Phil Koopman

Embodied AI Safety

PhilKoopman.Substack.com

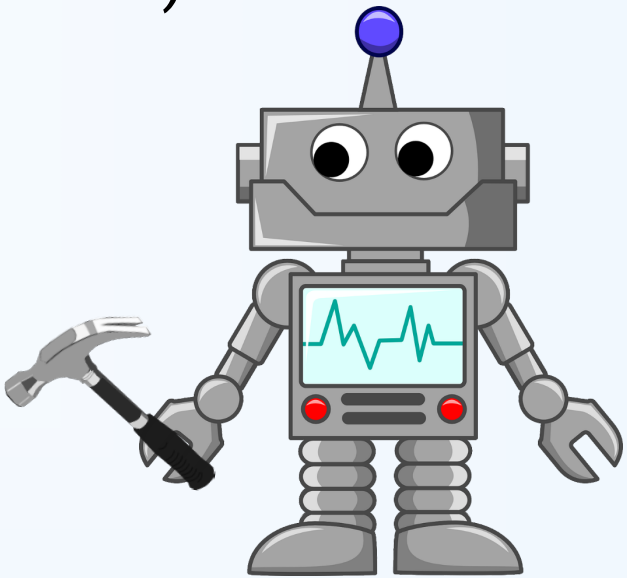
Embodied AI (eAI) Safety Overview

eAI = AI/ML + sensors + actuators

- **Key concepts in core areas:**
  - System Safety
  - Cybersecurity
  - AI based on Machine Learning (AI/ML)
  - Human/Computer Interaction
- **The journey to eAI safety**
  - Revisiting acceptable risk
  - Safe eAI challenges
  - Re-imagining safety engineering

Motivation for eAI Safety

- **Why use AI/ML?**
  - Perception tasks (e.g., object classification)
  - Natural language interface
  - Dealing with unstructured, open-world environment



- **Physical AI means physical safety**
  - How safe is safe enough?
  - Where is the accountability for harm?
  - How do we instill trust in the technology?

Computers Are Not Necessarily Safer

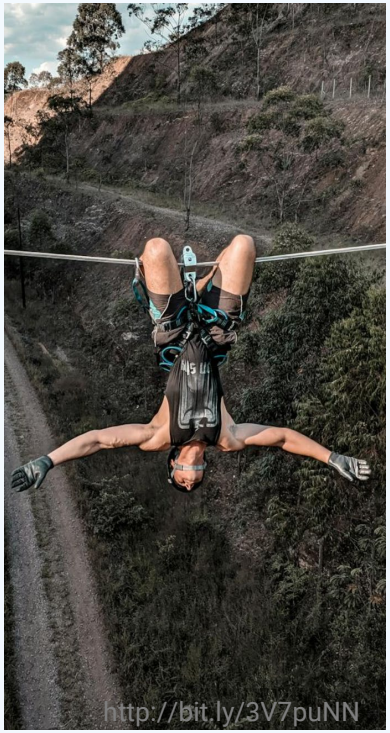
- **Human operators make mistakes**
- **Computers make mistakes too!**

Crash into utility pole

Crash into articulated bus

Safety Engineering Concepts

- **It's all about risk mitigation**
  - Identify hazards and risks
  - **Mitigate hazards** & validate mitigation
- **Technical areas**
  - Redundancy to help mitigate failures
  - Safety standards for engineering rigor
  - Design assurance beyond just testing
- **Need safety culture & independence**



Risk Analysis

- **Determine risk for each identified hazard**
  - Risk = Frequency * Severity

RISK TABLE

	LOW SEVERITY	HIGH SEVERITY
LOW FREQUENCY	LOW RISK	???
HIGH FREQUENCY	MEDIUM RISK	HIGH RISK

- **Assign a Safety Integrity Level (SIL) based on risk**

Safety Standard: Engineering Rigor

- **SIL-driven hypothetical example of rigor:**

Activity	SIL 1	SIL 2	SIL 3	SIL 4
Warning-Free Compilation	Required	Required	Required	Required
Conforms to MISRA C		Required	Required	Required
Comprehensive Static Analysis			Required	Required
Formal Proof of Correctness				Required
Informal Peer Review	Required	Required	– NO –	– NO –
Fagan-Style Peer Inspection			Required	Required
Computer Self-Test	Required	Required	Required	Required
Redundant Computers			Required	Required

➔ **Still a developing area for eAI engineering**

Safety Engineering Challenges

1. **Only the bad days matter**
   - 99,999,999 vs. 99,999,998 safe miles
   - 1 vs. 2 fatalities per 100,000,000 miles
   - A single 10-mile safe ride means little

2. **Rare, high severity-events**
   - What is zero probability * infinite cost?
   - Economics push toward low SIL
   - News headlines push toward high SIL

Oct, 2023:
Cruise Robotaxi
Drags Pedestrian

Cybersecurity Engineering Concepts

- **Security properties vary by application**
  - Confidentiality / Integrity / Availability
  - eAI emphasis on safety integrity & availability
- **Security attacks as “malicious faults”**
  - Introduce fault missing from hazard analysis
  - Violate safety analysis assumptions
- **Adversary often has physical access**

Machine Learning (ML) Concepts

- **Feed a system lots of data**
  - "Learn" **statistical properties**
  - Generate statistically **plausible** results

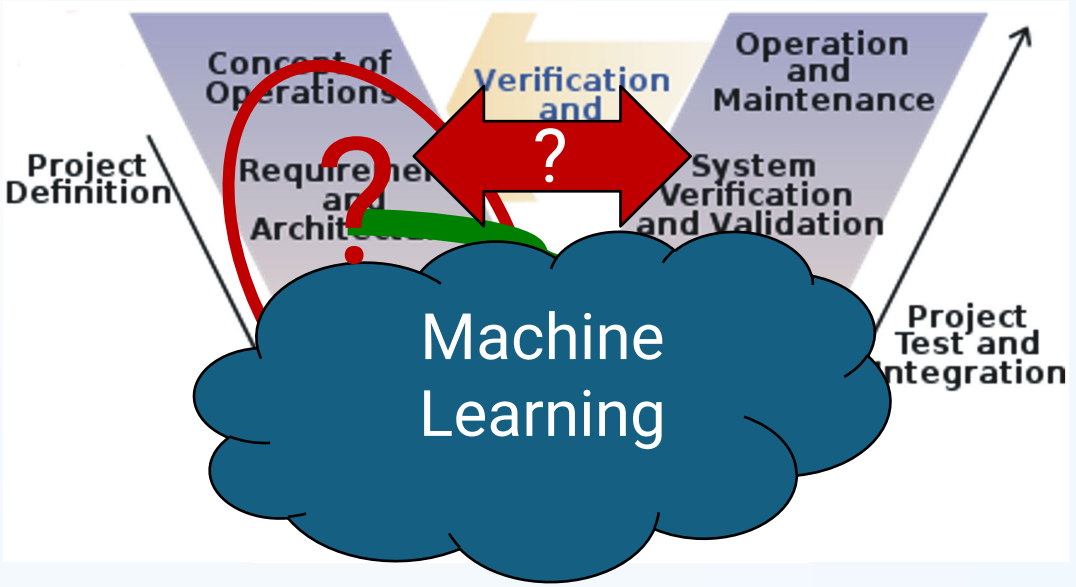
Mitchells vs. Machines

- **Different flavors:**
  - **Classification:** Car? Person?
  - **Generative art:** Randomly match statistics of the goal
  - **LLM/Foundation:** Predict next likely output in a sequence

- **Artificial General Intelligence? Nope.**
  - The Turing test turns out to be a measure of gullibility

Machine Learning Breaks the Vee

- **Vee model for safety**
  - Trace requirements to implementation
  - Testing traces to the engineering process
- **Testing validates engineering rigor**
  - Engineering rigor reduces testing burden
  - Broken traceability means more testing required



Humans and eAI Safety

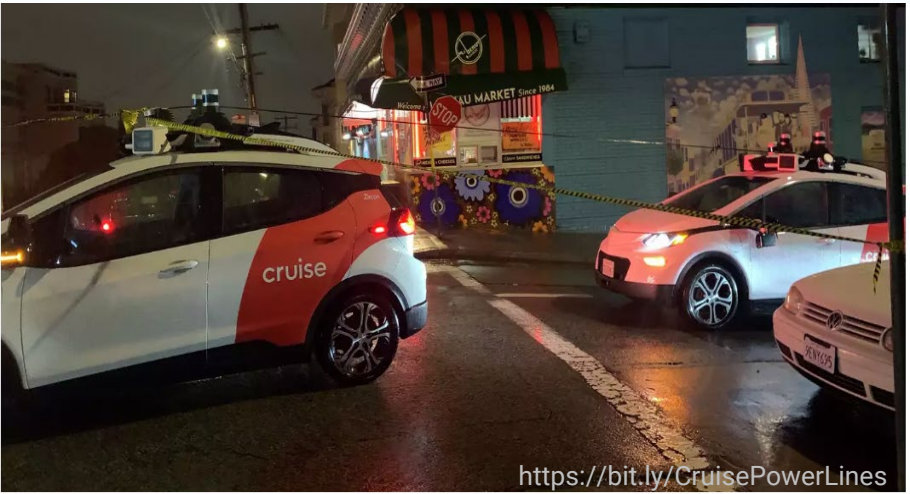
- **Inherent human limitations as supervisors**
  - Perception-Response Time (PRT)
  - Ironies of Automation lengthen PRT
    - Automation complacency & bias
    - Effective automation➔ineffective supervision
- **Serious ethical & legal issues:**
  - Who is responsible for eAI misbehavior?
  - Blaming non-zero PRT won't make it safe
    - But a Moral Crumple Zone strategy is common

eAI Safety Issues in the Wild

- **False alarms**
  - Phantom braking & driver controllability
- **Unpredictability**
  - How many tests if results differ?
- **Statistical approaches to safety**
  - ML sweet spot is often 90%-99%
  - How do you get 99.9999999% with ML?
- **Heavy tail edge cases**
  - All eAI systems will have incomplete training
  - All eAI systems need a human backup of some sort

http://bit.ly/41S0DBp

Safety Is More Than Net Harm



Two Cruise cars in San Francisco became wrapped in downed Muni wires and caution tape at Leavenworth Street and Clay Street on March 21, 2022.

Courtesy of John-Phillip Bettencourt

If nobody was harmed (this time) does that make it safe?

Safe Enough: Avoiding Risk Hot Spots

- **Safer than human operator…**

  … is only the starting point

- **Also consider risk hot spots:**
  - ➢Specific unsafe behaviors
  - ➢Risk transfer onto the vulnerable
  - ➢Harm due to eAI negligence
  - ➢Avoiding negative externalities
  - ➢Compensating for other's mistakes
  - ➢Personal & psychological safety

- **Need a multi-constraint satisfaction approach**

Duty of Care & eAI Negligence

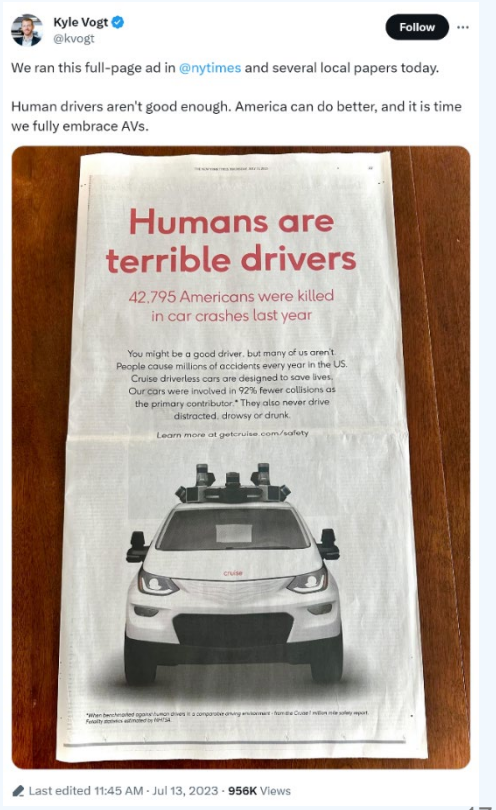
- **Duty of care for human operators**
  - Doing something potentially dangerous?
  - Act as a “reasonable person” would
  - Harm from breach of duty ➔ negligence
- **Duty of care for computers?**
  - Based on behaviors, not design defects
  - Mistakes treated as if by a human operator
    - Manufacturer should be responsible party
- **Statistical safety does not forgive negligence**

Misguided Messaging

- ***"We're Saving Lives!"*** **is all downside**

- **Proving *Saving Lives!* requires**
  - Exposure to ~10 expected fatalities
  - For robotaxis, perhaps 1 billion miles

- **Popular opinion won't last that long**
  - News photos undermine the narrative
  - People think in stories, not statistics



https://twitter.com/kvogt/status/1679517290847694848

People React To Stories

- **How did you feel about these stories?**

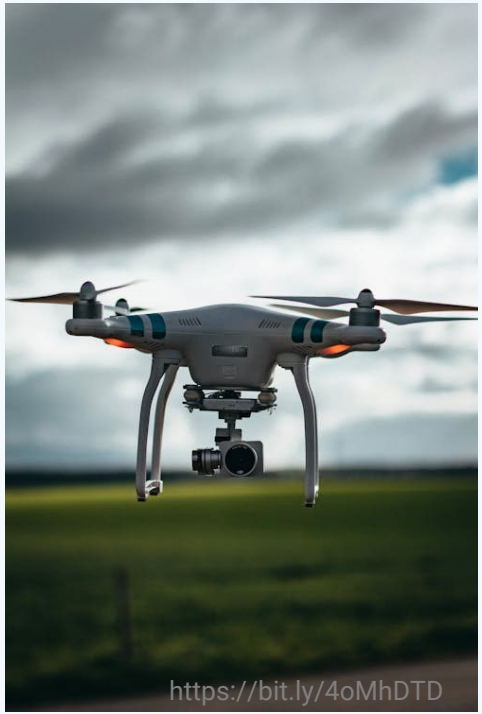
Crash into utility pole

Crash into articulated bus

- **61% of US drivers fear robotaxis** 2025 AAA survey http://bit.ly/4mVQqvQ

Re-imagining Safety Engineering

- **Societal: net risk won't be enough**
  - "Better than human" per incident
- **Technical: heavy tail edge cases**
  - Imperfect system in an imperfect world
- **Legal: AI accountability approach**
  - Apply human negligence standards to AI
  - Respect limits of human capabilities
- **Multi-constraint satisfaction approach**
  - Stakeholders contribute aspects of risk constraints



Justifiable Trust for Safe eAI

- **Promises beyond *Saving Lives!***
  - Measurable, responsible behaviors
- **Accountability**
  - Accept proportionate responsibility
  - Independent oversight
- **Safety constraint satisfaction approach**
  - Net risk reduction alone is not enough for safety

Embodied AI Safety: The Book

Amazon.com (US)

- **Country-specific Amazon web pages:**

  AU, CA, FR, DE, IT, JP, NL, PL, ES, SE, GB

  - ISBN: 9798292384410 Trade Paperback
  - ISBN: 9798292384618 Hardcover
  - 452 pages

Rest of series: