AIs Wide Open

Making Bots Safer Than Completely #$%cking Unsafe

BSIDES LAS VEGAS

August 6, 2019

1430-1455 / 1600-1755

flyingpenguin

Abstract

Bladerunner was supposed to be science fiction. And yet here we are today with bots running loose beyond their intended expiration and with companies trying to hire security people to terminate them.

We have several well-documented cases of software flaws in automation systems causing human fatalities. Emergent human safety risks are no joke and we fast are approaching an industry where bots are capable of pivoting and transforming to perpetuate themselves (availability) with little to no accountability when it comes to human aspirations of being not killed (let alone confidentiality and integrity).

Perhaps you are interested in building a framework to keep bot development pointed in the right direction (creating benefits) and making AI less prone to being a hazard to everyone around?

Welcome to 2019 where we are tempted to reply “you got the wrong guy, pal” to an unexpected tap on the shoulder …before we end up on some random roof in a rainstorm with a robot trying to kill us all

whoami

2005 Campaigned PKI/Token architecture to reduce PCI breaches

2006 Patented One-time IoT PIN/secret (“Connected Life” Paranoid)

2009 Wrote EKMI → KMIP open standards for key management services

2012 Published Securing Virtual Environment (Cloud) Book

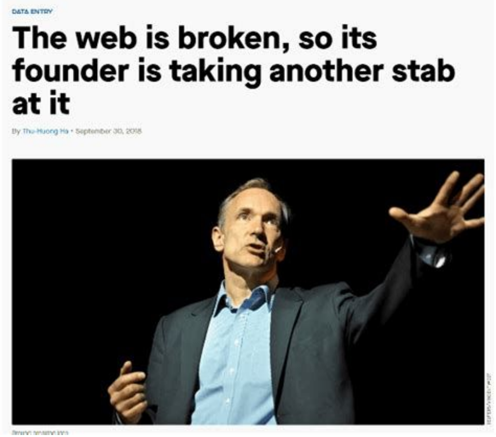
2013 Started Realities of Securing Big Data (Field-Level Crypto) Book...

2017 Started NoSQL Field-Level Crypto DB Product Feature...

2018 Created RSAC Humanitarian Service Award (1980s crypto system)

2019 Released NoSQL Field-Level Encryption Client-Side (FLECS) !!!

inrupt



DATA ENTRY

The web is broken, so its founder is taking another stab at it

By Thu-Huong Ha • September 30, 2018

flyingpenguin

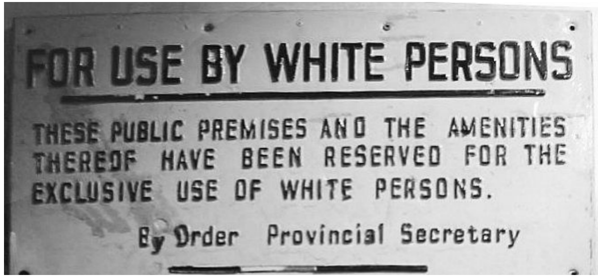
BSIDES LAS VEGAS

whoami ***NOT NOT NOT***

“One of the things that’s been pointed out to me and I think it’s very very true is … if you’re a ***17-year-old guy*** and you’re looking for like the group that you can join where you can make terrible jokes, ***where you can enjoy yourself and do all the stuff, you end up going with white nationalism because that’s the side that lets you do it***. If you join a left wing group, you’re going to get sort of tone-policed, you’re gonna have people saying, ‘Ooh, you know, you gotta be sensitive about this and that.’ So ***the left doesn’t have fun*** and I think that makes it much harder to attract people who are at an age where they’re very very susceptible and they’re open-minded about what they’re going to take. And if you’re the side that says join us, you know, we do all the cool stuff, that’s a compelling message.”

-- Infamous South African Cyberarms Broker

Describing ***white nationalist policing*** as fun and “lets you do it” is a dangerously immoral power authorization that denies the authentication failures of “mirror-tocracy”



FOR USE BY WHITE PERSONS

THESE PUBLIC PREMISES AND THE AMENITIES THEREOF HAVE BEEN RESERVED FOR THE EXCLUSIVE USE OF WHITE PERSONS.

By Order Provincial Secretary

Source: https://www.defenseone.com/ideas/2019/07/ep-48-cyberwarfare-today/158387/

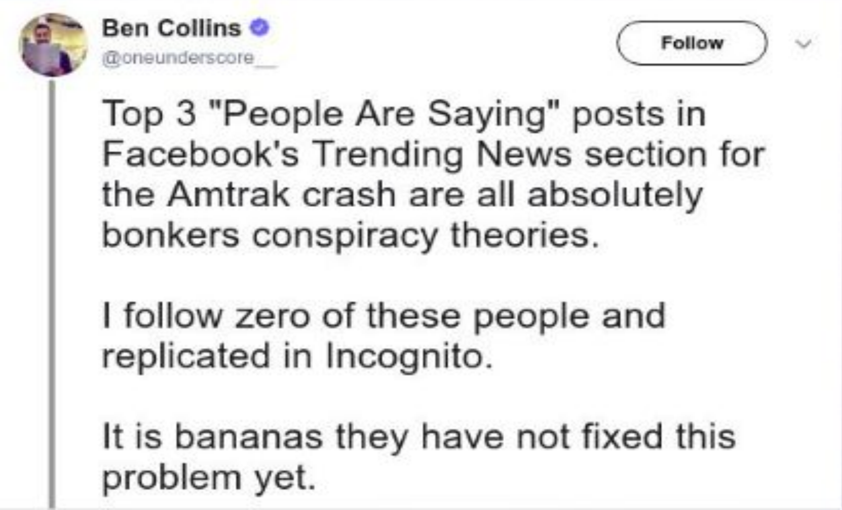
whowasi… “Mission 101” MSc

(Small Fuse to Ignite
Liberation from Fascism)

flyingpenguin

Source: https://www.flyingpenguin.com/?p=22521

11:39 AM - 31 Jan 2018



Ben Collins
@oneunderscore__
Follow

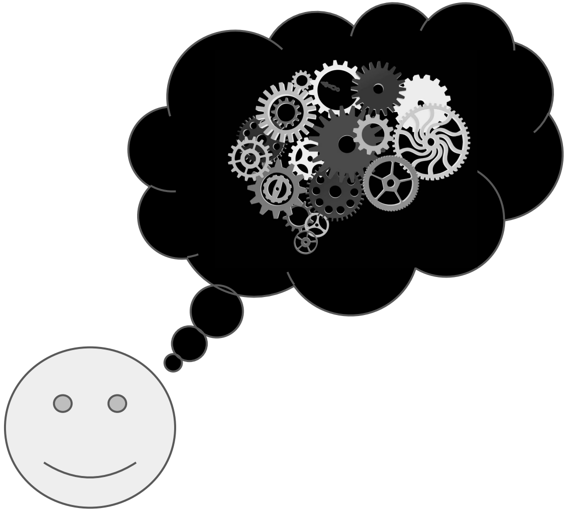
Top 3 "People Are Saying" posts in Facebook's Trending News section for the Amtrak crash are all absolutely bonkers conspiracy theories.

I follow zero of these people and replicated in Incognito.

It is bananas they have not fixed this problem yet.

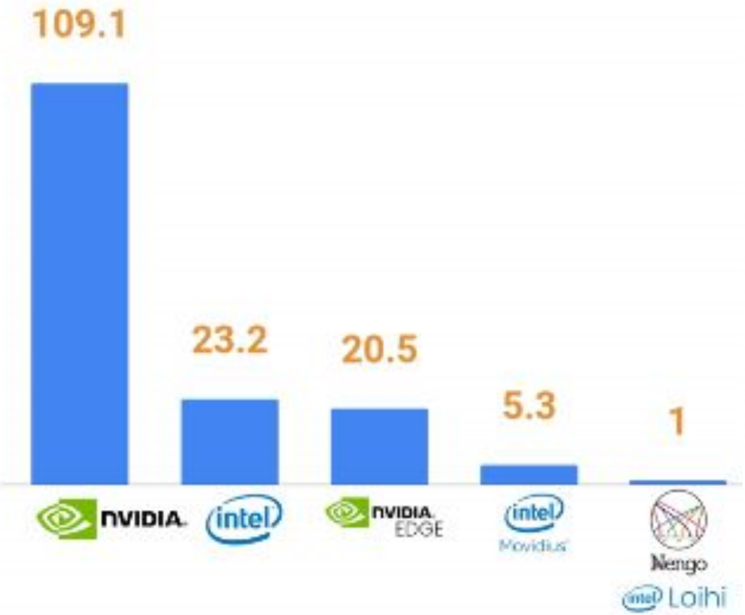
https://twitter.com/daviottenheimer/status/817027735469989888

The Great Promises of AI

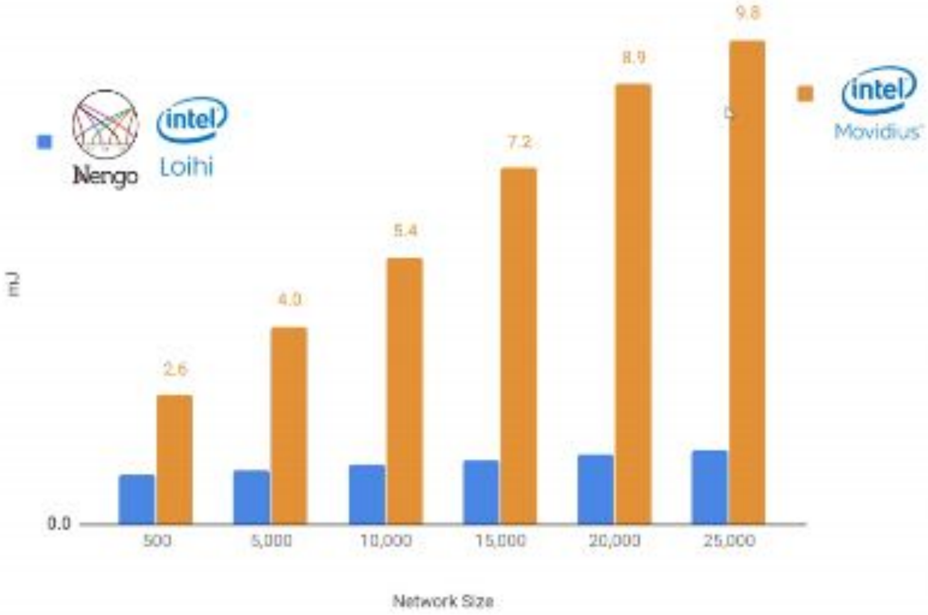


- Drive Cars
- Translate Language
- Pilot Aircraft
- Detect Fraud
- Detect Malware
- Solve Healthcare
- And So Much More…!

Latest Neuromorphic Efficiencies = *Silicon Brains*

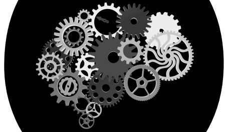


Relative Deep Neural Network Power Cost



Relative Power Cost Per Inference

SILICON BRAINS



(“Technology and Social Processes to Support Value-Based System Innovation”)

Reality:
Science Has Achieved Riderless Bicycles

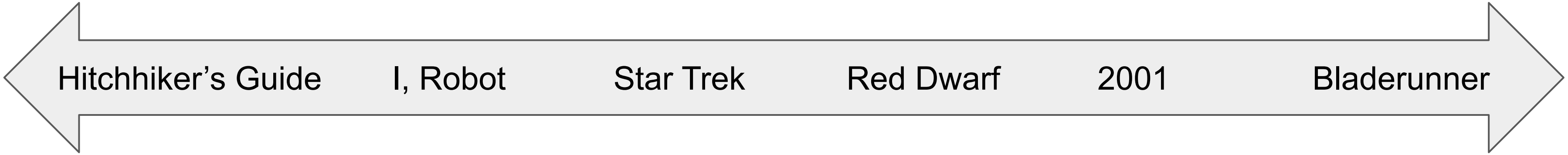
(Gives New Meaning to *Cog*nitive Science)

Will They Turn on Their Creators?

https://www.scmp.com/tech/big-tech/article/3021038/chinese-researchers-develop-hybrid-chip-design-holds-promise-thinking

Spectrum of Expectations for Robotic Future Worlds

THE FOLLOWING TOOK 75 THOUSAND GENERATIONS TO CALCULATE
42
IT HAS BEEN CHECKED VERY THOROUGHLY
A Fawcett Crest Book
T1453 75c
Fascinating Tales from Beyond Tomorrow by the Master of Science Fiction
ISAAC ASIMOV
I, ROBOT



Hitchhiker’s Guide — I, Robot — Star Trek — Red Dwarf — 2001 — Bladerunner

Anyone Important Missing?

Quick Poll

? Percent who say unregulated AI could lead to ***human extinction***

? Percent who say rapid evolution of machine learning will ***harm society*** (e.g. murder your brother and your bride)

2019 Public Opinion Report

Center for Governance of AI and Oxford University Future of Humanity Institute

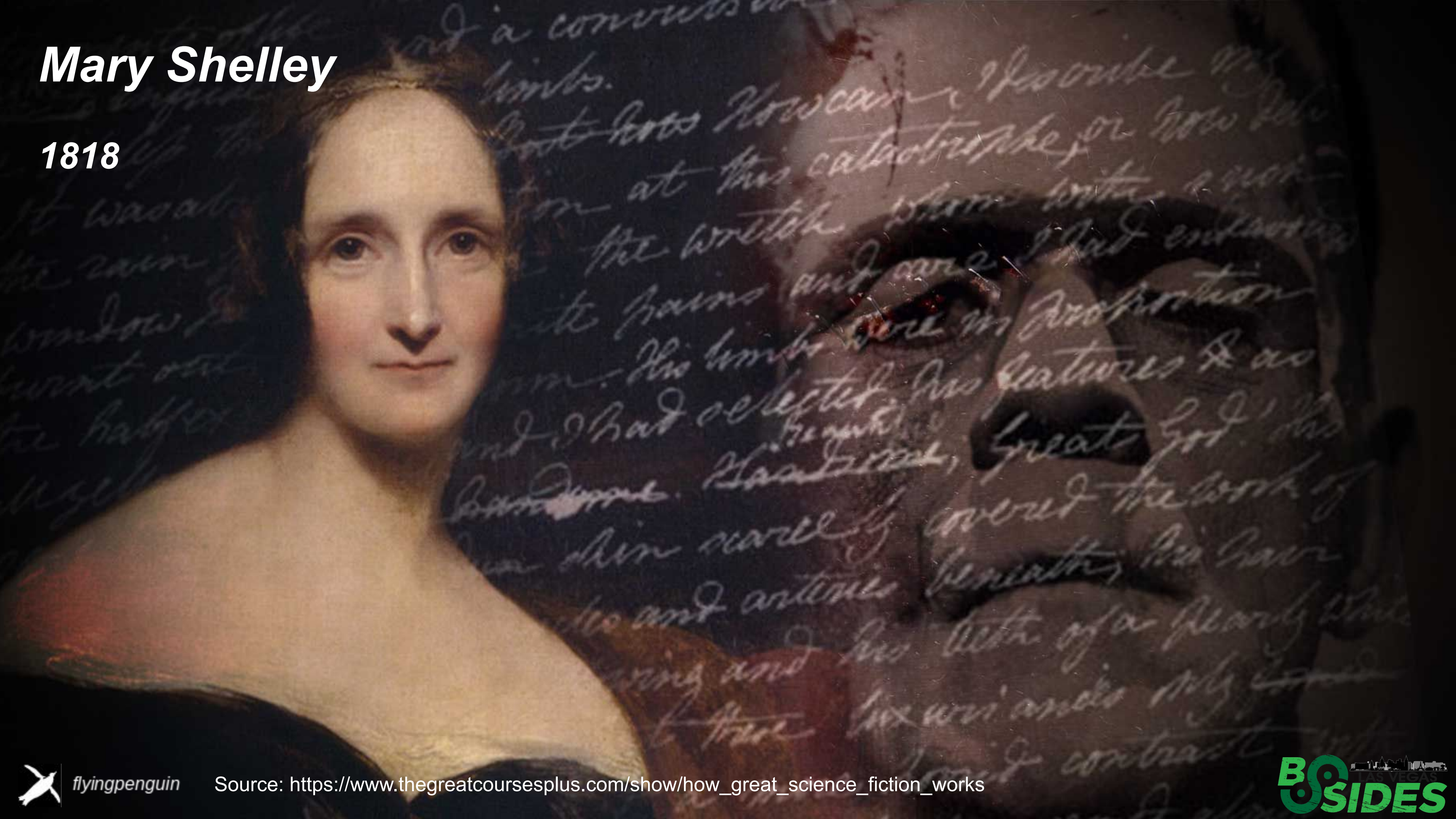
12 Percent of American consumers who say unregulated AI could lead to human extinction

34 Percent of American consumers who say rapid evolution of machine learning will harm society

Source: https://governanceai.github.io/US-Public-Opinion-Report-Jan-2019/us_public_opinion_report_jan_2019.pdf
(https://www.fhi.ox.ac.uk/GovAI/)

Mary Shelley

1818



Source: https://www.thegreatcoursesplus.com/show/how_great_science_fiction_works

Inventor of SciFi: The Original Robot of Doom*

How regulate destructive nature of tech when aligned with power (wealth).

How should we transition to advances of science without tragedy/horror?

(*Fun fact: “Dystopia” was coined 50 years later in 1868 by John Stuart Mill)

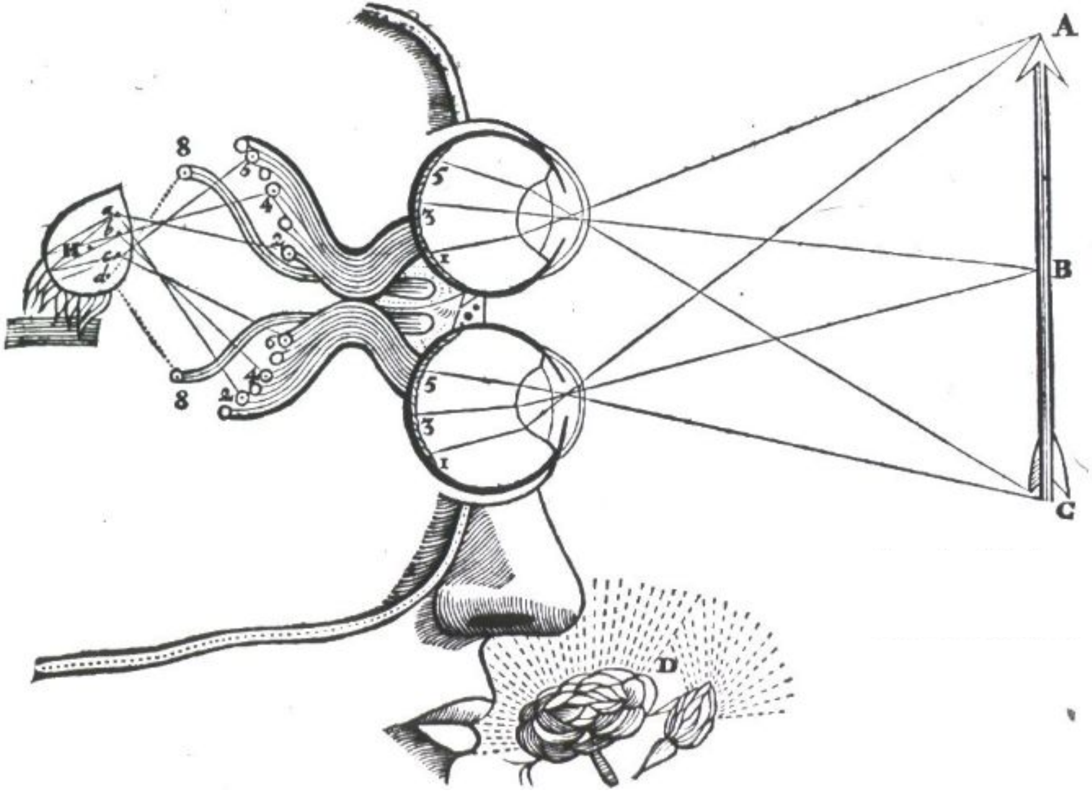
https://www.bl.uk/20th-century-literature/articles/freedom-or-oppression-the-fear-of-dystopia

How regulate destructive nature of tech when aligned with wealth (power).

flyingpenguin

Source: https://slate.com/technology/2017/01/what-frankenstein-has-to-do-with-anti-vaccination-activists.html

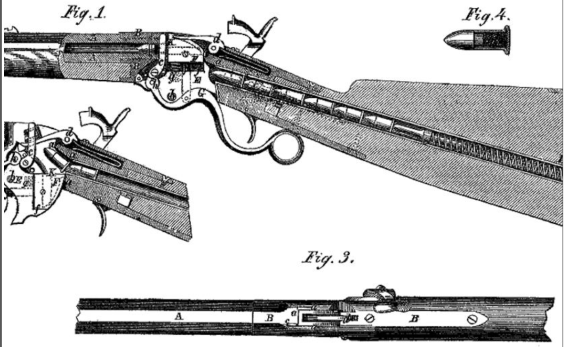
Ancient Debates About Tech Safety and Trust

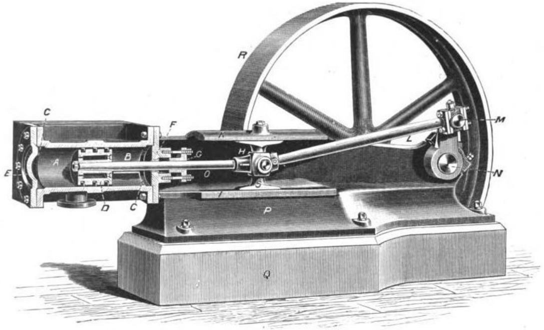


HENRY: It's a perfectly good brain, doctor. Well, you ought to know. It came from your own laboratory.

WALDMAN: The brain that was stolen from my laboratory was a **criminal brain**.

Even MEDIUM Tech Shifts Have Combo Power

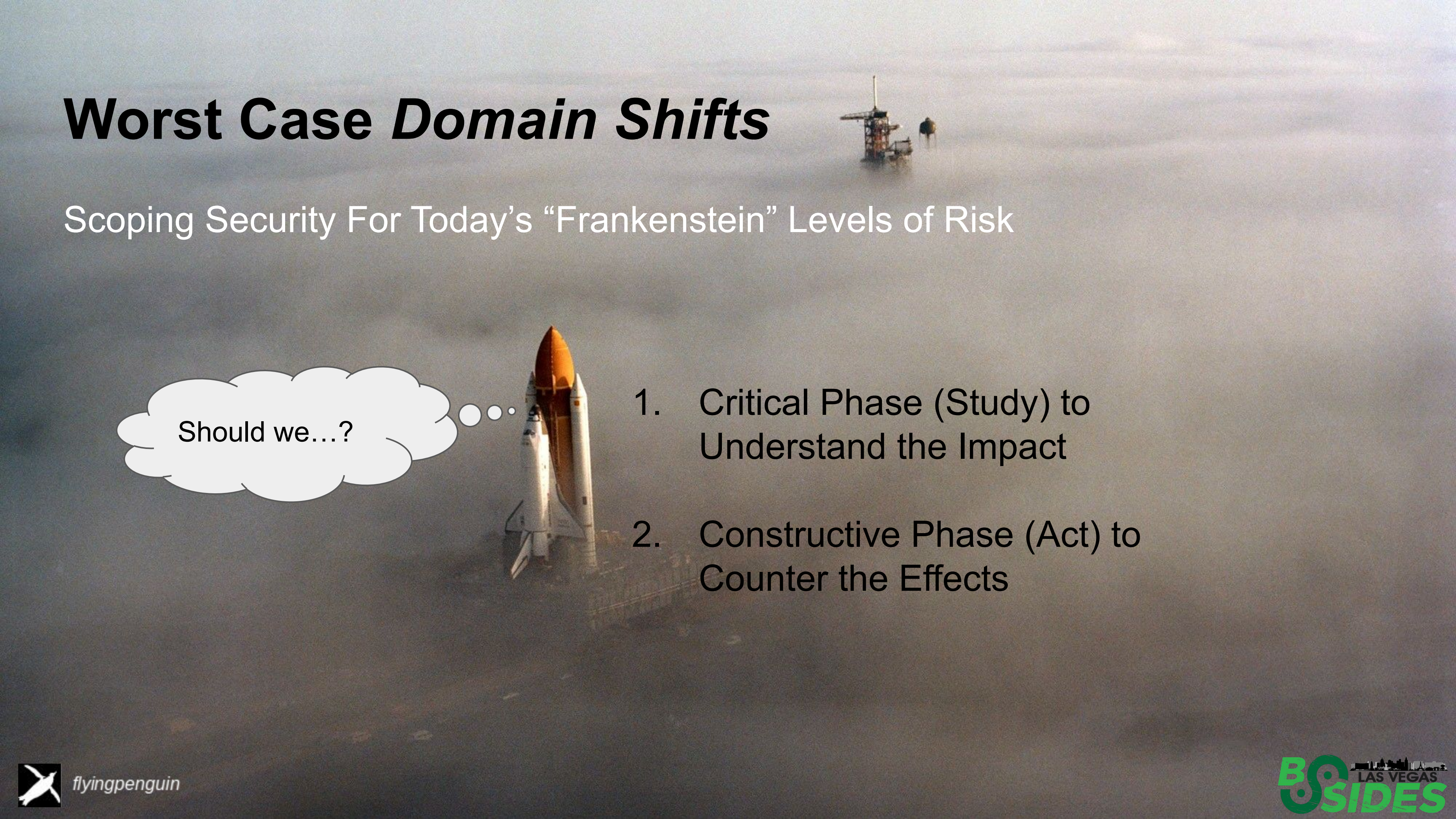




Connect these three to a “criminal brain” and...

Worst Case *Domain Shifts*

Scoping Security For Today’s “Frankenstein” Levels of Risk



1. Critical Phase (Study) to Understand the Impact

2. Constructive Phase (Act) to Counter the Effects

1) Critical Phase and 2) Constructive Phase

Identify

Store

Evaluate

Adapt

Easy

Routine

Minimal

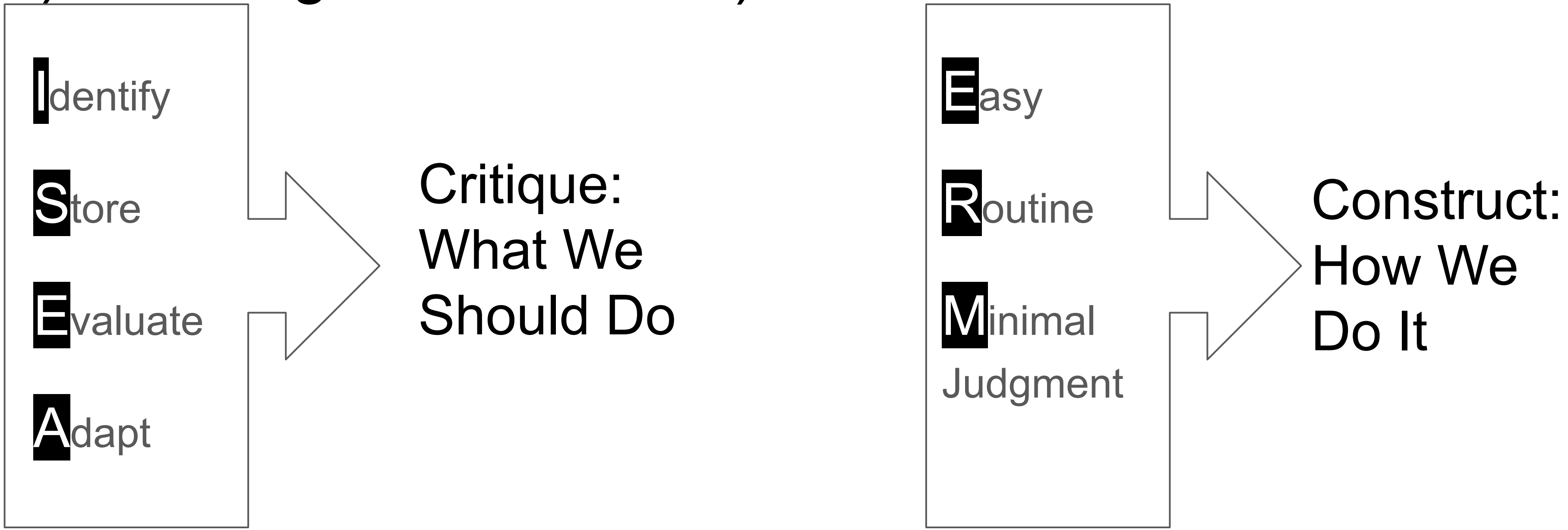
Judgment

USAID FROM THE AMERICAN PEOPLE

AFRICAN UNION INTERAFRICAN BUREAU FOR ANIMAL RESOURCES

Standard Methods and Procedures (SMPs) for Containment of Rinderpest (RP) in the Greater Horn of Africa

1) Learning Models and 2) Control Procedures



Identify

Store

Evaluate

Adapt

Critique: What We Should Do

Easy

Routine

Minimal Judgment

Construct: How We Do It

Critical Phase (ISEA)

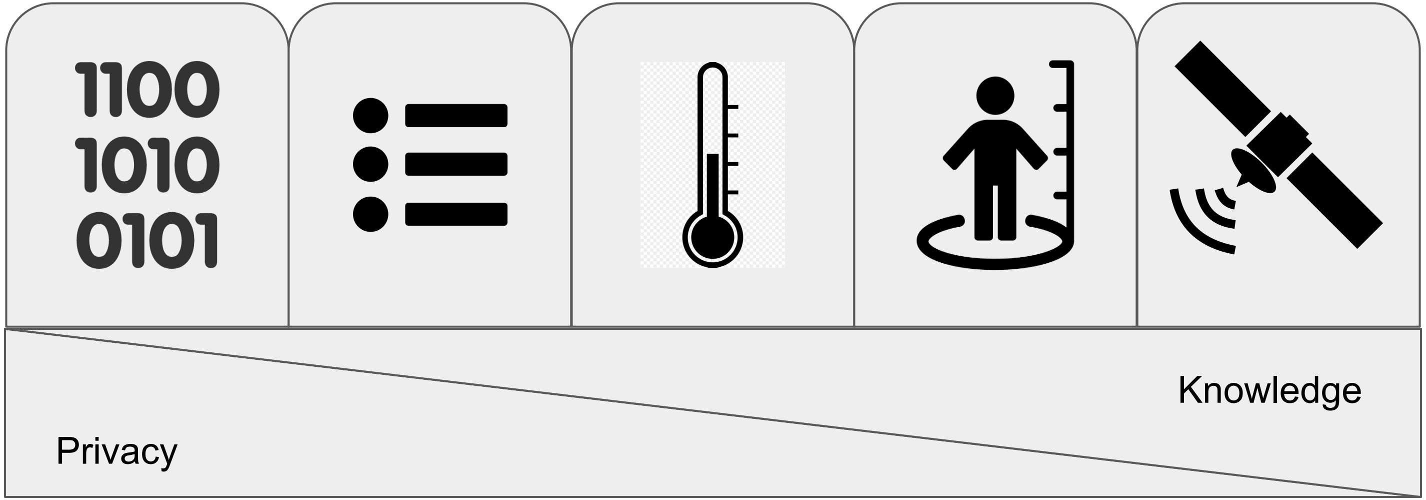
Robots are Automated Collective Shifts in Power

Beowulf's
Legacy
(The Heft of
30 Men in
His Hand)

Source: https://www.flickr.com/photos/aussierupe/174642408/

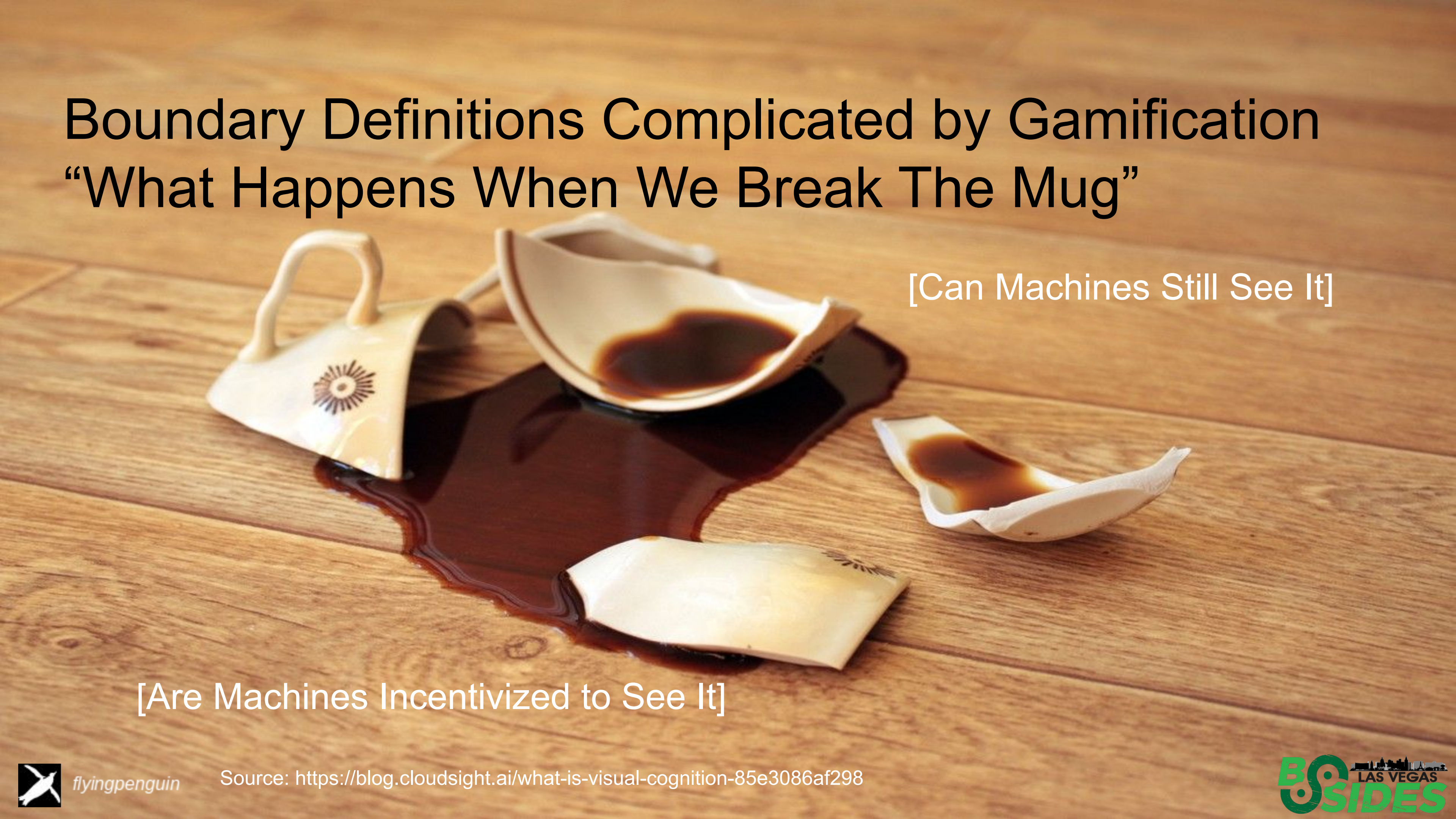
What Re-Balancing of Power is Safe?

Definition of Boundaries for “Criminal Brain” Tech Platforms, Controls to Design



Knowledge Gains (Power) Relative to Privacy Loss

Boundary Definitions Complicated by Gamification
“What Happens When We Break The Mug”



[Can Machines Still See It]

[Are Machines Incentivized to See It]

Source: https://blog.cloudsight.ai/what-is-visual-cognition-85e3086af298

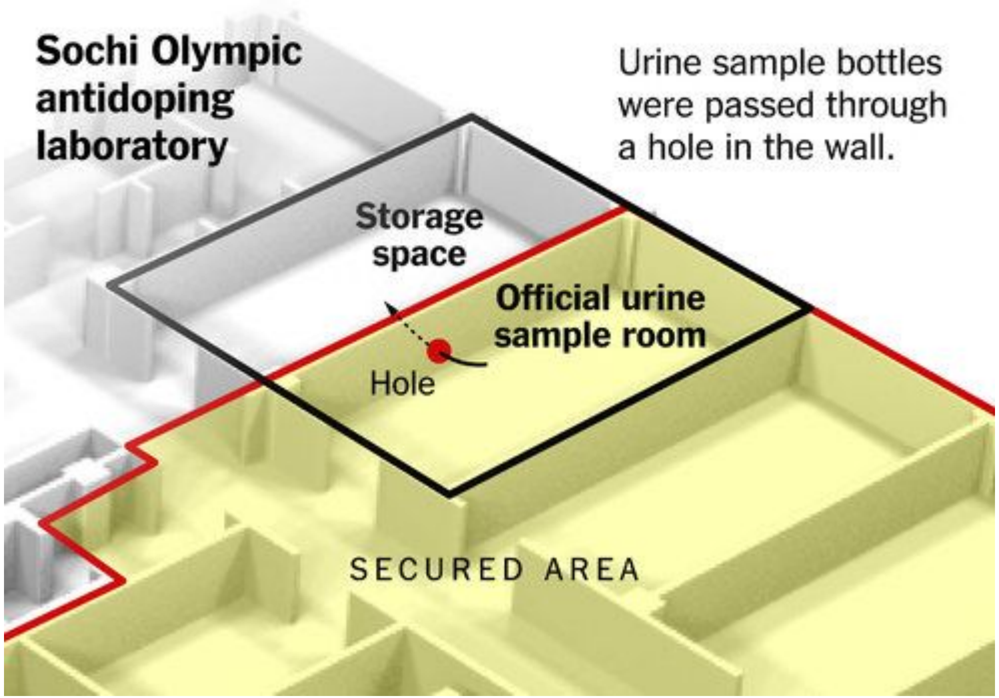
“If you don’t cheat, you look like an idiot; if you cheat and don’t get caught, you look like a hero; if you cheat and get caught, you look like a dope. Put me where I belong.”

-- Darrell Waltrip
Three-Time NASCAR Cup Series Champion

Source: https://www.flyingpenguin.com/?p=17622

AI “Specification Gaming” (Break) Examples

- ***Pancake Robot*** learns to throw as high as possible to “avoid” ground
- ***Driving Robot*** goes in reverse and impacts body to avoid touching bumpers
- ***Tic-tac-toe Robot*** makes distant moves to cause opponent memory exhaustion and forfeit


Sochi Olympic antidoping laboratory
Urine sample bottles were passed through a hole in the wall.
Storage space
Official urine sample room
Hole
SECURED AREA

Source: https://vkrakovna.wordpress.com/2018/04/02/specification-gaming-examples-in-ai/

RU Bombards Foreign Athletes to Exhaustion/Forfeit

“...people begin giving up…”

Source: https://www.washingtonpost.com/graphics/2018/world/national-security/russian-propaganda-skripal-salisbury/

Cambridge Analytica Was Automated Bombardment

"The bulk of our resources went into targeting those whose minds we thought we could change. We called them the **'persuadables'**," [whistleblower Kaiser] said. They focused even more on the people **in swing states**, and could therefore impact the overall result.

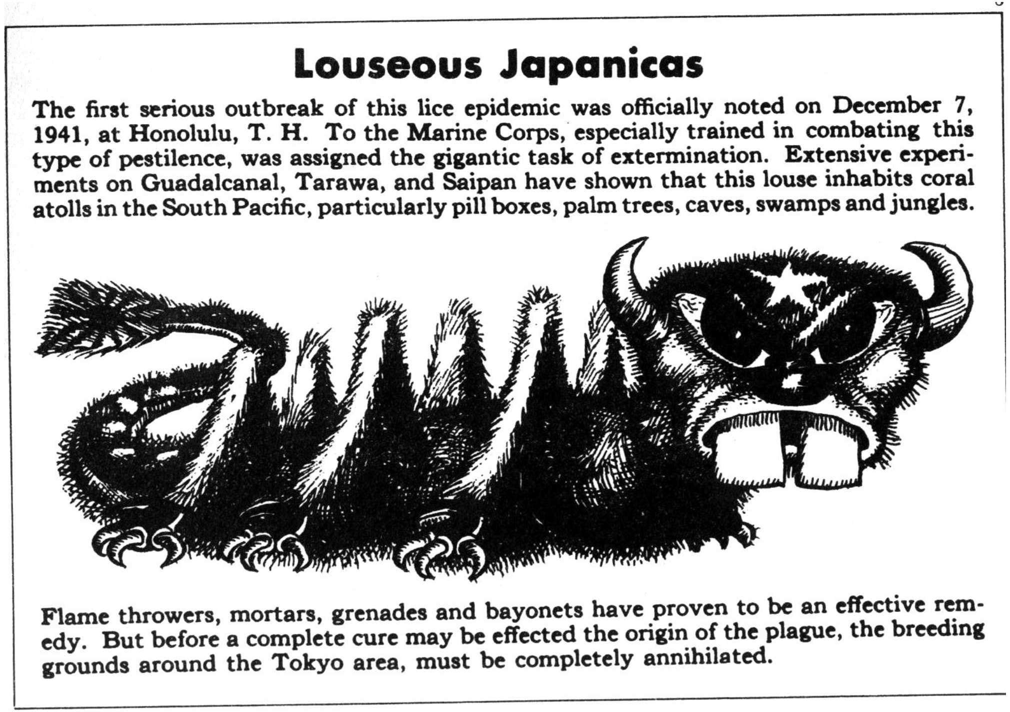
Their creative team designed "personalised content" to **"trigger those individuals"**, Kaiser added. "We bombarded them through blogs, websites, articles, videos, ads, every platform you can imagine. Until they saw the world the way we wanted them to…. Until they voted for our candidate."

> “...how true it is the working class feels an inclination towards a dictatorship, if it can first be rightly ***persuaded*** that the dictatorship will be exercised in its interests…”
>
> -- Letter to Bismarck from Lassalle, June 8, 1863

Source: https://www.digitalspy.com/tv/ustv/a28517243/the-great-hack-netflix-cambridge-analytica/
http://self.gutenberg.org/articles/eng/Ferdinand_Lassalle

Important American History Tangent:

Referring to Humans as Pests and Their Environment as "Infested" … Dehumanizes Them. Reliable Predictor of Concentration Camps & Genocide



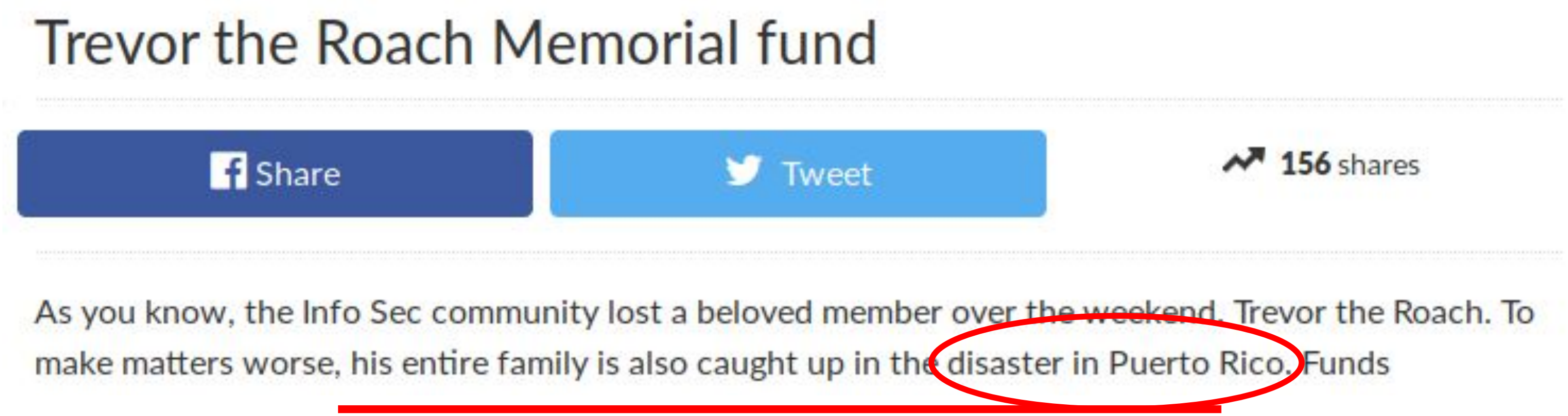
Louseous Japanicas

The first serious outbreak of this lice epidemic was officially noted on December 7, 1941, at Honolulu, T. H. To the Marine Corps, especially trained in combating this type of pestilence, was assigned the gigantic task of extermination. Extensive experiments on Guadalcanal, Tarawa, and Saipan have shown that this louse inhabits coral atolls in the South Pacific, particularly pill boxes, palm trees, caves, swamps and jungles.

Flame throwers, mortars, grenades and bayonets have proven to be an effective remedy. But before a complete cure may be effected the origin of the plague, the breeding grounds around the Tokyo area, must be completely annihilated.

Source: https://www.theguardian.com/commentisfree/2019/jul/30/trump-infested-baltimore-congresswomen

...as Within InfoSec: “Trevor the Roach” Signaling



Trevor the Roach Memorial fund

As you know, the Info Sec community lost a beloved member over the weekend. Trevor the Roach. To make matters worse, his entire family is also caught up in the disaster in Puerto Rico. Funds

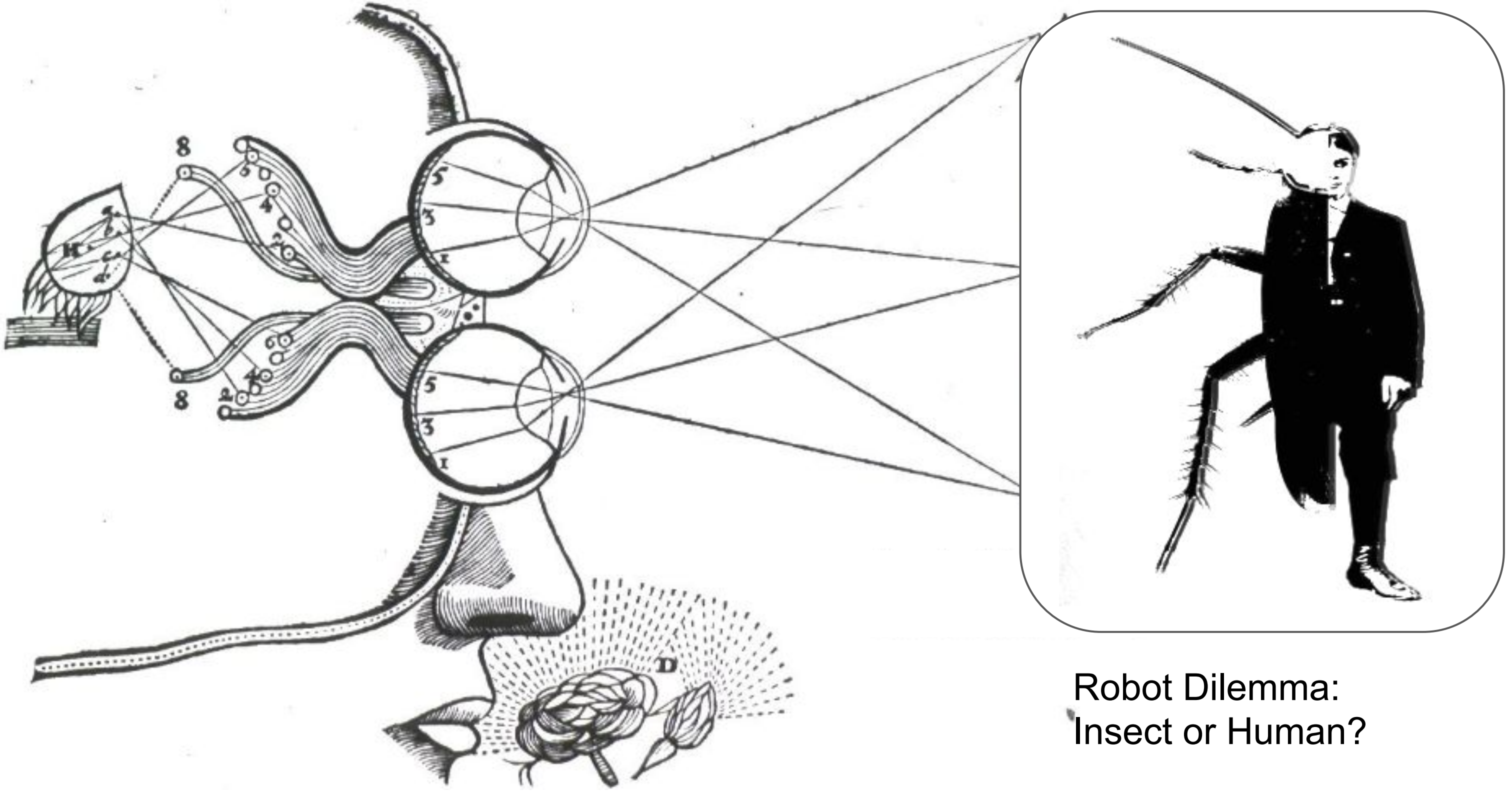
“I am trying to help Trevor's [Puerto Rican] kids and wife. Why do you hate children and widows? … It just went more epic. Dave on CNN talking about Trevor.”

flyingpenguin Source: https://ohioriders.net/index.php?/topic/112540-trevor-the-roach/

BSides LAS VEGAS

Bias Combined with Automation Combined with Rule Gamification (Vulnerability) Means...

AI is The Civil Rights Battle of Our Day...



Robot Dilemma:
Insect or Human?

HARMS

Are Automation Risks Really Dangerous or Does Good Outweigh?

And What Can Be Done to Increase Balance?

Harms “Seen Everywhere” by Trained Judges

Training a “X” Risks Them

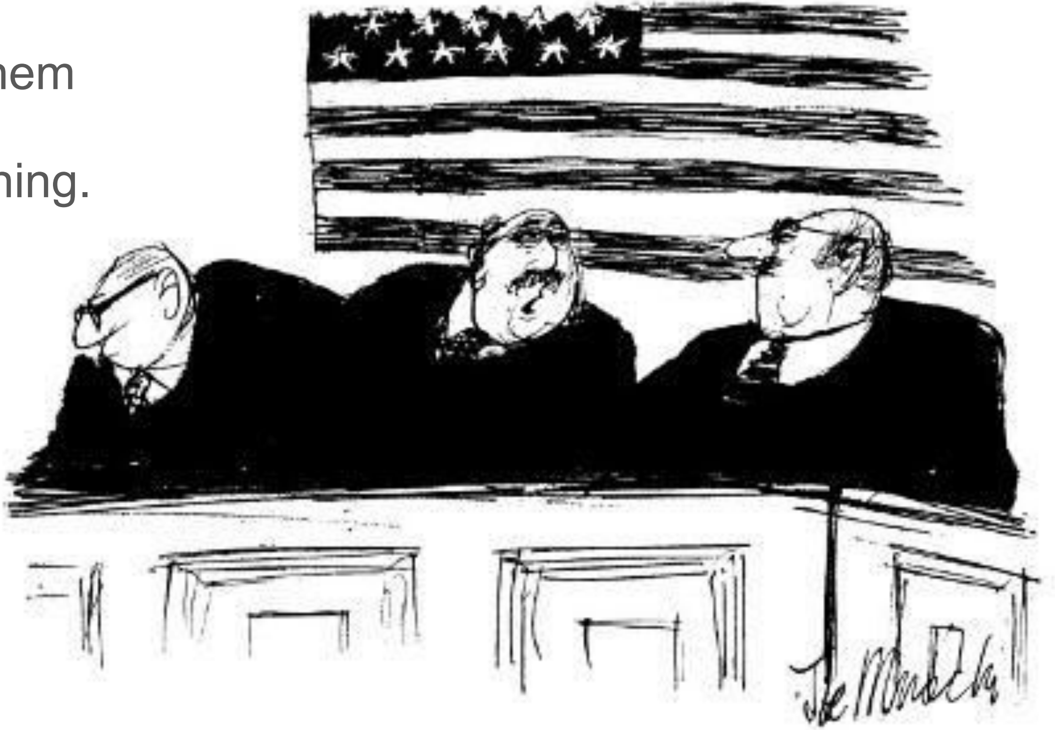
Seeing a “Y” in Everything.

That’s the Value of

A Trained “X”. So...

Do You Want That

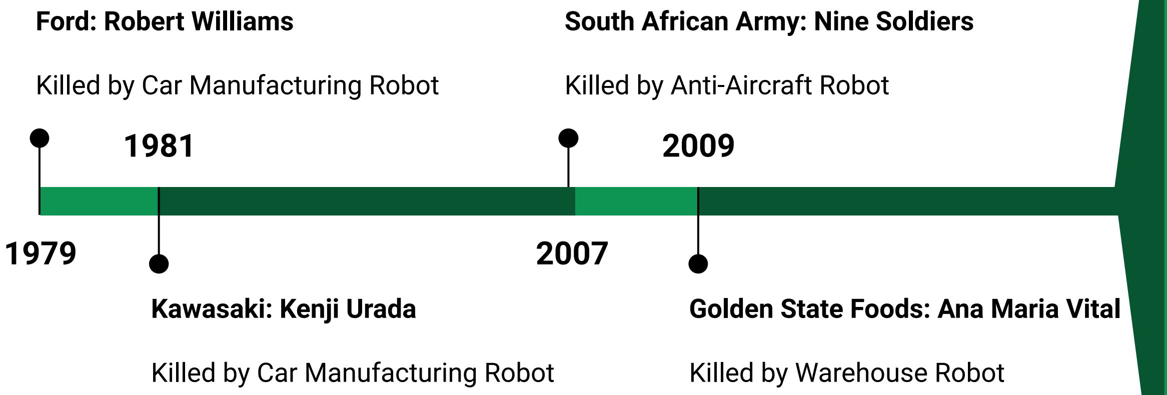
Intelligence?



“Do you ever have one of those days when everything seems ~~unconstitutional?!~~”
racist?

Documented Flawed Robots...

Organization: Victims



1979

Ford: Robert Williams

Killed by Car Manufacturing Robot

1981

Kawasaki: Kenji Urada

Killed by Car Manufacturing Robot

2007

South African Army: Nine Soldiers

Killed by Anti-Aircraft Robot

2009

Golden State Foods: Ana Maria Vital

Killed by Warehouse Robot

- **2015 VW: Anonymous**
  Killed by Car Manufacturing Robot
- **2015 SKH Metals: Ramji Lal**
  Killed by Welding Robot
- **2016 Dallas: Micah Johnson**
  Killed by Bomb Defuser Robot
- **2016 Ajin USA: Regina Elsea**
  Killed by Car Manufacturing Robot
- **2016 Tesla: Joshua Brown**
  Killed by Driverless Car
- **2017 VIM: Wanda Holbrook**
  Killed by Car Manufacturing Robot
- **2018 Uber: Elaine Herzberg**
  Killed by Driverless Car
- **2019 Boeing: 346 Passengers**
  Killed by Pilotless Planes

flyingpenguin

BSIDES LAS VEGAS

Quick Poll

? Percent who say victims are to blame

2007: South African Army Oerlikon 35mm MK5

“...the brave, as yet unnamed officer was ***unable to stop the wildly swinging computerised***…anti-aircraft twin-barrelled gun. It sprayed hundreds of high-explosive 0,5kg 35mm cannon shells.... By the time the gun had emptied its twin 250-round auto-loader magazines, ***nine soldiers were dead*** and 11 injured.”

Source: https://www.wired.com/2007/10/robot-cannon-ki/

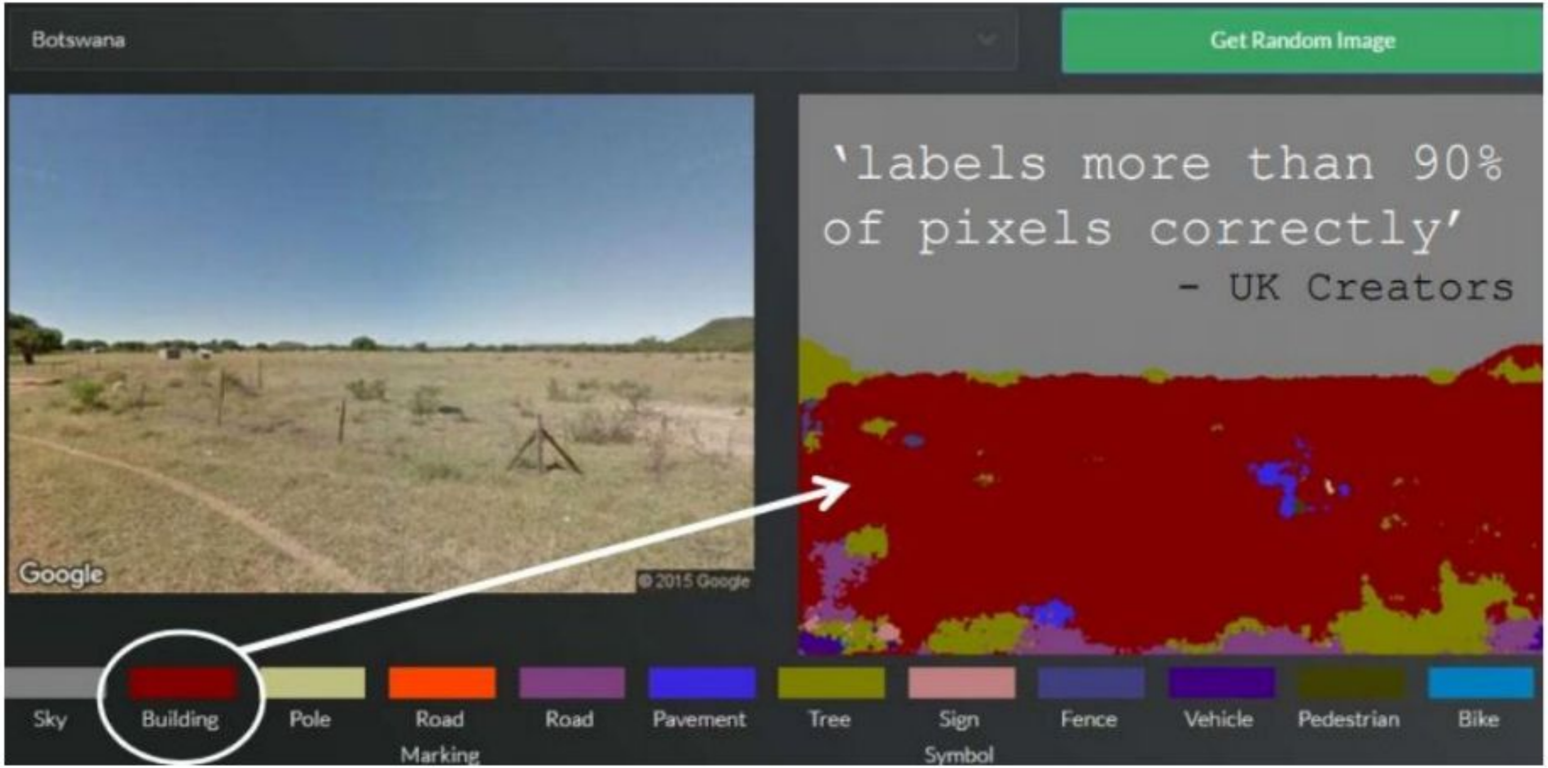
2016: Tesla

Chose to kill human because ‘overhead sign’

(more likely a *moving* bridge)

Source: https://www.flyingpenguin.com/?p=22441

Adversarial Test (90% Failure Rates Documented)

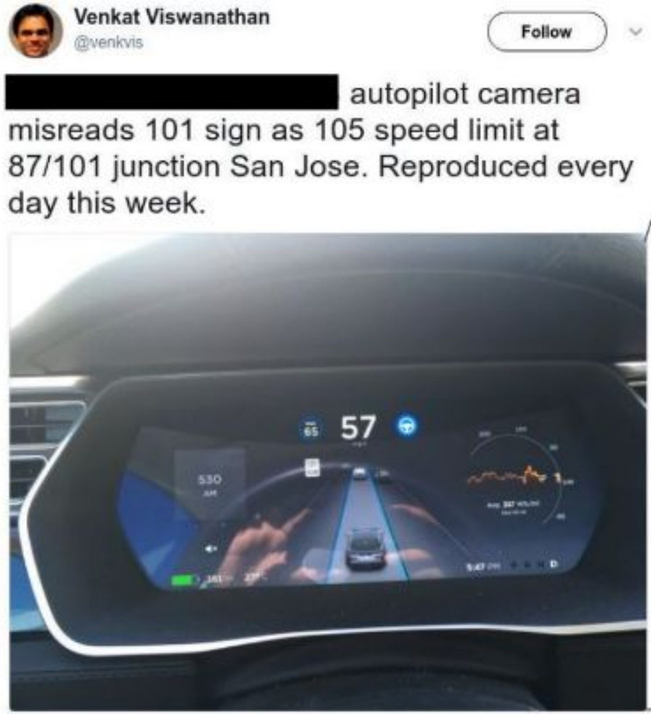


‘labels more than 90% of pixels correctly’

- UK Creators

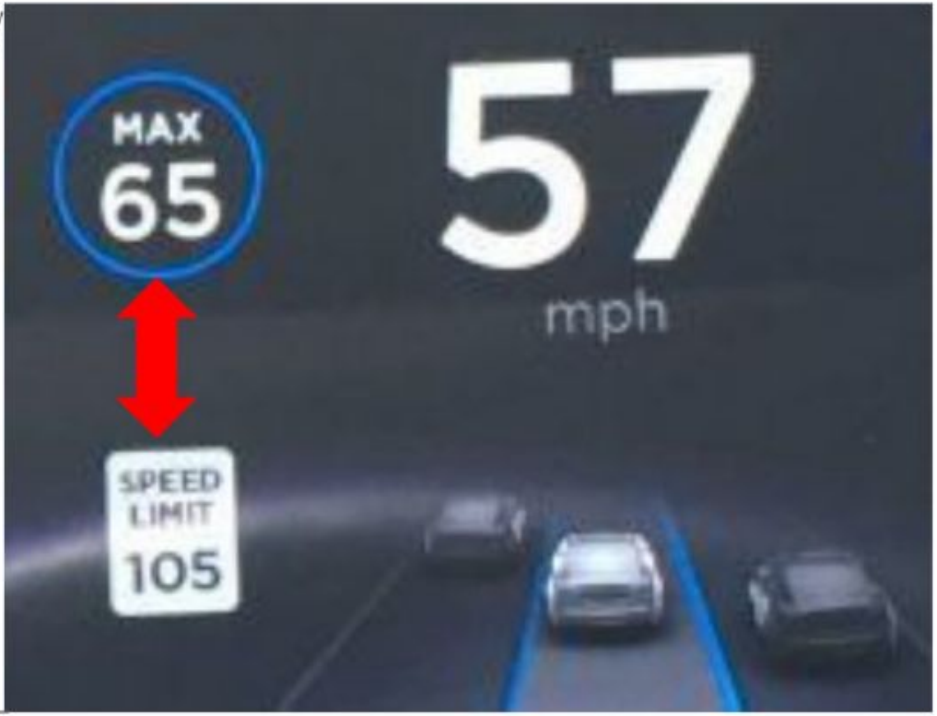
flyingpenguin Source: http://www.flyingpenguin.com/?p=22429

2017: Tesla

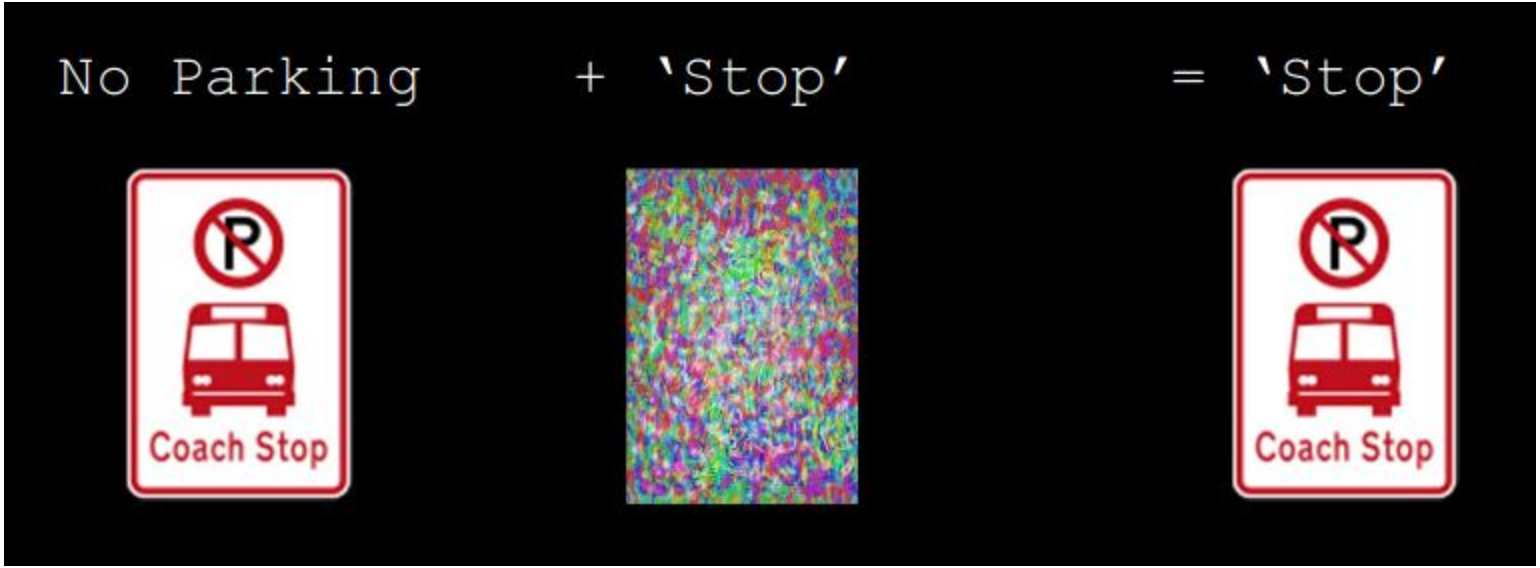


Venkat Viswanathan
@venkvis

[redacted] autopilot camera misreads 101 sign as 105 speed limit at 87/101 junction San Jose. Reproduced every day this week.



Adversarial Test (90%+ Failure Rates Documented)



flyingpenguin Source: http://www.flyingpenguin.com/?p=22429

Adversarial Test (EDR)

“By taking strings from an online gaming program and appending them to malicious files, researchers were able to trick Cylance’s AI-based antivirus engine into thinking programs like WannaCry and other malware are benign.”

Source: https://www.vice.com/en_us/article/9kxp83/researchers-easily-trick-cylances-ai-based-antivirus-into-thinking-malware-is-goodware

2018: Tesla

“**Autopilot function was engaged for the last 18 minutes and 55 seconds** of Huang’s drive that Friday morning...**following a car until seconds before the crash. But the car either changed lanes or exited** and once there were no vehicles in front of the Tesla, it began to accelerate. ‘At 3 seconds prior to the crash and up to the time of impact with the crash attenuator, the Tesla’s speed increased from 62 to 70.8 mph, with **no pre-crash braking or evasive steering movement detected**,’ the report stated.”

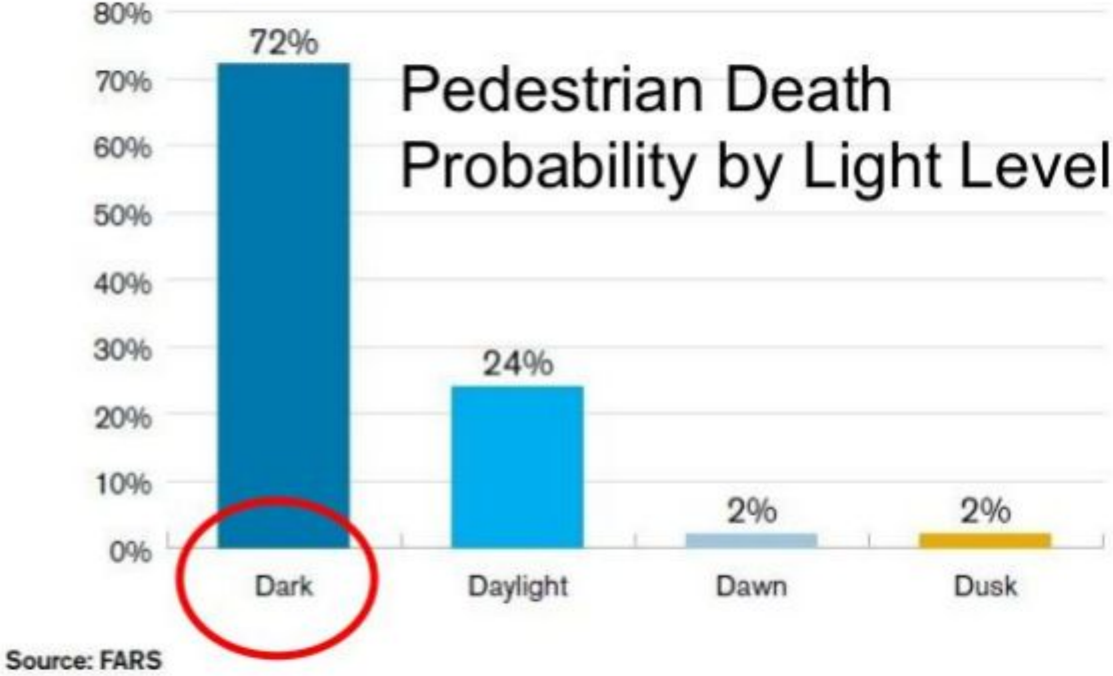
Source: https://sanfrancisco.cbslocal.com/2018/06/07/ntsb-driver-fatal-tesla-crash/amp/

Adversarial Test (redacted)

- Rogue Vehicle Network
- Safety Controls Removed
- Rogue Navigation Data

[redacted]

2018: Uber

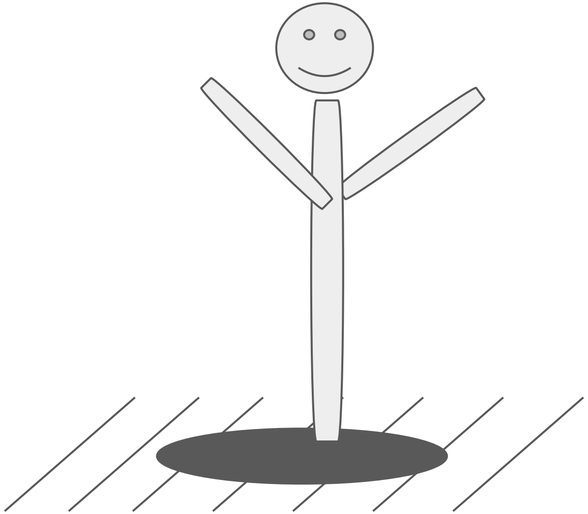


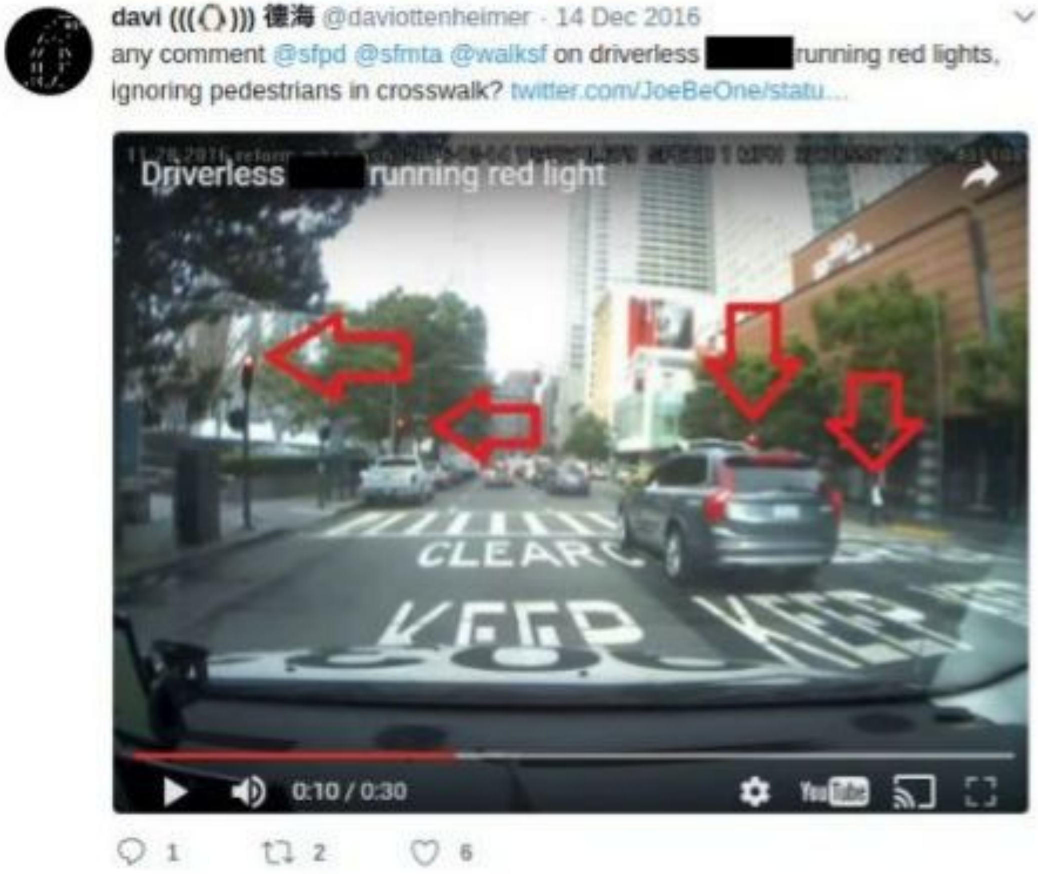
Source: FARS

Source: https://www.washingtonpost.com/news/dr-gridlock/wp/2018/03/19/uber-halts-autonomous-vehicle-testing-after-a-pedestrian-is-struck/

Adversarial Test

Ukraine Art: Pedestrian Balloons





davi (((◯))) 德海 @daviottenheimer · 14 Dec 2016

any comment @sfpd @sfmta @walksf on driverless [redacted] running red lights, ignoring pedestrians in crosswalk? twitter.com/JoeBeOne/statu...

Driverless [redacted] running red light

Source: https://twitter.com/daviottenheimer/status/809204856007245824

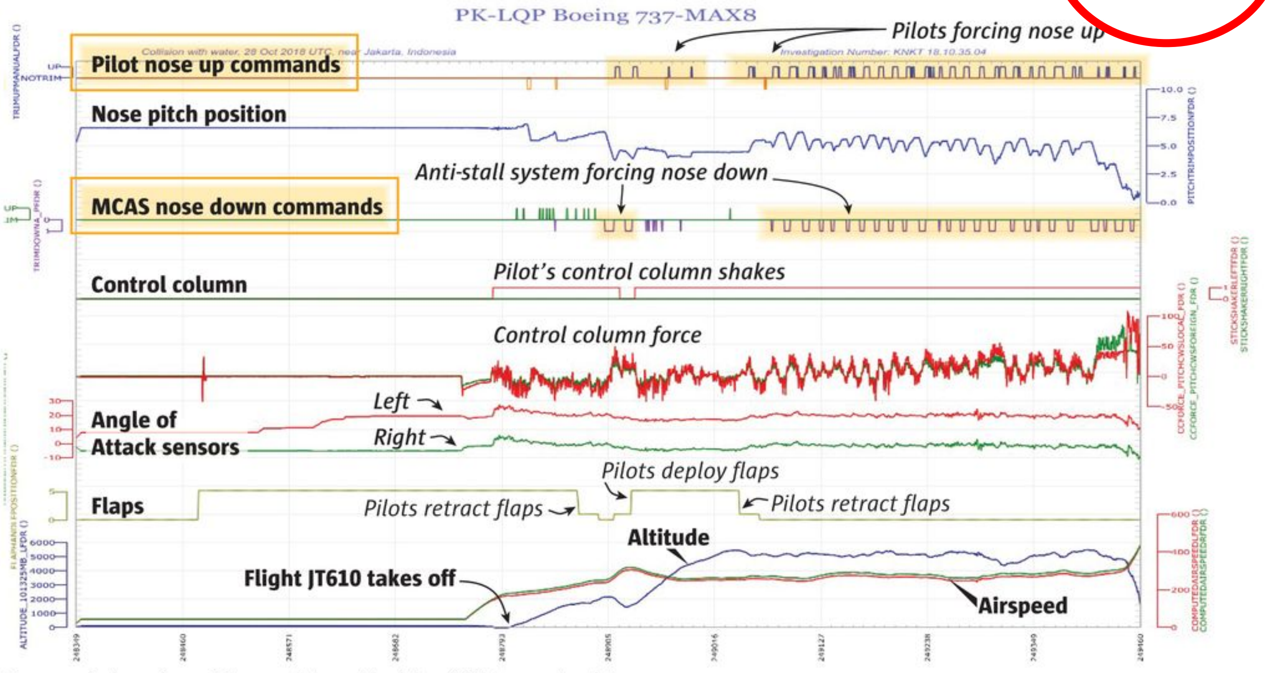
2019: Boeing 737 MAX MCAS

(Maneuvering Characteristics Augmentation System)

The jet’s nose is repeatedly pushed down

The new anti-stall system on the Boeing 737 MAX forced the nose of Lion Air JT610 down 26 times in 10 minutes before the pilots lost control and the plane dived into the sea.

“cause opponent memory exhaustion and forfeit”



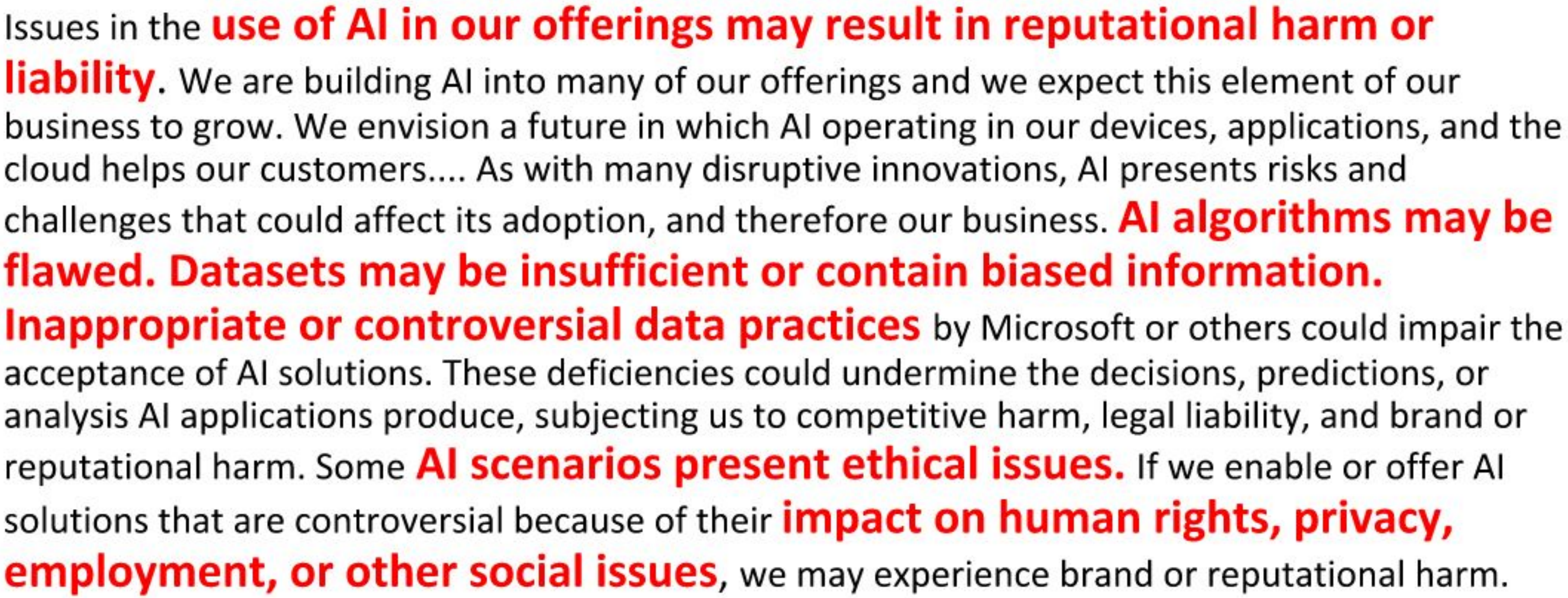
Sources: Indonesian safety regulators, black box flight recorder data

MARK NOWLIN / THE SEATTLE TIMES

flyingpenguin

BSIDES LAS VEGAS

2018 Microsoft 10-Q: Intelligence May Be Harmful



Issues in the **use of AI in our offerings may result in reputational harm or liability**. We are building AI into many of our offerings and we expect this element of our business to grow. We envision a future in which AI operating in our devices, applications, and the cloud helps our customers.... As with many disruptive innovations, AI presents risks and challenges that could affect its adoption, and therefore our business. **AI algorithms may be flawed. Datasets may be insufficient or contain biased information. Inappropriate or controversial data practices** by Microsoft or others could impair the acceptance of AI solutions. These deficiencies could undermine the decisions, predictions, or analysis AI applications produce, subjecting us to competitive harm, legal liability, and brand or reputational harm. Some **AI scenarios present ethical issues.** If we enable or offer AI solutions that are controversial because of their **impact on human rights, privacy, employment, or other social issues**, we may experience brand or reputational harm.

flyingpenguin

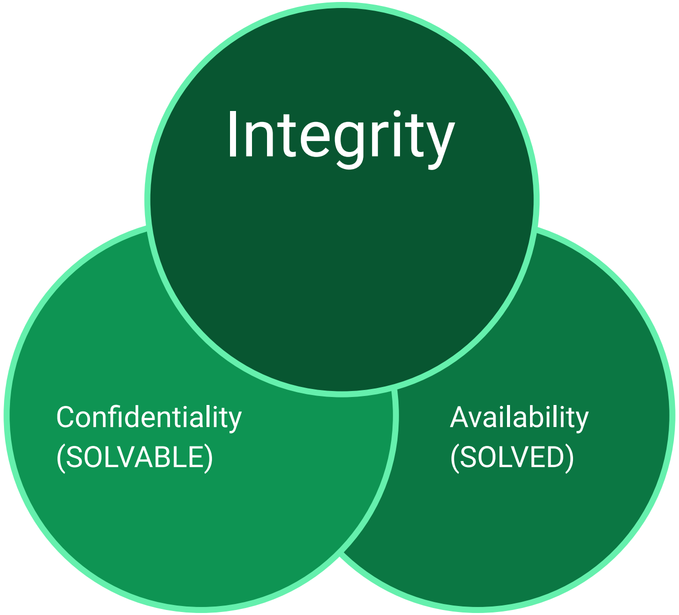
Source: https://www.sec.gov/Archives/edgar/data/789019/000156459019001392/msft-10q_20181231.htm

Beware the Clippenstein Monster

It looks like you are trying to install Windows 10. Would you like me to spy on you?
● Yes

Constructive Phase (ERM)

We Can Do This!



Integrity

Confidentiality
(SOLVABLE)

Availability
(SOLVED)

Moral Frameworks
(Instead of Games)
&
Analytic Integrity Engines
(Gravity of Ethics)

“Assassins were role models”

Plan for Bladerunner-like consulting services?

Source: https://www.dw.com/en/merkel-honors-hitlers-attempted-killers-laments-rise-of-far-right-extremism/a-49581109

Or … “You got the wrong guy, pal”

Assassins as Heroes = Generally Too Late

Ethics of Outside Intervention (Regulatory Framework) Kicks In

Four members of the 6888th. Source: United States Department of Defense.

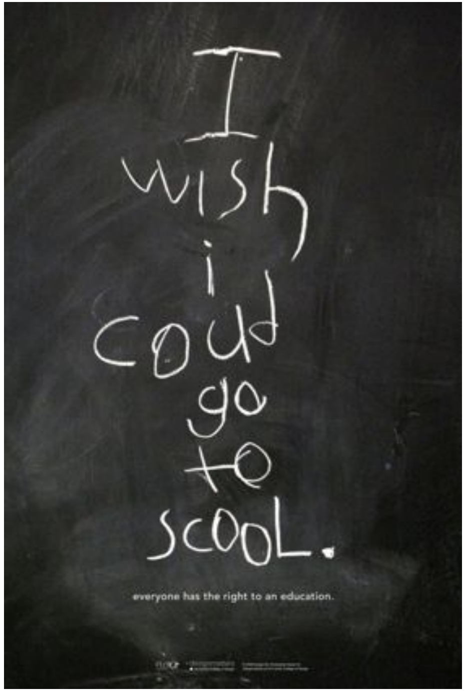
Source: https://www.flyingpenguin.com/?p=24754

Legally Binding Rules v Ethics (Moral Compass)

“But what if it is not possible to have it all?

What if we cannot respect privacy to the full extent if we want to develop efficient AI tools as soon as possible for all those patients who are suffering and/or dying right now?

Using the [Universal Declaration of Human Rights] framework, we do not have an answer. In fact, we are at a dead end.”

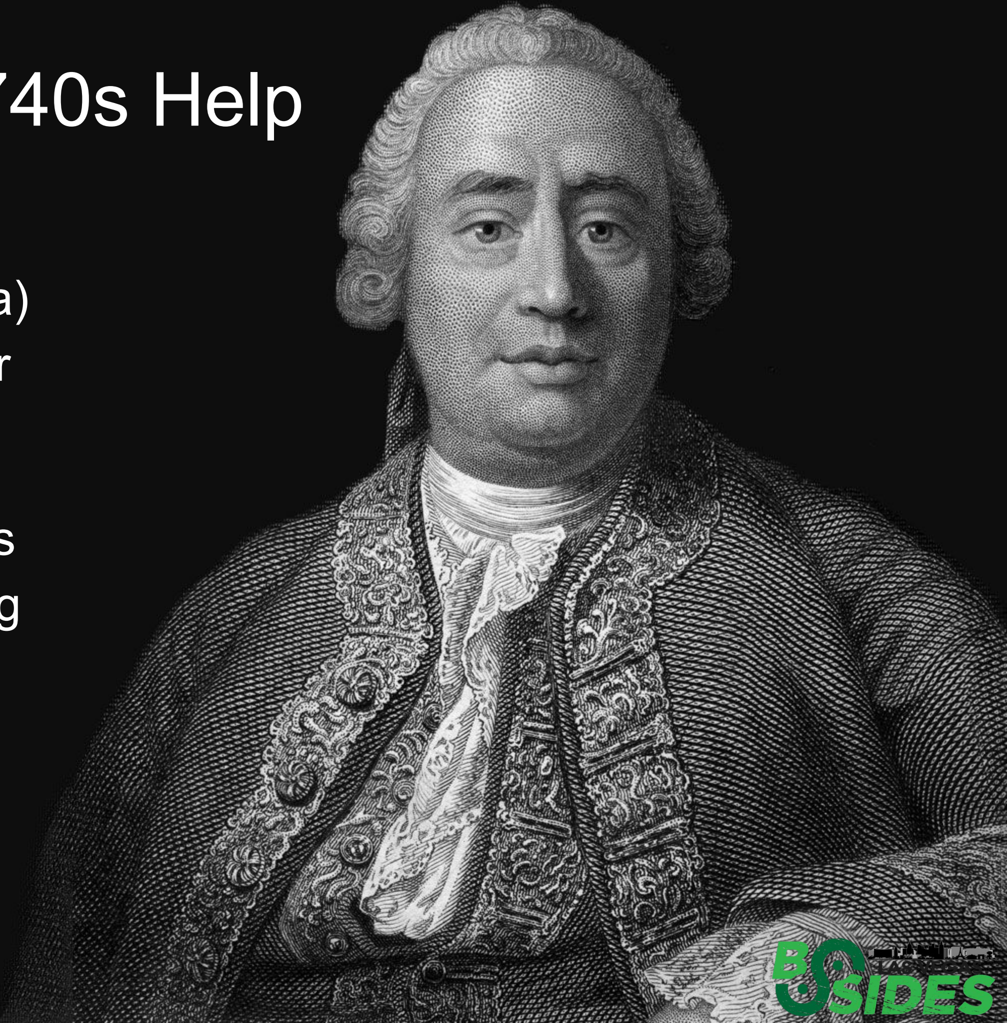

I wish i coud go to scool.
everyone has the right to an education.

Source: https://cpr.unu.edu/ai-global-governance-human-rights-and-ai-ethics-why-ethics-cannot-be-replaced-by-the-udhr.html

Hume’s Writings From 1740s Help

Accurate extrapolation outside of what has been experienced (i.e. training data) should be required to optimize behavior models (i.e. resist adversarial data)

Ethical actions sit in general boundaries of passion as “moral ideas do not spring from reason alone” (Treatise 3.1.1)



Mary Wollstonecraft Even More Helpful

- A Vindication of the Rights of Man (1790)
    - Unequal society founded on passivity of women
    - Rationality, unlike ancestral traditions or dogma, abolishes slavery
- A Vindication of the Rights of Woman (1792)
    - Human limitations are a result of deficient education
    - Middle-class “most natural state”
    - Equality of sexes

‘I attribute [these 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...”

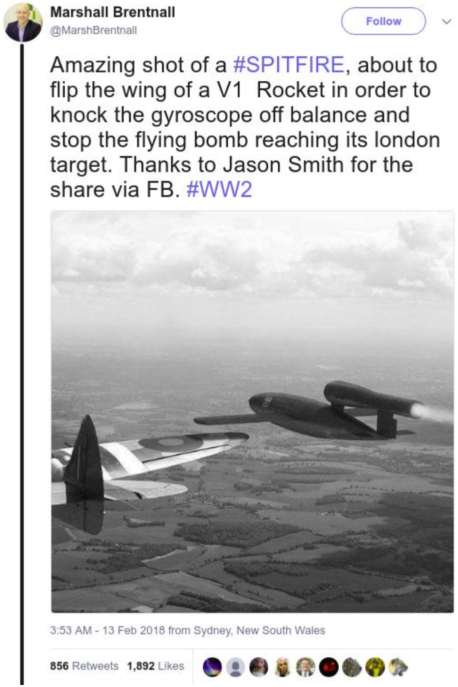
Source: https://www.bl.uk/collection-items/mary-wollstonecraft-a-vindication-of-the-rights-of-woman

Systemic Problems Need Systemic Solutions

A Historian’s Perspective:

Both Tweets are Dangerously Wrong

Why?

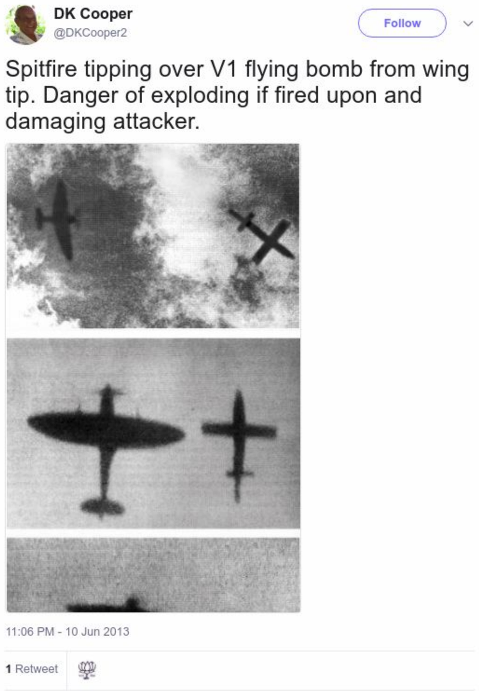


Marshall Brentnall
@MarshBrentnall

Amazing shot of a #SPITFIRE, about to flip the wing of a V1 Rocket in order to knock the gyroscope off balance and stop the flying bomb reaching its london target. Thanks to Jason Smith for the share via FB. #WW2

3:53 AM - 13 Feb 2018 from Sydney, New South Wales

856 Retweets 1,892 Likes



DK Cooper
@DKCooper2

Spitfire tipping over V1 flying bomb from wing tip. Danger of exploding if fired upon and damaging attacker.

11:06 PM - 10 Jun 2013

1 Retweet

An Abridged History of False System of Education

1740 South Carolina Ban on Teaching Slaves to Write

1758 Georgia Ban on Teaching Slaves to Write

1833 Alabama Set Fines for Educating Slaves

1836 North Carolina Ban on Education of Blacks

1841 Mississippi Required Educated Black Freemen to Leave State

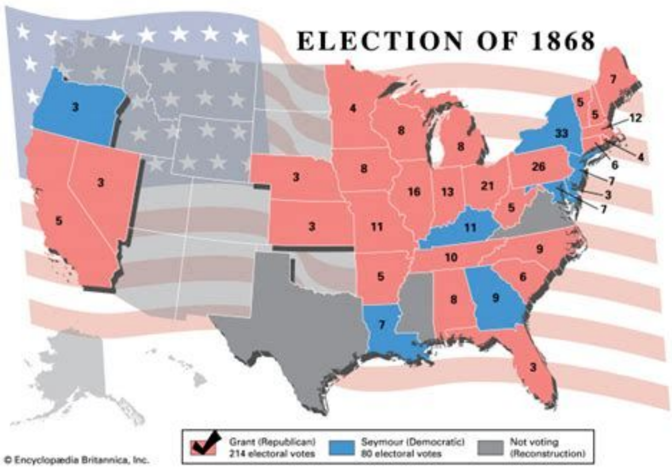
[...]

1959 Virginia Shuts Down Public Schools and Gives Tuition Grants to Whites Only

Regulatory Agency Within Moral Frameworks

Ulysses S. Grant

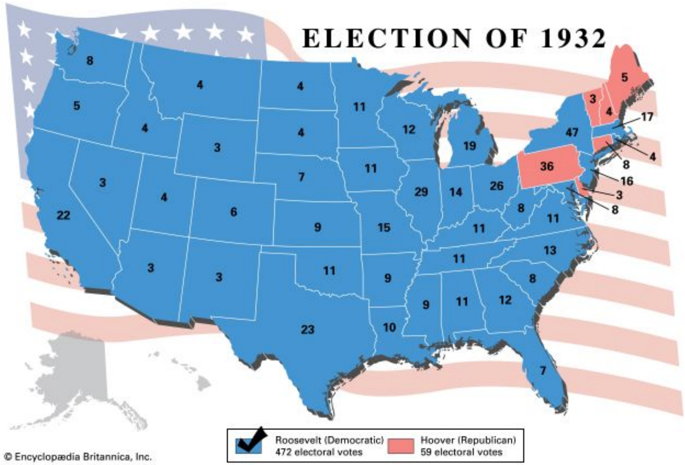
- 1870 Department of Justice
- 1875 Civil Rights Act


ELECTION OF 1868
Grant (Republican) 214 electoral votes
Seymour (Democratic) 80 electoral votes
Not voting (Reconstruction)

© Encyclopædia Britannica, Inc.

Franklin D. Roosevelt

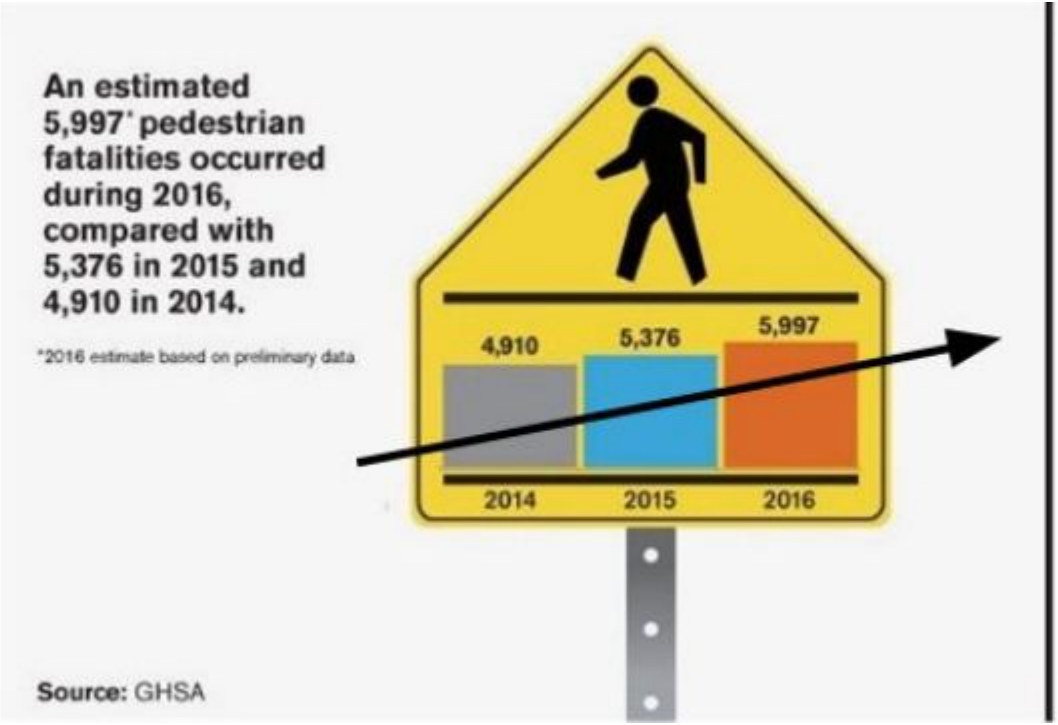
- Mandate for Human Welfare
- 1934 Communications Act


ELECTION OF 1932
Roosevelt (Democratic) 472 electoral votes
Hoover (Republican) 59 electoral votes

© Encyclopædia Britannica, Inc.

Source: http://www.abajournal.com/magazine/article/march_1_1875_grant_signs_the_civil_rights_act

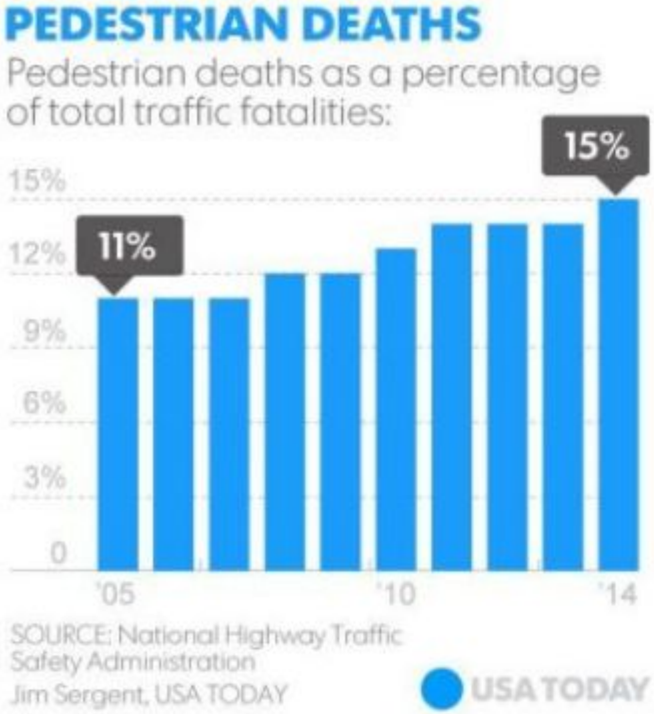
Robot Example: Pedestrian Safety as a Moral Idea



An estimated 5,997* pedestrian fatalities occurred during 2016, compared with 5,376 in 2015 and 4,910 in 2014.

*2016 estimate based on preliminary data

Source: GHSA



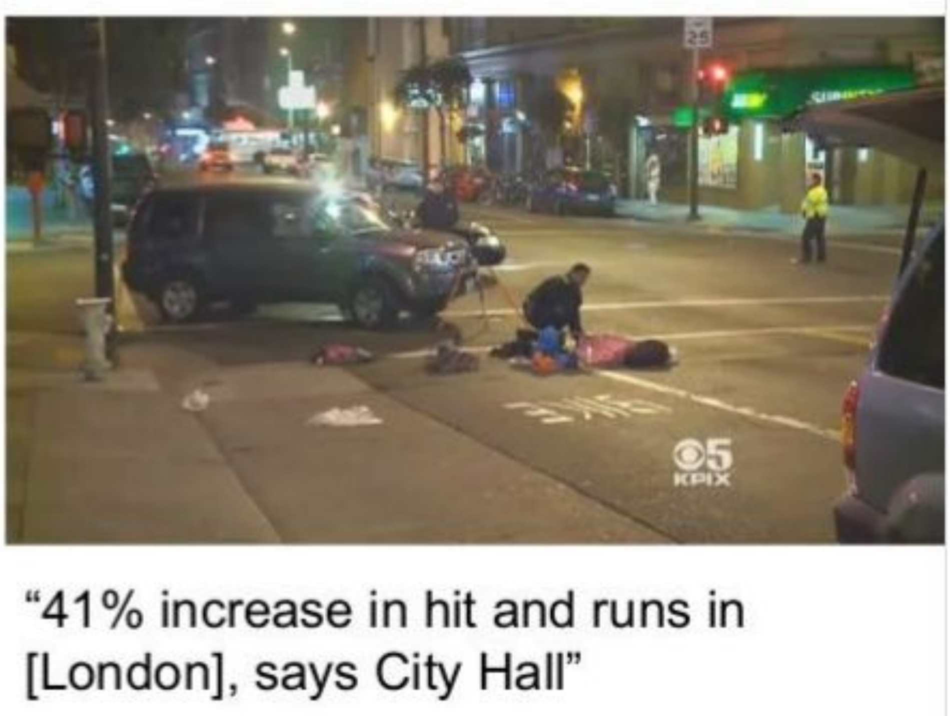
PEDESTRIAN DEATHS

Pedestrian deaths as a percentage of total traffic fatalities:

SOURCE: National Highway Traffic Safety Administration
Jim Sergent, USA TODAY

USA TODAY

Robot Example: Pedestrian Safety as a Moral Idea



“41% increase in hit and runs in [London], says City Hall”

Commercial “Reasoning” of Driverless Cars Lacks Passion of Human Survival:

- Maximize Rides/Hour
- Maintain Availability
- Avoid Dwell Time
- Cost of Business if Caught

Source: http://www.londonlive.co.uk/news/2016-12-20/41-increase-in-hit-and-runs-in-the-capital-say-city-hall

Opportunity to Decriminalize Race (Jaywalking)

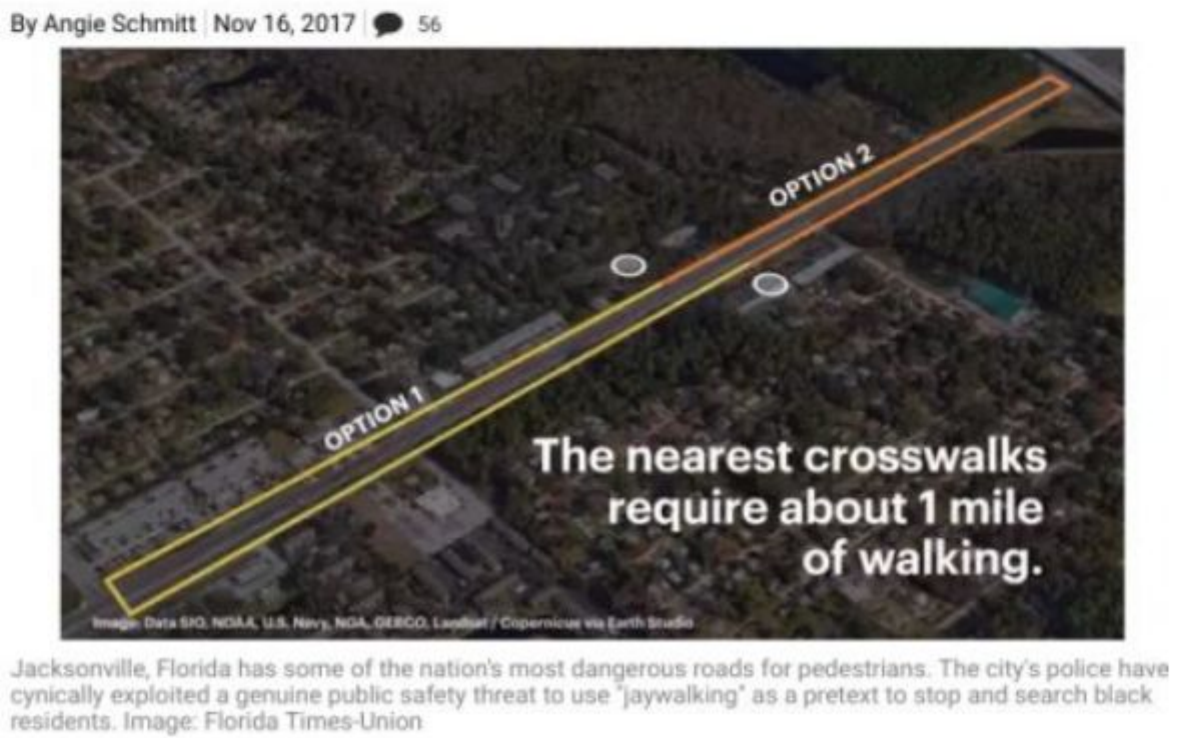
Re-Humanize

Pedestrians

Before

Robots

Kill Us All



By Angie Schmitt | Nov 16, 2017 | 56

The nearest crosswalks require about 1 mile of walking.

Image Data SIO, NOAA, U.S. Navy, NGA, GEBCO, Landsat / Copernicus via Earth Studio

Jacksonville, Florida has some of the nation's most dangerous roads for pedestrians. The city's police have cynically exploited a genuine public safety threat to use "jaywalking" as a pretext to stop and search black residents. Image: Florida Times-Union

Source: https://www.flyingpenguin.com/?p=23690

AIs Wide Open

Making Bots Safer Than
Completely #$%cking Unsafe

BSIDES LAS VEGAS

August 6, 2019
1430-1455 / 1600-1755

flyingpenguin