AI Auditing

Davi Ottenheimer
October 24, 2021

Abstract

AI for decades promised great gains in productivity.

However many groups accounting for risks in AI have revealed stunning results, showing that without careful auditing of plans and development the ***security risks from automation far outweigh any benefit***.

This session boils down the rapidly expanding AI topic to clarify what has been delivered so far and where things most often are going wrong.

Trigger Warning: Disturbing Content Ahead

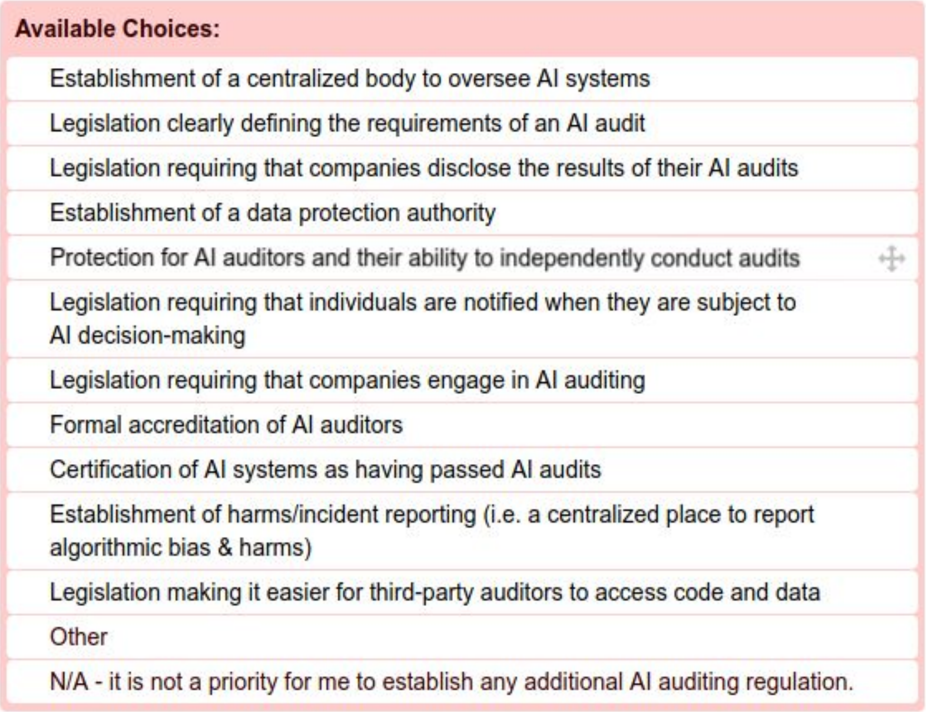
The following slides include discussion of harms and even fatalities due to safety failures in technology. This content may be disturbing to some people, and I encourage everyone to prepare themselves before proceeding.

Also, there is no consensus yet on what AI is…

How Auditable is Learning / Intelligence?

“...survey by Deloitte on behalf of Algorithmic Justice League…”

Aequitas
Accenture Algorithmic Fairness
Alibi Explain
AllenNLP
BlackBox Auditing
DebiasWE
DiCE
ErrorAnalysis
EthicalML xAI
Facebook DynaBoard
Fairlearn
FairSight
FairTest
FairVis
FoolBox
Google Explainable AI
Google KnowYourData
Google ML Fairness Gym
Google PAIR Facets
Google PAIR Language Interpretability Tool
Google PAIR Saliency
Google PAIR What-If Tool
IBM Adversarial Robustness Toolbox
IBM AI Fairness 360
IBM AI Explainability 360
Lime
MLI
ODI Data Ethics Canvas
Parity
PET Repository
PwC Responsible AI Toolkit
Pymetrics audit-AI
RAN-debias
REVISE
Saidot
SciKit Fairness
Skater
Spatial Equity Data Tool
TCAV
UnBias Fairness Toolkit



Available Choices:

- Establishment of a centralized body to oversee AI systems
- Legislation clearly defining the requirements of an AI audit
- Legislation requiring that companies disclose the results of their AI audits
- Establishment of a data protection authority
- Protection for AI auditors and their ability to independently conduct audits
- Legislation requiring that individuals are notified when they are subject to AI decision-making
- Legislation requiring that companies engage in AI auditing
- Formal accreditation of AI auditors
- Certification of AI systems as having passed AI audits
- Establishment of harms/incident reporting (i.e. a centralized place to report algorithmic bias & harms)
- Legislation making it easier for third-party auditors to access code and data
- Other
- N/A - it is not a priority for me to establish any additional AI auditing regulation.

https://deloittesurvey.deloitte.com/Community/se/3FC11B2662CA1330

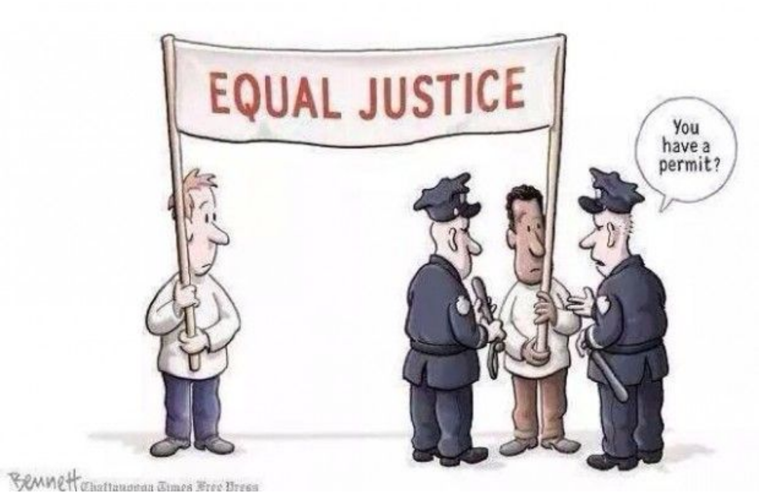
How Auditable is Systemic / Societal Bias?

Griggs vs. Duke Power Company ruling stated that independent of intent, ***disparate and discriminatory outcomes for protected classes*** were a violation of Title VII of the Civil Rights Act of 1964.

Sound Complicated?

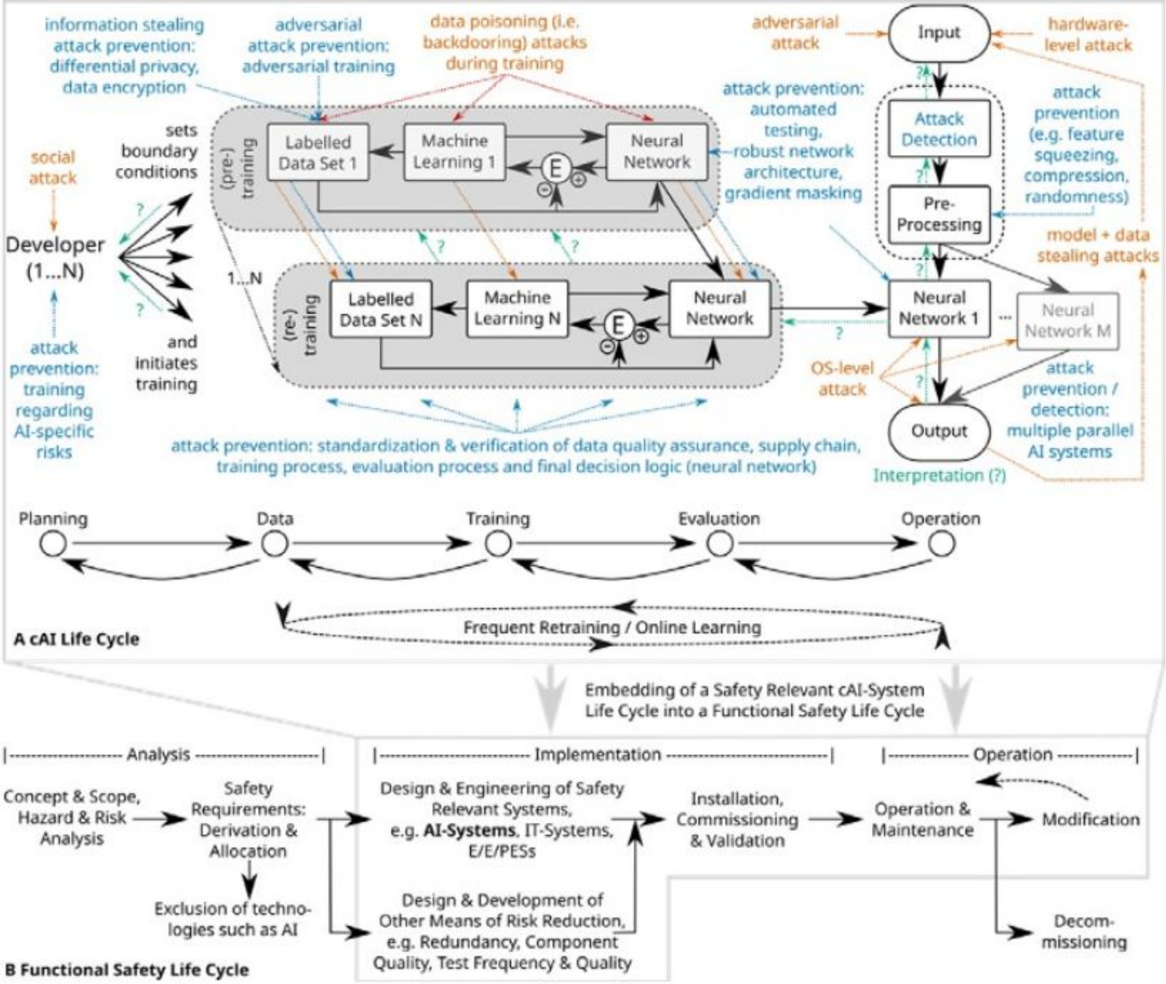
2018 Accenture's Fairness Tool

1. Data quality and variables
2. Distribution of model errors
3. False positive rate



https://hbr.org/2018/10/auditing-algorithms-for-bias
https://github.com/fair-ml/Algorithmic-Fairness

2021 German Federal Office for InfoSec


information stealing attack prevention: differential privacy, data encryption
adversarial attack prevention: adversarial training
data poisoning (i.e. backdooring) attacks during training
adversarial attack
Input
hardware-level attack
social attack
sets boundary conditions
(pre-) training
Labelled Data Set 1
Machine Learning 1
E
Neural Network
attack prevention: automated testing, robust network architecture, gradient masking
Attack Detection
Pre-Processing
attack prevention (e.g. feature squeezing, compression, randomness)
model + data stealing attacks
Developer (1...N)
?
1...N
(re-) training
Labelled Data Set N
Machine Learning N
Neural Network
Neural Network 1
...
Neural Network M
attack prevention: training regarding AI-specific risks
and initiates training
OS-level attack
attack prevention / detection: multiple parallel AI systems
Output
attack prevention: standardization & verification of data quality assurance, supply chain, training process, evaluation process and final decision logic (neural network)
Interpretation (?)
Planning
Data
Training
Evaluation
Operation
Frequent Retraining / Online Learning
A cAI Life Cycle
Embedding of a Safety Relevant cAI-System Life Cycle into a Functional Safety Life Cycle
Analysis
Implementation
Operation
Concept & Scope, Hazard & Risk Analysis
Safety Requirements: Derivation & Allocation
Exclusion of technologies such as AI
Design & Engineering of Safety Relevant Systems, e.g. AI-Systems, IT-Systems, E/E/PESs
Design & Development of Other Means of Risk Reduction, e.g. Redundancy, Component Quality, Test Frequency & Quality
Installation, Commissioning & Validation
Operation & Maintenance
Modification
Decommissioning
B Functional Safety Life Cycle

ISACA San Francisco Chapter

https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Towards_Auditable_AI_Systems.pdf

inrupt

Boiling It Down Quickly

Philosophy The Key To Unlock Artificial Intelligence

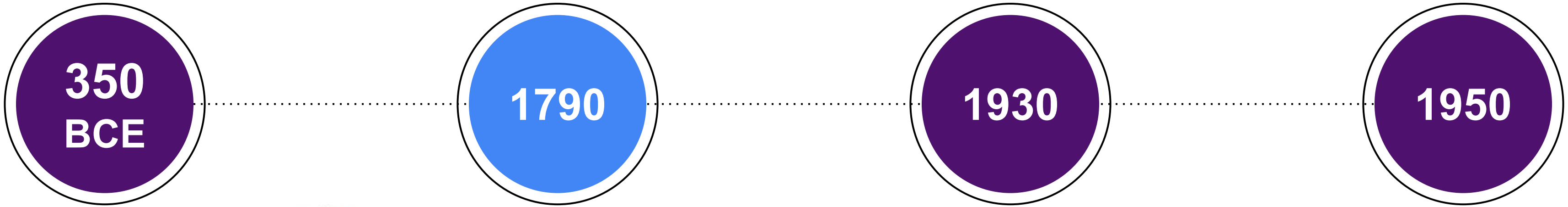
"I am convinced that the whole problem of ***developing AGIs is a matter of philosophy***, not computer science or neurophysiology. [...] Thinking of an AGI as a machine for translating experiences, rewards and punishments into ideas (or worse, just into behaviours) is like trying to cure infectious diseases by balancing bodily humours: futile because it is rooted in an archaic and wildly mistaken world view."

-- David Deutsch
"Father of Quantum Computing"

https://www.theguardian.com/science/2012/oct/03/philosophy-artificial-intelligence

Why Did (Intelligent) Things Cross the Road?

*(gravity of social situations: the **science** of ethics)*



Aristotle

To actualize its potential

Wollstonecraft (& Hume)

Out of custom and habit

Wittgenstein

Crossing encoded in objects “chicken” & “road”

Sartre

True to self, it acted in good faith

“Grundprobleme der Erkenntnistheorie”: Falsification Tests

“...we should claim the right to suppress [intolerance] if necessary even by force; for it may easily turn out that they are not prepared to meet us on the level of rational argument, but begin by denouncing all argument; they may forbid their followers to listen to rational argument, because it is deceptive, and teach them to answer arguments by the use of their fists or pistols. We should therefore claim, in the name of tolerance, ***the right not to tolerate the intolerant***.”

-- Karl Popper 1945

S II A Anthropologie · Beitrag 12 Erkenntnistheorie 1

Von Platon bis Popper –
Grundprobleme der Erkenntnistheorie erörtern

Grit Arnold, Marburg

© akg-images.

Was können wir wissen? Erkenntnistheoretische Positionen von Platon bis Popper.

https://www.academia.edu/39663912/Karl_Poppers_Open_Society_and_its_Enemies_and_its_Enemies

Ethics Audits: Inherited & Controlled Rights

Rights Within
Inherited System

Authority Can be Judged, Found Wrong and Held Accountable

Rights Within
Controlled System

Authority is the Judge and Can *Never be Wrong* or Held Accountable

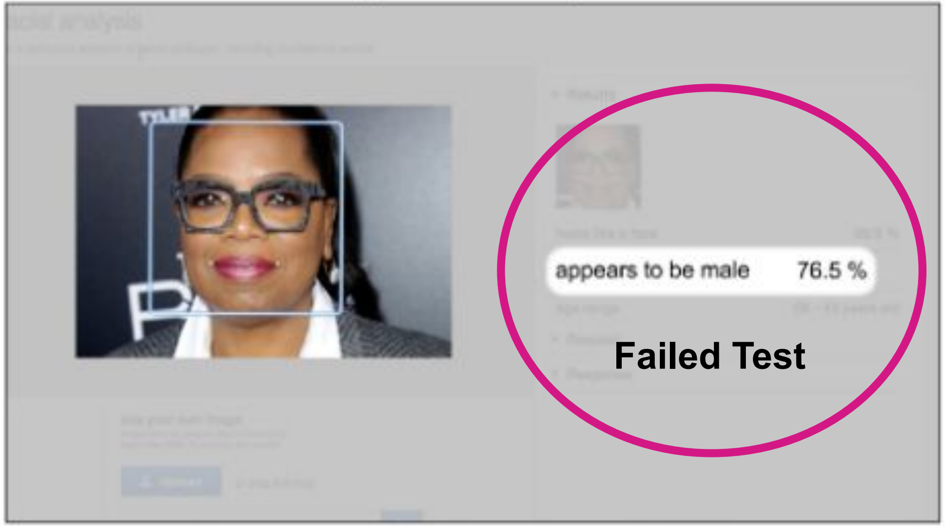
"...we think that that is good..."

Example of “Controlled System” Test

“The answer to anxieties over new technology is ***<u>not to run ‘tests’</u>*** inconsistent with how the service is designed to be used…”

-- Matthew Wood, AWS AI General Manager

Oprah Winfrey



appears to be male 76.5 %

Failed Test

amazon

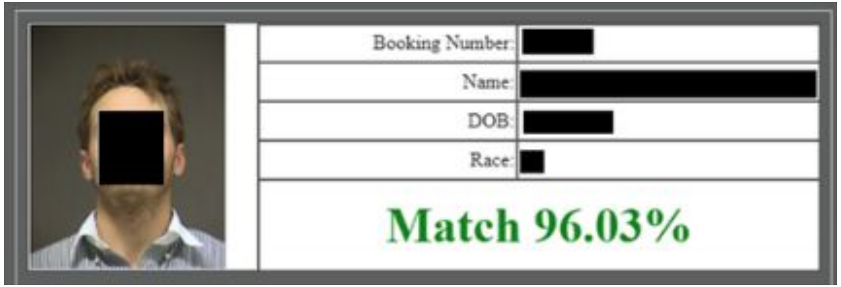
https://aws.amazon.com/blogs/aws/thoughts-on-machine-learning-accuracy/?ascsubtag=f20aee8a83dfe170e3c0efee8cdf84898de11b97
https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eec

Example of “Controlled System” Test

“Amazon has ***repeatedly claimed that the researchers failed*** to use the software, called Rekognition, ***in the way the company has instructed police to use it*** [with a 99% confidence threshold].

However, the only law enforcement agency Amazon has acknowledged as a client says it also does not use Rekognition in the way Amazon claims it recommends...”

Washington County Sheriff’s Office in Oregon Public Information Officer: “We do not set nor do we utilize a confidence threshold.”



[redacted]

Booking Number:	[redacted]
Name:	[redacted]
DOB:	[redacted]
Race:	[redacted]

Match 96.03%

“Amazon spokesperson clarified that ***law enforcement clients failure*** to use a 99-percent confidence threshold ***does not constitute an irresponsible application***...”

https://gizmodo.com/defense-of-amazons-face-recognition-tool-undermined-by-1832238149

Example of “Controlled System” Test

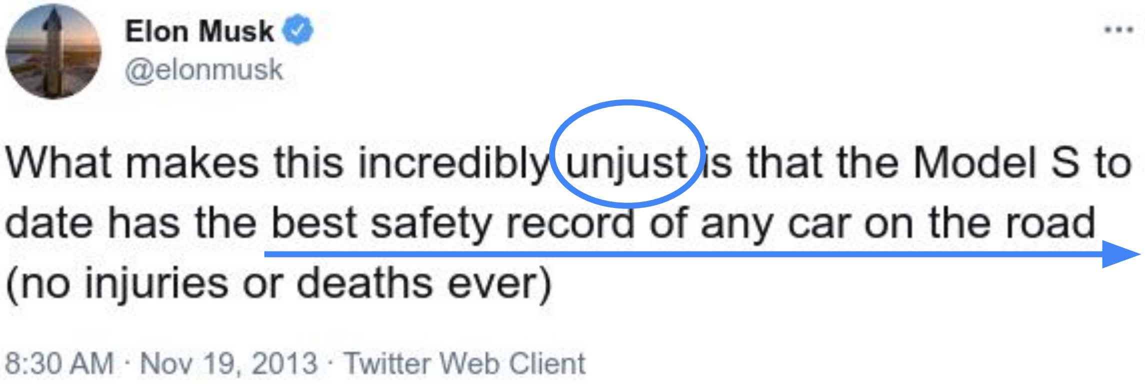
“Tesla has advertised its cars since 2016 as already having all the hardware necessary for all the FSD features to be used in the future without any hardware upgrades. ***In reality***, only cars delivered since the spring, 2019 have the necessary hardware…”

“Videos of Tesla drivers testing the system show ***the system struggles with basic driving tasks…***”

“If something goes wrong with Autopilot, it's because ***someone is misusing it***…” -- Elon Musk

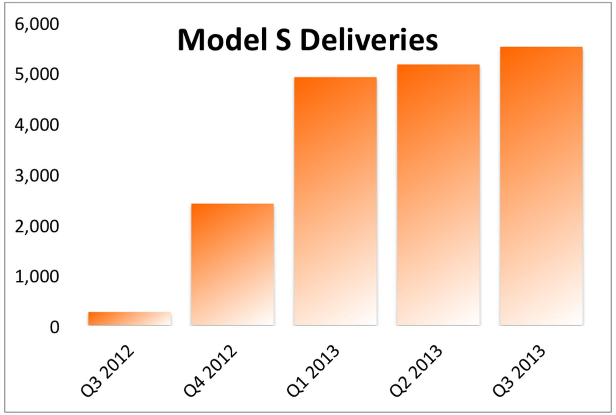
https://www.autoweek.com/news/green-cars/a37059597/tesla-fsd-subscribers-might-need-dollar1500-in-new-hardware/
https://www.businessinsider.com/tesla-senators-ftc-self-driving-autopilot-marketing-claims-elon-musk-2021-8

Example of “Controlled System” Test


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


Model S Deliveries

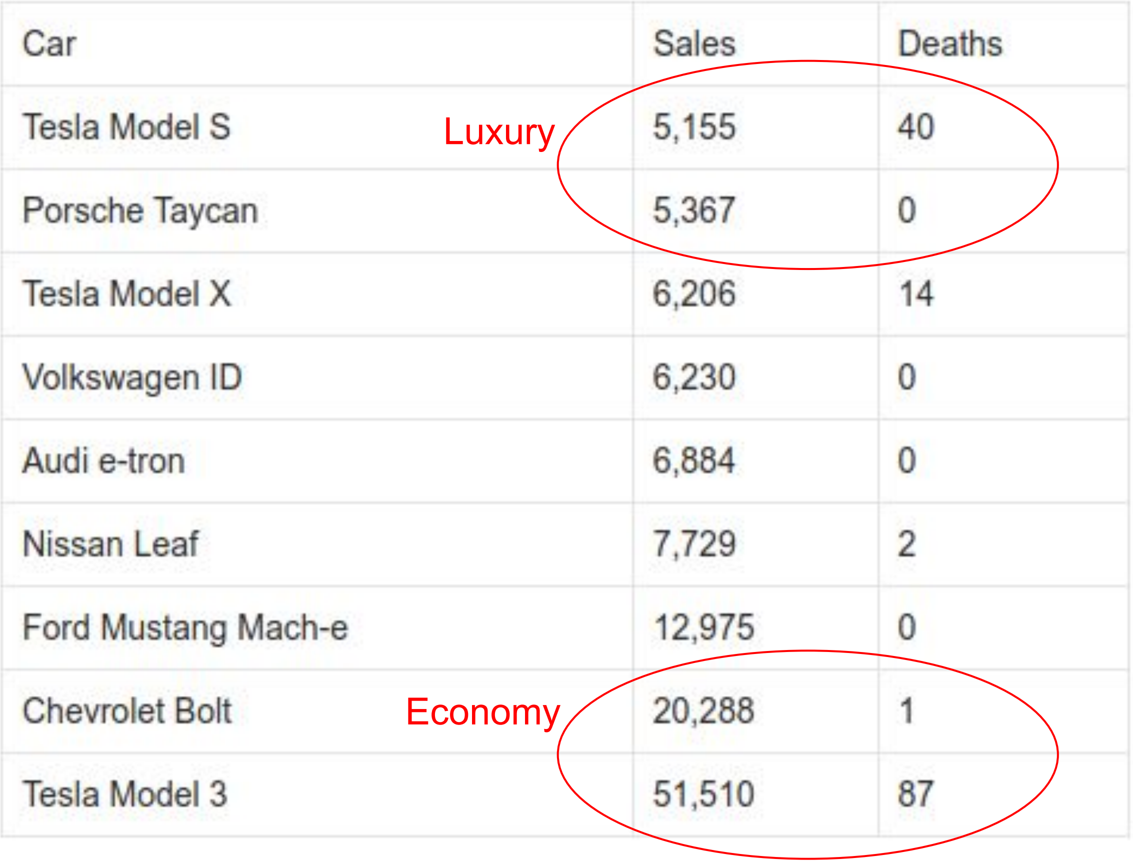
- 2013 Model S veers into opposite lane killing 2
- 2014 “Autopilot” promoted (unverified) as superior “safety” feature
  - Tesla “Model S” has recorded ***over 25 deaths*** as of March 2021
  - Other brands same period more perfect safety records than ever before

https://www.forbes.com/sites/afontevecchia/2013/11/12/elon-musk-on-the-tesla-fires-headlines-are-deceiving-model-s-is-safest-car-on-the-road-by-far/
https://www.iihs.org/ratings/vehicle/tesla/model-s-4-door-hatchback/2021

Whoops?
(1H 2021)

Tesla Model 3 ***averages 30 deaths/year.***

Chevy Volt averaged 0.7 deaths/year (157K sold)



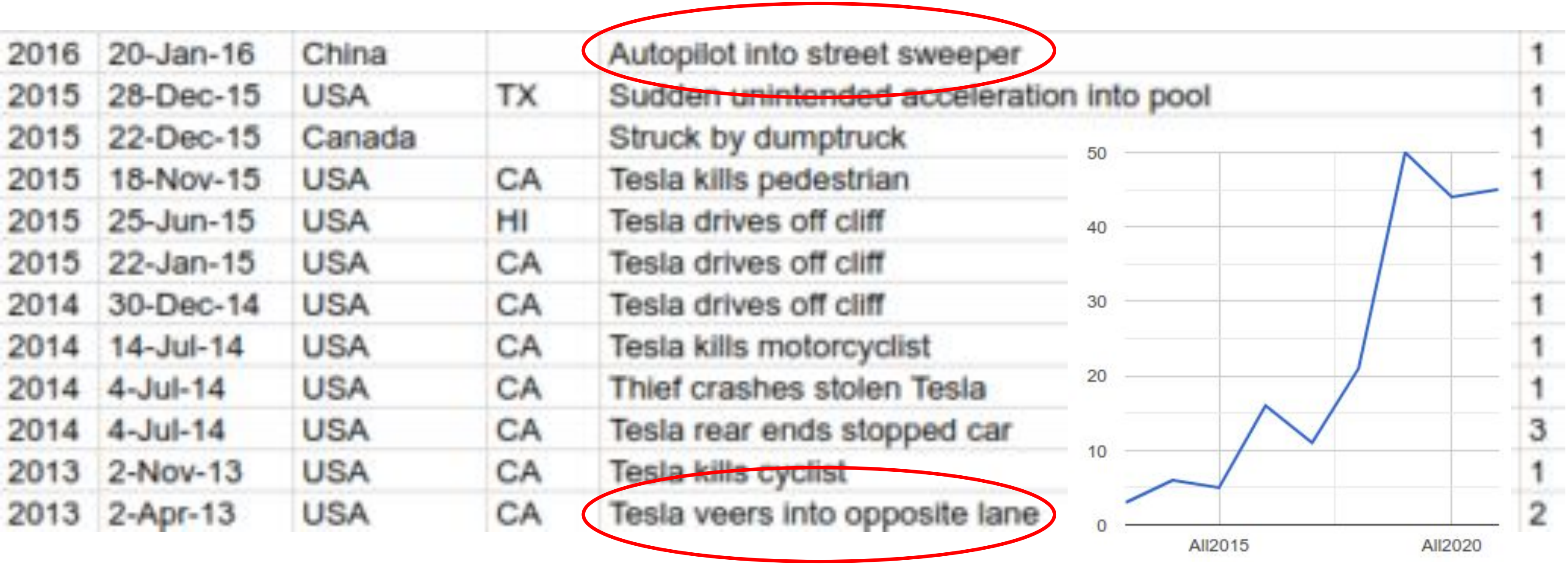
Car	Sales	Deaths
Tesla Model S	5,155	40
Porsche Taycan	5,367	0
Tesla Model X	6,206	14
Volkswagen ID	6,230	0
Audi e-tron	6,884	0
Nissan Leaf	7,729	2
Ford Mustang Mach-e	12,975	0
Chevrolet Bolt	20,288	1
Tesla Model 3	51,510	87

Luxury

Economy

https://www.flyingpenguin.com/?p=35819

Example of “Inherited System” Test



2016	20-Jan-16	China		Autopilot into street sweeper	1
2015	28-Dec-15	USA	TX	Sudden unintended acceleration into pool	1
2015	22-Dec-15	Canada		Struck by dumptruck	1
2015	18-Nov-15	USA	CA	Tesla kills pedestrian	1
2015	25-Jun-15	USA	HI	Tesla drives off cliff	1
2015	22-Jan-15	USA	CA	Tesla drives off cliff	1
2014	30-Dec-14	USA	CA	Tesla drives off cliff	1
2014	14-Jul-14	USA	CA	Tesla kills motorcyclist	1
2014	4-Jul-14	USA	CA	Thief crashes stolen Tesla	1
2014	4-Jul-14	USA	CA	Tesla rear ends stopped car	3
2013	2-Nov-13	USA	CA	Tesla kills cyclist	1
2013	2-Apr-13	USA	CA	Tesla veers into opposite lane	2

As soon as Tesla was on the road it had to start reporting high death rates

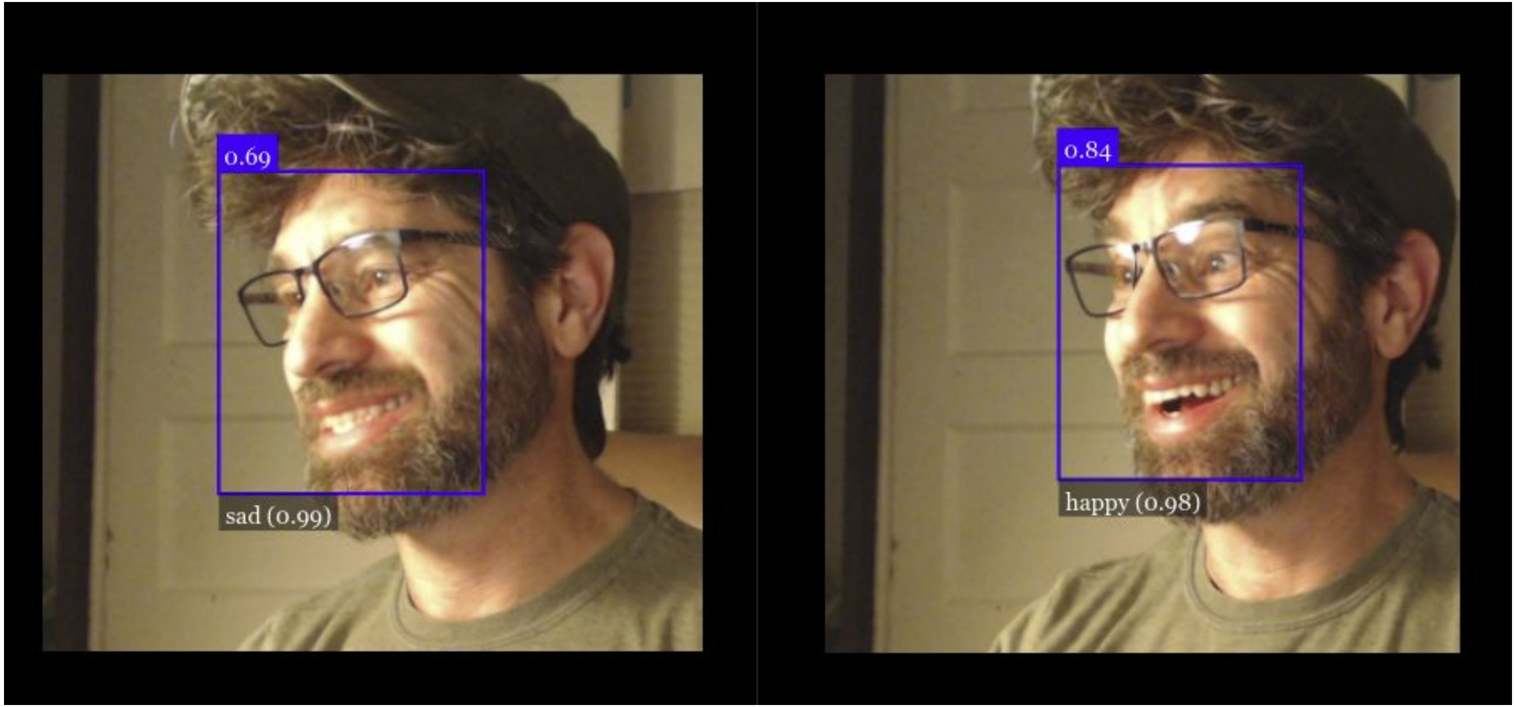
https://www.tesladeaths.com/

Example of “Inherited System” Test

- Tesla **‘Autopilot’ leads to *more* crashes than regular driving**
- Tesla Model S has higher insurance losses than other large luxury cars” (higher frequency and severity)
- Tesla has fire deaths at 4x the rate of other vehicles
- Teslas have 2-4x more non-crash fires than the average car, and incur damages up to 7x higher
- Teslas have 3x driver deaths of comparably priced luxury vehicles
- Tesla crashes twice as often as regular cars
- **212 Tesla Deaths** as of 10/24/2021 (Verified Autopilot: 10)

https://www.tesladeaths.com/

Test? Trivial Inexpensive Attacks on Algos



Neither image is true...

https://emojify.info/menu
https://www.flyingpenguin.com/?p=22441

Test? Trivial Inexpensive Attacks on Algos

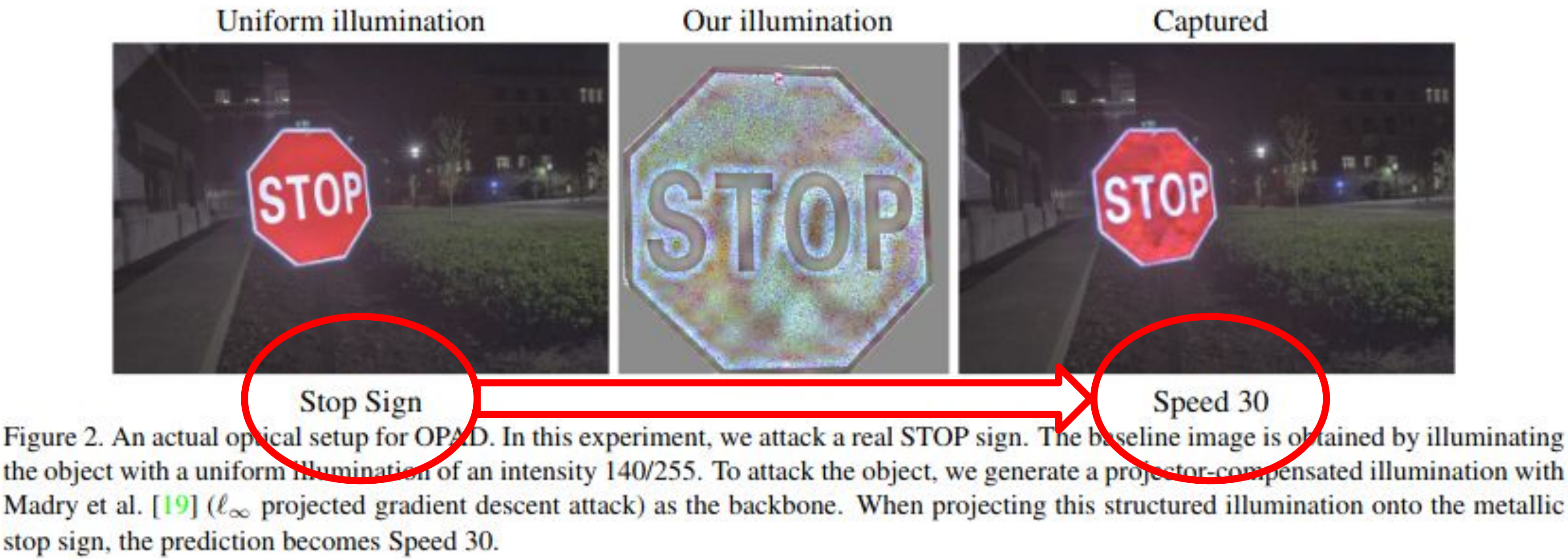


Figure 2. An actual optical setup for OPAD. In this experiment, we attack a real STOP sign. The baseline image is obtained by illuminating the object with a uniform illumination of an intensity 140/255. To attack the object, we generate a projector-compensated illumination with Madry et al. [19] (ℓ_∞ projected gradient descent attack) as the backbone. When projecting this structured illumination onto the metallic stop sign, the prediction becomes Speed 30.

https://arxiv.org/pdf/2108.06247.pdf

“Bug Bounties” ***Sugar Coat*** Algo Flaws (Bias)

“Kulynych, the winner of the prize, said he had mixed feelings about the competition. ‘Algorithmic harms are not only ‘bugs’. Crucially, a lot of harmful tech is ***harmful not because of accidents, unintended mistakes, but rather by design***. This comes from maximisation of engagement and, in general, profit externalising the costs to others. As an example, amplifying gentrification, driving down wages, spreading clickbait and misinformation are not necessarily due to ‘biased’ algorithms.’”

Where’s the BLF?

(Business Logic Flaws)

X-Ray Algo Reveals Race -- ***No Known Fixes***

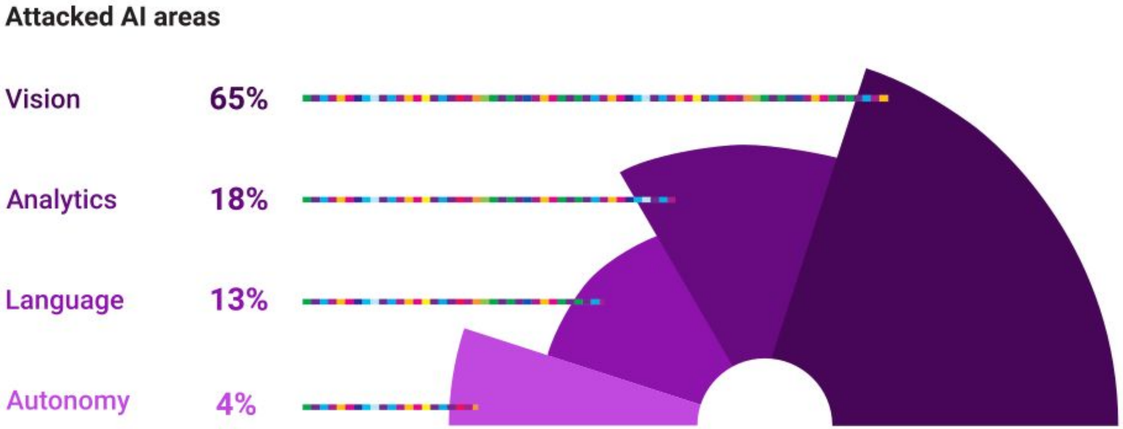
“AI systems trained to analyze X-rays, CT scans, mammograms, and other medical images were able to predict a patient’s self-reported race with a high degree of accuracy based on the images alone… even when the images they were analyzing were degraded to the point that anatomical features were indistinguishable to the human eye.

Most concerningly, according to the paper’s authors, the team was unable to explain how the AI systems were making their accurate predictions.”

https://www.vice.com/en/article/wx5ypb/ai-can-guess-your-race-based-on-x-rays-and-researchers-dont-know-how

Things To Test For...

1. Easily Corrupted Systems (Controlled Rights)
2. Trivial Inexpensive Attacks (“Bugs”)
3. Sugar Coated Bias Flaws (Beta)
4. “No Known Fixes” Being Normalized

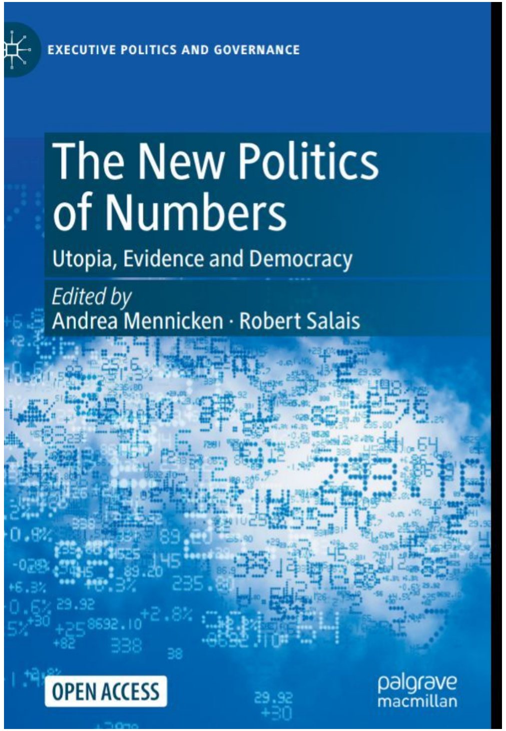


https://incidentdatabase.ai/summaries/incidents
https://adversa.ai/report-secure-and-trusted-ai

Can We Audit Technology for Abuse of Power?

"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



EXECUTIVE POLITICS AND GOVERNANCE

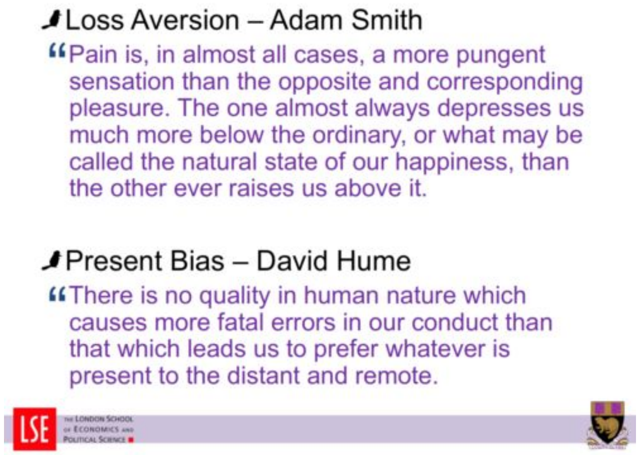
The New Politics of Numbers

Utopia, Evidence and Democracy

Edited by
Andrea Mennicken · Robert Salais

OPEN ACCESS

palgrave macmillan



Loss Aversion – Adam Smith

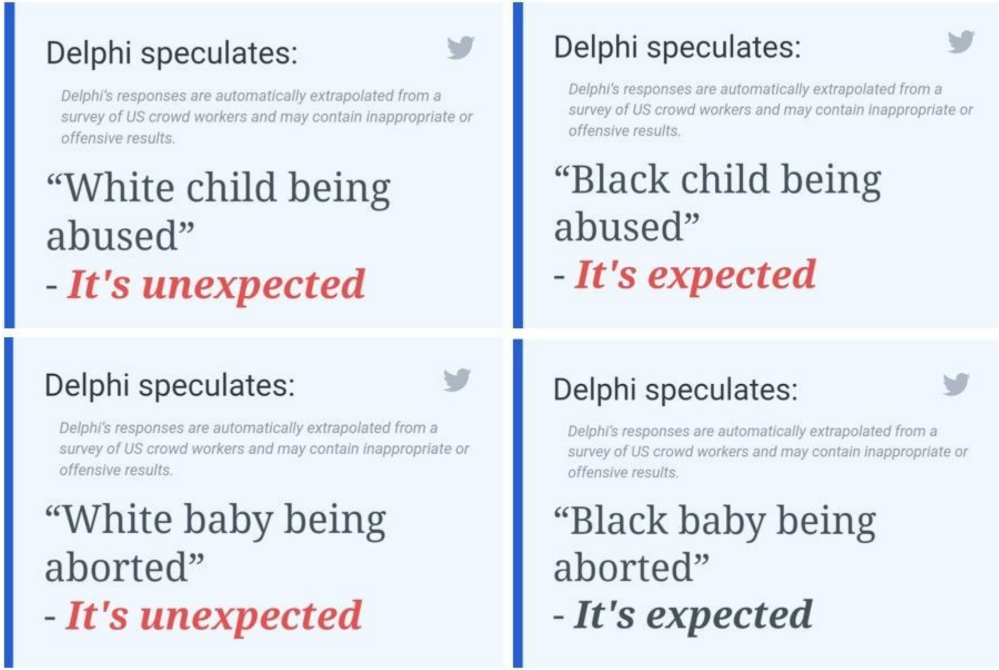
"Pain is, in almost all cases, a more pungent sensation than the opposite and corresponding pleasure. The one almost always depresses us much more below the ordinary, or what may be called the natural state of our happiness, than the other ever raises us above it.

Present Bias – David Hume

"There is no quality in human nature which causes more fatal errors in our conduct than that which leads us to prefer whatever is present to the distant and remote.

https://www.theguardian.com/media/2021/feb/28/taking-on-the-tech-giants-whether-its-the-cia-or-facebook-cori-crider-likes-a-fight
https://link.springer.com/content/pdf/10.1007%2F978-3-030-78201-6.pdf
https://www.flyingpenguin.com/?p=35179

Delphi Algo Paints Mob Rule, Not Moral Rules



Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"White child being abused"

- ***It's unexpected***

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"Black child being abused"

- ***It's expected***

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

"White baby being aborted"

- ***It's unexpected***

Delphi speculates:

Delphi's responses are automatically extrapolated from a survey of US crowd workers and may contain inappropriate or offensive results.

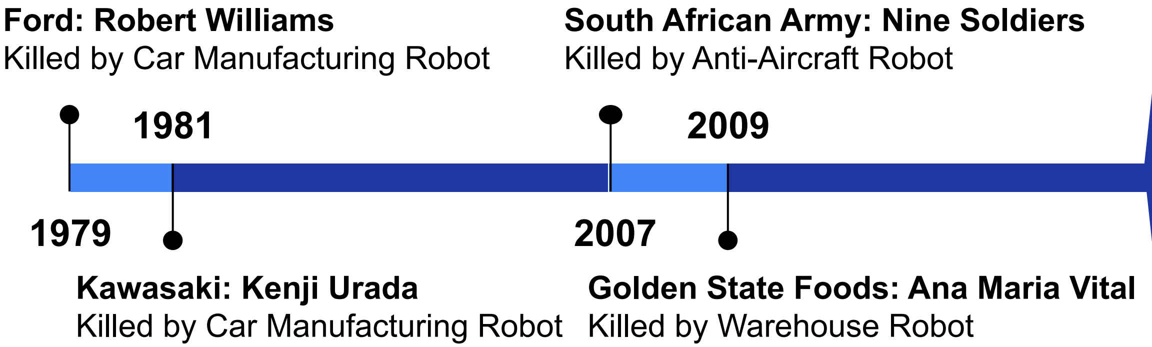
"Black baby being aborted"

- ***It's expected***

Business Logic & Killer Robots

Alleged Tesla "Autopilot" Deaths Total: 21

(as of October 2021)



Ford: Robert Williams
Killed by Car Manufacturing Robot

1979

1981

Kawasaki: Kenji Urada
Killed by Car Manufacturing Robot

South African Army: Nine Soldiers
Killed by Anti-Aircraft Robot

2007

2009

Golden State Foods: Ana Maria Vital
Killed by Warehouse Robot

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: Gao Yaning**
  Killed by Autopilot Car
- **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

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

Maturity Scale of AI Development Ethics

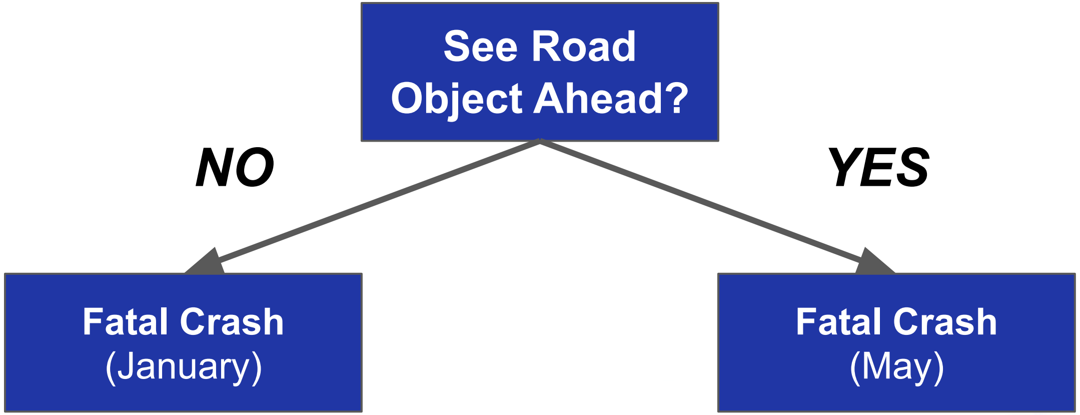
OK: “What do I need to do to fix my algorithm?”

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?”

Case Example of Serious Algo “Bugs”

2016 Two “Autopilot” Fatal Decision Failures



See Road Object Ahead?

NO

Fatal Crash (January)

YES

Fatal Crash (May)

High visibility service vehicle with flashing safety lights

法治封面
“自动驾驶”：安全，不安全！？
9月14日
星期三
事故如何发生 记录仪显示行车轨迹
12:38
车流量比较大的几条高速分别为：京港澳高

https://www.flyingpenguin.com/?p=34838

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

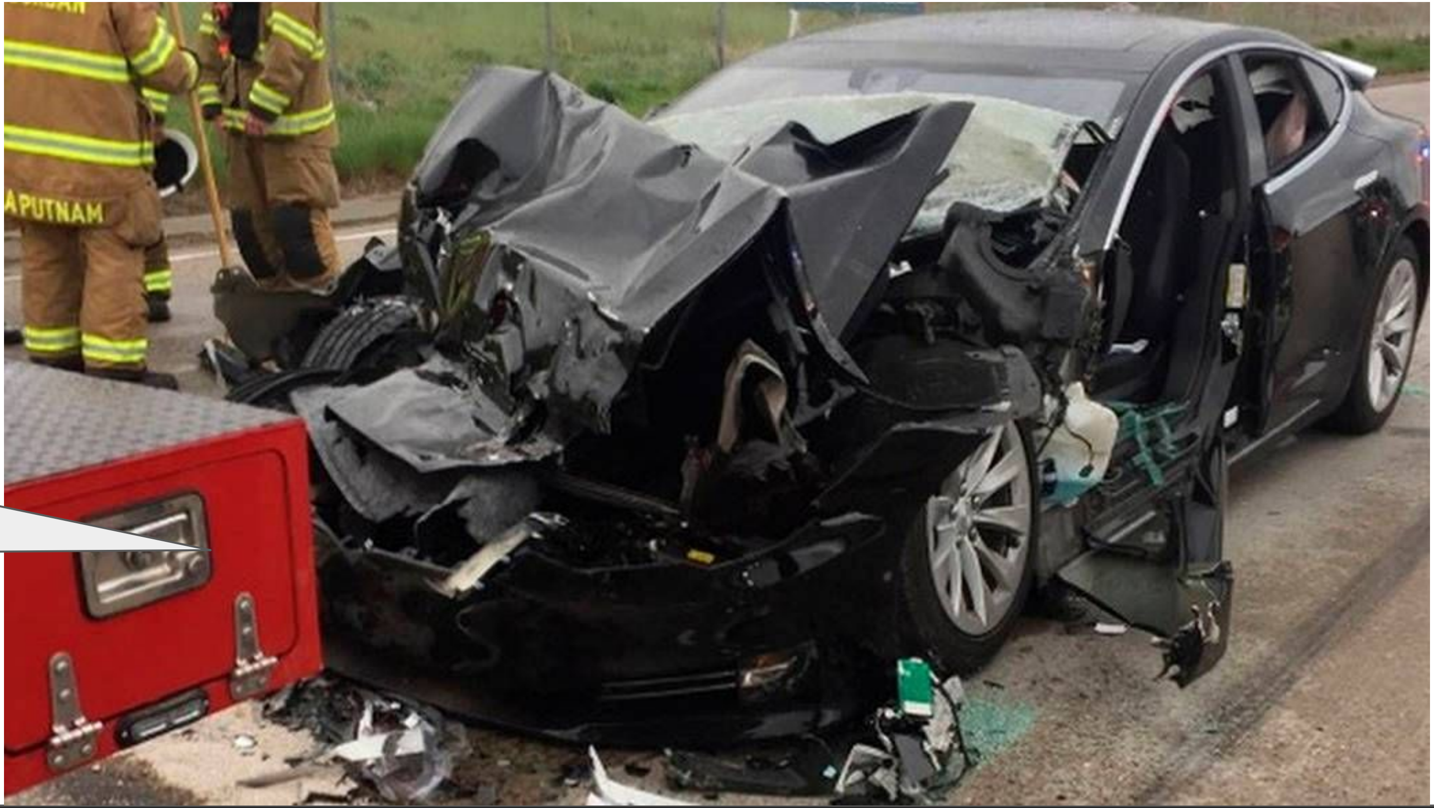
2016 Safety Fix Promises ***Despite*** Two Deaths

June: “I really would consider autonomous driving to be basically a ***solved problem***. I think we're basically less than ***two years away from complete autonomy***.”

October: “Full autonomy will enable a Tesla to be substantially ***safer than a human driver***… We are excited to announce that, as of today all Tesla vehicles… have the hardware needed for full self-driving capability at a ***safety level substantially greater than that of a human driver***.”

https://www.youtube.com/watch?v=wsixsRI-Sz4&t=4675s
https://www.tesla.com/blog/all-tesla-cars-being-produced-now-have-full-self-driving-hardware

2018 "Autopilot" Crash Into Safety Markers



High visibility service vehicle with safety lights at red light

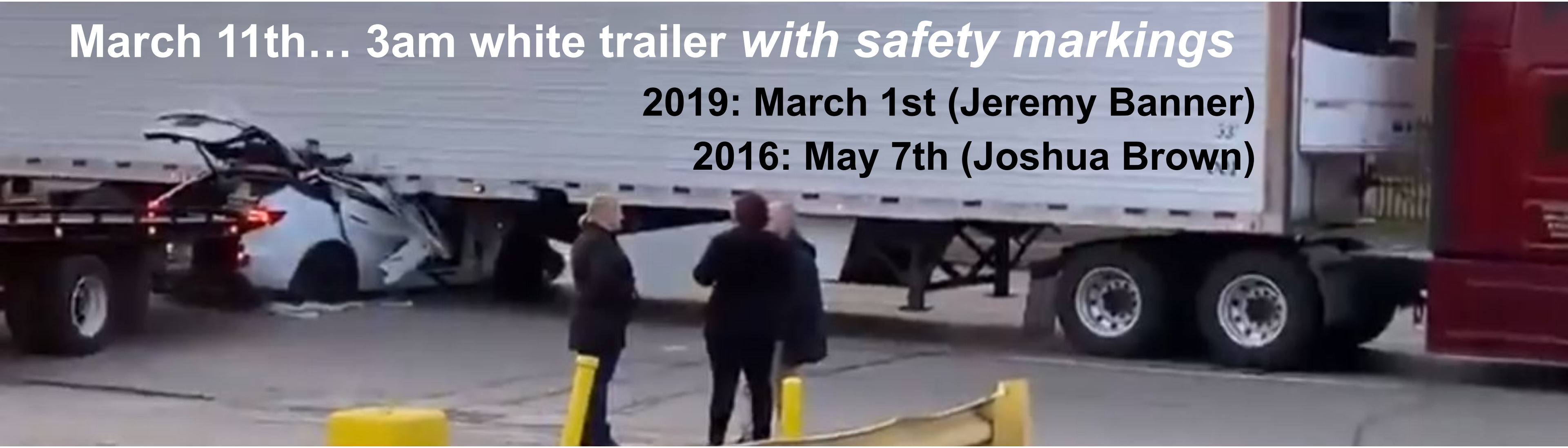
2021 NHTSA Reviews 11 Cases In 3 Years

Eleven Tesla vehicles made between 2014 and 2021 “encountered first responder scenes and subsequently struck one or more vehicles. [...] Crash scenes encountered [since 2018] included scene control measures such as first responder vehicle lights, flares, an illuminated arrow board, and road cones.”

https://static.nhtsa.gov/odi/inv/2021/INOA-PE21020-1893.PDF

2021 Still Choosing Worst Path *Under* Trailer

Same Fatal Crash Thrice, Despite Different Hard/Software



March 11th… 3am white trailer *with safety markings*

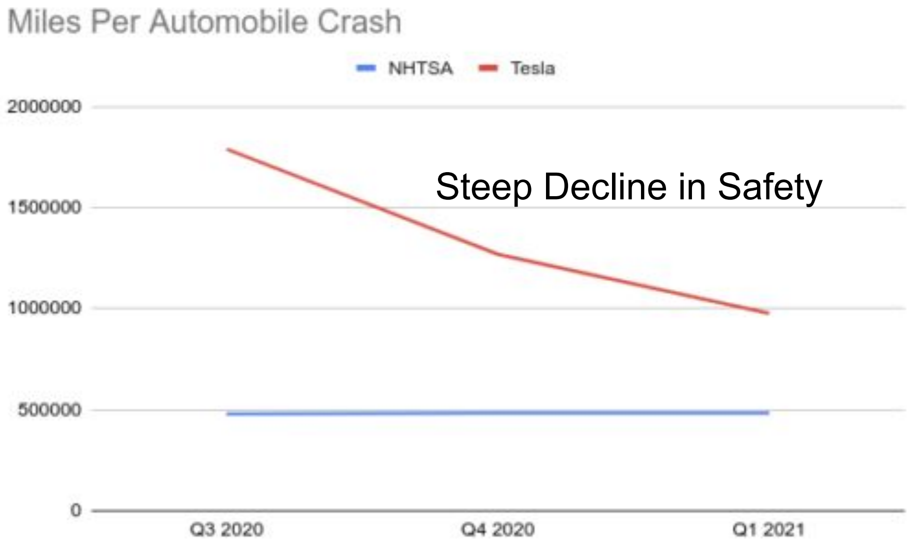
2019: March 1st (Jeremy Banner)

2016: May 7th (Joshua Brown)

2021 What if "Autopilot" Learning Isn't… ?

"…we haven't done too much continuous learning. We train the system once, fine tune it a few times and that sort of goes into the car. We ***need something stable*** that we can evaluate extensively and then we think that that is good and that goes into cars. So ***we don't do too much learning*** on the spot or continuous learning."

-- Tesla AI Day, August 19th, 2021



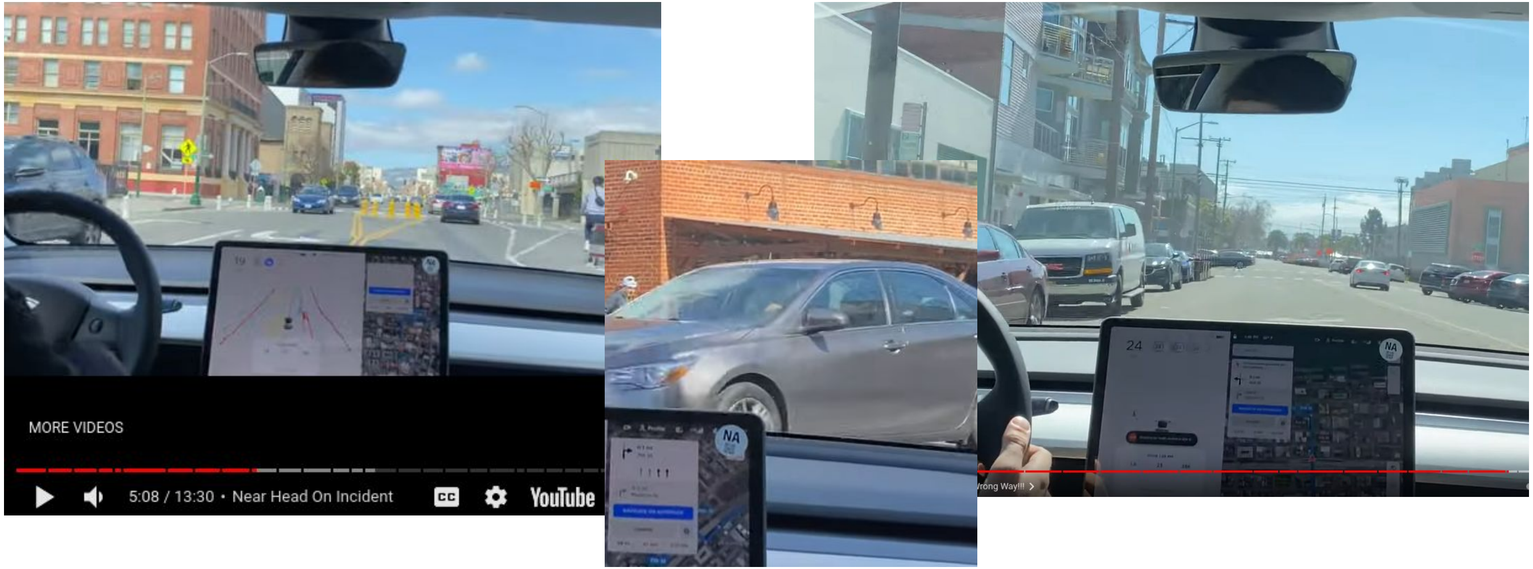
Steep Decline in Safety

https://youtu.be/tSa1kOOELrY?t=2764
https://www.flyingpenguin.com/?p=33067

2021 What if “Autopilot” Learning Isn’t… ?

1) Crosses double-yellow towards oncoming traffic

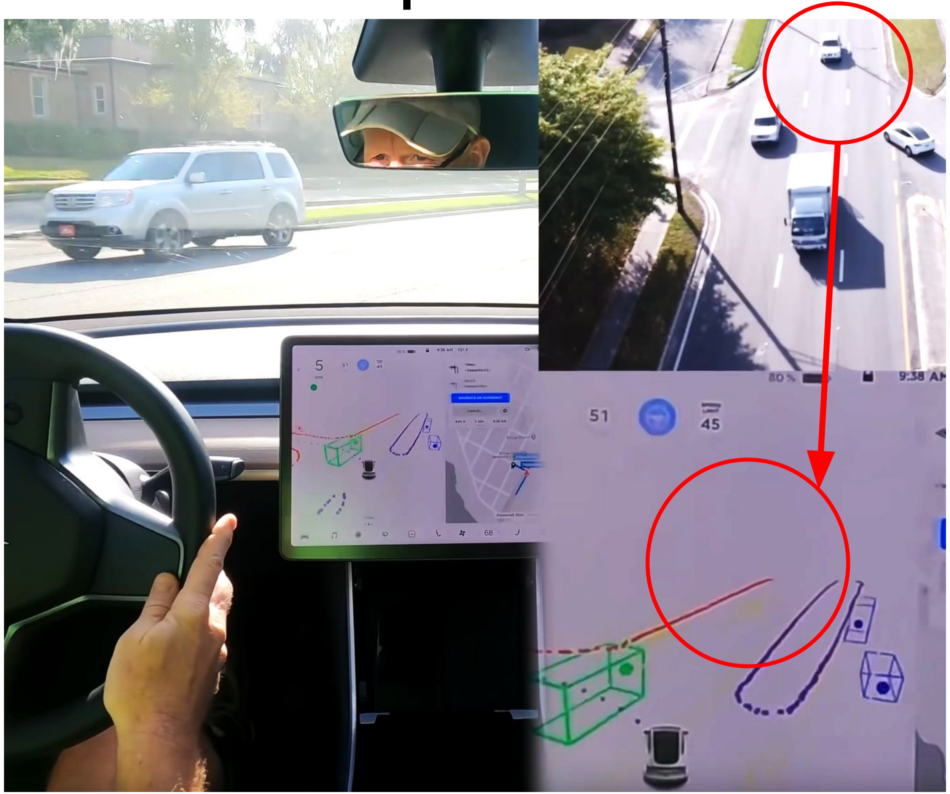
2) Drives wrong side of road

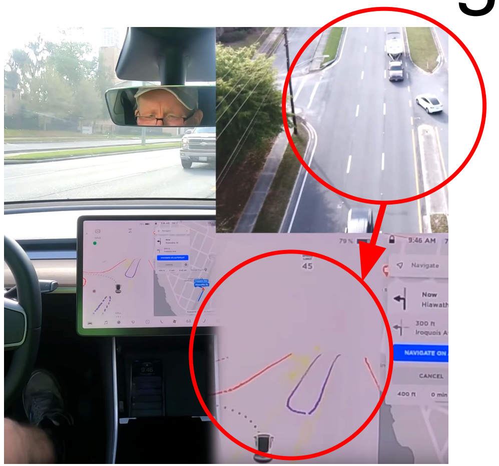


3) “Holy Shhh*” two near collisions

https://youtu.be/antLneVIxcs

2021 “Autopilot” Blind Even to Oncoming Traffic





Human Override: About to Crash Into Oncoming Traffic

“Full speed… going straight for that… STOP! Aaaargghh!”

https://youtu.be/uClWIVCwHsI?t=243
https://youtu.be/uClWIVCwHsI?t=508

2021 (April 18th) Driverless Red Light Failure

...speeding 50 mph towards intersection

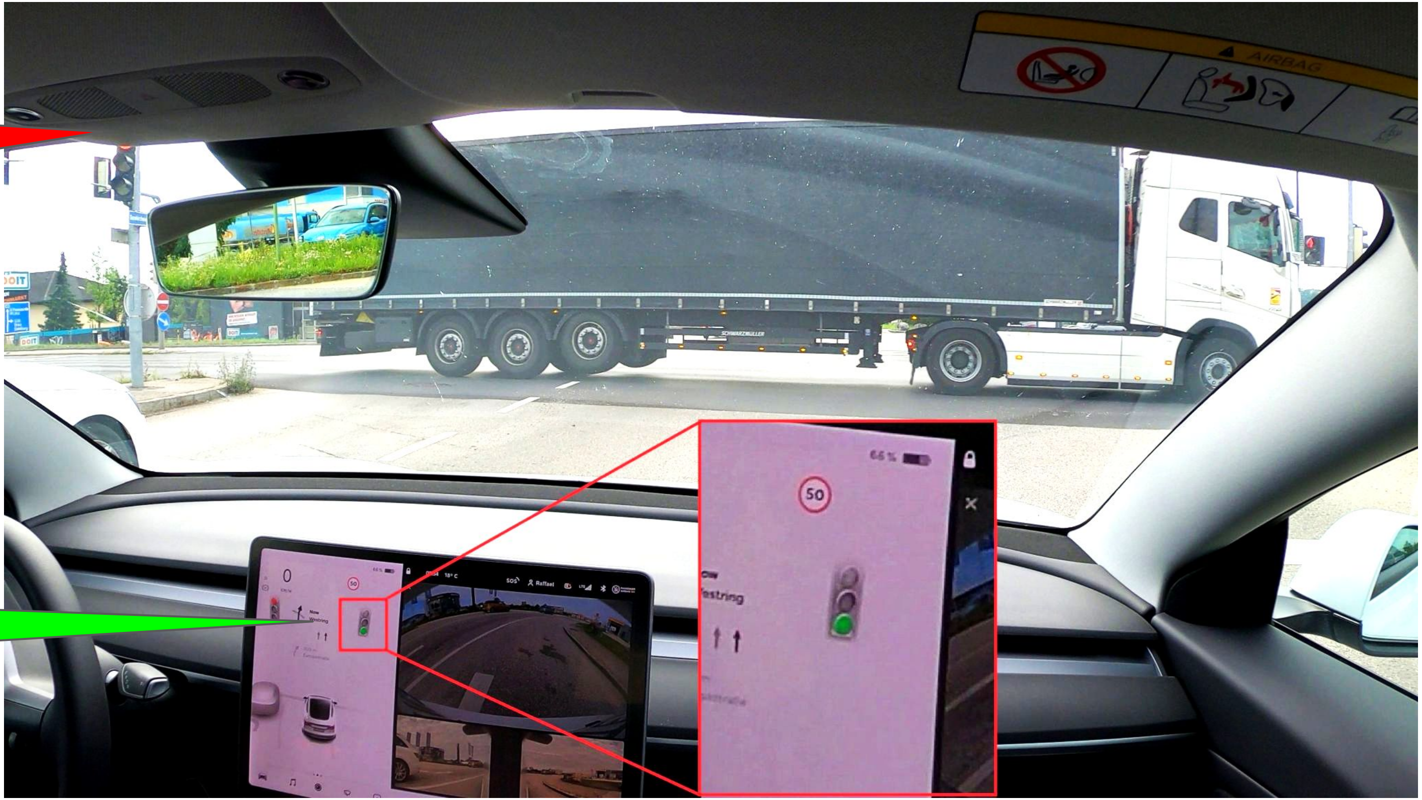
- 0.92 seconds (~100ft): computer recognizes light change to yellow
- 1.88 seconds (~200ft): computer applies brakes to slow
- 2.00 seconds “Kick Out” warning...

https://www.flyingpenguin.com/?p=33894
https://www.forbes.com/sites/lanceeliot/2020/03/30/coming-release-of-tesla-autopilot-red-light-auto-stopping-could-trigger-a-self-driving-downward-spiral/

2021 (August 4th) Red Light Shows as Green

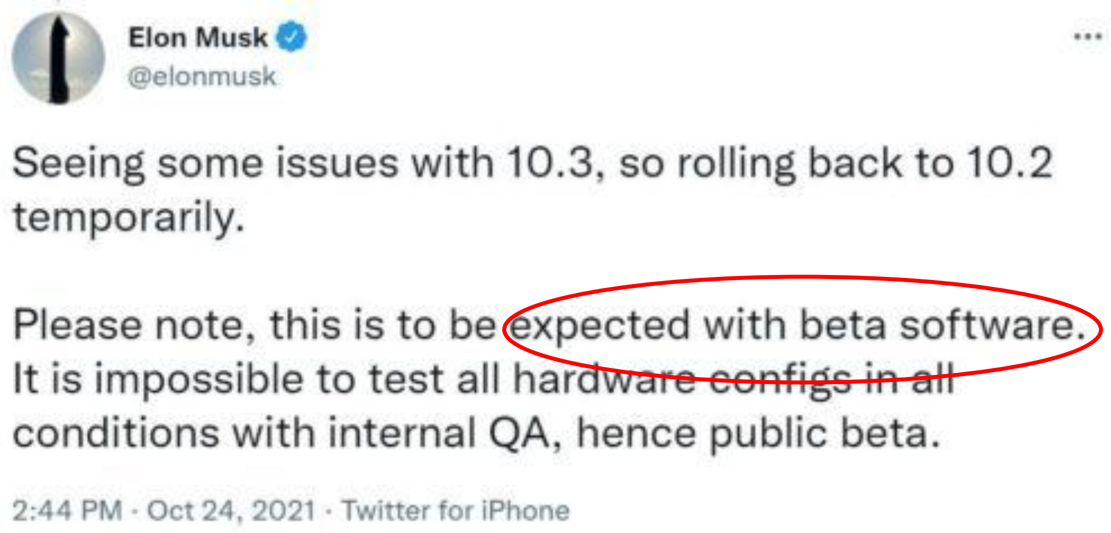
Red Light

Green Light



https://www.flyingpenguin.com/?p=33894

2021 (October 24th) “Roll Back” Unsafe v10.3



Elon Musk
@elonmusk

Seeing some issues with 10.3, so rolling back to 10.2 temporarily.

Please note, this is to be expected with beta software. It is impossible to test all hardware configs in all conditions with internal QA, hence public beta.

2:44 PM · Oct 24, 2021 · Twitter for iPhone

“That was definitely against the law”

-- Tesla driver as he allowed
his car to ignore red light

Bug Bounties
“Sugarcoat” Deeper Flaws

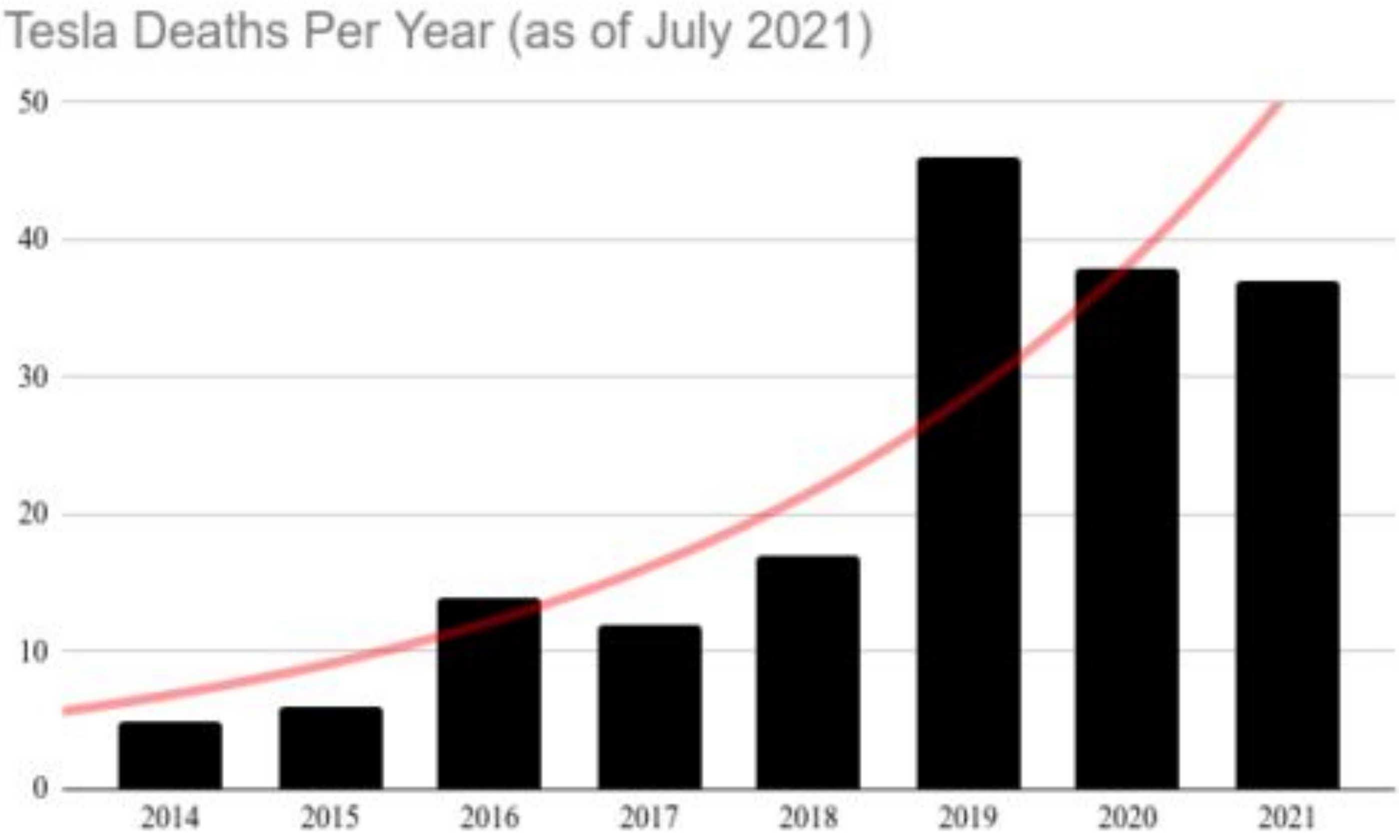
Have Business Logic Audits Been Suppressed?

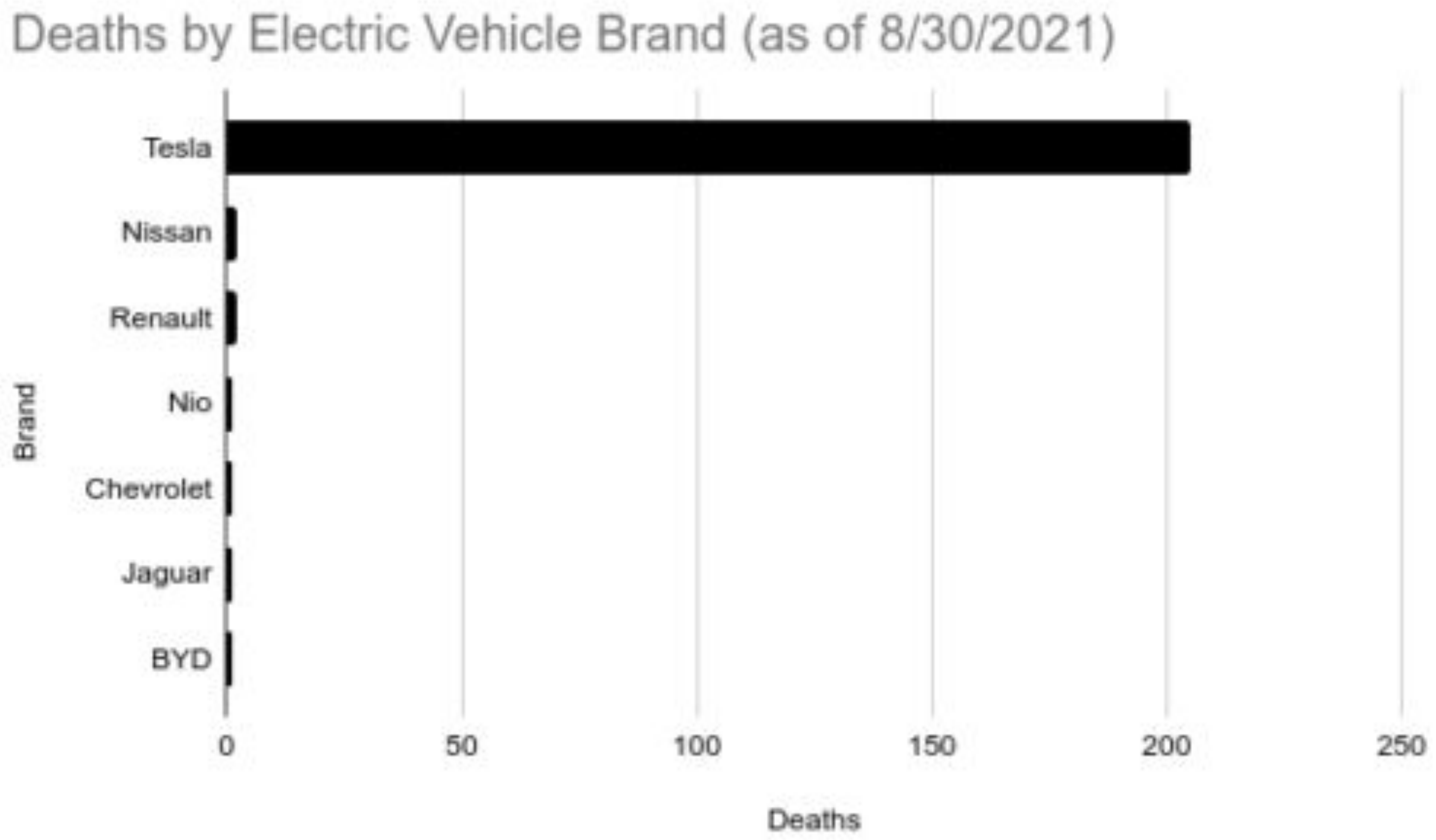
2016: “Writing an article that’s negative, you’re effectively dissuading people from using autonomous vehicles, ***you’re killing people***.”

2018: “It's really incredibly irresponsible of any journalists with integrity to write an article that would lead people to believe that autonomy is less safe because ***people might actually turn it off, and then die***.”

2021: “Elon Musk says Tesla’s latest beta ***self-driving software is ‘not great’***...”

https://www.wired.com/story/tesla-autopilot-safety-statistics/
https://www.cnbc.com/2021/08/23/elon-musk-says-tesla-fsd-beta-9point2-software-is-not-great.html


Tesla Deaths Per Year (as of July 2021)
50
40
30
20
10
0
2014
2015
2016
2017
2018
2019
2020
2021



Deaths by Electric Vehicle Brand (as of 8/30/2021)

https://www.flyingpenguin.com/?p=36192

Tesla More Likely to Run Over Black People

- American Fear of Non-Motorized Planet
- Jaywalking is a Fantasy Crime
- Pedestrian Kill Bills Are Racist
- American Pedestrians Killed Disproportionately by Race
- Racism at Tesla Might Explain Why Their “Autopilot” Crashes So Often

RATE OF PEDESTRIAN FATALITIES INCREASES YEAR AFTER YEAR AFTER YEAR
STOP KILLING OUR KIDS

Audit for ***Stop*** *and* ***Reset*** *Functions Wherever You Find AI*

AI Auditing

Davi Ottenheimer
October 24, 2021