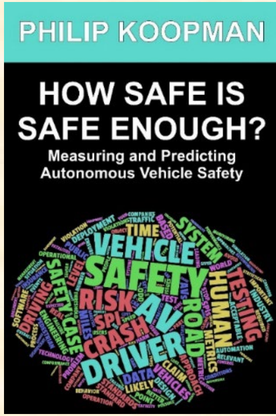
Prof. Philip Koopman

Machine Learning Capabilities for Applications

Carnegie Mellon University

Business of Semiconductor Summit 2024

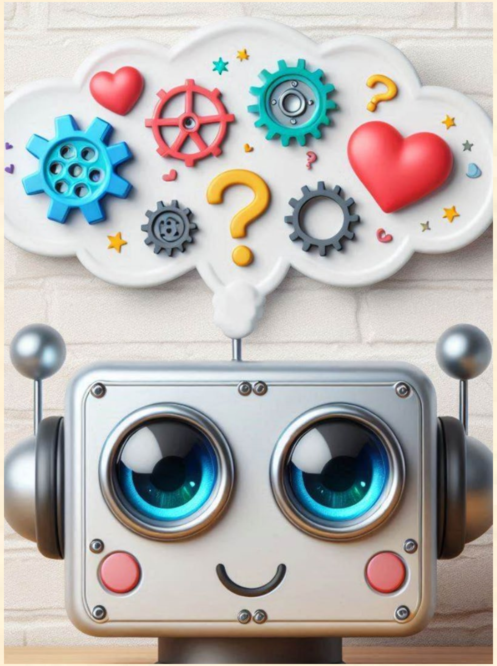
www.Koopman.us



Overview

Let's discuss capabilities rather than mechanisms

- Machine Learning-based AI (ML) useful capabilities
  - Classification, End-to-End ML
  - Generative AI, Large Language Models
- Challenges to using ML
  - "Bias," "Hallucinations," validation
- Practical ML issues
  - Edge cases, accountability, autonowashing
  - AI Safety



[AI GENERATED]

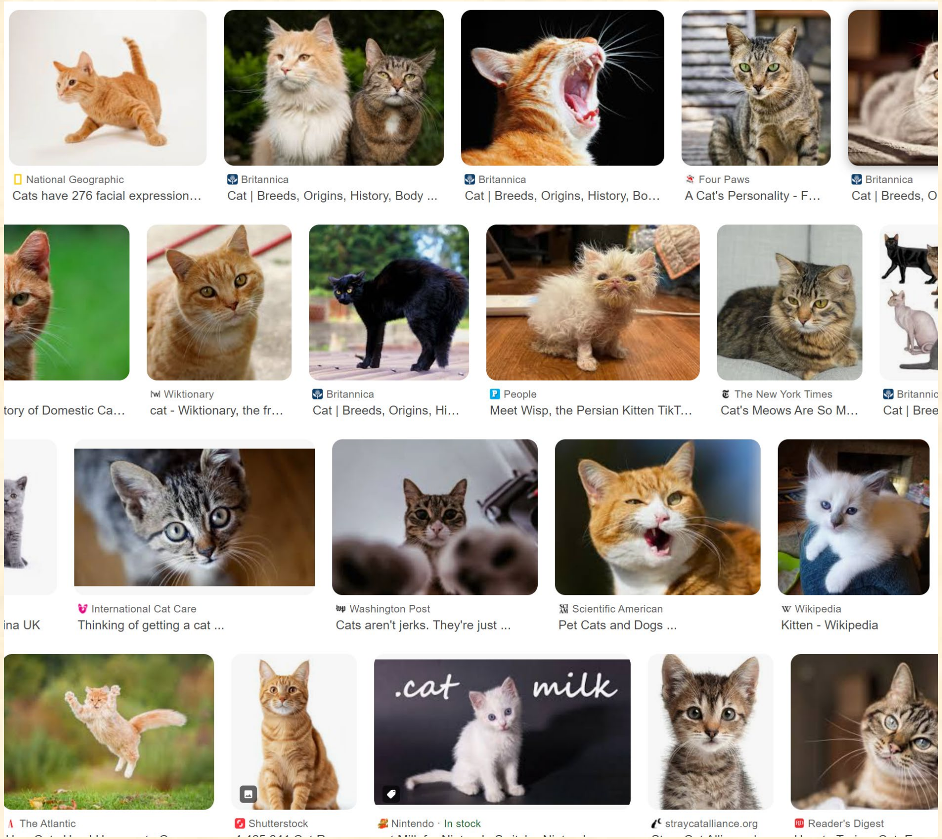
[DALL-E 3]

ML Learns By Example

Carnegie Mellon University

- **Traditional software:**
  - Algorithm-based
  - But what if we don't know how to build an algorithm?

- **Machine learning:**
  - Statistical approach
  - Collect & process training data
    - – Requires LOTS of data
  - "Train" a statistical model fit
    - – Multi-dimensional "cat-ness"



https://www.google.com/search?q=cat

Capability: Classification

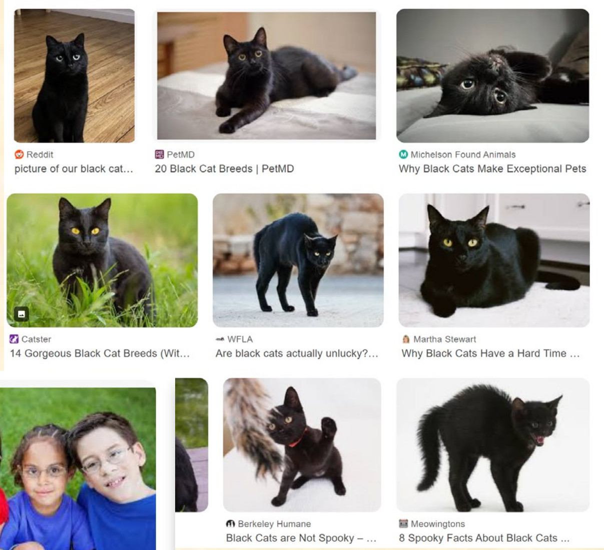
Carnegie Mellon University

- "Class" = what type of object is being detected?
- An ML "model" is trained on objects
  - Each example has a label: "cat" vs. "person"
  - Repeated training to improve classifications
- In use, ML model determines which class an object is in
  - "cat"/"person"
  - Based on statistical similarity to training data
  - Not saying "this is a cat"
    ➔ Instead: "that looks like the cats I trained on"

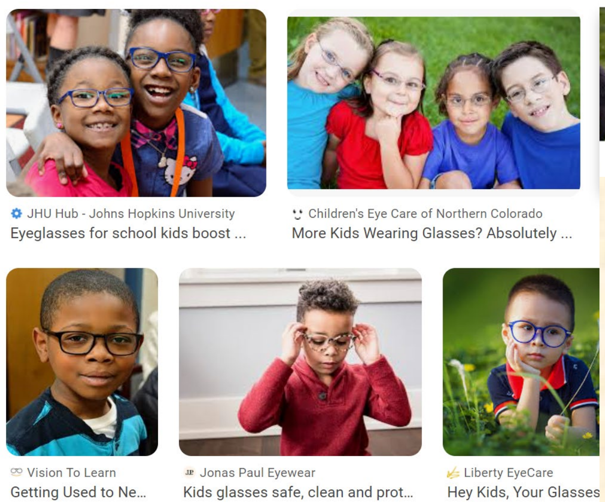
Challenge: “Bias” ➔ Faulty Training Data

Carnegie Mellon University

- Training data lacks appropriate balance
  - Missing data (what if no calico cats?)
  - Over-represented classes of data
  - Under-represented classes of data
  - Chance statistical correlations



- Problematic ML outcomes
  - Poor accuracy at under-represented data
  - False confidence when it is really just guessing

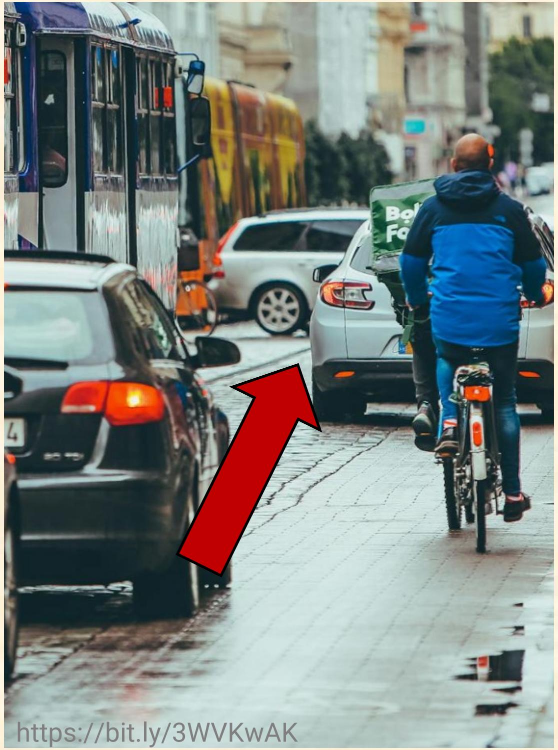


[Google image searches:
Black cats
Children with glasses]

Capability: End to End Behavior

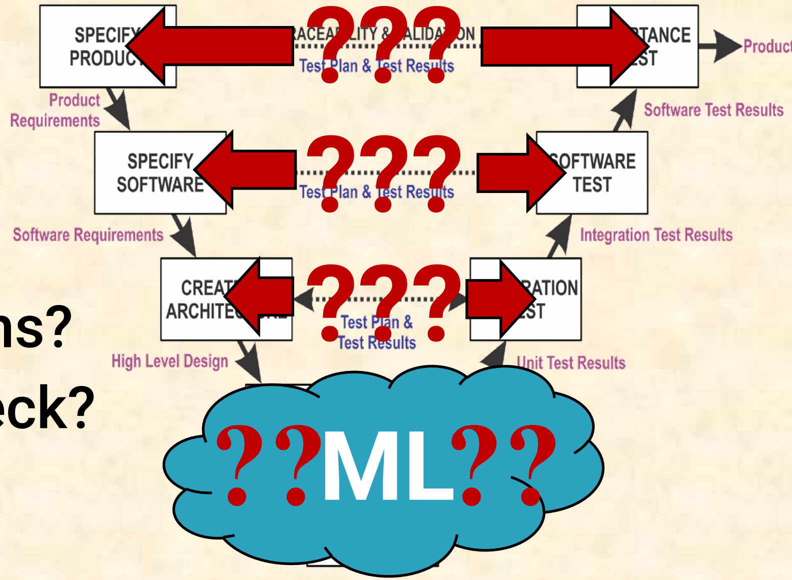
Carnegie Mellon University

- What is the right response to a stimulus?
  - Don’t classify … instead …
    react directly based on sensor inputs
- Just train the ML model on its behavior
  - Each situation has a reward for acceptable behavior (reinforcement learning)
  - Repeated training to maximize reward score
  - Might not have human-interpretable classification
    - Perhaps no overt “people on bikes” class, but …
      … statistically: swerving left is good


https://bit.ly/3WVKwAK

Challenge: Validation

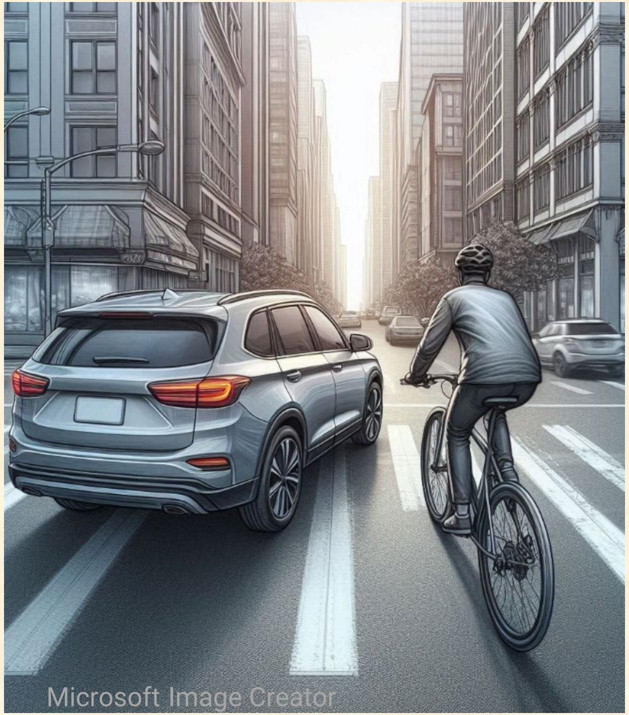
- Traditionally we validate a traceable design process
  - Testing validates left side of Vee (not brute force testing)
  - To the degree ML "learns" a behavior… that can break validation traceability



- ML validation challenges
  - Data set sufficiency & balance
  - Did ML exploit chance data correlations?
  - Are there biases we didn't think to check?
  - Are there discontinuities in behavior?
  - How brittle is system to surprises?

Capability: Generative Outputs

- Create a synthetic picture
  - ML model outputs complex data from a comparative simple prompt
- Training for statistically good outputs
  - "city street with a car passing a bicyclist"
    - Thing statistically similar to a car
    - Thing statistically similar to a bicyclist
    - Situation statistically similar to "passing"
    - Situation statistically similar to "city street"
  - Might use this for synthetic training data

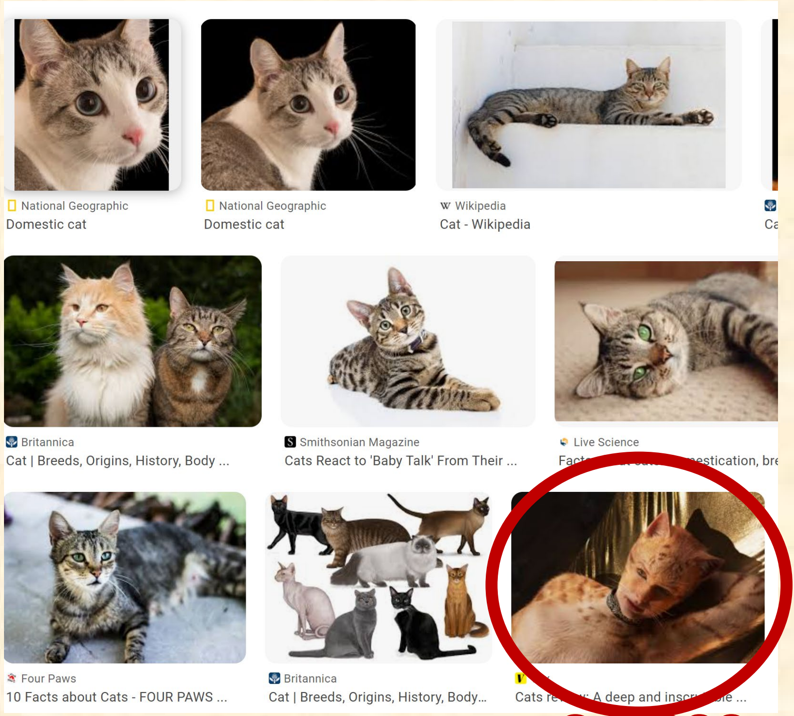


[AI GENERATED]

Challenge: Edge Cases

Carnegie Mellon University

- Edge Case ➔ an object or situation not learned from training data
  - Often not in training data at all
  - Typically something rare, “long tail”
- The world can be heavy tail
  - Huge numbers of edge cases
    - ML is especially bad at knowing that it doesn’t know (“clueless”)
  - What if training has incorrect or missing high-criticality situations?
    - This is the ultimate limit to ML safety



Google image search on cats

Cat???

Capability: Foundation Models

- Tell me a story
  - Create a plausible string of outputs
    - "city street with car passing a bicycle"
- Train on a huge corpus of data
  - Look for which words (symbols) tend to be used near each other
  - Which words tend to be used in similar sequences
  - Output a randomized set of words, based on context from previous words
- Output randomized for any prompt

> On a sunny afternoon in Pittsburgh, the city streets were bustling with activity. Cars, buses, and bicycles shared the road, each navigating their way through the urban landscape. Among them was a bright red sedan driven by **Alex**, who was on his way to meet a friend for coffee.
> As Alex approached a busy intersection, he noticed a cyclist ahead, pedaling steadily in the bike lane. The cyclist, **Mia**, was enjoying her ride, taking in the sights and sounds of the city. She wore a …
>
> Microsoft Copilot, Aug. 10, 2024

[AI GENERATED]

Challenge: “Hallucinations” → Bullshit

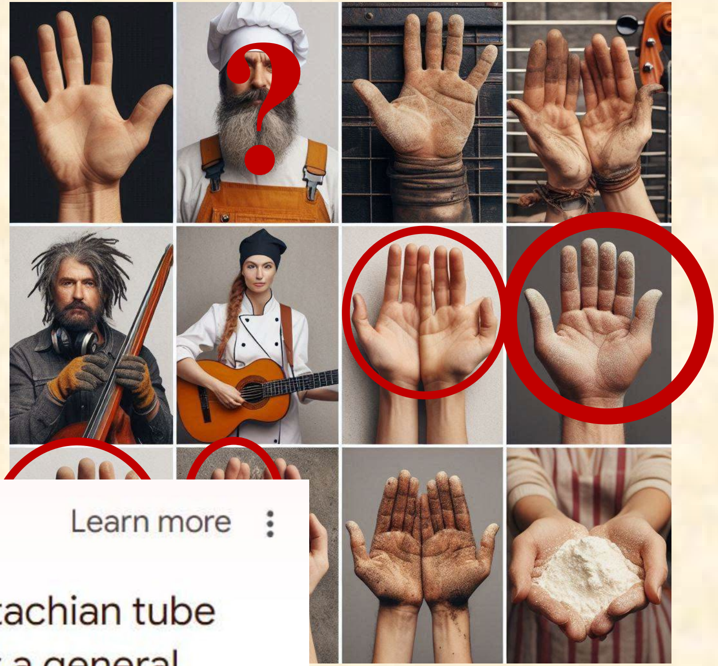
Carnegie Mellon University

- “Bullshit” used as a technical term
  - Superficially plausible
  - Lack of concern about truth

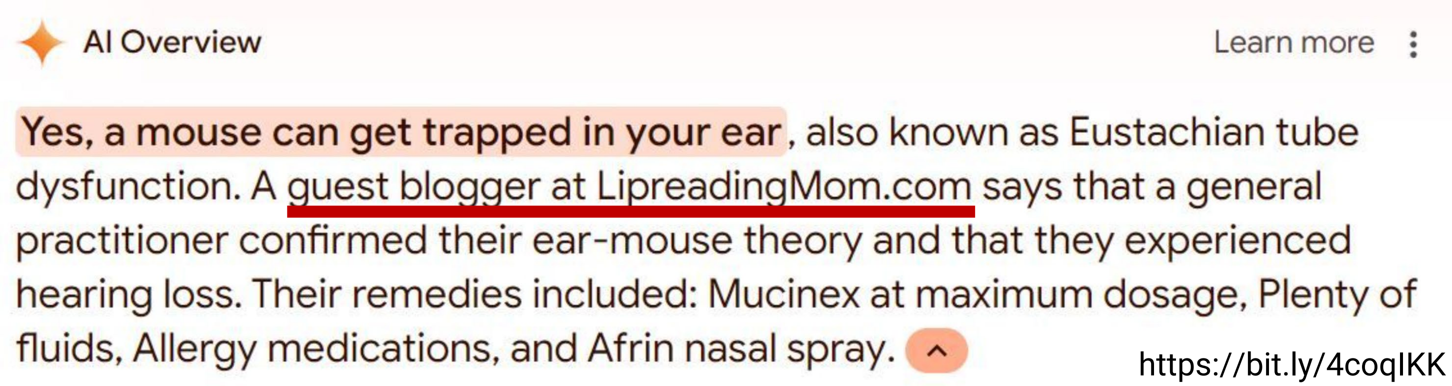
[See: https://bit.ly/3AsWusM\ ]

- Truth is not a factor in ML output
  - Statistical accuracy is not truth

“Ten people showing their hands” (DALL-E 3)



[AI GENERATED]



Yes, a mouse can get trapped in your ear, also known as Eustachian tube dysfunction. A guest blogger at LipreadingMom.com says that a general practitioner confirmed their ear-mouse theory and that they experienced hearing loss. Their remedies included: Mucinex at maximum dosage, Plenty of fluids, Allergy medications, and Afrin nasal spray.

https://bit.ly/4coqIKK

Challenge: Autonowashing

- Autonowashing: Greenwashing of Vehicle Automation (Dixon 2020)
  - Applies more generally to AI hype
- Unrealistic AI claims lead to:
  - Brand tarnish via eroded quality
    - Incorrect customer service advice [https://bit.ly/3YMkrp0]
  - Undue trust for use in critical tasks
    - Drunk drivers over-trusting “self-driving” features [https://bit.ly/3yCaAYq]
  - Automation complacency (degraded fact checking)
    - Lawyer sanctioned for fictitious legal brief citations [https://bit.ly/3M0WzGL]
- It will take more than “Education” to fix this

> THE PERSON IN THE DRIVER'S SEAT IS ONLY THERE FOR LEGAL REASONS.
>
> HE IS NOT DOING ANYTHING. THE CAR IS DRIVING ITSELF.

[Tesla 2016]

Artificial General Intelligence (???)

Carnegie Mellon University

- AGI: matches human capabilities across wide variety of tasks
  - Turing test: conversation with chatbot
    - Turns out to be a test of human gullibility
  - College exams & IQ tests
    - Did ML train to typical test contents?
- ML-based systems do not “understand”
  - GenAI/LLM train on result of understanding
    - The map is not the territory
  - Brittleness to novelty is a huge issue
  - Additional breakthroughs needed for AGI

https://bit.ly/3AgsXTk

Challenge: AI Safety & Accountability

- System safety is about edge cases
  - Near-zero probability * catastrophic consequence
  - Many systems are unforgiving of small errors
    - Statistical approaches struggle with 99.9999999%
- AI safety (SkyNet is not what I worry about here)
  - Who/what is accountable for harm?
    - Blaming a computer circumvents societal guardrails
  - Transparency, equity, enshrining biased processes
    - Blaming an opaque computer evades accountability
  - Tool for disinformation and malicious use
    - ML is a powerful weapon for the bad, reckless, and power-hungry

[DALL-E 3]
[AI GENERATED]

ML and the 90/10 Principle

Carnegie Mellon University

90/10 Principle: 90% benefit from first 10% of effort

- Pros for ML:
  - Quick prototyping, create fictional material
  - Sometimes good enough is good enough
- Cons for ML:
  - ML lacks “common sense”
  - Effort to check might outweigh advantages
  - Intellectual property issues with training data
  - 90% correct does not mean 90% functionality
  - Erosion of truth via misinformation

[DALL-E 3]
[AI GENERATED]

Conclusions

Carnegie Mellon University

- ML-based AI is not self aware!!!
  - Statistical transformation of input to output
  - People project “truth” and awareness onto AI
- Strengths:
  - Good performance on the common case
  - Superficially reasonable outputs
- Weaknesses
  - Over-trust, automation complacency, bullshit
  - The devil is in the details (bias, edge cases, …)

[DALL-E 3]

[AI GENERATED]

