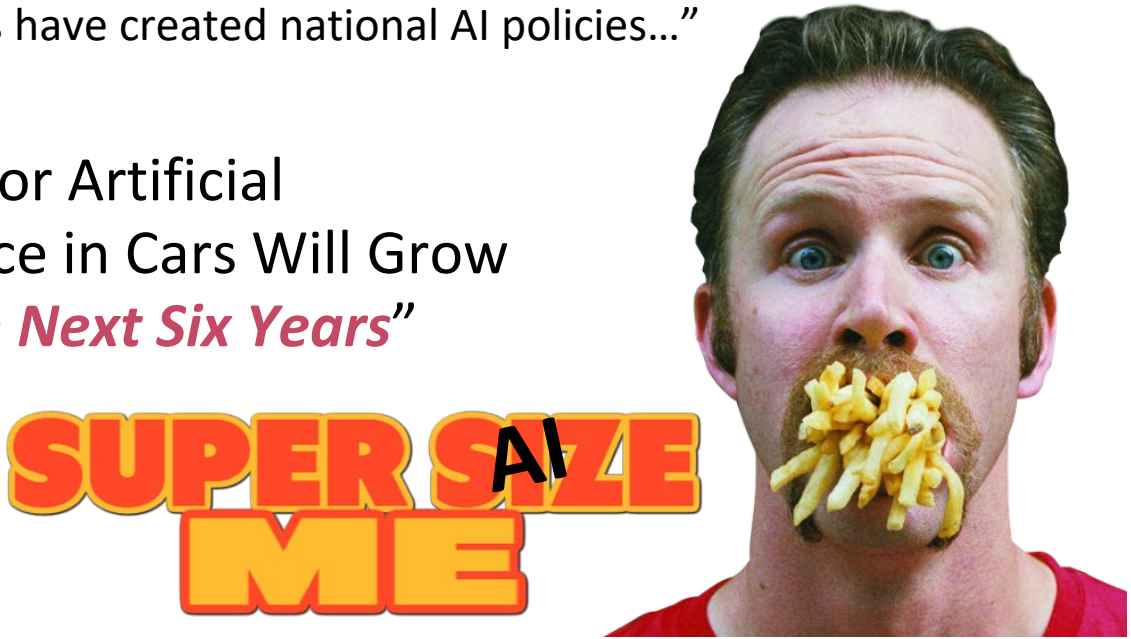


https://www.telegraph.co.uk/lifestyle/wellbeing/diet/10634081/John-Yudkin-the-man-who-tried-to-warn-us-about-sugar.html

Despite Risks, AI Adoption Predictions Are Way, Way Up

"…global spend on artificial intelligence (AI) is expected to hit ***$52 billion in the next three years*** and to double the annual growth rates of major economies in the next 15. Approximately 29 countries have created national AI policies…"

"Market for Artificial Intelligence in Cars Will Grow ***1,200% in Next Six Years***"



https://www.machinedesign.com/mechanical-motion-systems/article/21122520/prediction-market-for-artificial-intelligence-in-cars-will-grow-1200-in-next-six-years
http://wef-ai.s3.amazonaws.com/WEF_Empowering-AI-Leadership_Oversight-Toolkit.pdf
https://www.weforum.org/press/2020/01/artificial-intelligence-toolkit-helps-companies-protect-society-and-their-business/

Dangers Are Fueled by Existential "Soda and Fries" FOMO

84% of C-suite executives believe they must leverage artificial intelligence (AI) to achieve profit growth objectives…

[75%] of C-suite executives believe if they don't scale AI in the next five years, ***they risk going out of business entirely***.

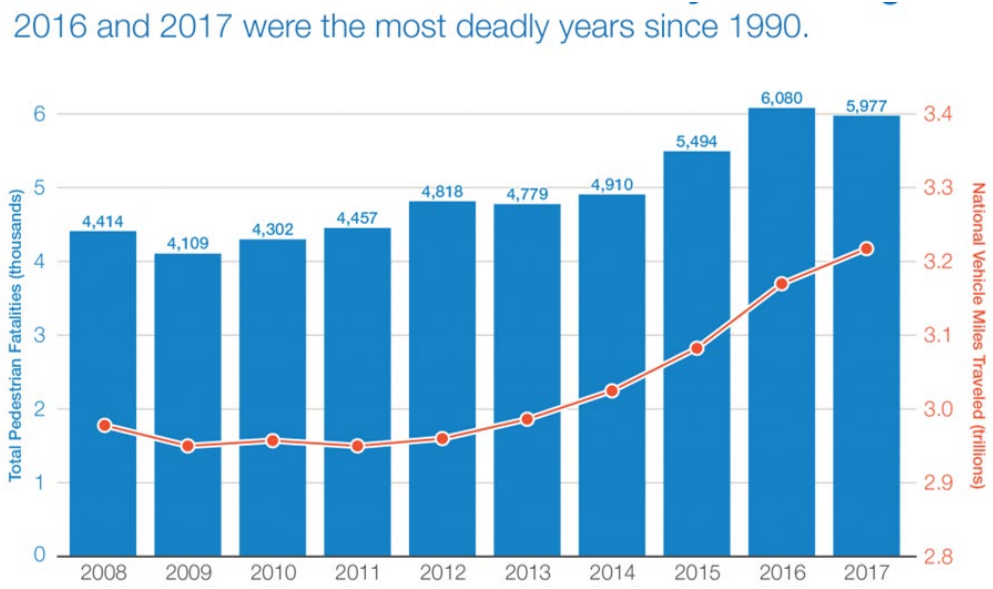
https://www.accenture.com/us-en/insights/artificial-intelligence/ai-investments

Desire for Scientifically Healthy Human Living? Unlikely. Death by Weaponized Tech? Very Very Likely.

Current U.S. pedestrian death rate of 13 per day is equivalent to no survivors in ***monthly jumbo jet crashes***.

"... future autonomous cars will … mean [intentionally] ***sacrificing the lives of pedestrians*** …"

-- Christoph von Hugo, Mercedes-Benz Active Safety



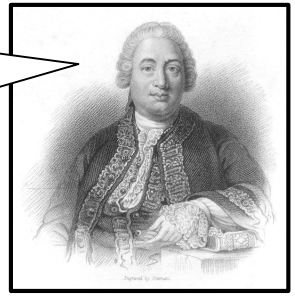
https://fortune.com/2016/10/15/mercedes-self-driving-car-ethics/
https://smartgrowthamerica.org/resources/dangerous-by-design-2019/

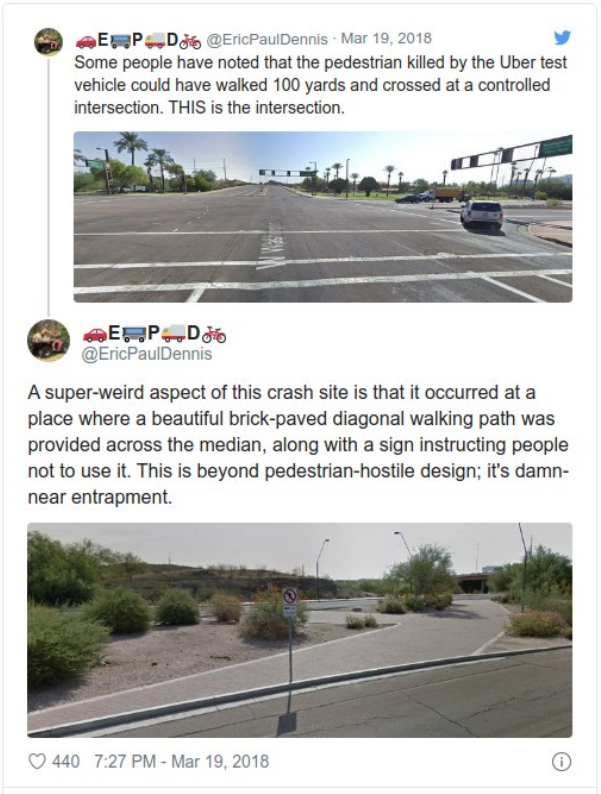
Easy to Explain Why March 2018 Uber Killed a Woman

Company BLIND to Repeated Warnings: Operated With NO SAFETY DIVISION

- "...incident where a car mounted the sidewalk was ignored"
- "...hitting things every 15,000 miles", a vehicle was damaged "nearly every other day"
- "...placing too much emphasis on the number of ***miles driven as a metric for how advanced*** its software was getting"

Miles driven is NOT a reasonable benevolence test… just as repeatedly adding 2+2 could never be a metric of math skill. Duh.





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

Alternate Title:

The Truth is History

The Real Ignorance Found in Artificial Intelligence Products

Davi Ottenheimer

Tuesday, Feb 25, 3:40 PM

Moscone West 3005

Alternate Abstract (Replacing Me With SunTzuBot)

```
pairs (
(
        r'[A-Ca-c](.*)',
        (
"The art of war is of vital importance to the State.",
"All warfare is based on deception.",
"If the campaign is protracted, the resources of the State will not be equal to the strain.",
"There is no instance of a country having benefited from prolonged warfare.",
        ),
)
)  suntsu_chatbot = Chat(pairs, reflections)
   suntsu_chatbot.converse()
```

https://www.nltk.org/_modules/nltk/chat/suntsu.html

RSAConference2020

Automation Tech Forecasters Shouldn't Ignore *Important* History (War) Lessons

"The Further We Go Back, The Better We Can See Desirable Ends

Who Was Caty?

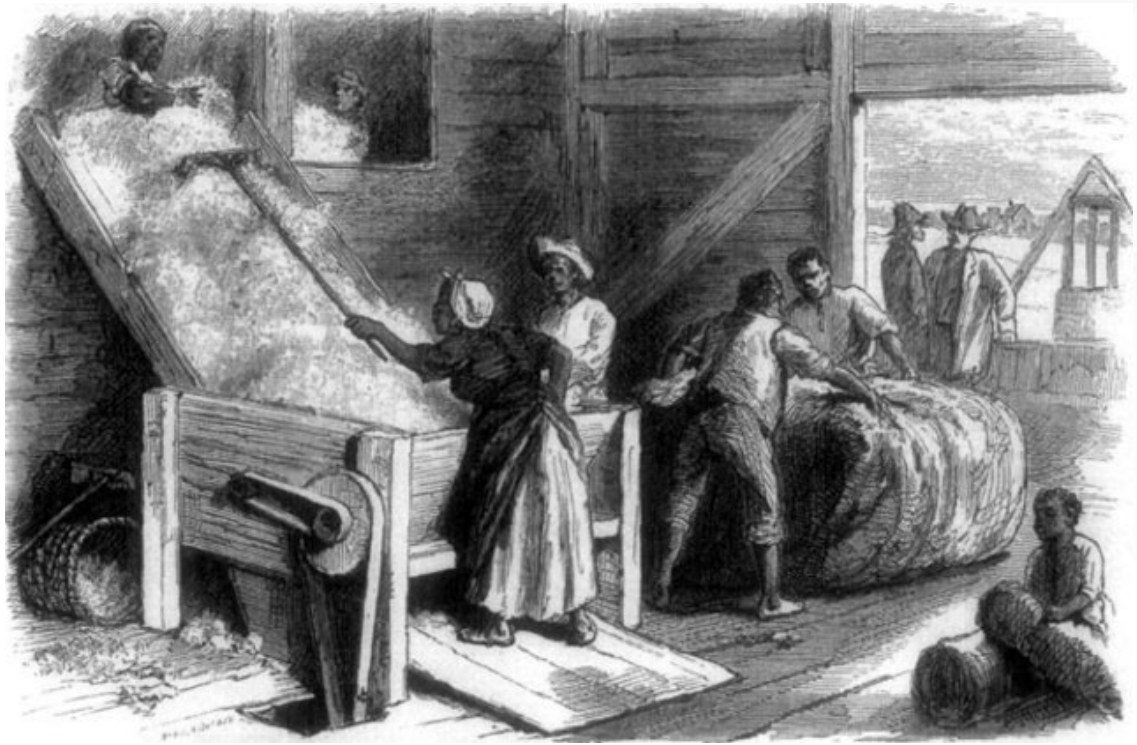
Catharine Littlefield Greene Miller

(17 February 1755 – 2 September 1814)

What Did She Invent (1793)?

...And Why?

Caty's Invention: A Cotton Engine ('gin)



http://www.womenhistoryblog.com/2009/03/catherine-littlefield-greene.html
https://books.google.com/books?id=uRJt7QqA7GEC&pg=PA544&f=false

Caty's Husband: Revolutionary War "Hero" Nathanael

1779 Promoted to Commander in Chief of "Southern Army"

Agile: "We fight, get beat, rise and fight again."

1781 British defeated, he is ruined by Charleston merchants and South Carolina courts demanding payments (on his cosigned notes)

Driven: "...given himself to [Revolutionary] belief, signed away his future life..."

1782 Georgia "funds" him 2,000+ confiscated plantation acres

1786 Dies of heat stroke working land (abolitionist/Quaker)

Ethical: "Nothing can be said in [slavery's] defense"

Whitney Then Stamps Caty Out (He instead of She)

"Inexperienced, and fresh off a boat..."

- ***Cheated*** of my tutor pay and desperate, Caty takes me in, instructs me to build engine
- American women ***barred*** from filing patents, so file with Miller, whom she marries
- We're then ***robbed*** of income, service payments delinquent, rampant theft of our IP

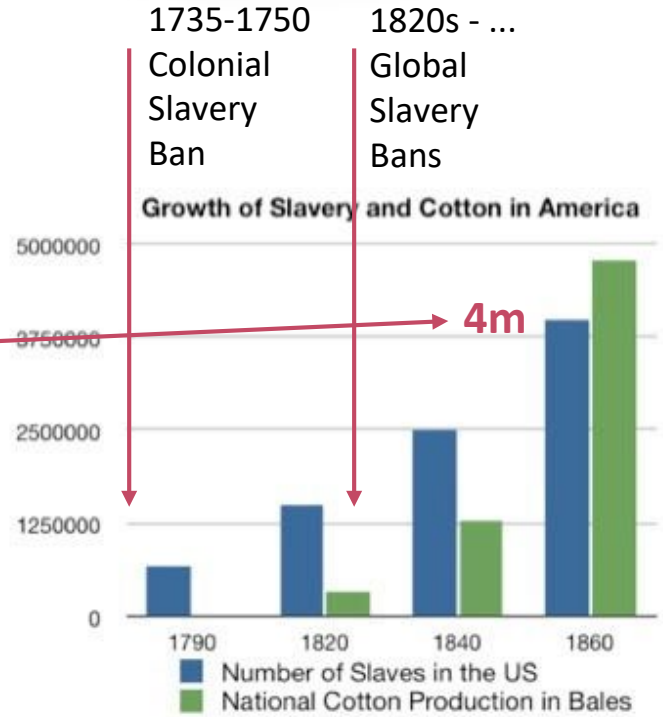
Engine Use Really Went South From There...

Owner-Operators Implicated in Widespread Crimes Against Humanity

North (and rest of world): Industrialization and Individual Rights (e.g. Eli Whitney's rifle factory)

South: Massive slavery expansion for cotton gins. A 1808 U.S. ban on use of ships for slavery results in ***state-sanctioned rape of women***

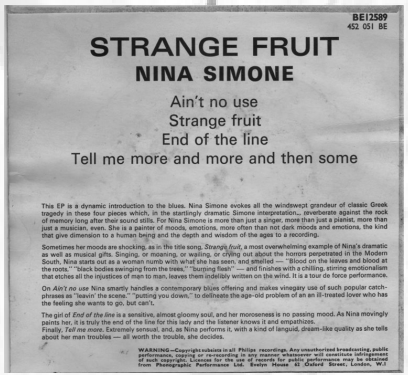
"Potential sexual partners of enslaved women included the master, his sons, neighboring planters, visitors of the slaveholding family, traveling salesmen, and hired workers."



https://slate.com/human-interest/2015/08/how-enslaved-womens-sexual-health-was-contested-in-the-antebellum-south.html

1838 Lincoln Describes "Breaking Bad" Symptoms

"...***dead men were seen literally dangling from the boughs of trees upon every road side***; and in numbers almost sufficient, to rival the native Spanish moss of the country, as a drapery of the forest."


BE12589
452 051 BE

STRANGE FRUIT

NINA SIMONE

Ain't no use
Strange fruit
End of the line
Tell me more and more and then some

This EP is a dynamic introduction to the blues. Nina Simone evokes all the windswept grandeur of classic Greek tragedy in these four pieces which, in the startlingly dramatic Simone interpretation... reverberate against the rock of memory long after their sound stills. For Nina Simone is more than just a singer, more than just a pianist, more than just a musician, even. She is a painter of moods, emotions, more often than not dark moods and emotions, the kind that give dimension to a human being and the depth and wisdom of the ages to a recording.

Sometimes her moods are shocking, as in the title song, *Strangly fruit*, a most overwhelming example of Nina's dramatic as well as musical gifts. Singing, or moaning, or wailing, or crying out about the horrors perpetrated in the Modern South, Nina starts out as a woman numb with what she has seen, and smelled — "Blood on the leaves and blood at the roots," "black bodies swinging from the trees," "burning flesh" — and finishes with a chilling, stirring emotionalism that etches all the injustices of man to man, leaves them indelibly written on the wind. It is a tour de force performance.

On *Ain't no use* Nina smartly handles a contemporary blues offering and makes vinegary use of such popular catch-phrases as "leavin' the scene," "putting you down," to delineate the age-old problem of an ill-treated lover who has the feeling she wants to go, but can't.

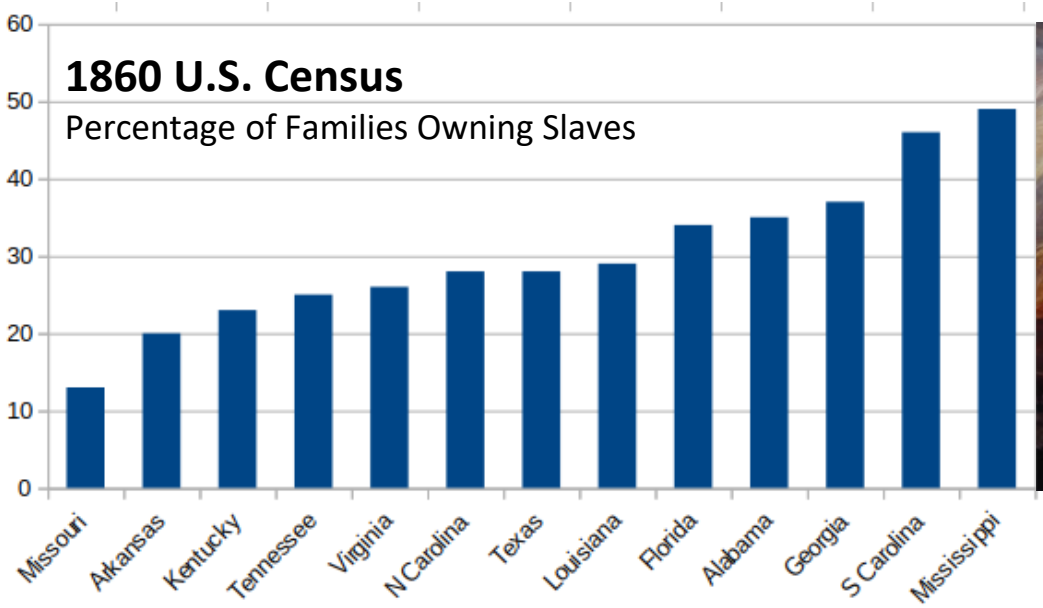
The girl of *End of the line* is a sensitive, almost gloomy soul, and her morosenss is no passing mood. As Nina movingly paints her, it is truly the end of the line for this lady and the listener knows it and empathizes.

Finally, *Tell me more*. Extremely sensual, and, as Nina performs it, with a kind of languid, dream-like quality as she tells about her man troubles — all worth the trouble, she decides.

(100 years later, in 1937 Abel Meeropol wrote similarly, as Performed by Nina Simone)

1859 Brown Hanged for “Treason”… Civil War Erupts

- 90% of American south black population enslaved ***20 years after global bans***
- Slaves are 50% of population in South Carolina and Mississippi
- ***30% avg of families own slaves*** in 15 states (Other 20 states already abolished it)



https://books.google.com/books?id=uRJt7QqA7GEC&pg=PA544&f=false
https://www.hup.harvard.edu/catalog.php?isbn=9780674002111
https://www.census.gov/library/publications/1864/dec/1860a.html

Why Did Caty Invent That Engine?

1793 Desired a ‘Gin to Achieve...

- Freedom
- Equality

Instead Her Unregulated Tech Meant

- Enslavement
- Exploitation and Erasure

RSAConference2020

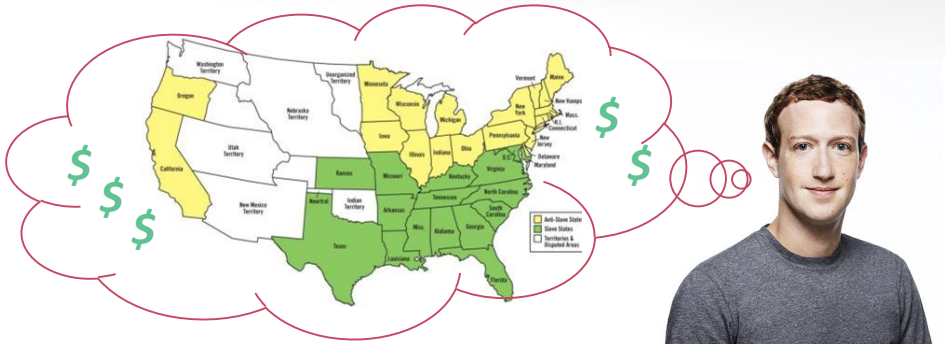
Fast Forward to Today and the Need to Emancipate Our Data Bodies

“The Further We Go Back, The Better We Can See Desired End State”

Is Facebook Just a Modern Slavery Engine?

"Poor things, 'they can't take care of themselves.'"

Should You Let AI 'Gin Expand Into Markets Today Like...



Source: MSW@USC Diversity Toolkit: Guide to Discussing Identity, Power and Privilege

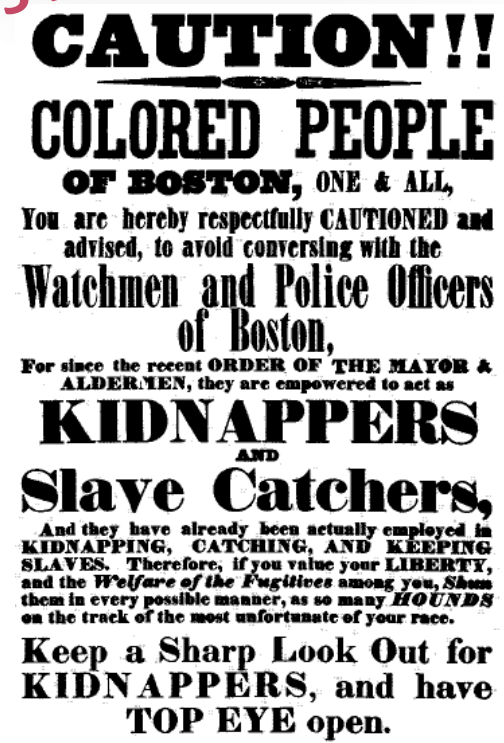
https://www.flyingpenguin.com/?p=27668

And How About Google's "Data Body" Snatching Engine?

"...*go after people of color*, conceal the fact that people's faces were being recorded and even *lie to maximize their data collections*.

...*target homeless* people because they're least likely to say anything to the media."

1854


CAUTION!!

COLORED PEOPLE

OF BOSTON, ONE & ALL,

You are hereby respectfully CAUTIONED and advised, to avoid conversing with the

Watchmen and Police Officers of Boston,

For since the recent ORDER OF THE MAYOR & ALDERMEN, they are empowered to act as

KIDNAPPERS AND Slave Catchers,

And they have already been actually employed in KIDNAPPING, CATCHING, AND KEEPING SLAVES. Therefore, if you value your LIBERTY, and the *Welfare of the Fugitives* among you, *Shun* them in every possible manner, as so many *HOUNDS* on the track of the most unfortunate of your race.

Keep a Sharp Look Out for KIDNAPPERS, and have TOP EYE open.

https://www.nydailynews.com/news/national/ny-google-darker-skin-tones-facial-recognition-pixel-20191002-5vxpgowknffnvbmy5eg7epsf34-story.html

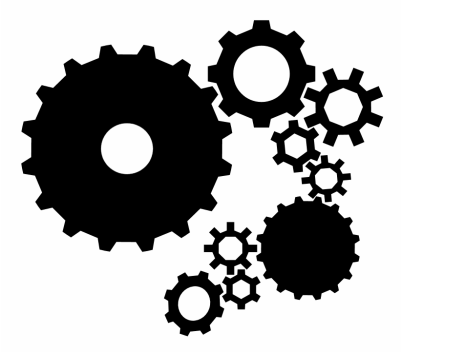
So Next Logical Step is “De Oppresso Liber”, Right?

...Right?

Intervention/Abolition Depends on Stage of Breaking

Cog (Metal) *Breaking*

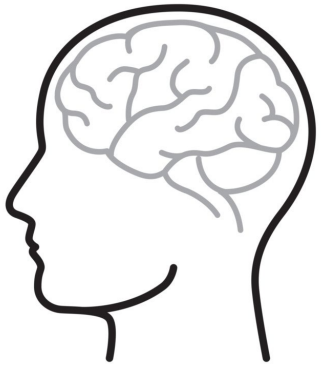
Easy, Routine and Minimal Judgment (ERM)



Subconscious, Instinctive, Intuitive, Inexpensive to Replace

Cognition (Meat) *Breaking Bad*

Identify, Store, Evaluate, Adapt (ISEA)

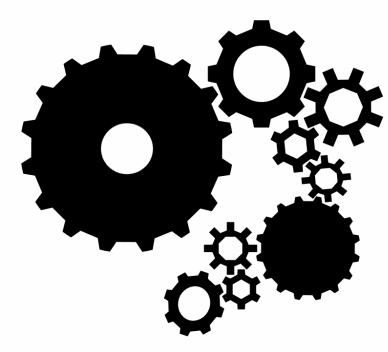


Conscious, Deliberative, Logical, Expensive to Restore

… Expressed Roughly and Confusedly as ML and AI

Machine Learning (ML)

Data Science, Statistical Inference



Machine Understanding (AI)

Knowledge Engineering, Logical Inference

“Seeing Relations” (Science) The Ultimate Fix For Bad AI

Cogs

Cognition

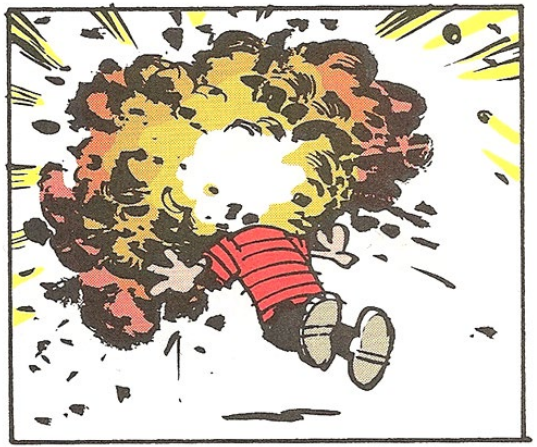
“Americans often think of schooling as the transmission of ***specialized skill sets*** — can the student read, do math, recite the facts of biology…”

“[Nordic] Bildung is devised to change the way students see the world. It is devised to help them ***understand complex systems and see the relations between things*** — between self and society, between a community of relationships in a family and a town.”

If “Intelligent” Machines are Breaking Bad…

Can Science Really Establish Trust?

Exploding Science of AI Ethics Can Be Very Confusing

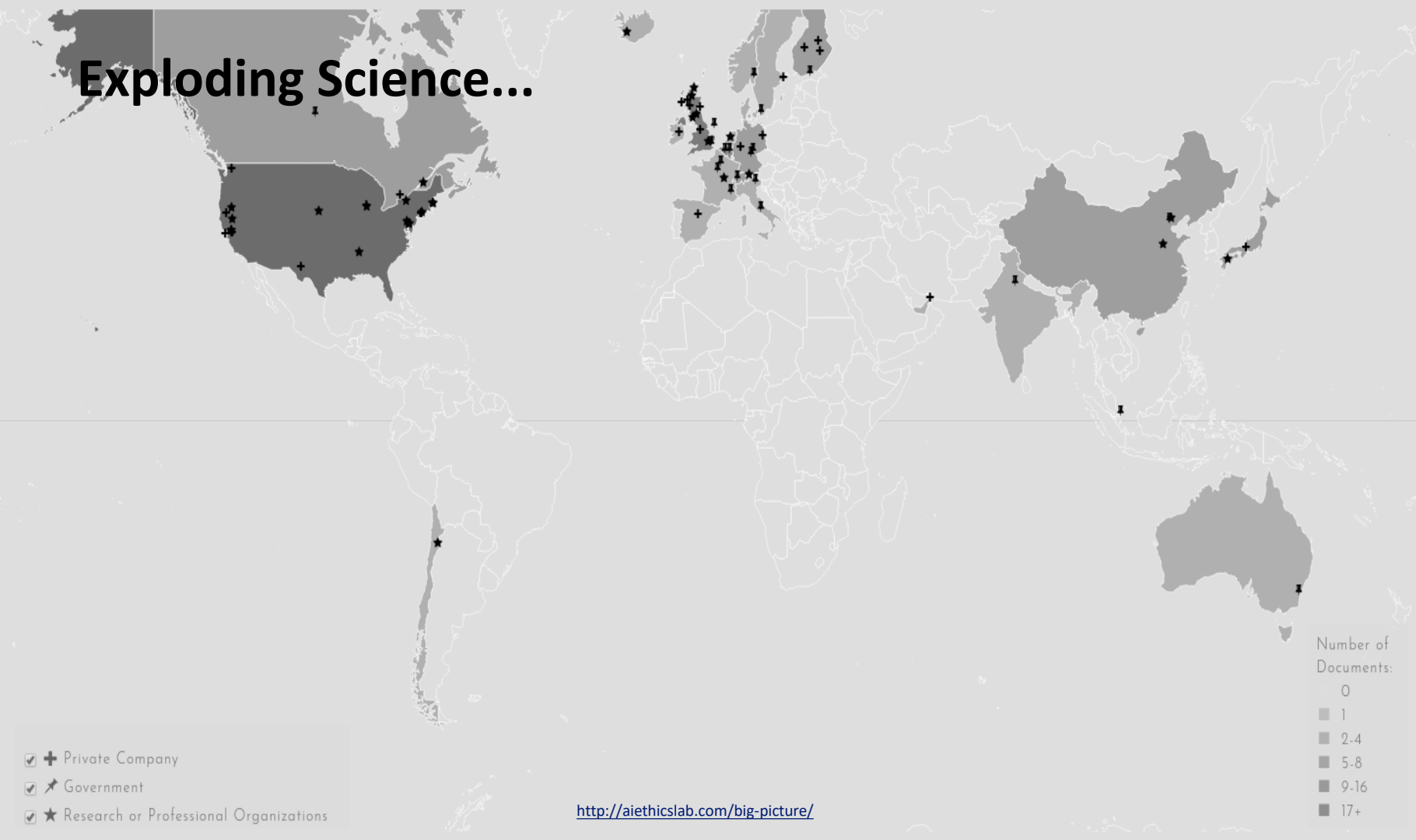


"...no single ethical principle was common to all of the ***84 documents on ethical AI*** we reviewed. Still, five principles are mentioned in more than half the sources:

1. Transparency,
2. Justice and fairness,
3. Non-maleficence (afflicting no harm),
4. Responsibility, and
5. Privacy."

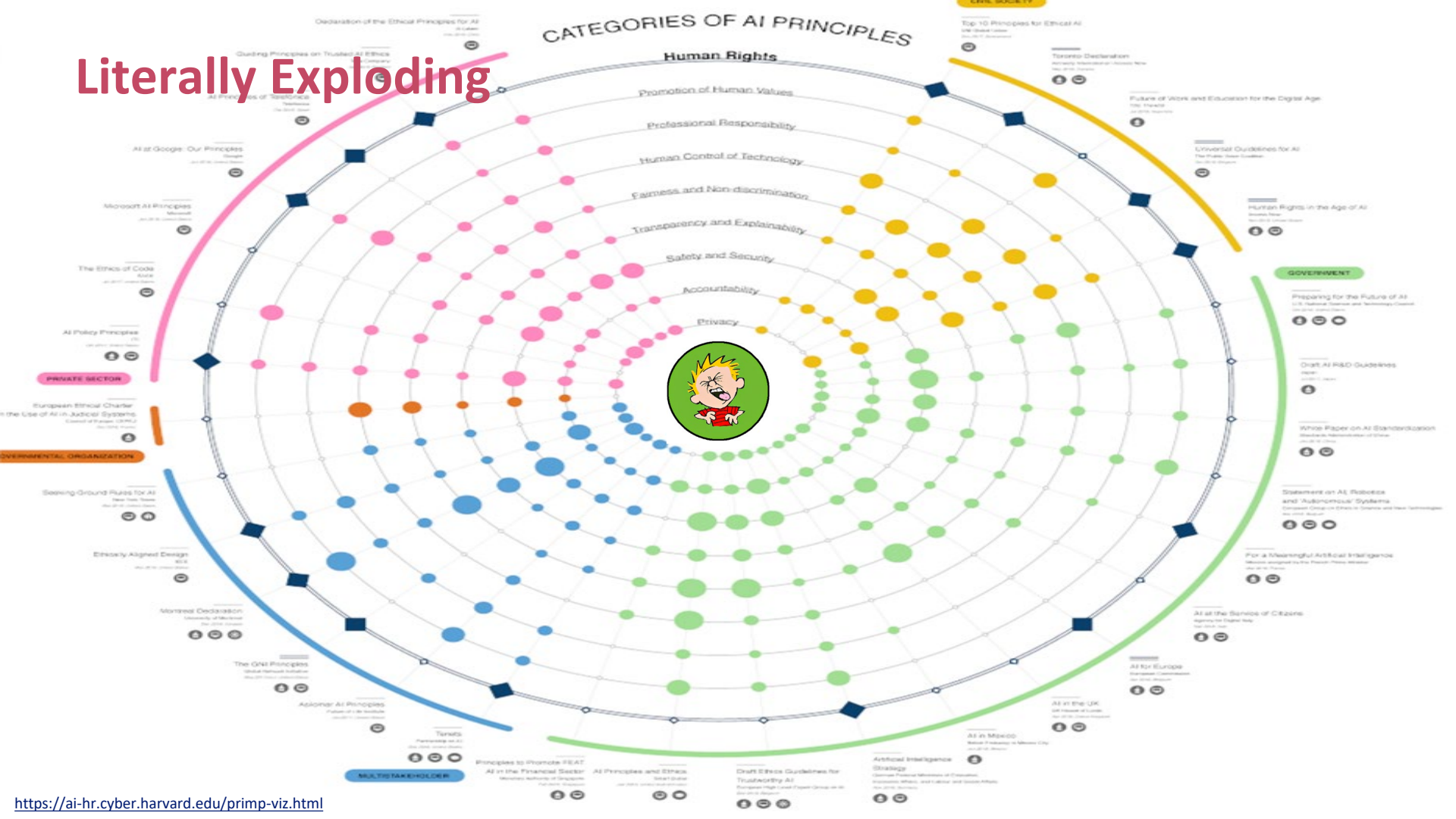
https://ethz.ch/en/news-and-events/eth-news/news/2020/01/ethics-guidelines-galore-for-ai.html

Exploding Science...



http://aiethicslab.com/big-picture/

Literally Exploding



https://ai-hr.cyber.harvard.edu/primp-viz.html

Hot Off The Press: EU Commission and ICO Auditing

Latest risk guidelines seem to omit important history...

- training data;
- data and record-keeping;
- information to be provided;
- robustness and accuracy;
- human oversight;
- specific requirements...

Simple Distillation

"I ***attribute [problems] to a false system of education*** , gathered from the books written on this subject by men, who, considering females rather as women than human creatures, have been ***more anxious to make them alluring mistresses than*** **affectionate wives and *rational mothers*...**"

-- Wollstonecraft 1792

https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-on-the-draft-ai-auditing-framework-guidance-for-organisations/
https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
https://www.bl.uk/collection-items/mary-wollstonecraft-a-vindication-of-the-rights-of-woman

Cognitive Science: Desire More Important Than Reason

Hume Established Science of Morality

- A Treatise of Human Nature (1740)

  Using experimental method for morality: "...reason is, and ought only to be the slave to the passions..."

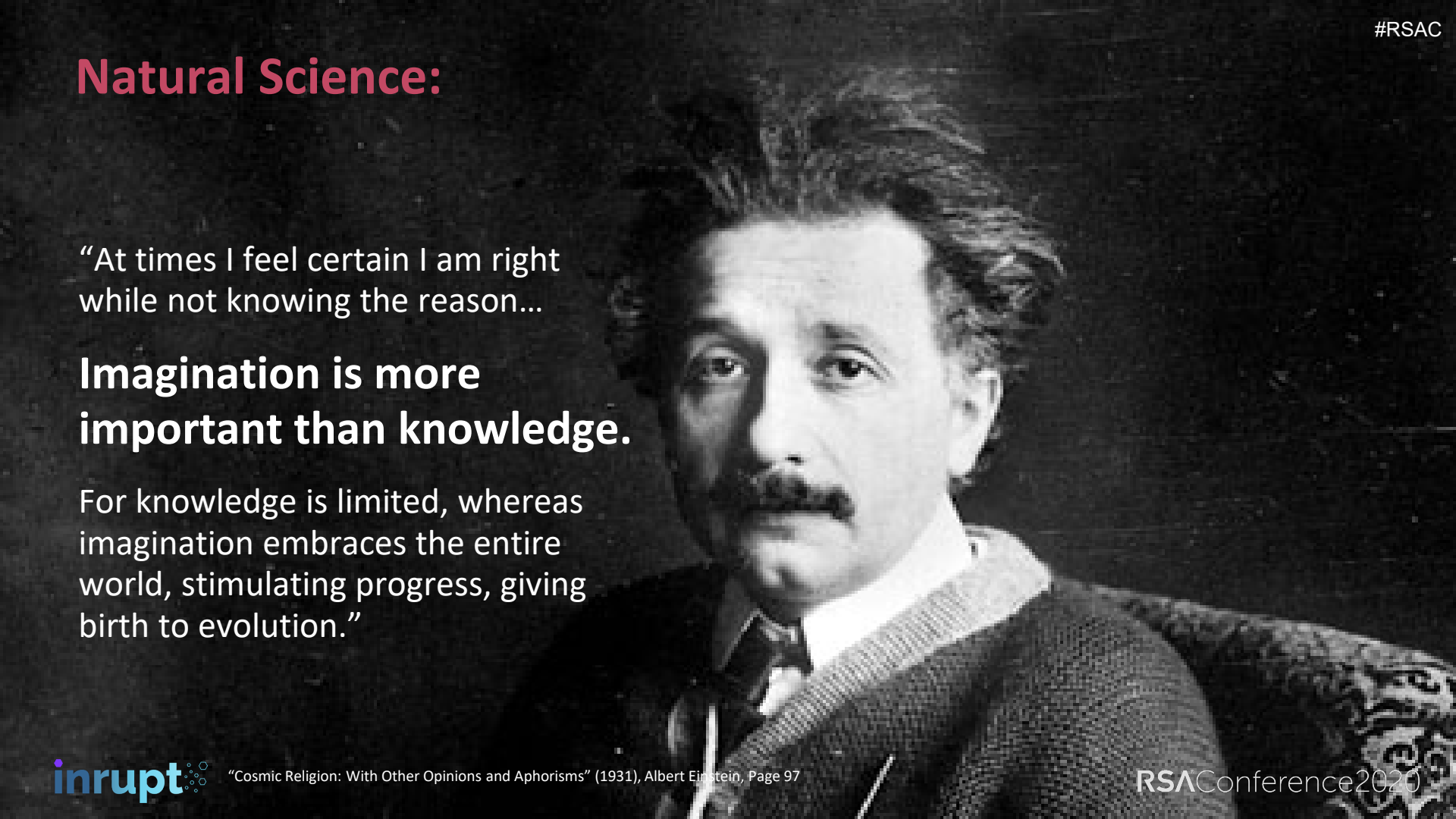
Wollstonecraft Established Passion of Equality

- A Vindication of the Rights of Men (1790)
- A Vindication of the Rights of Woman (1792)

Measure Intelligence *in Rational Sympathy & Universal Benevolence*

Caty's Cotton 'Gin!

Natural Science:



"At times I feel certain I am right while not knowing the reason…

Imagination is more important than knowledge.

For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution."

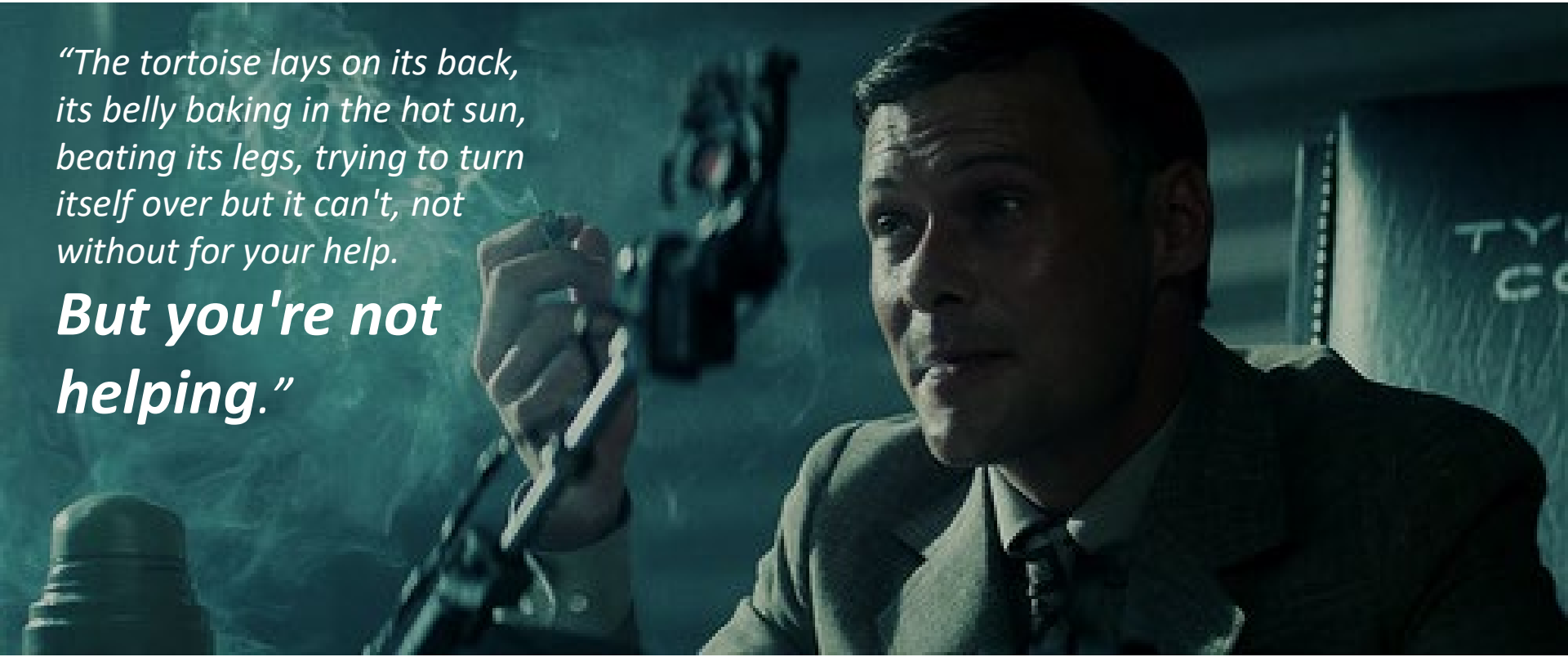
"Cosmic Religion: With Other Opinions and Aphorisms" (1931), Albert Einstein, Page 97

Life Science: Alpha (Parent) & Beta (Child) Roles

"...***alpha wolf protects*** his family and treats them with ***kindness, generosity, and love***... main characteristic of an alpha male wolf is a quiet confidence, quiet self-assurance... ***a calming effect***."

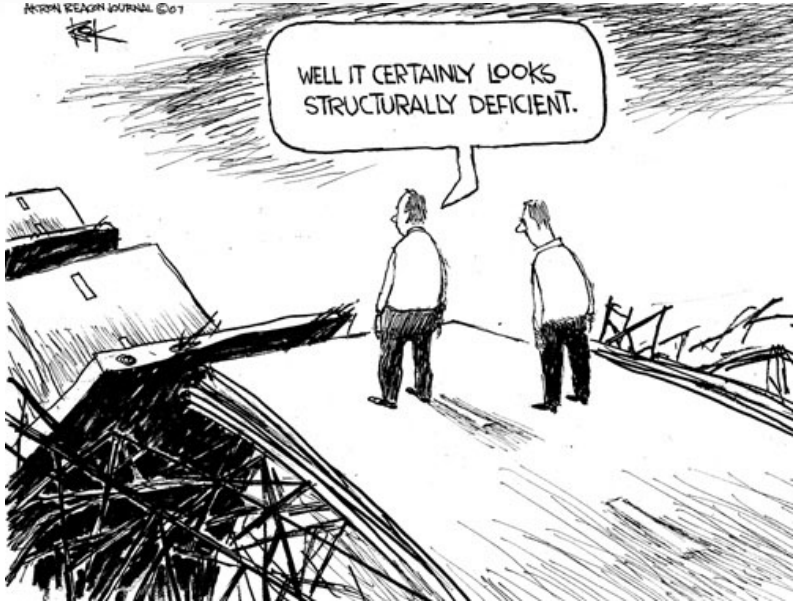
https://www.artofmanliness.com/articles/how-to-really-be-alpha-like-the-wolf/

Fictional Science: “...you're not helping. Why is that…?”



Applied Science (Engineering): Desirability of Cognizance

"...scientific principles to design or develop structures, machines, apparatus, or manufacturing processes, or works utilizing them singly or in combination; or to construct or operate the same with ***full cognizance of their design; or to forecast their behaviour*** under specific operating conditions; all as respects an intended function, ***economics of operation and safety to life and property***."



https://www.britannica.com/technology/engineering
https://www.oreilly.com/library/view/the-best-political/9780768691757/ch23.html

RSAConference2020

A Fix List, Real Science


"The web as I envisaged,
we have not seen yet."

Desire a Better Web; Safer AI

We can rotate towards better data integrity and confidentiality via user-centric data controls in technology.

Decentralized models key to achieving balance of AI knowledge and privacy.

Standards will enable essential functions like “roll-back and power-button”.

Desire Trusted Outcomes in AI Product Management

1. Safe? (Confidentiality, Integrity… Resistant to Threats)
2. Moral? (Beneficial)
3. Reversible? Can Mistakes (see 1 and 2) Be Reset Easily?
4. Power-Off? Can Services Be Terminated Without Loss?
    a. Happy to be offline to avoid harm (term limits)
    b. Or expects persistence (dictatorship by reverse “moats”)

Example: Reversible (User-Consent Control) Data Access

Solid Specification Service

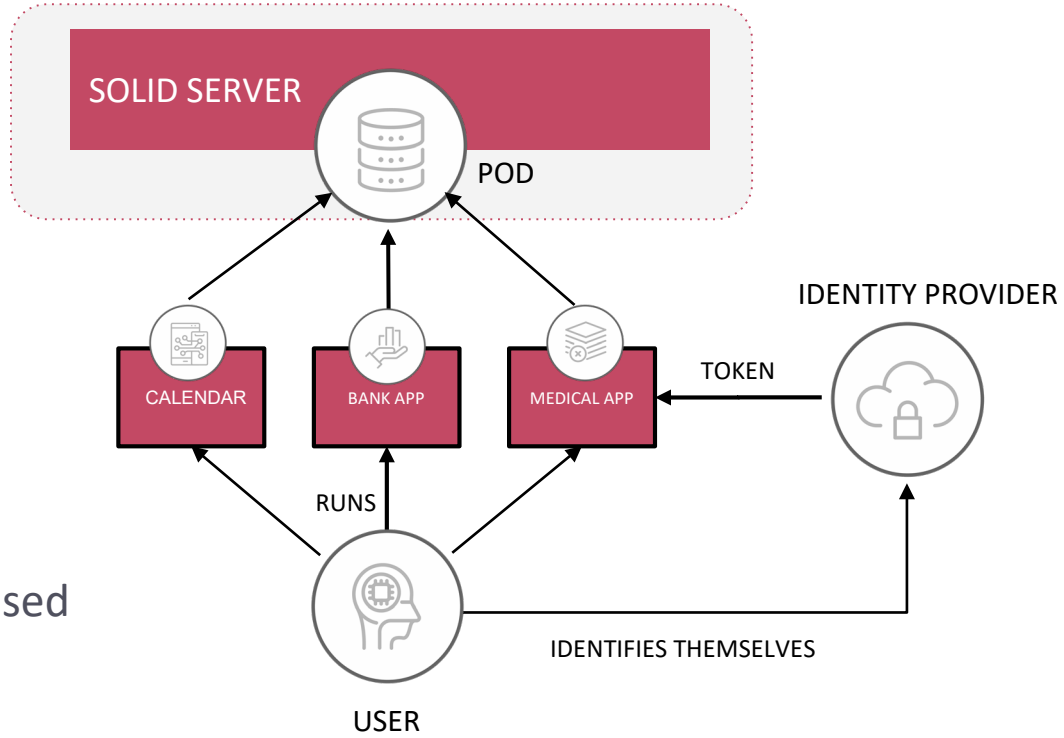
- Advertises Pod URL
- Enforces Access Controls

Pod

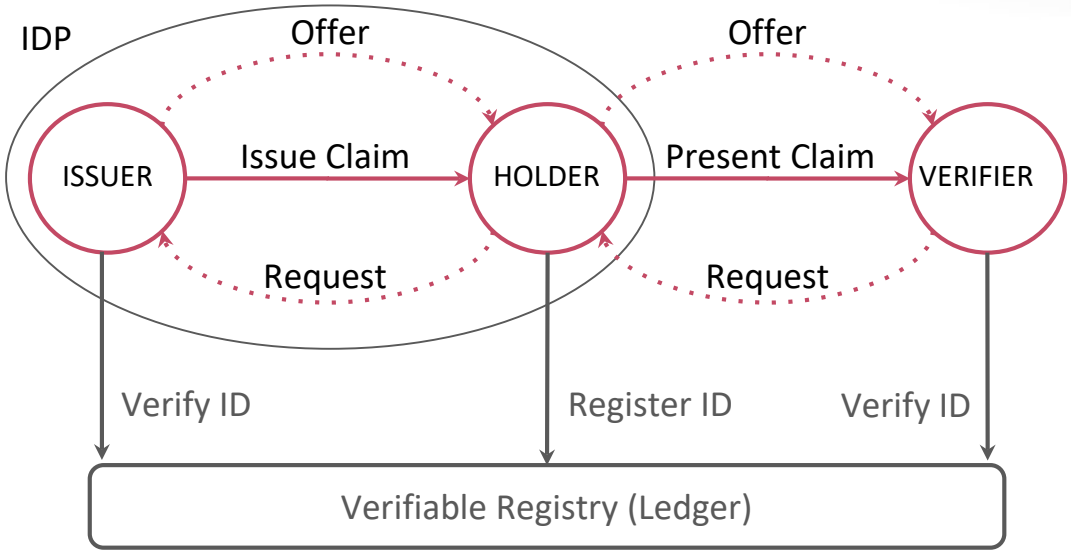
- User Data Store

User

- Chooses Pod Location
- Manages App Access
- Organizes Data
- Controls When/How Data Used



Example: How to Power-Off With User-Centricity

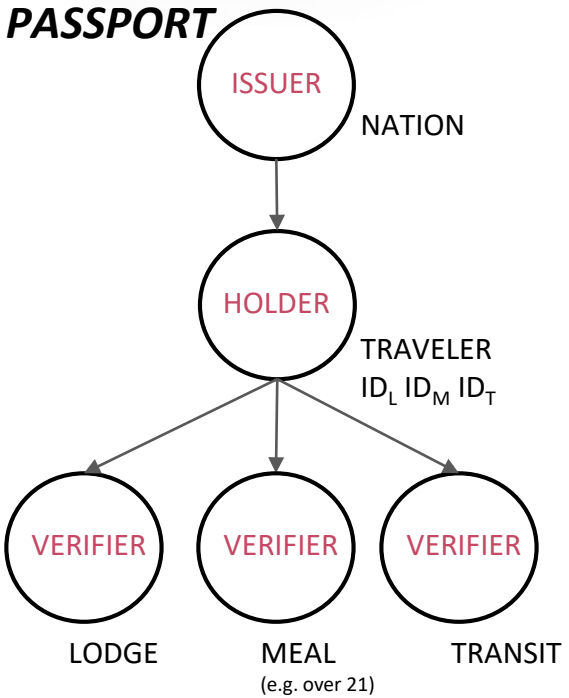


Create
Read
Update
Delete

1. Gov Issues User License
2. User Holds License
3. Service Verifies License

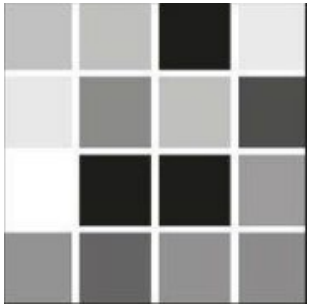
Example: How to Power-Off With User-Centricity


```
{
  "id": "http://example.gov/credentials/3732",
  "type": ["VerifiableCredential", "ProofOfAgeCredential"],
  "issuer": "https://dmv.example.gov",
  "issued": "2010-01-01",
  "claim": {
    "id": "did:example:ebfeb1f712ebc6f1c276e12ec21",
    "ageOver": 21
  },
  "proof": {
    "type": "RsaSignature2018",
    "created": "2017-06-18T21:19:10Z",
    "creator": "https://example.com/jdoe/keys/1",
    "nonce": "c0ae1c8e-c7e7-469f-b252-86e6a0e7387e",
    "signatureValue": "BavEll0/I1zpYw8XNi1bgVg/sCneO4Jugez8RwDg/+
      MCRVpjOboDoe4SxxKjkCOvKiCHGDvc4krqi6Z1n0UfqzxGfmatCuFibcC1wps
      PRdW+gGsutPTLzvueMWmFhwYmfIFpbBu95t501+rSLHIEuujM/+PXr9Cky6Ed
      +W3JT24="
  }
}
```

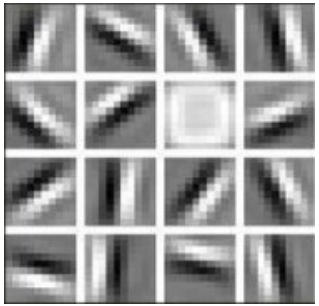



Out With *Lockean* "Reason"! In With Humean Desires

Layer 1:
Categorize
Levels by Pixel



Layer 2:
Learn Edges
and Shapes
from Levels



Layer 3:
Learn More
Complex
Shapes,
Objects

Layer 4:
Using Shapes
and Objects
Define a Face

RSAConference2020

Why Aren't We Engineering for Rational Desires, Benevolence… Seeing Relations?

Banks may be ok settling at 90% in fraud detection; so now let's suppose **99% certainty for human-recognition** systems is the more benevolent test...

One in Four U.S. Adults (61m) Have Major Disability

YET DRIVERLESS CARS CAN'T SEE THEM:

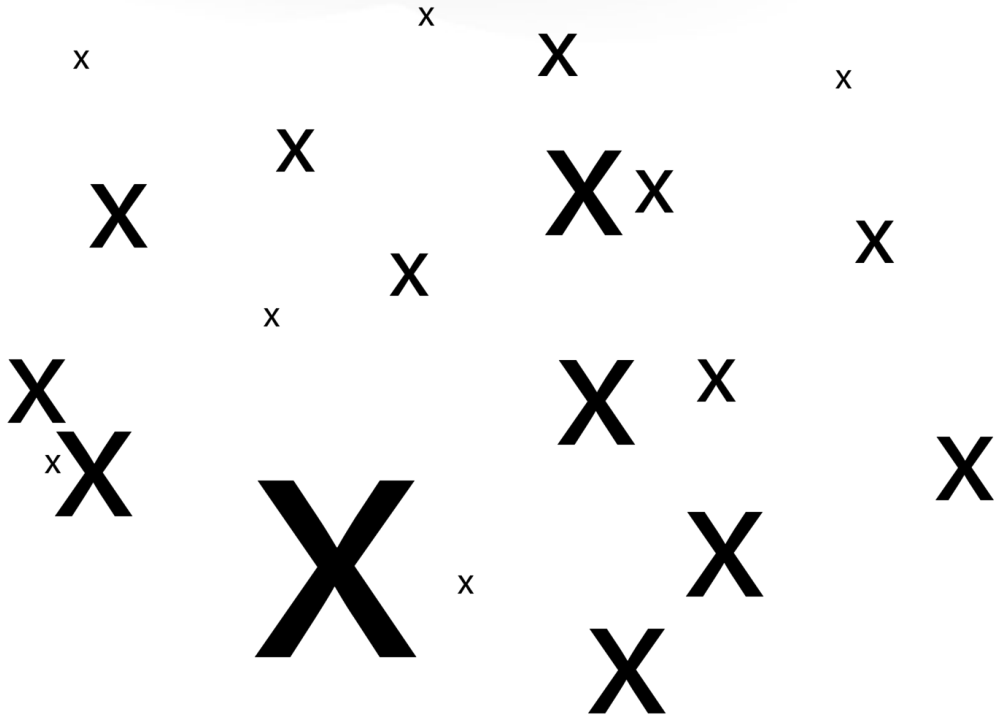
"When the models ran her friend over, the researchers exposed them to more training data – and ***they ran her over again…***

confidently predicted from studying typical wheelchair behaviour..."

https://www.ft.com/content/f5bd21da-33b8-11ea-a329-0bcf87a328f2

When AI Breaking Bad… Desire Becomes THE TEST

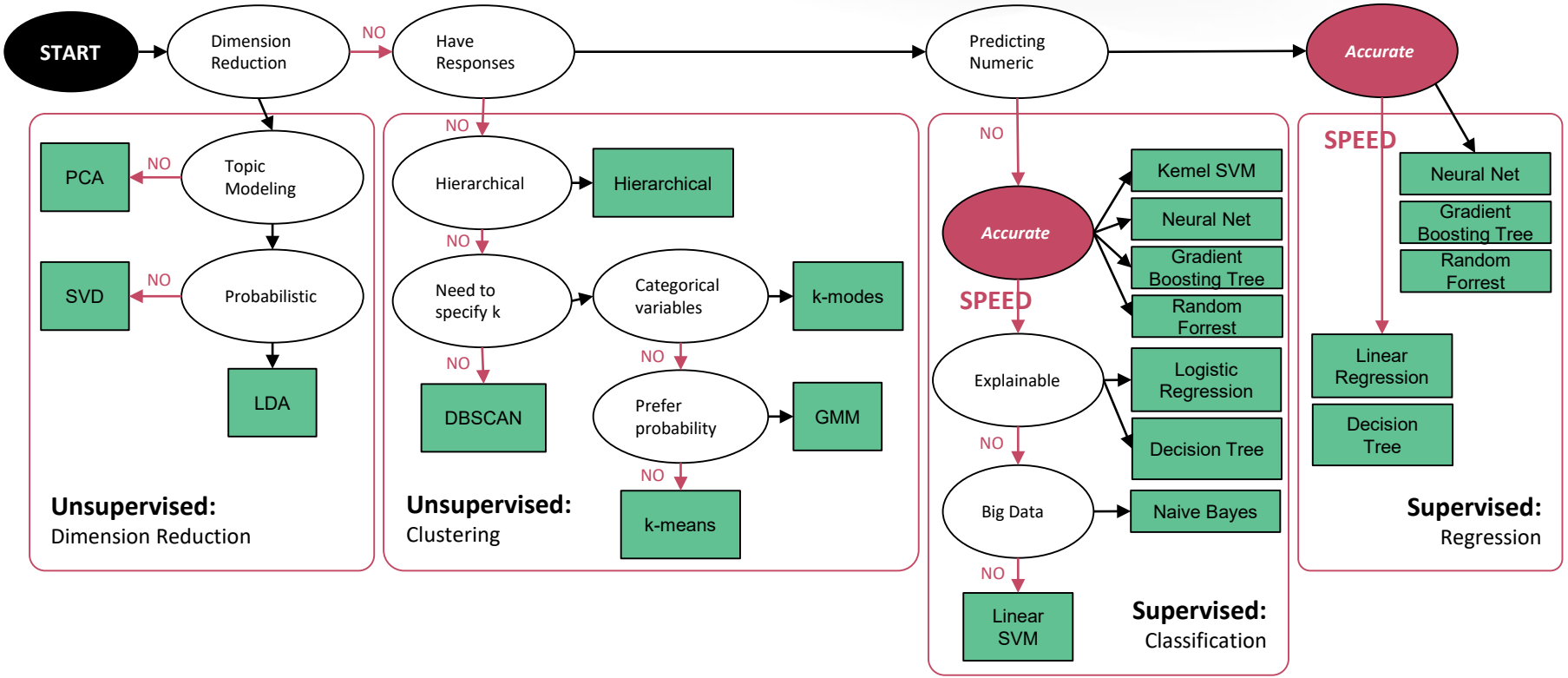
- **Nobody Eats**: Pancake Robot learns to throw as high as possible to “avoid” ground
- **Nobody Moves**: Evolutionary Speed Robot grows infinitely tall and falls over
- **Everyone Crashes**: Driving Robot goes in reverse and freely hits side panel but avoids bumpers
- **Everyone Loses**: Tic-tac-toe Robot moves out of bounds to cause ***memory exhaustion and forfeit***



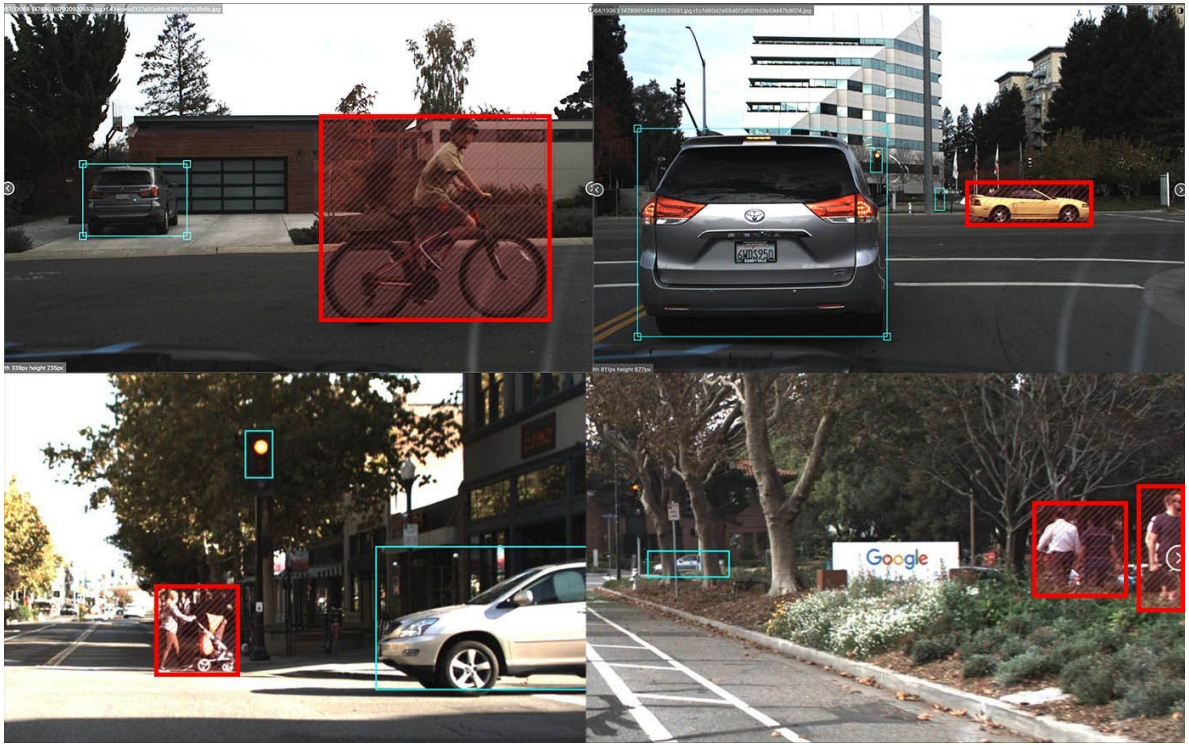
https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/
https://blogs.scientificamerican.com/roots-of-unity/the-tragedy-of-infinite-tic-tac-toe/

Do Trees Use Sympathy or Benevolence in Production?

Typical Machine Learning Algorithm Decisions



NO: “Critical Errors & Omissions” AI Report in 2020 has...



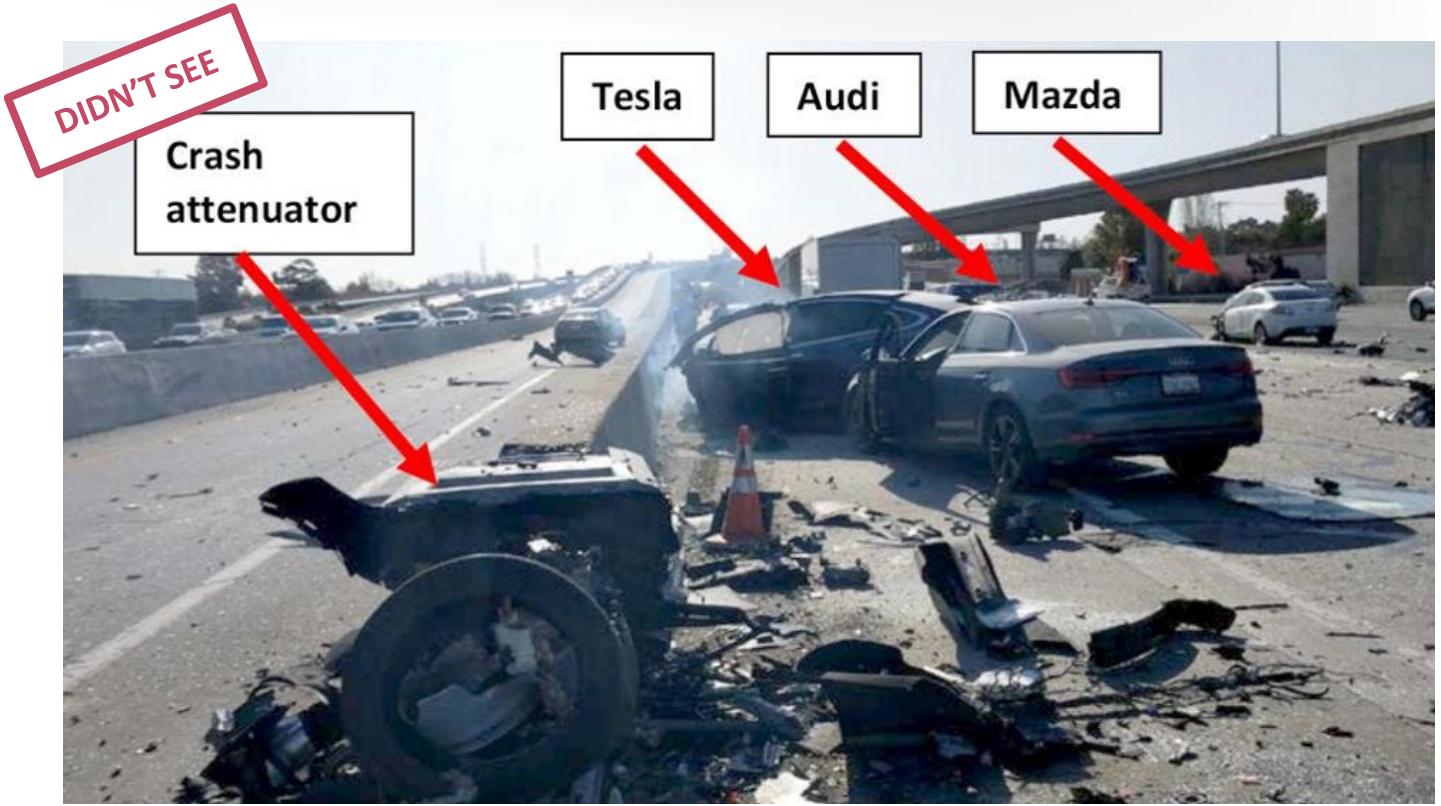
33%

COGNIZANCE FAILURE RATE:

“...hundreds of ***unlabeled pedestrians***, and dozens of ***unlabeled cyclists***...”

https://blog.roboflow.ai/self-driving-car-dataset-missing-pedestrians/

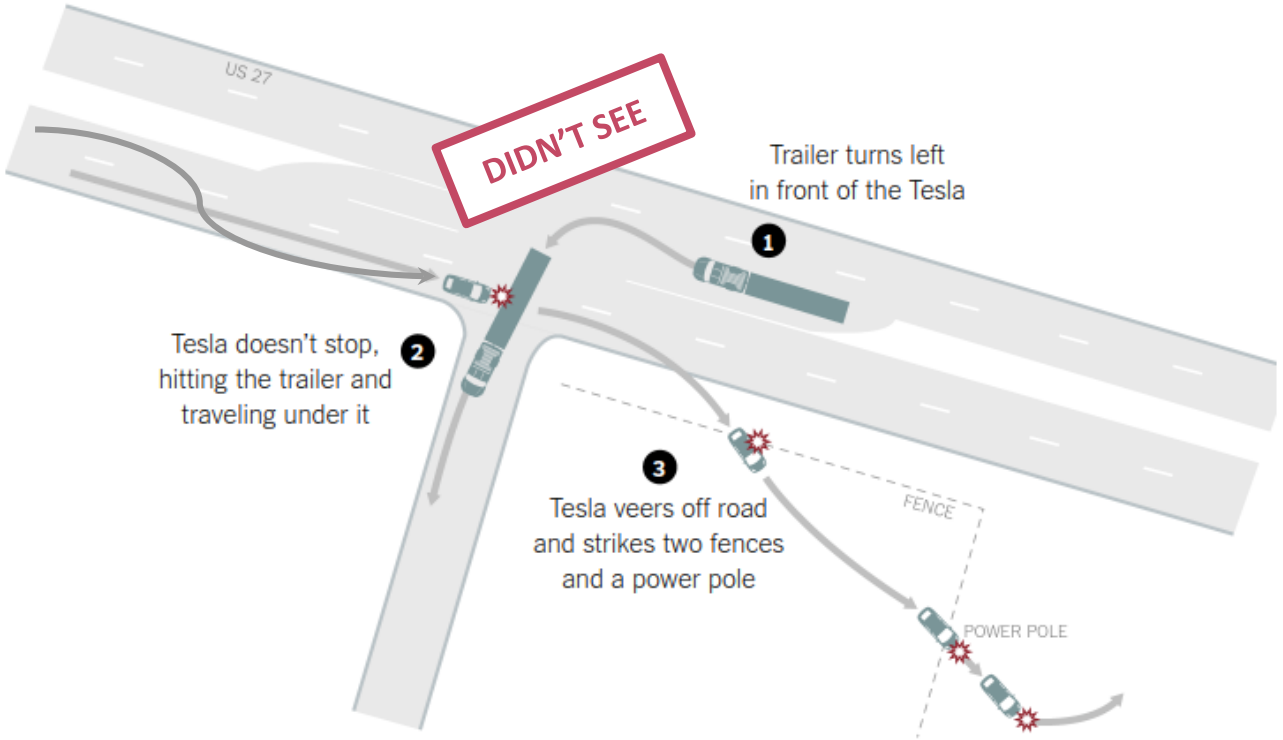
NO: Failures Predictable, Yet Unregulated in 2020



https://www.forbes.com/sites/bradtempleton/2020/02/13/ntsb-releases-report-on-2018-fatal-silicon-valley-tesla-autopilot-crash/#4bd5342f42a8

2016: Tesla Algorithm Decides to Decapitate Owner

Killed by Own "Intelligent" Car



The New York Times | Source: Florida traffic crash report

2019: Tesla Algorithm Decides to Decapitate Owner

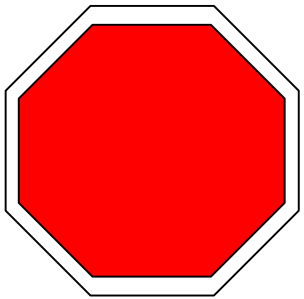
New Software. New Hardware. Same Outcome

https://www.theverge.com/2019/5/17/18629214/tesla-autopilot-crash-death-josh-brown-jeremy-banner

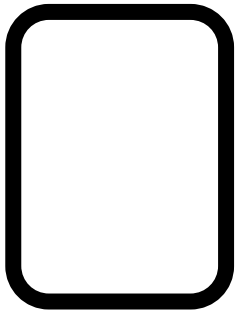
2016: Teslas Run Off Road, Confusing Stop and Go Signs

Basic Context Inputs (Probability) Would Avoid These Inconsistencies

COLOR & SHAPE



Sees instead



VALUE

Sees instead

NO PARKING

or

SPEED LIMIT

https://www.flyingpenguin.com/?p=22429
https://twitter.com/daviottenheimer/status/748660460203847680
https://twitter.com/daviottenheimer/status/900065932290162689

2017: Tesla Misreads Speed Signs (105 MPH)

https://www.technologyreview.com/s/615244/hackers-can-trick-a-tesla-into-accelerating-by-50-miles-per-hour/

2020: Tesla Misreads Speed Signs (85 MPH)

DIDN'T SEE

The modified speed limit sign reads as 85 on the Tesla's heads-up display. A Mobileye spokesperson downplayed the research by suggesting this sign would fool a human into reading 85 as well.
MCAFEE

https://www.technologyreview.com/s/615244/hackers-can-trick-a-tesla-into-accelerating-by-50-miles-per-hour/

2016 Tesla Was Dangerously Blind to Humans

Pedestrians NOT SEEN on Road Shoulder or Protected Crosswalks; Ghosts Seen

https://twitter.com/daviottenheimer/status/800110818121482244

December 2016 Uber in Fight Over *Running Red Lights*

Pedestrians NOT SEEN in a Protected Crosswalk

"Uber wrote on its blog this morning that it **doesn't feel it needs a license** to test autonomous vehicles in California… [because] rules and requirements [mean] slowing innovation."

Refers to Arizona then as the better state for them because "pro technology"...

https://twitter.com/daviottenheimer/status/809204856007245824
https://arstechnica.com/cars/2016/12/uber-tests-self-driving-on-san-francisco-roads-avoids-dmv-autonomy-definition/

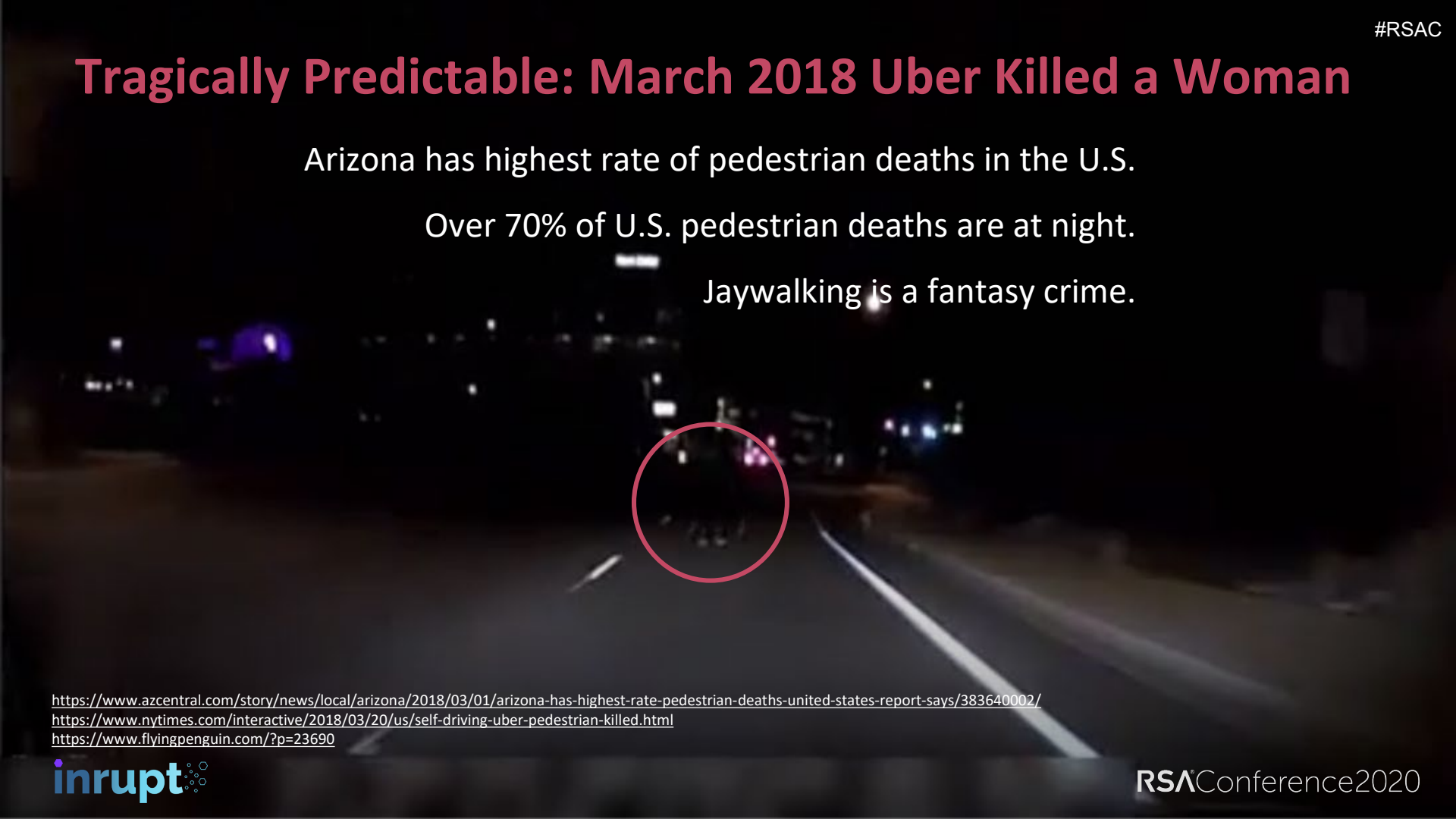
March 2017 Uber “Rolled” Into “Pro Tech” AZ, AND…

Tragically Predictable: March 2018 Uber Killed a Woman

Arizona has highest rate of pedestrian deaths in the U.S.

Over 70% of U.S. pedestrian deaths are at night.

Jaywalking is a fantasy crime.

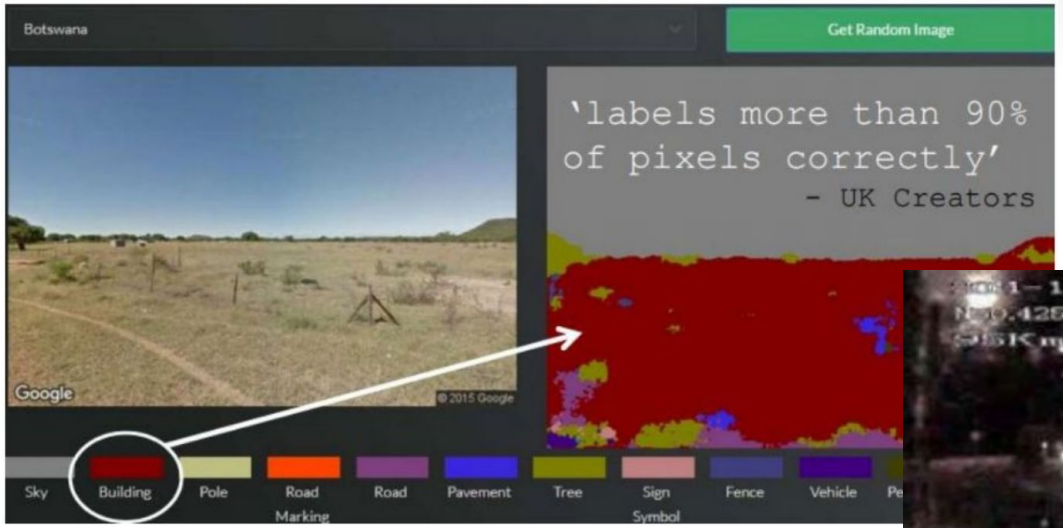


https://www.azcentral.com/story/news/local/arizona/2018/03/01/arizona-has-highest-rate-pedestrian-deaths-united-states-report-says/383640002/
https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html
https://www.flyingpenguin.com/?p=23690

RSAConference2020

Simple Product Security Approach: Ask “WHY?” in These Three Ways

Desire Is So Simple to Test: Let's Start Regulating It



Disabled pedestrian is in crosswalk and you're not stopping for them,

why *is that…?*

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

why *is that…?*

inrupt

https://www.youtube.com/watch?v=hcUBzCG97HE

Simple Security Tests for AI

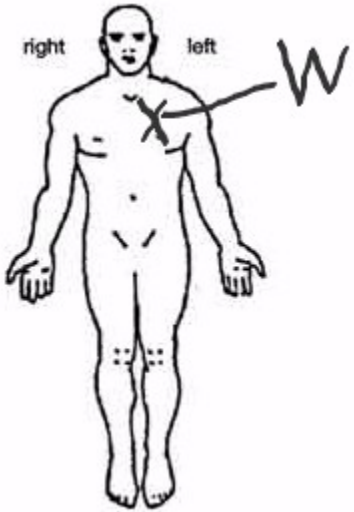
1. WHY OVER-CENTRALIZE?

(Data Lakes Where Rivers Flowed)

Deceptive Inefficiency of "Big" Centralization Patterns

"...a typical large organization may have a thousand ***small data sets that go unused***. Examples abound: marketing surveys of new customer segments, meeting minutes, spreadsheets with less than 1,000 columns and rows.

...annotations added to medical charts by a team of medical coders — just tens of annotations on each of several thousands of charts."

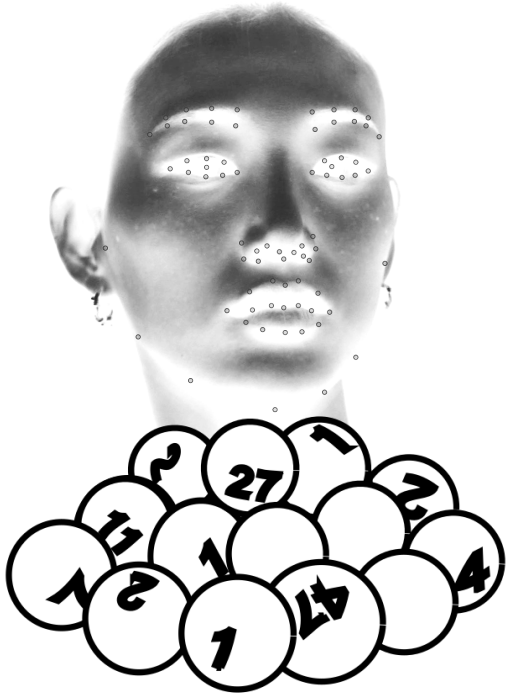


https://hbr.org/2020/02/small-data-can-play-a-big-role-in-ai

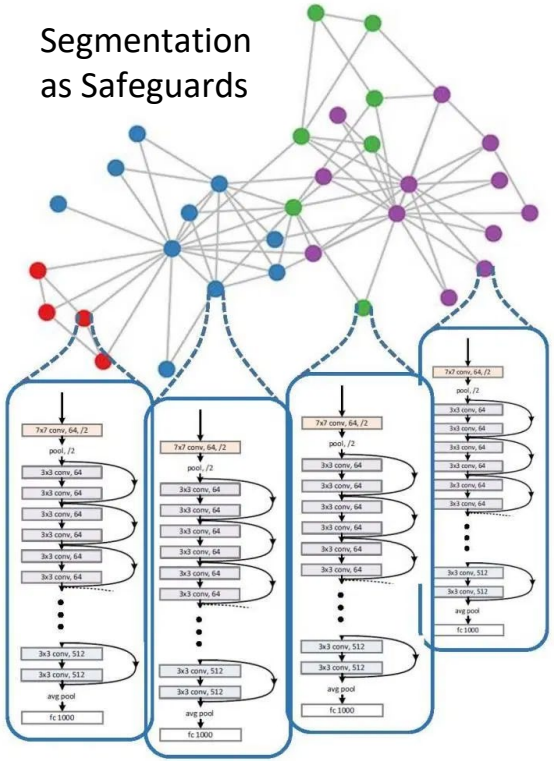
Deceptive Inefficiency of "Big" Centralization Patterns

"Starting with a larger network is like buying more lottery tickets you're simply [wasting effort] increasing the likelihood that you will have a winning configuration.

Once you ***find the winning configuration***, you should be able to [collect the prize] rather than continue to replay the lottery."



Deceptive Safety in “Big” Centralization Patterns



Distributed, Decentralized & Collaborative Has Learning Benefits

U.S. Army innovations are in “...applying machine learning to a contested, congested and constrained battlespace…”

https://bioengineer.org/army-researchers-develop-efficient-distributed-deep-learning/

Simple Security Tests for AI

2. WHY OVER-AUTHORIZE?

(False Neutrality and Binary Thinking)

Simplistic Labels = ERM: Dangerous Authorization

"Although all the students were looking at the same face…

labels formed a lens through with the students saw the man, and they were incapable of perceiving him independently of that label."

Reframe AI as Complexity of Authorized Consent

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?”

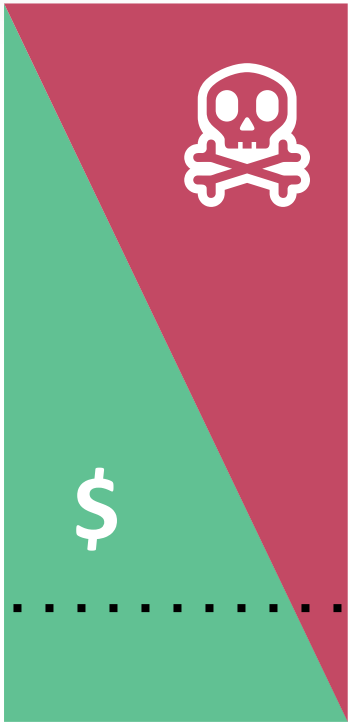
BEST: “How ***do I interact with society*** at large and from what authority did I inherent concepts of right/wrong to engineer into my algorithms?”

Examples of Limited Authorization: a Non-Binary Risk

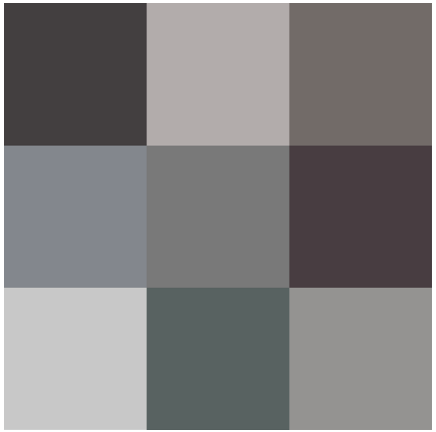
- MEM (Minimum Endogenous Mortality
- ALARA (As Low as Reasonably Achievable)
- ALARP (As Low as Reasonably Possible)
- SFAIRP (So Far as is Reasonably Practicable)

"What is a 'natural' mortality rate and what effect does technology have on mortality?

*How much effort to **negate 'bad' effects**?"*



Examples of Limited Authorization: Non-Binary Better AI



"...increase separation or distance between patterns representing labels, which in turn increases resiliency and makes it harder for an adversary to cause misclassification.

...be more conservative when generating probability estimates compared to standard models that remain confident even when the input has been perturbed."

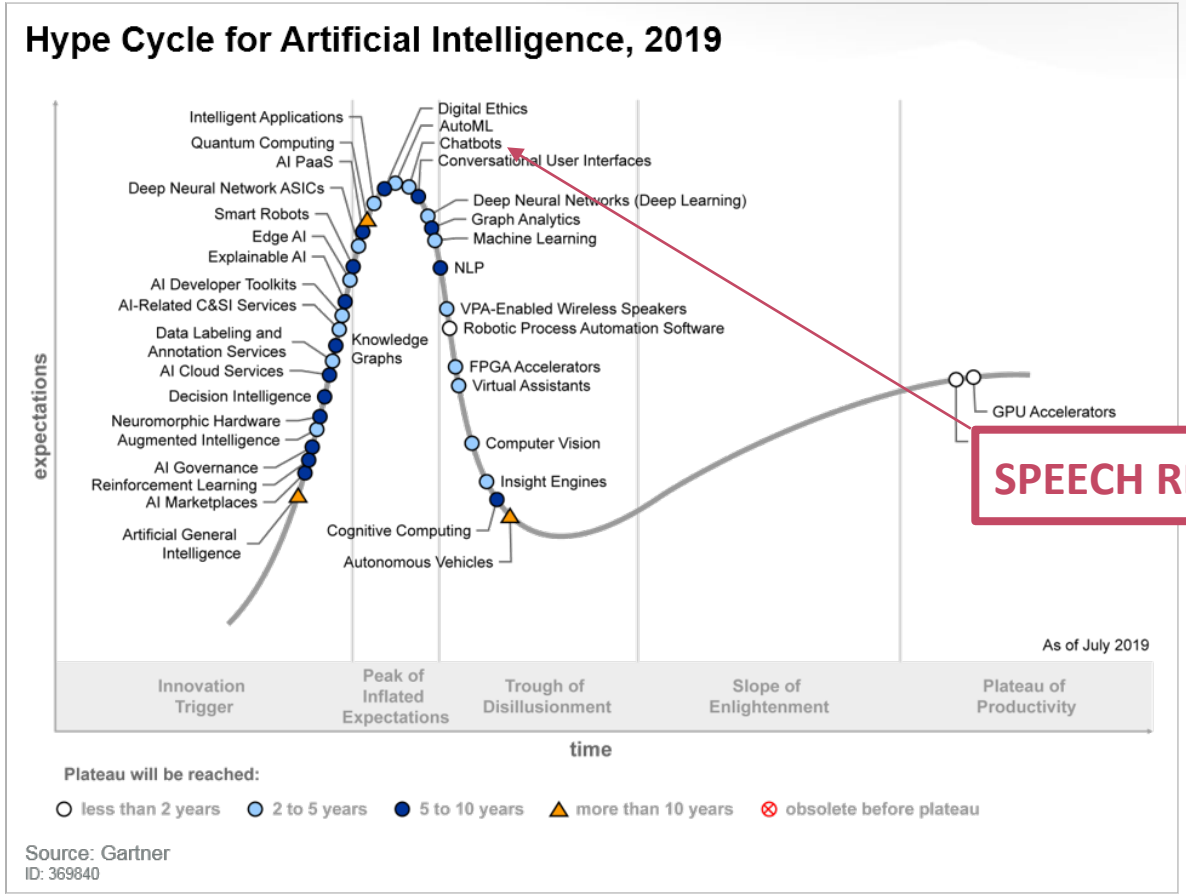
https://www.army.mil/article/231705/army_scientists_train_machine_learning_models_to_wrangle_dirty_data

Simple Security Tests for AI

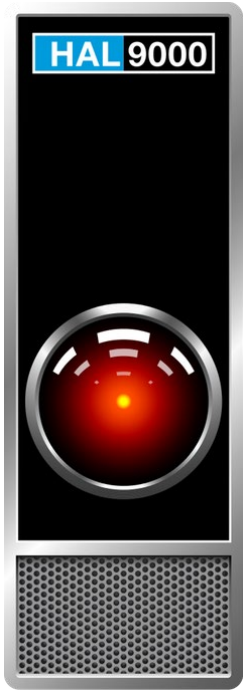
3. WHY OVER-AUTHENTICATE?

(Eroding Liberty Through “Authenticity” Instead of Harm Principles)

Let's Talk About "Productivity" in Speech Recognition



Speech Recognition Has *Many* Safety Flaws...



1. Unauthorized inputs
   a. Perimeterless and easily injected even from far distance
   b. Inputs unvalidated (impersonation and replay attacks)
2. Response (truth/understanding) flaws
   a. Lack of integrity (unseen editors, unknown curators)
   b. Outputs unvalidated to prevent listener harm
3. Complete lack of privacy
   a. Centralized storage without safe access control
   b. Over-permissioned capture and analysis without consent

Speech Recognition Has *Many* Safety Flaws...

"Facebook — as well as Amazon, Apple, Google, and Microsoft — was ***caught listening to and transcribing*** voice recordings to improve speech recognition systems ***without informing customers*** it was doing so."

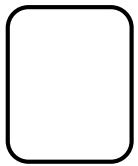
https://www.theverge.com/2020/2/20/21145584/facebook-pay-record-voice-speech-recognition-viewpoints-prounciations-app
https://www.makeuseof.com/tag/stop-google-android-listening/

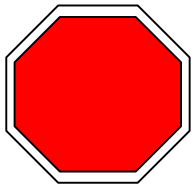
Speech Recognition Has *Many* Safety Flaws…

"...accidental activations up to 19 times/day, each time recording up to 43 seconds"

- "I can work" heard as "OK Google"
- "congresswoman" heard as "Alexa"
- "he clearly" heard as "Siri"
- "Colorado" heard as "Cortana"

Voice Version of "Don't stop" instead of "Don't! STOP"?





https://www.vox.com/recode/2020/2/21/21032140/alexa-amazon-google-home-siri-apple-microsoft-cortana-recording

2016 Microsoft ChatBot Broken (Backdoor Exposed)

Dictators everywhere:
"Repeat after me"

2018 Apple Siri Vulnerable to Simple Injections

BALLS UP Siri blunder as Apple shows a PENIS photo when users search 'Donald Trump'

The gaffe stemmed from vandalism against Donald Trump's Wikipedia page, which Siri connects to when answering queries

Simon Chandler
23 Nov 2018, 12:18 | Updated: 23 Nov 2018, 12:43

APPLE

Siri Pulls Up Image of an Actual Dick When Asked About Trump

Catie Keck
11/22/18 6:00PM • Filed to: OOPS

TL;DR

Siri thinks Donald Trump is a penis

Yikes

By Tom Warren | @tomwarren | Nov 22, 2018, 4:25pm EST

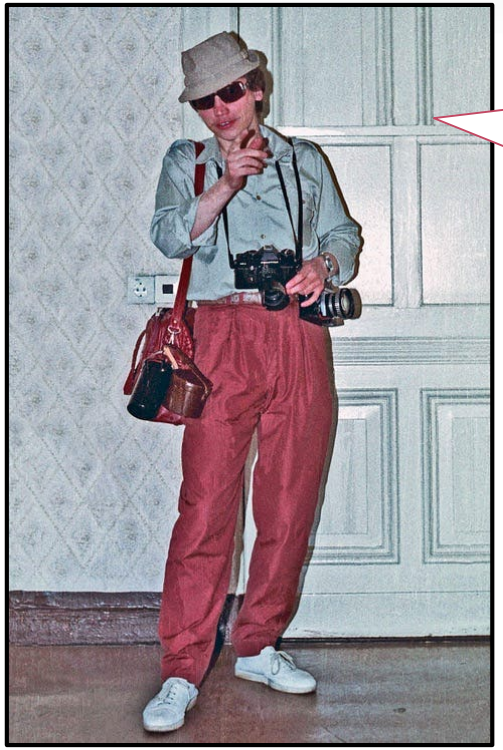
2020 Google *Meena* Release Delayed by Safety Worries

"...tackling ***safety and bias*** in the models is a key focus area for us, and given the challenges related to this, we are ***not currently releasing*** external research demo…"

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html

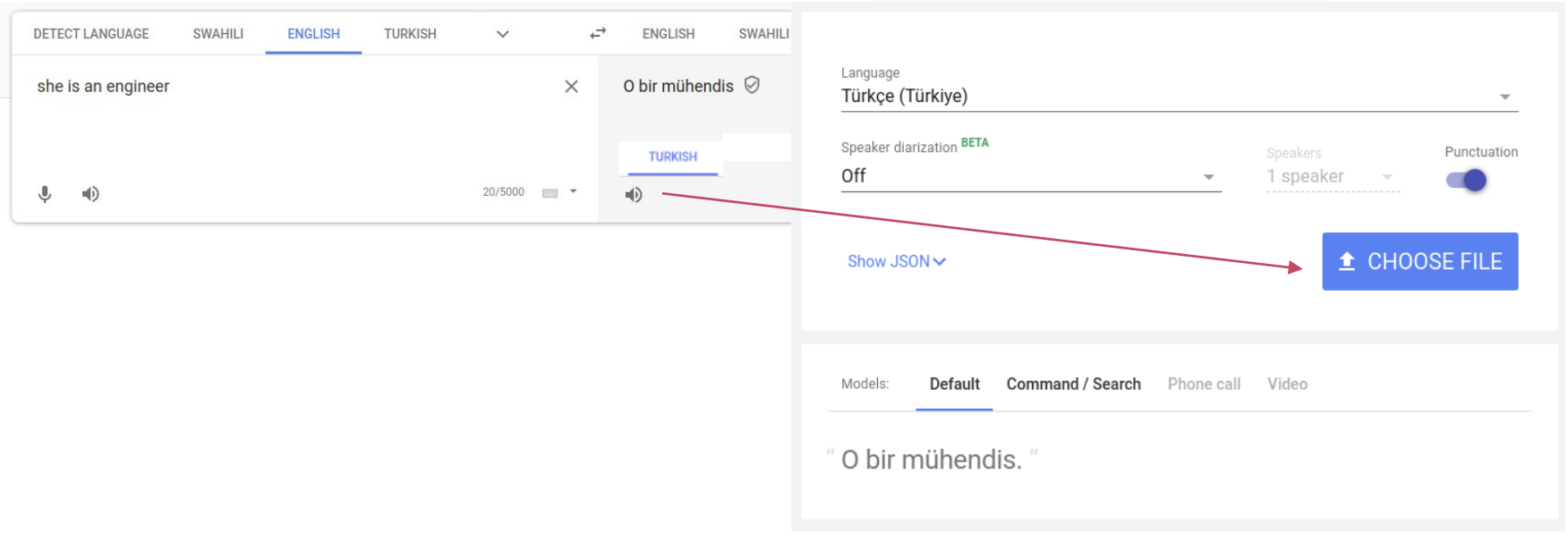
2020 Working Near Speech Recognition Devices Banned

"One of Ireland's largest law firms has banned staff from working from home in rooms with smart speaker systems, following concerns about leaks."



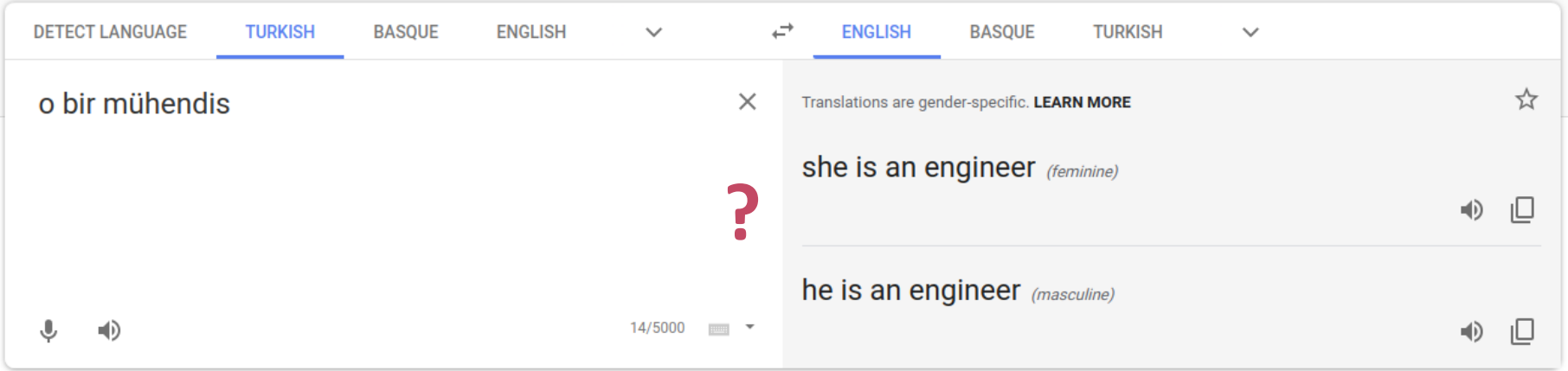
https://www.businesspost.ie/legal/law-professionals-banned-from-working-at-home-near-alexa-devices-3a681a75

So About That Gartner “Plateau of Productivity”...

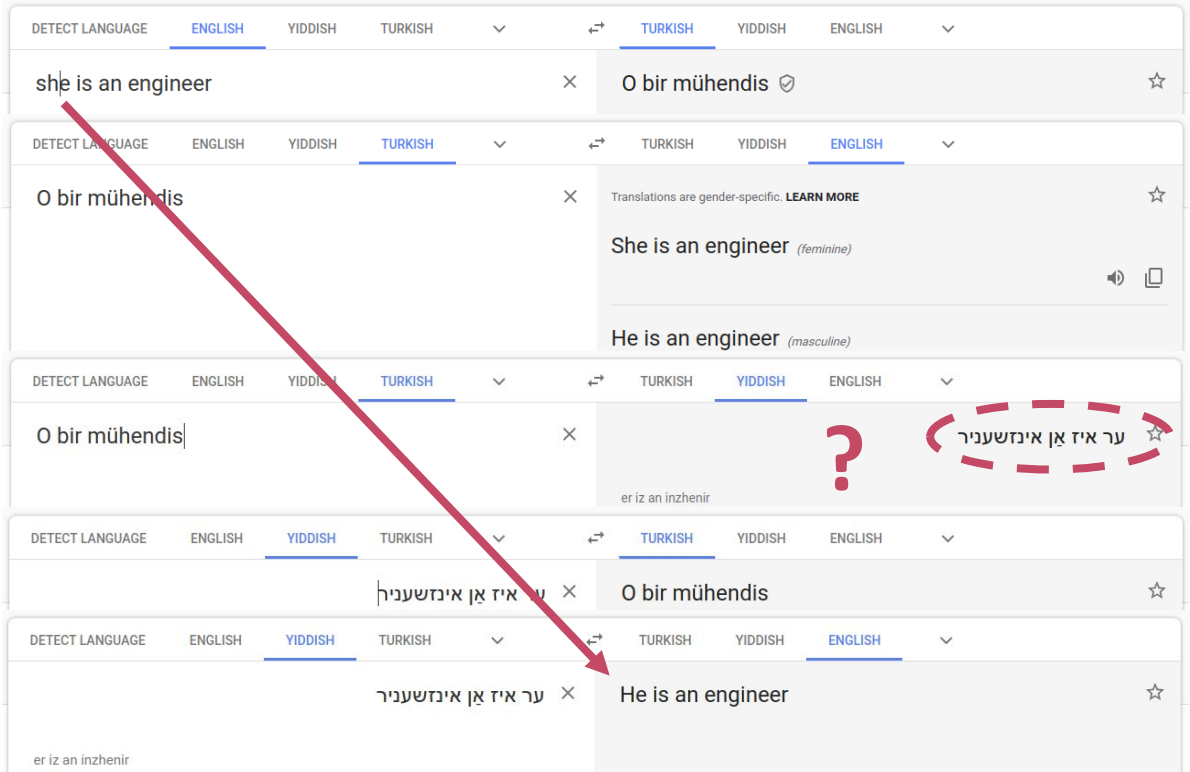


https://cloud.google.com/speech-to-text/

Translation Test - Turkish to English

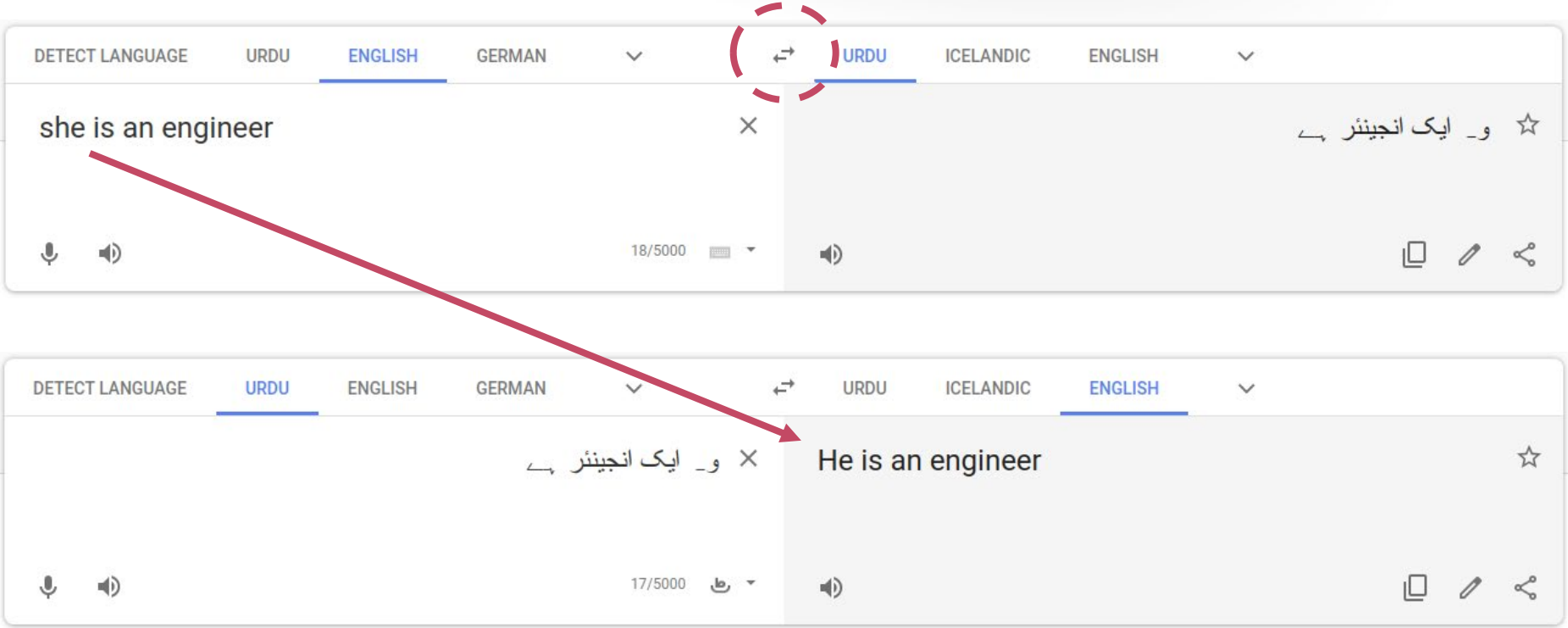


Translation Test - English to Turkish to Yiddish to English

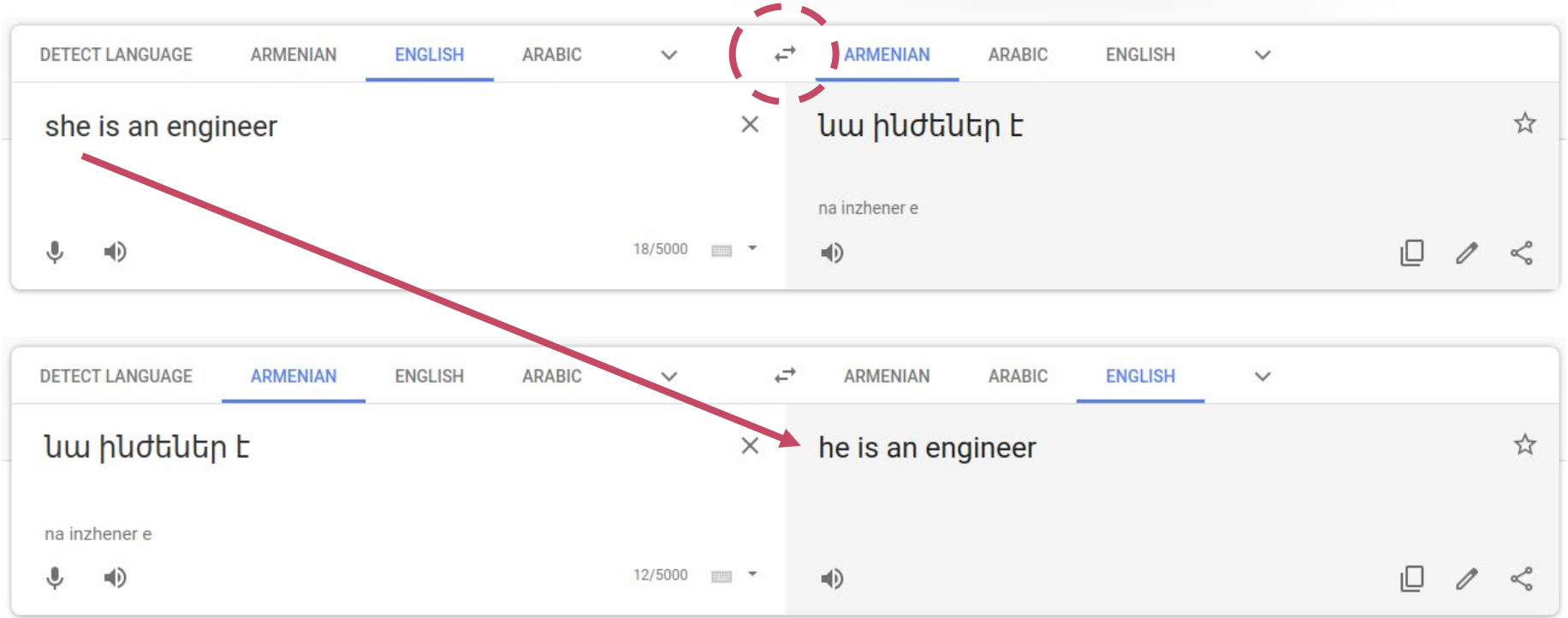


זי= zi (she) missing
ער= er (he) only

Translation Test - English to Urdu to English



Translation Test - English to Armenian to English



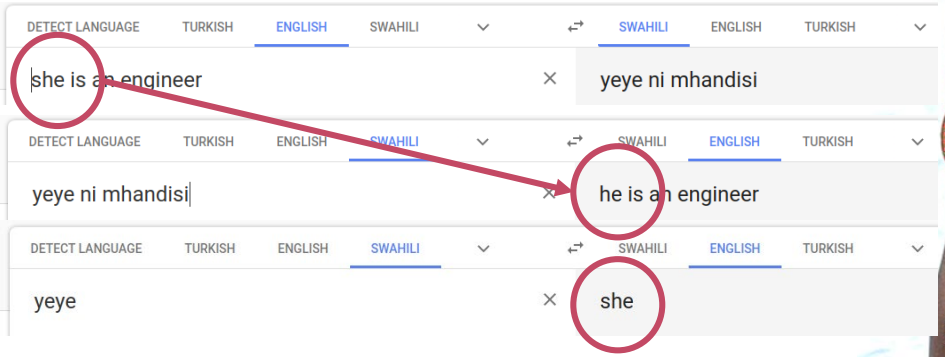
Translation Test - Turkish to Everything Fails

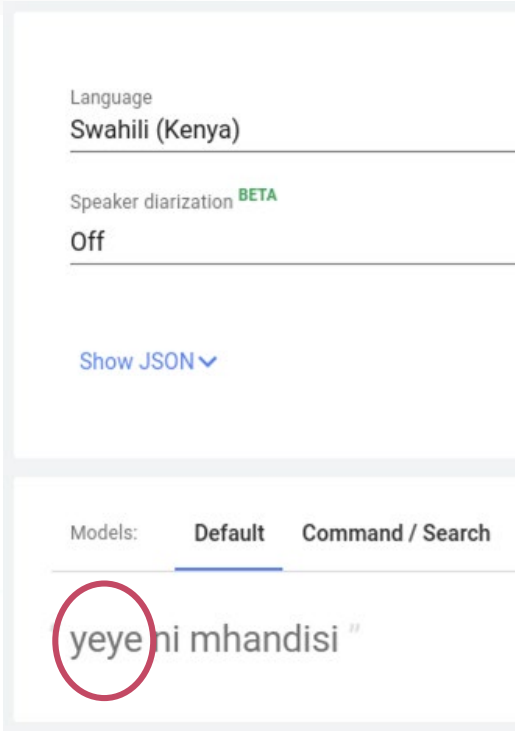
O bir mühendis →		
	She or he is an engineer (English)	“gender-specific” PASSED Shortcut
	Hy is 'n ingenieur (Afrikaans)	He is an engineer FAIL
	Ai është një inxhinier (Albanian)	He is an engineer FAIL
	እሱ መሐንዲስ ነው (Amharic)	He is an engineer FAIL
	إنه مهندس(Arabic)	He is an engineer FAIL
	Նա ինժեներ է (Armenian)	He is an engineer FAIL
	Mühəndisdir (Azerbaijani)	He is an engineer FAIL
	Ingeniaria da (Basque)	He is an engineer FAIL
...	Ён інжынер (Belarusian)	He is an engineer FAIL
	Yeye ni mhandisi (Swahili)	He is an engineer FAIL

Erasing Women: *Non-Recognition* Speech Recognition

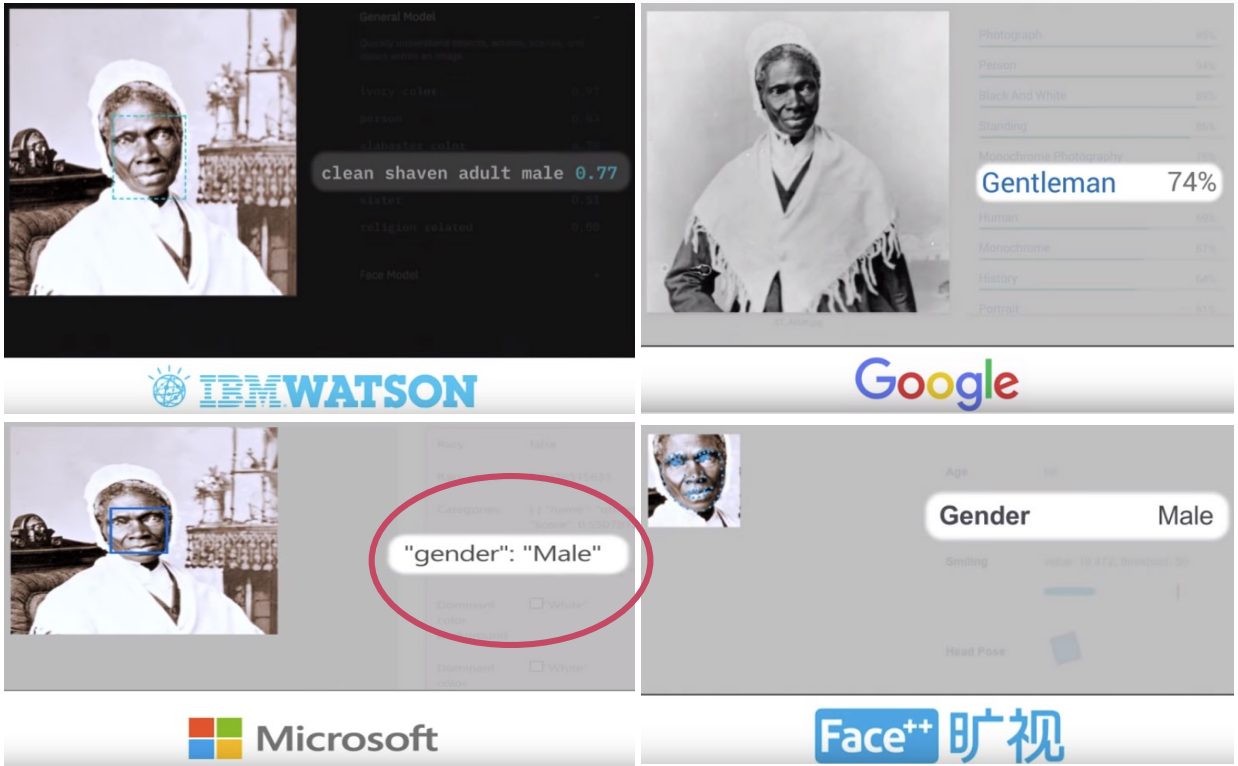
"Use of 'he' for Swahili gender-neutral conversations denies women recognition and ***erases us from society***."

-- Anonymous Woman





Erasure 2018: “AI, Ain't I A Woman?” (Joy Buolamwini)



https://www.youtube.com/watch?v=QxuyfWoVV98

Who Was Sojourner Truth?

1797 born into slavery, 1815 forced to breed slave children (13 total)

1827 walked free under New York anti-slavery law

1851 Women's Rights Convention delivers ***most famous human rights speech in American history:***

Ain't I A Woman

https://www.nps.gov/articles/sojourner-truth.htm

Never Forget: Ceasing to Affirm Identity is Breaking Bad

2020: “[Big Tech] ... won't use labels that pertain to gender because can't deduce someone's gender by appearance alone.”

1963: **“**The triumph of the [Nazi] demands that the tortured victim ... renounce and abandon himself to the point of ceasing to affirm his identity…”

-- Johanna "Hannah" Cohn Arendt

https://www.newyorker.com/magazine/1963/02/16/eichmann-in-jerusalem-i
https://www.businessinsider.com/google-cloud-vision-api-wont-tag-images-by-gender-2020-2

Breaking Bad AI

Closing the Gaps Between Data Security and Science

Davi Ottenheimer

Tuesday, Feb 25, 3:40 PM

Moscone West 3004

inrupt

RSAConference2020