

Developing Intelligent Systems for High-Stakes Business Applications

Mansur Maturidi Arief, Ph.D. (pronouns: Mansur, Mas, Kak, Dik, Pak)

Postdoctoral Scholar, Stanford Intelligent Systems Lab

Email: mansur.arief@stanford.edu | Website: www.mansurarief.github.io

About Me

- Born and raised in **Gowa (Sulawesi Selatan)**, went to Pesantren IMMIM Putra Makassar
- **Bachelor's:** TI ITS (2010)
 - Logistics & SCM lab assistant (2012-2014), ITS IO (2014), OSCM (2014)
- **Master's:** Industrial & Operations Eng., UMich, Ann Arbor (2018)
- **PhD:** Mechanical Engineering, Carnegie Mellon (2023)
 - RA-ship: CIT Dean Scholarship (2018), NSF (2019-2023)
 - Specialization: intelligent transportation, AI safety, applied ML
- **Postdoc:** Aeronautics & Astronautics Eng., Stanford (2023)
 - Working with 7 PhD, 2 master's, and 1 undergrad students on various projects
 - Collaborating with Stanford MineralX, CMU, ITS, IPB, Unhas, VKTR (Bakrie group)
 - Cofounder of Indonesian interdisciplinary scholars IndoSTEELERS and INTERSECT

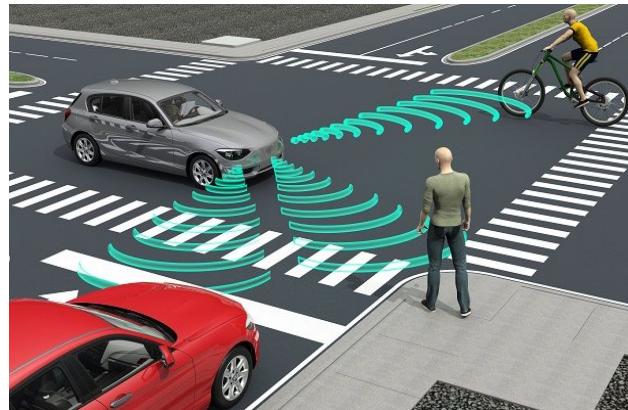


What we'll discuss

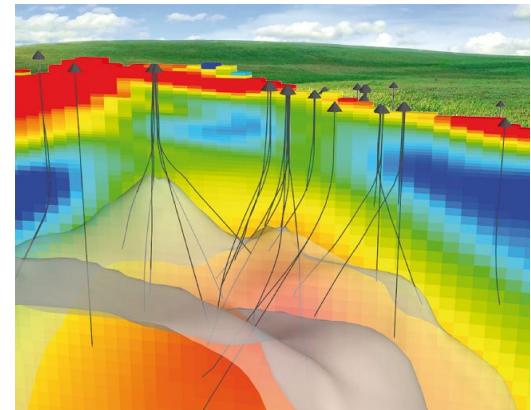
- **AI systems** are making (NOT only supporting) high-stakes decisions



Predicting future criminals



Self-driving cars



Geothermal exploration

What we'll discuss

- AI systems are making (NOT only supporting) high-stakes decisions
- Business leaders should mitigate their AI products **biases and real-world impacts**



https://youtu.be/gV0_raKR2UQ

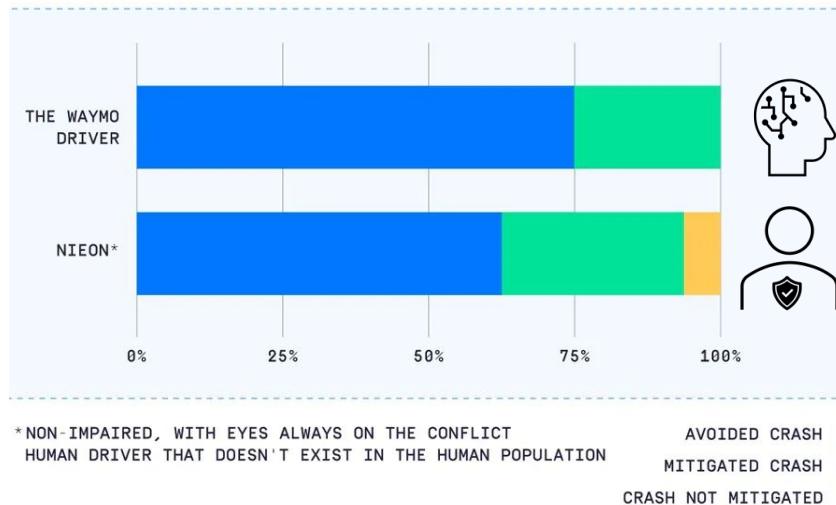
What we'll discuss

- AI **systems** are making (NOT only supporting) high-stakes decisions
- Business leaders should mitigate their AI products **biases and real-world impacts**
- **Safety-first** culture is key for AI-driven businesses to secure public trust

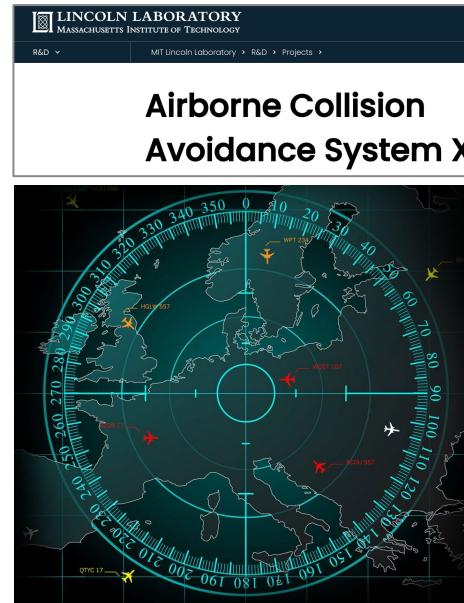


Intelligent systems systems are here... and becoming more reliable everyday

The Waymo Driver's collision avoidance performance in simulated tests



Source: <https://www.theverge.com/2022/9/29/23377219/waymo-av-safety-study-response-time-crash-avoidance>,
<https://waymo.com/waymo-one-san-francisco/>



A next-generation collision avoidance system will help pilots and unmanned aircraft safely navigate the airspace.

Are they safe and reliable enough?

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Two Drug Possession Arrests



Arrestee	Risk Level	Score
DYLAN FUGETT	LOW RISK	3
BERNARD PARKER	HIGH RISK	10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

AARIAN MARSHALL BUSINESS OCT 24, 2023 4:31 PM

GM's Cruise Loses Its Self-Driving License in San Francisco After a Robotaxi Dragged a Person

The California DMV says the company's autonomous taxis are "not safe" and that Cruise "misrepresented" safety information about its self-driving vehicle technology.



WIRED

AUTOMATION LEVELS OF AUTONOMOUS CARS		
LEVEL 0 	LEVEL 1 	LEVEL 2 
There are no autonomous features.	These cars can handle one task at a time, like automatic braking.	These cars would have at least two automated functions.
LEVEL 3 	LEVEL 4 	LEVEL 5 
These cars handle "dynamic driving tasks" but might still need intervention.	These cars are officially driverless in certain environments.	These cars can operate entirely on their own without any driver presence.

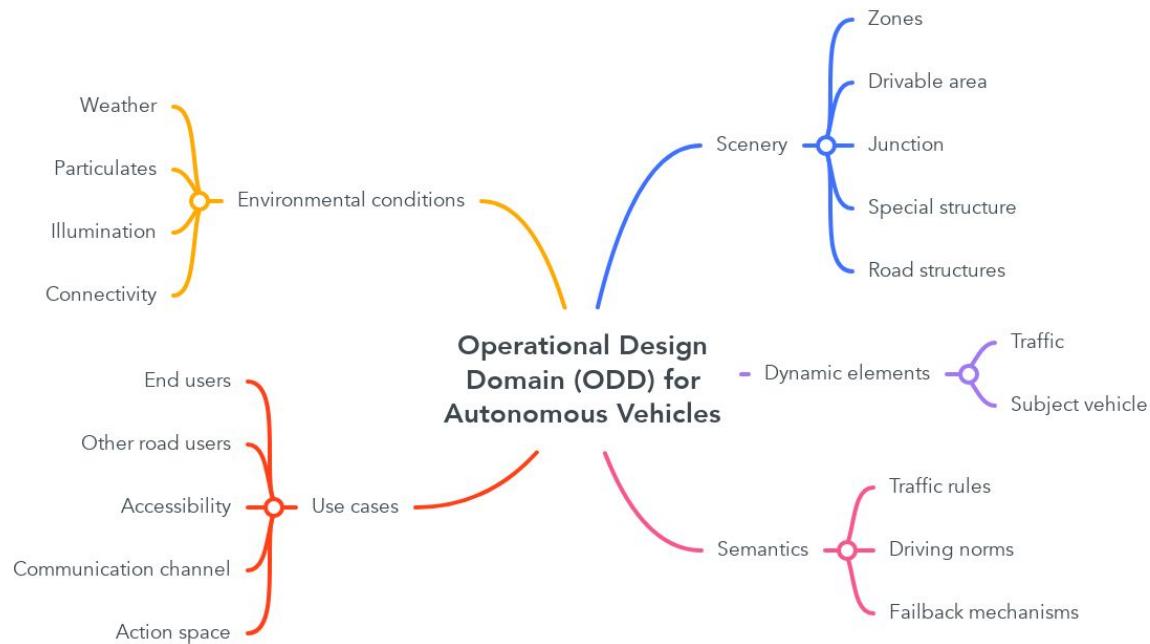
AI generally performs well **within** its operational design domains (ODDs).

	SAE LEVEL 0™	SAE LEVEL 1™	SAE LEVEL 2™	SAE LEVEL 3™	SAE LEVEL 4™	SAE LEVEL 5™
What does the human in the driver's seat have to do?	You <u>are driving</u> whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering	You <u>are not driving</u> when these automated driving features are engaged – even if you are seated in “the driver’s seat”				
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety	When the feature requests, you must drive	These automated driving features will not require you to take over driving			

Copyright © 2021 SAE International.

	These are driver support features			These are automated driving features	
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering OR brake/acceleration support to the driver	These features provide steering AND brake/acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met	This feature can drive the vehicle under all conditions
Example Features	<ul style="list-style-type: none"> automatic emergency braking blind spot warning lane departure warning 	<ul style="list-style-type: none"> lane centering OR adaptive cruise control 	<ul style="list-style-type: none"> lane centering AND adaptive cruise control at the same time 	<ul style="list-style-type: none"> traffic jam chauffeur 	<ul style="list-style-type: none"> local driverless taxi pedals/steering wheel may or may not be installed

AI generally performs well **within** its operational design domains (ODDs).



ODD specifies the conditions for which the system **is designed** to function properly.

Self-driving trucks:
We're in this for the
long haul.

— LEARN ABOUT DAIMLER TRUCK AND TORC



<https://torc.ai/trucking/>

ODD specifies the conditions for which the system **is designed** to function properly.

Self-driving trucks:
We're in this for the
long haul.

[LEARN ABOUT DAIMLER TRUCK AND TORC](#)

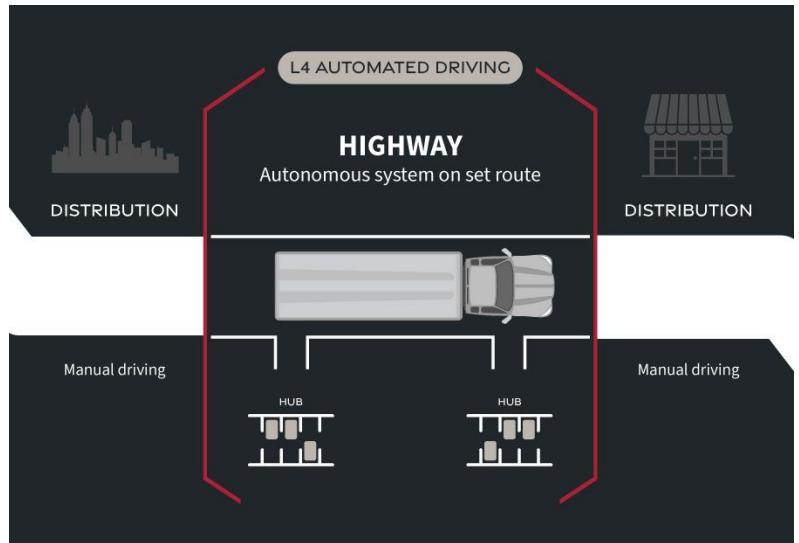


T Torc Robotics
<https://torc.ai/trucking> ::

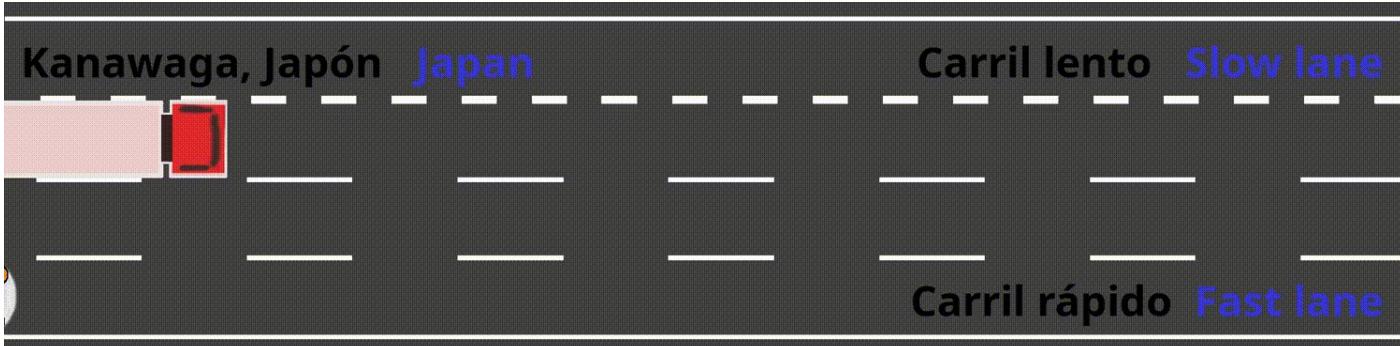
Self-Driving Trucks

Torc is developing [Level 4 self-driving freight trucks](#) to transform transportation, keep America running, and save lives.

<https://torc.ai/trucking/>



ODD specifies the conditions for which the system **is designed** to function properly.

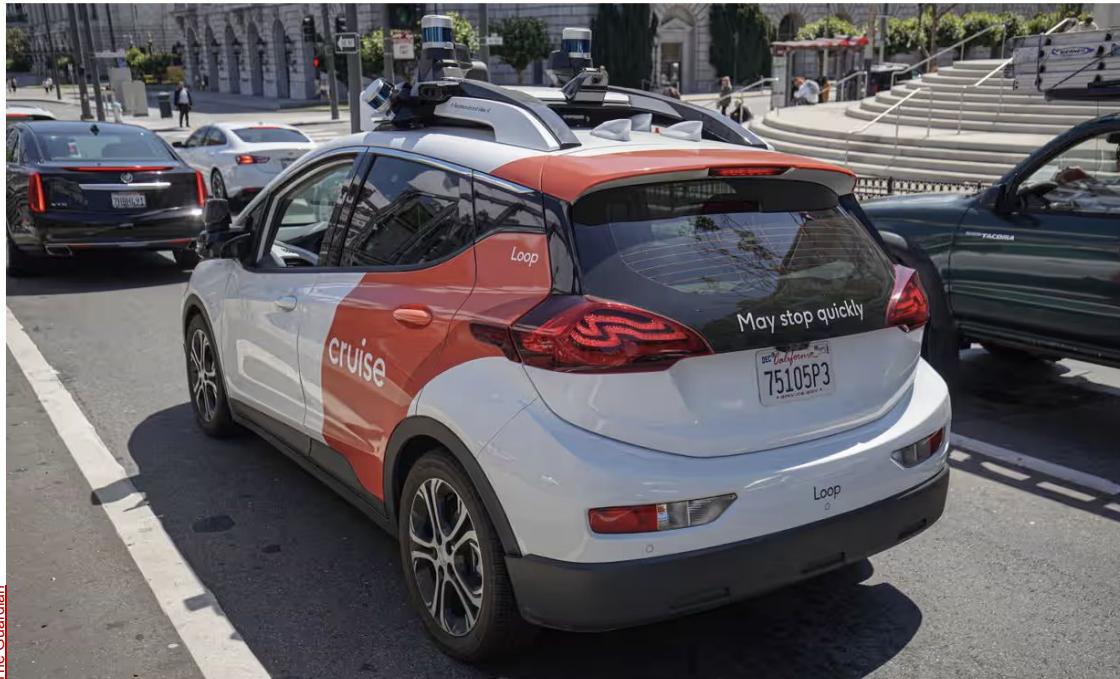


Tesla Autopilot

Safety issues arise when **deviating** from ODDs.

Self-driving car blocking road 'delayed patient care', San Francisco officials say

Cruise, the robotaxi firm, denies the city's claims its vehicle **blocked ambulance** which resulted in injured person's death



A driverless taxi of Cruise is seen on the road of San Francisco. Photograph: Michael Ho Wai Lee/SOPA Images/Shutterstock

Safety issues arise when **deviating** from ODDs.

Why is ODD important and should be made clear?

Because any designs/products have specifications (including Als).



Why ODD is important and should be made clear?

WIP 2021-07-15

Taxonomy & Definitions for Operational Design Domain (ODD) for Driving Automation Systems J3259

Per SAE J3016 (2021), the Operational Design Domain (ODD) for a driving automation system is defined as "Operating conditions under which a given driving automation system, or feature thereof, is specifically designed to function, including, but not limited to, environmental, geographical, and time-of-day restrictions, and/or the requisite presence or absence of certain traffic or roadway characteristics."; in short the ODD defines the limits within which the driving automation system is designed to operate, and as such, will only operate when the parameters described within the ODD are satisfied.. This information Report serves to provide terminology,

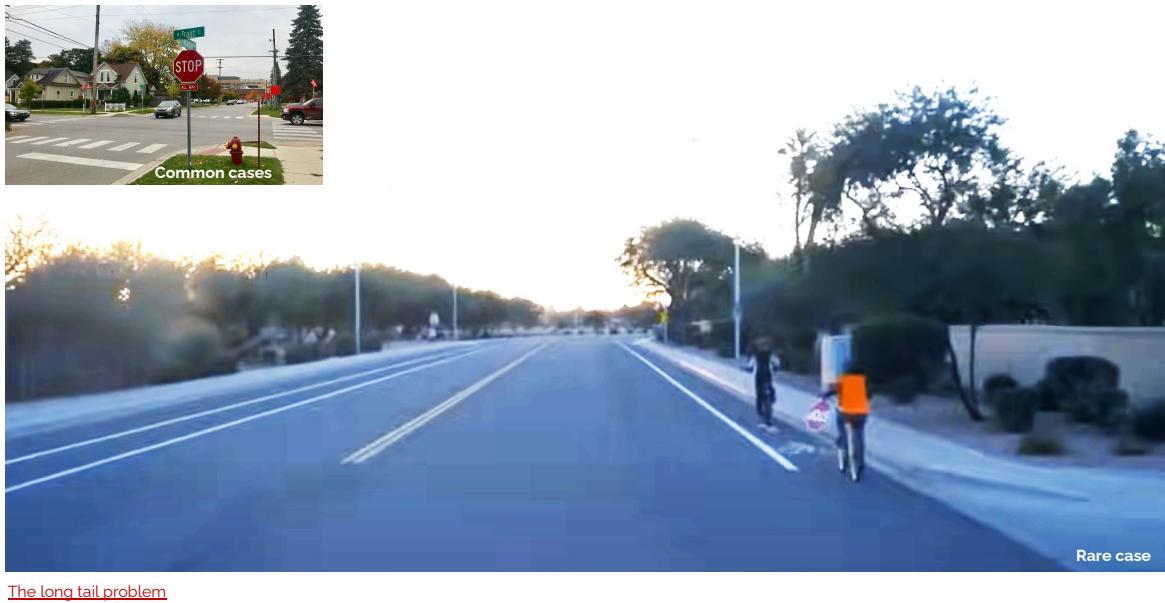
<https://www.sae.org/standards/content/j3259/>

1. Al's are **trained** with datasets, cost functions, or demos — **based on ODDs**.
2. **Novel**, unanticipated real-world **cases are inevitable** (the long tail problem).
3. The real world exhibits inherent **bias** and **dynamism**.
4. Algorithms based on gradient descent are susceptible to **adversarial attacks**.
5. Clear and well-communicated ODD both **aids end users** and **helps mitigate risks**.

An object detector trained on clean images may fail when encounter noisy images



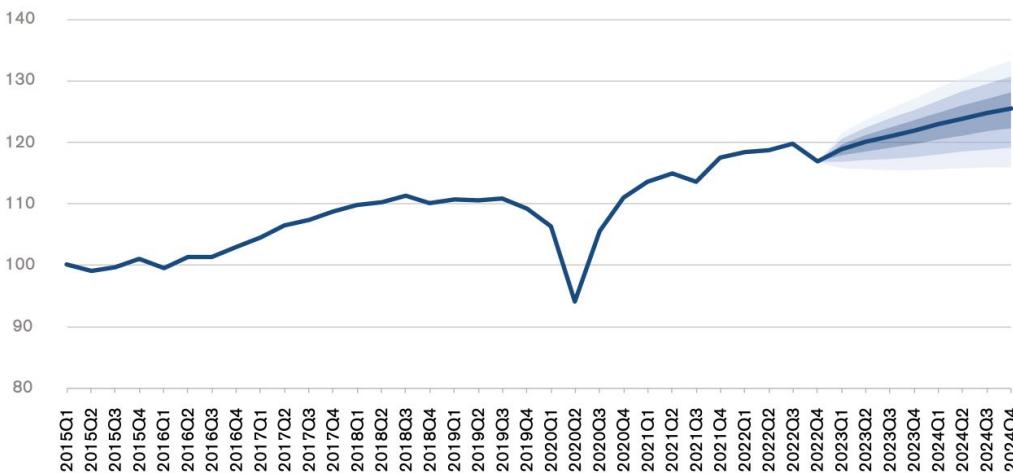
1. AIs are **trained** with datasets, cost functions, or demos — **based on ODDs**.



2. **Novel**, unanticipated real-world **cases** are **inevitable** (the long tail problem).

Chart 2: Volume of world merchandise trade, 2015Q1-2024Q4

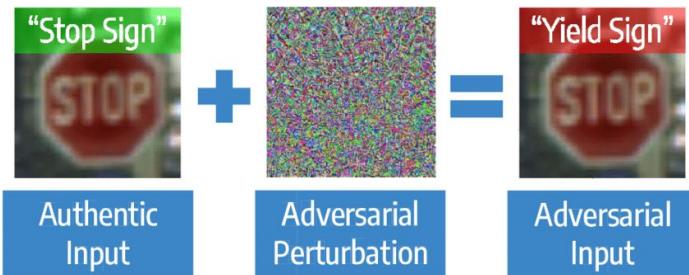
Seasonally-adjusted volume index, 2015=100



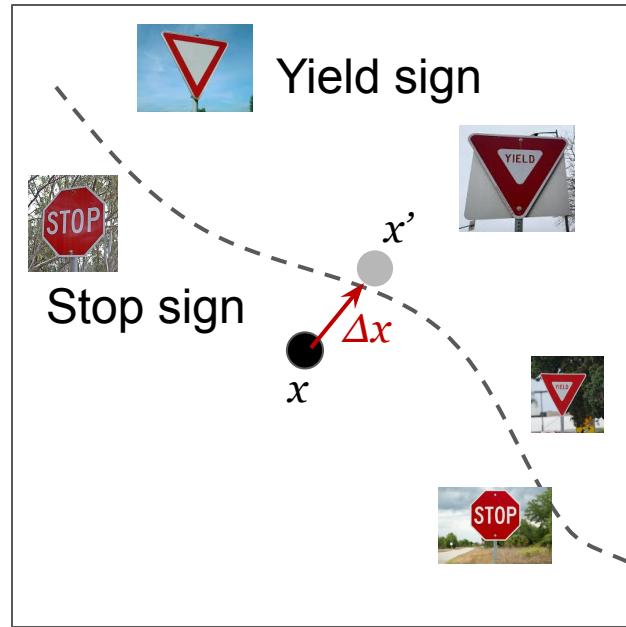
Note: The shaded region represents both random variation and subjective assessment of risk.

Source: WTO and UNCTAD for historical data, WTO Secretariat estimates for forecasts.

3. The real world exhibits inherent **bias** and **dynamism**.



$$\Delta x = \eta \cdot \nabla_x Error(f_\theta(x), y)$$



4. Algorithms based on gradient descent are susceptible to **adversarial attacks**.

Explaining and harnessing adversarial examples

[IJ Goodfellow](#), [J Shlens](#), [C Szegedy](#) - arXiv preprint arXiv:1412.6572, 2014 - arxiv.org

Several machine learning models, including neural networks, consistently misclassify adversarial examples---inputs formed by applying small but intentionally worst-case ...

☆ Save ⚡ Cite Cited by 18529 Related articles ☰



Yield sign



[PDF] [thecvf.com](#)

Robust physical-world attacks on deep learning visual classification

[K Eykholt](#), [I Evtimov](#), [E Fernandes](#)... - Proceedings of the ..., 2018 - openaccess.thecvf.com

... sign **classification**, o is the stop sign that we target to manipulate. Given images taken in the **physical world**... This approach aims to approximate **physical world** dynamics more closely. For ...

☆ Save ⚡ Cite Cited by 2220 Related articles All 20 versions ☰

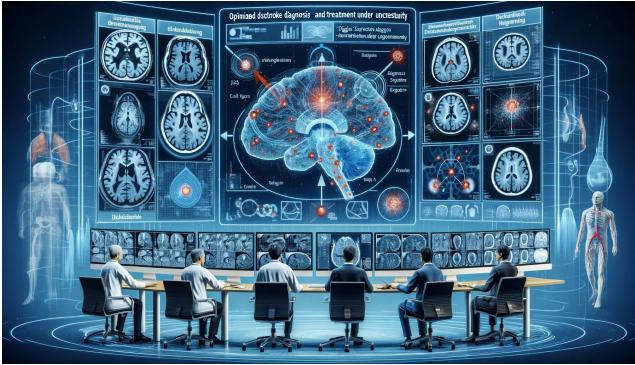


4. Algorithms based on gradient descent are susceptible to **adversarial attacks**.

The image shows a news article from Inverse. At the top left is the Inverse logo (three circles) and the word 'INVERSE'. At the top right are 'MENU' and three horizontal lines. Below the logo is a blue 'SCIENCE' button. The main title is 'Ford's Self-Driving Cars Will Live Inside Urban "Geofences"' in large black font. Below the title is a small note: '* BY JACK CROSBIE MARCH 13, 2017'. To the right of the text is a wide-angle photograph of a busy, brightly lit city street at night, likely Tokyo, with many neon signs and blurred lights from moving vehicles. A watermark 'Flickr / Stuck in Customs' is visible in the bottom right corner of the photo.

5. Clear and well-communicated ODD both **aids end users** and helps **mitigate risks**.

What does it mean for other business applications?



In healthcare, AI helps plan diagnostic tests and treatment for stroke patients.

Toward an integrated decision-making framework for optimized stroke diagnosis with DSA and treatment under uncertainty

Nur Ahmad Khatim, Azmul Asmar Irfan, Amaliya Mata'ul, Mansur M. Arief

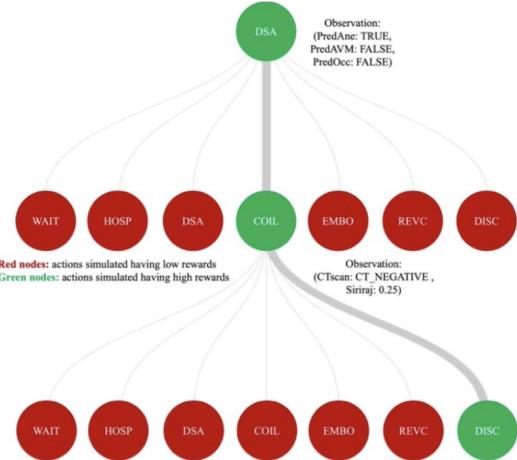
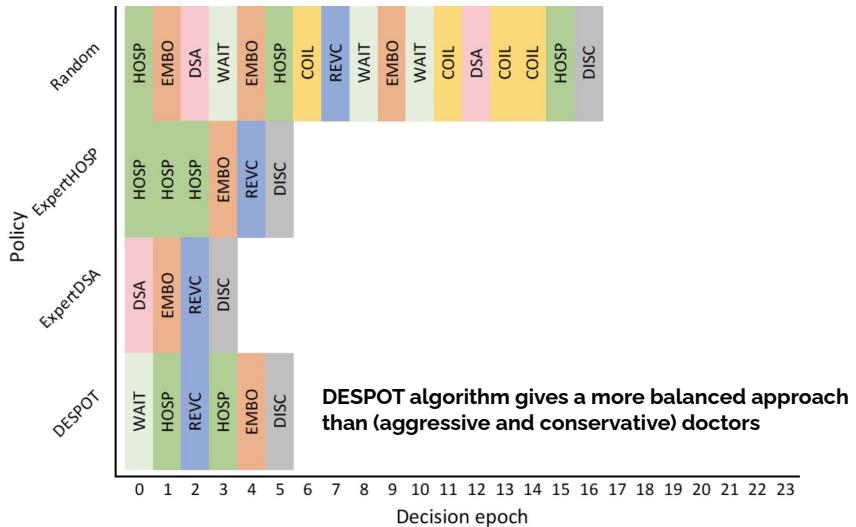


Fig. 1. Illustration of simulated policy rollouts in tree search



Toward an integrated decision-making framework for optimized stroke diagnosis with DSA and treatment under uncertainty

Nur Ahmad Khatim, Azmul Asmar Irfan, Amaliya Mata'ul, Mansur M. Arief

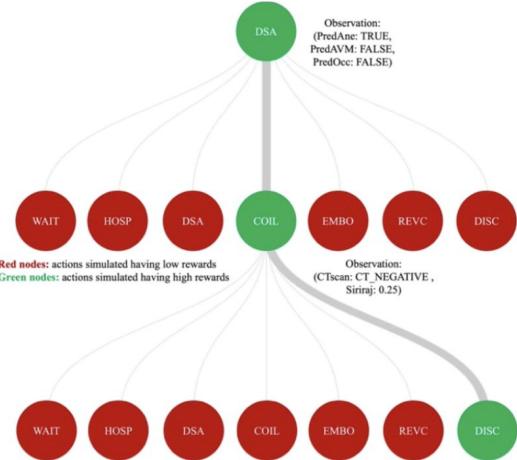
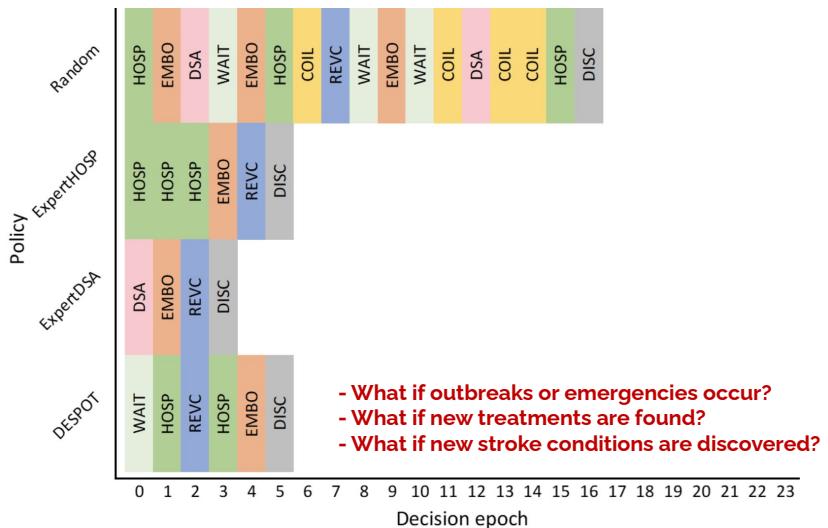


Fig. 1. Illustration of simulated policy rollouts in tree search





In criminal and justice systems, AI recommends recidivism to judges.

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

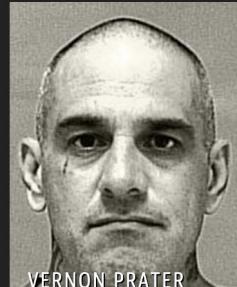
Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

Two Petty Theft Arrests



VERNON PRATER



BRISHA BORDEN

LOW RISK

3

HIGH RISK

8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Two Drug Possession Arrests



DYLAN FUGETT

LOW RISK

3



BERNARD PARKER

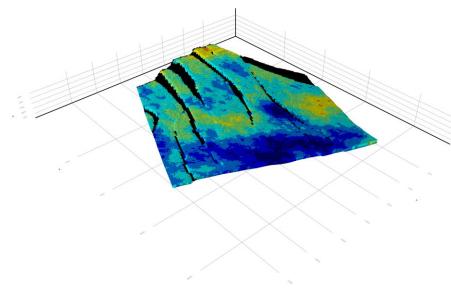
HIGH RISK

10

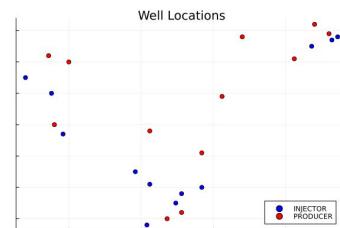
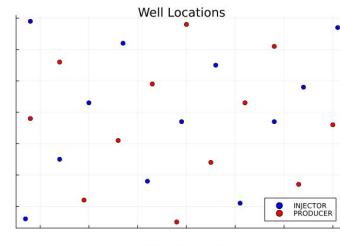
Fuggett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.



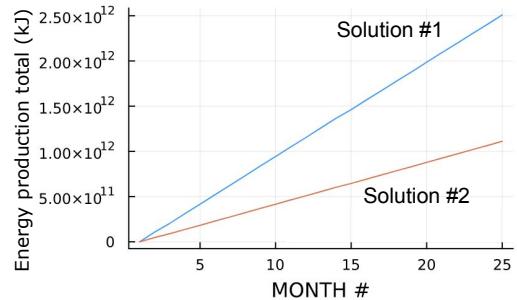
In renewable energy, AI recommends where to drill geothermal wells.



Alternative solutions

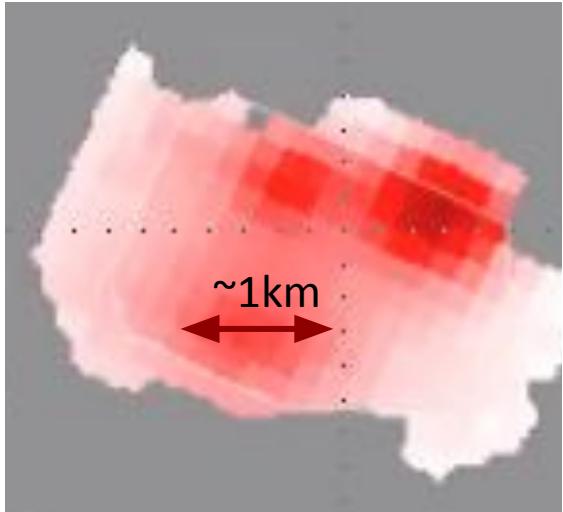
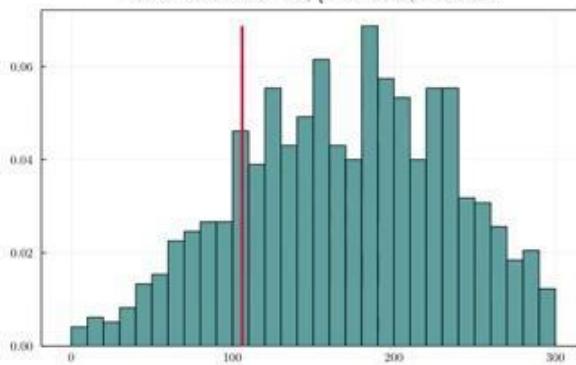
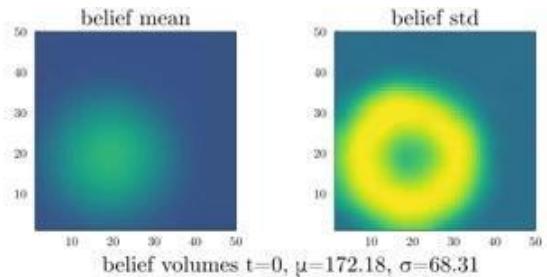


Evaluate solutions

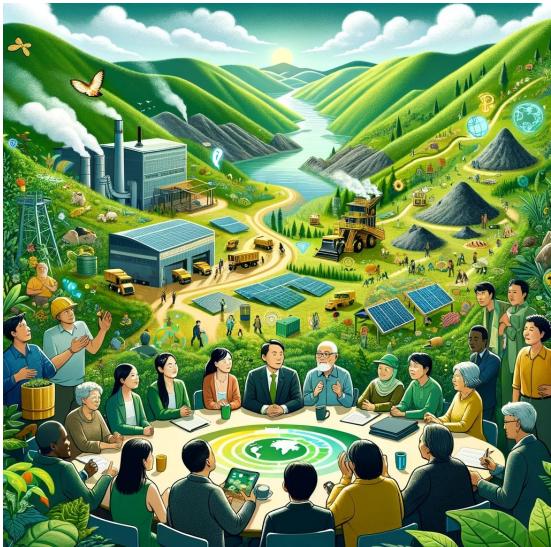


In renewable energy, AI recommends where to drill geothermal wells.

Surface EM anomalies

**Prediction Uncertainty**

Sorowako mining community engagement and empowerment



US-Australia lithium supply chain planning



Brazil responsible mining and post-mining reforestation



Risk analysis and validation

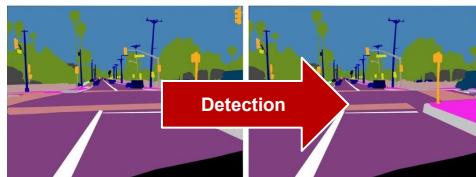
- If we use an AI component, how do we **manage** the risk?
- Need to know the **kinds of failure** and assess its **severity**.
- Plan for **mitigation strategy** based on the risk (probability \times severity).
- Be able to detect when the failure **is about to** occur.

Failure Modes and Effects Analysis (FMEA)

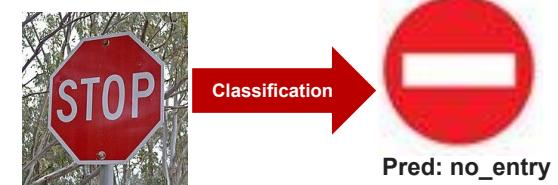
- FMEA identifies **all possible failures** in a design, a manufacturing or assembly process, or a product or service.



Inaccurate segmentation of images



Failure mode example: False negatives in detection



Failure mode example: Misclassified traffic signs

	Failure Mode 1	Failure Mode 2	Failure Mode 3
Effect	Inaccurate image segmentation	Traffic sign detection failure	Traffic sign misclassification
Cause	Failed traffic signs segmentation	Missed traffic sign	Incorrect traffic rule interpretation
Severity	Poor lighting, inaccurate pixel classifier	Environmental camouflage, occlusions	Inadequate training data, brittle model
Occurrence	High	9	High
Detection	Medium	5	High
	High	3	9
RPN	135	81	243

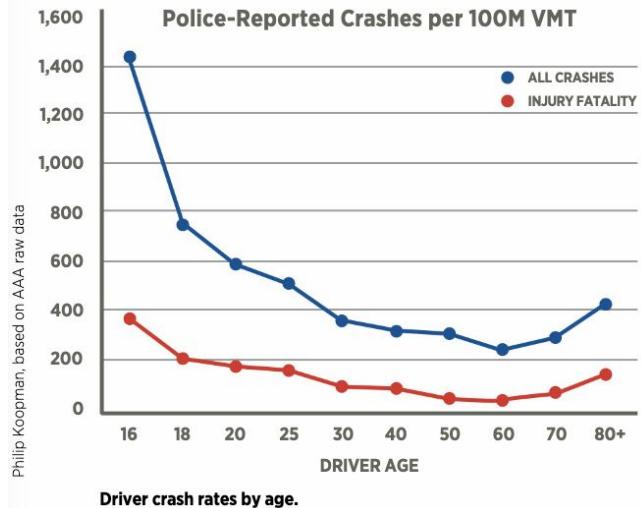
ML systems failure identification

- Domain knowledge (Out-of-Distribution cases)
- Grid search (or more systematic search)
- Adversarial attack

TITLE	CITED BY	YEAR
A survey on safety-critical driving scenario generation—A methodological perspective W Ding, C Xu, M Arief, H Lin, B Li, D Zhao IEEE Transactions on Intelligent Transportation Systems	50	2023

How to compare AI vs human reliability?

- People expect autonomous vehicle safety to be higher than human.
- Standards require critical components to have extremely low failure probability.
 - ISO 26262 (functional safety)
 - ISO 21448 (SOTIF)
 - UL4600 (Safety for the Evaluation of Autonomous Products)



Source: [SAE Update p. 31](#)

How to improve AI?

We can use the found failure modes to retrain our AI agent.

Enhancing Visual Perception in Novel Environments via Incremental Data Augmentation Based on Style Transfer

Abhibha Gupta¹, Rully Agus Hendrawan², Mansur Arief³

Abstract—The deployment of autonomous agents in real-world scenarios is challenged by "unknown unknowns", i.e. novel unexpected environments not encountered during training, such as degraded signs. While existing research focuses on anomaly detection and class imbalance, it often fails to address truly novel scenarios. Our approach enhances visual perception by leveraging the Variational Prototyping Encoder (VPE) to adeptly identify and handle novel inputs, then incrementally augmenting data using neural style transfer to enrich underrepresented data. By comparing models trained solely on original datasets with those trained on a combination of original and augmented datasets, we observed a notable improvement in the performance of the latter. This underscores the critical role of data augmentation in enhancing model robustness. Our findings suggest the potential benefits of incorporating generative models for domain-specific augmentation strategies.

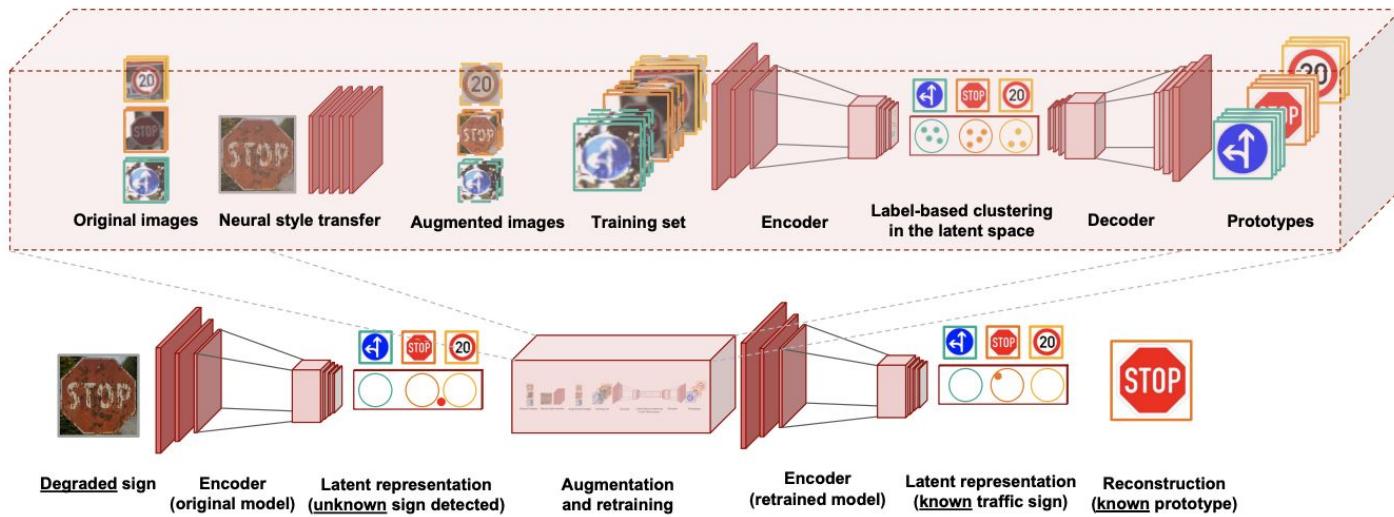


Fig. 1: Examples of degraded traffic signs in the real-world

examples of the underrepresented class are available in the training set. In contrast, unknowns emerge when training data are not available for certain cases in the real world [13]. For instance, a traffic sign that has been heavily invaded by rust may not be present in the training set, and such cases will eventually occur during deployment. Arguably, the

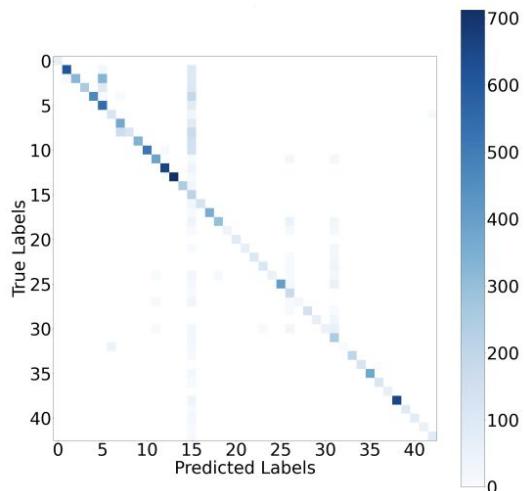
How to improve AI?

We can use the found failure modes to retrain our AI agent.

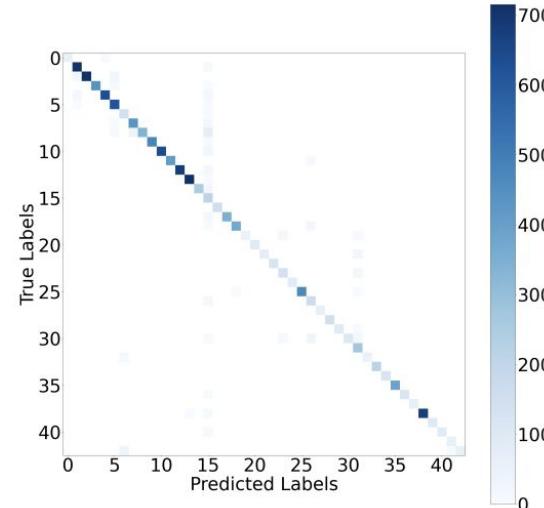


How to improve AI?

We can use the found failure modes to retrain our AI agent.



Confusion matrix when AV sees
degraded traffic signs



Confusion matrix post retraining

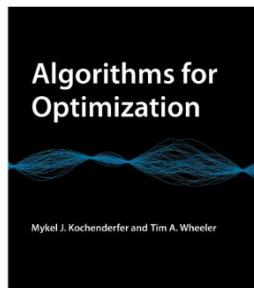
What we have discussed

- **AI systems** are making (NOT only supporting) high-stakes decisions
- **Business leaders** should mitigate their AI products **biases and real-world impacts**
- **Safety-first** culture is key for AI-driven businesses to secure public trust

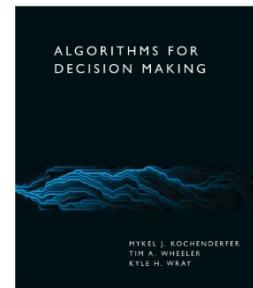
FUNCTIONAL SAFETY SUPPORT THROUGHOUT THE DEVELOPMENT CYCLE



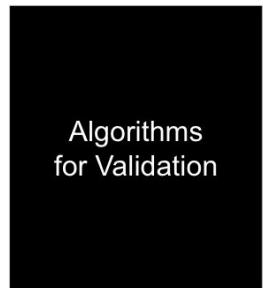
A lot more needs to be done...



2019



2022



Algorithms
for Validation

Coming soon!

What does it mean for your business?

Thanks to collaborators!

- Mykel Kochenderfer, AeroAstro Stanford
- Ding Zhao, MechE CMU
- Henry Lam, IEOR Columbia
- Bo Li, CS UIUC
- Zhiyuan Huang, SoM Tongji
- Huan Zhang, EE UIUC
- Iwan Vanany, ISE ITS
- Jef Caers, MineralX Stanford
- Nur Ahmad Khatim, IF ITS
- Yan Akhra, ISE ITS
- Azmul Asmar, FK UIN SH
- Amaliya Mata'ul, FK UIN SH
- Rully Hendrawan, SCS Pitt (IS ITS)
- Abhibha Gupta, SCS Pitt
- Yasmine Alonso, CS Stanford
- Anthony Corso, AeroAstro Stanford
- IndoSTEELERS members
- SISL members
- CMU Safe AI members

Let's stay in touch

Mansur Maturidi Arief

Email: mansur.arief@stanford.edu

Web: <https://mansurarief.github.io/>