Hazard Contribution Modes of Machine Learning Components

Colin Smith
Oak Ridge National Laboratory

Ewen Denney & Ganesh Pai
KBR / NASA Ames Research Center

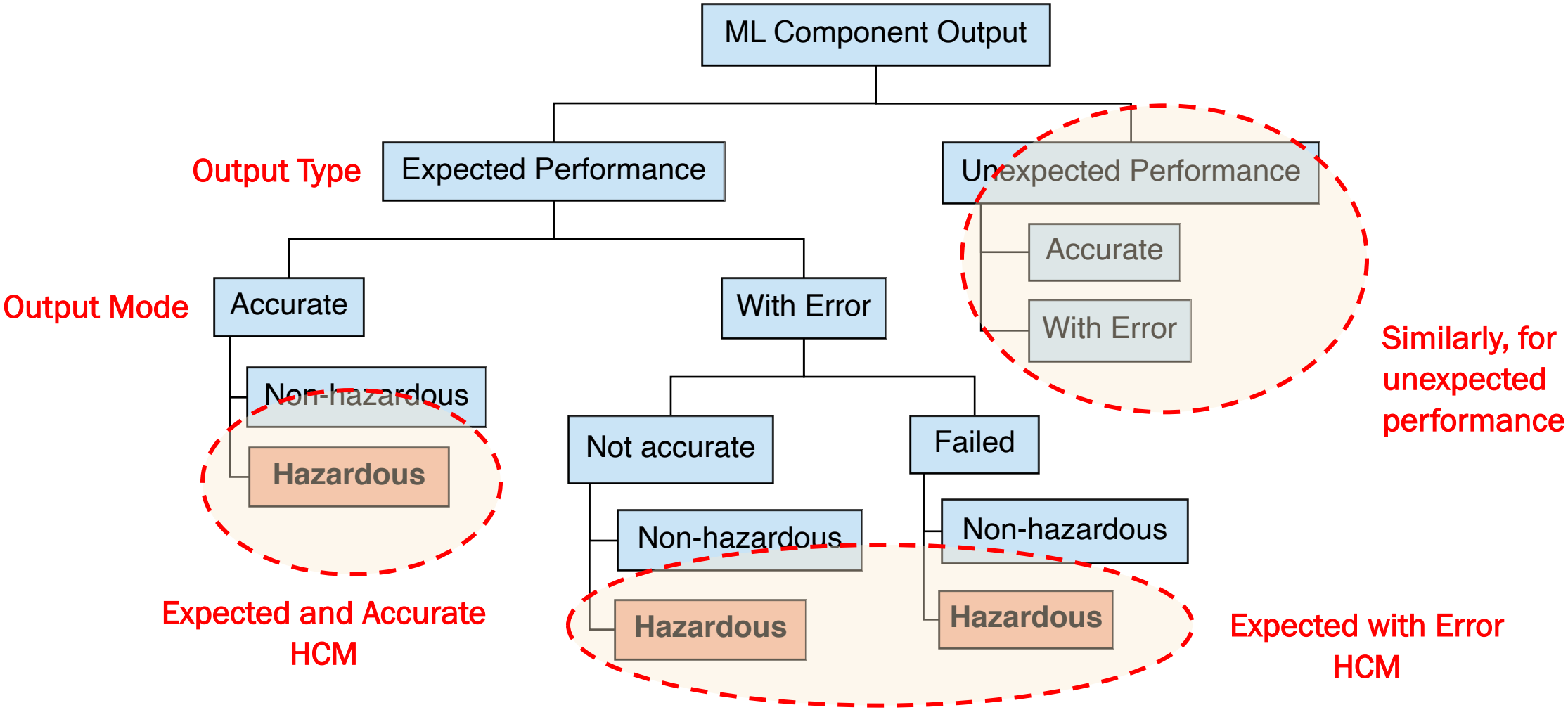
AAAI-20 Workshop on Artificial Intelligence Safety (SafeAI 2020). New York, NY. Feb. 7, 2020

Summary

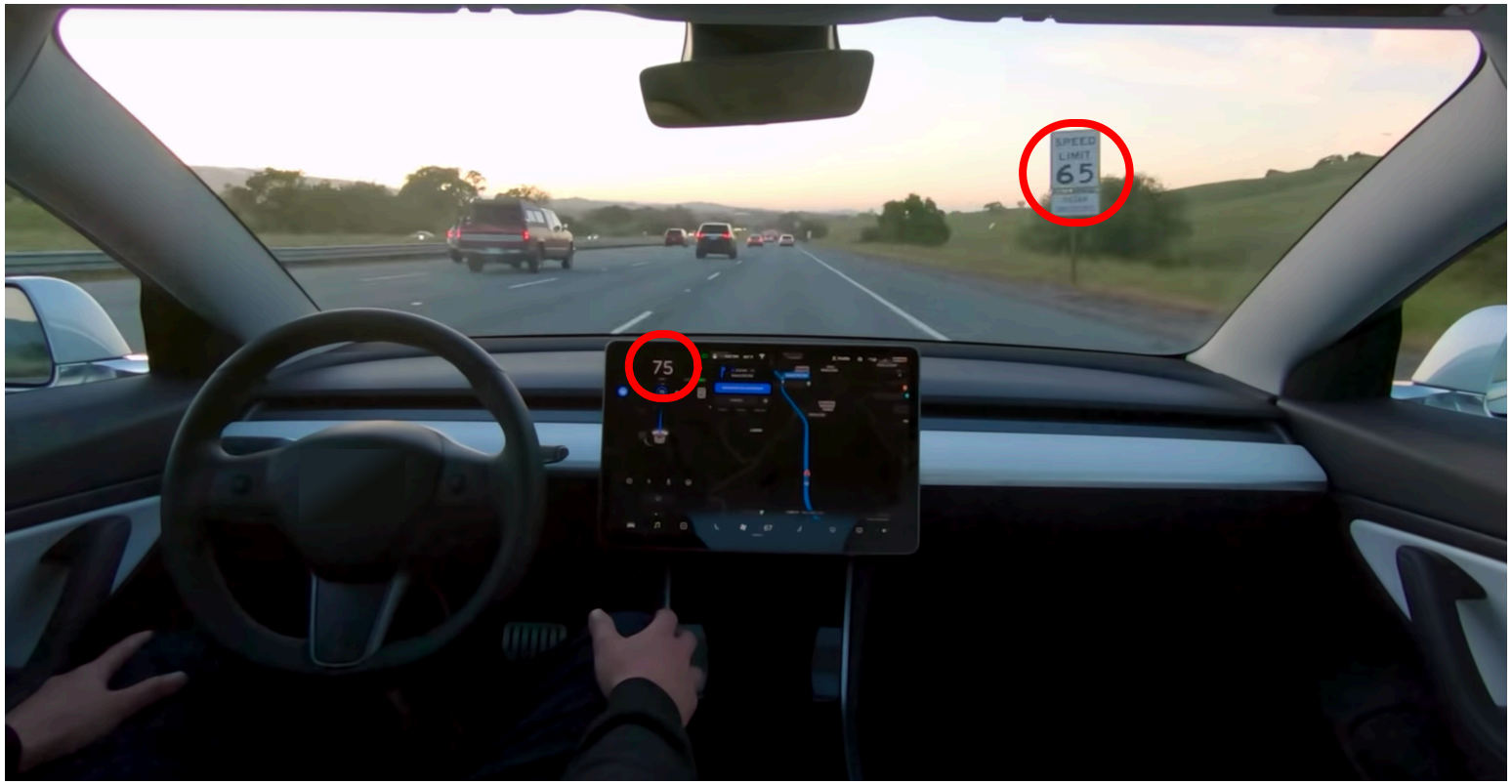
- Supporting (HazAn) of machine learning (ML) components
  - Bottom-up hazard analysis similar to (but not the same as) FMEA
  - Identifying candidate mitigations
  - Encoding mitigations as assurance argument patterns, justifying rationale for their use

- Hazard Contribution Mode (HCM)
  - The way in which a ML output contributes to a system hazard
  - 4 broad categories

Hazard Contribution Modes



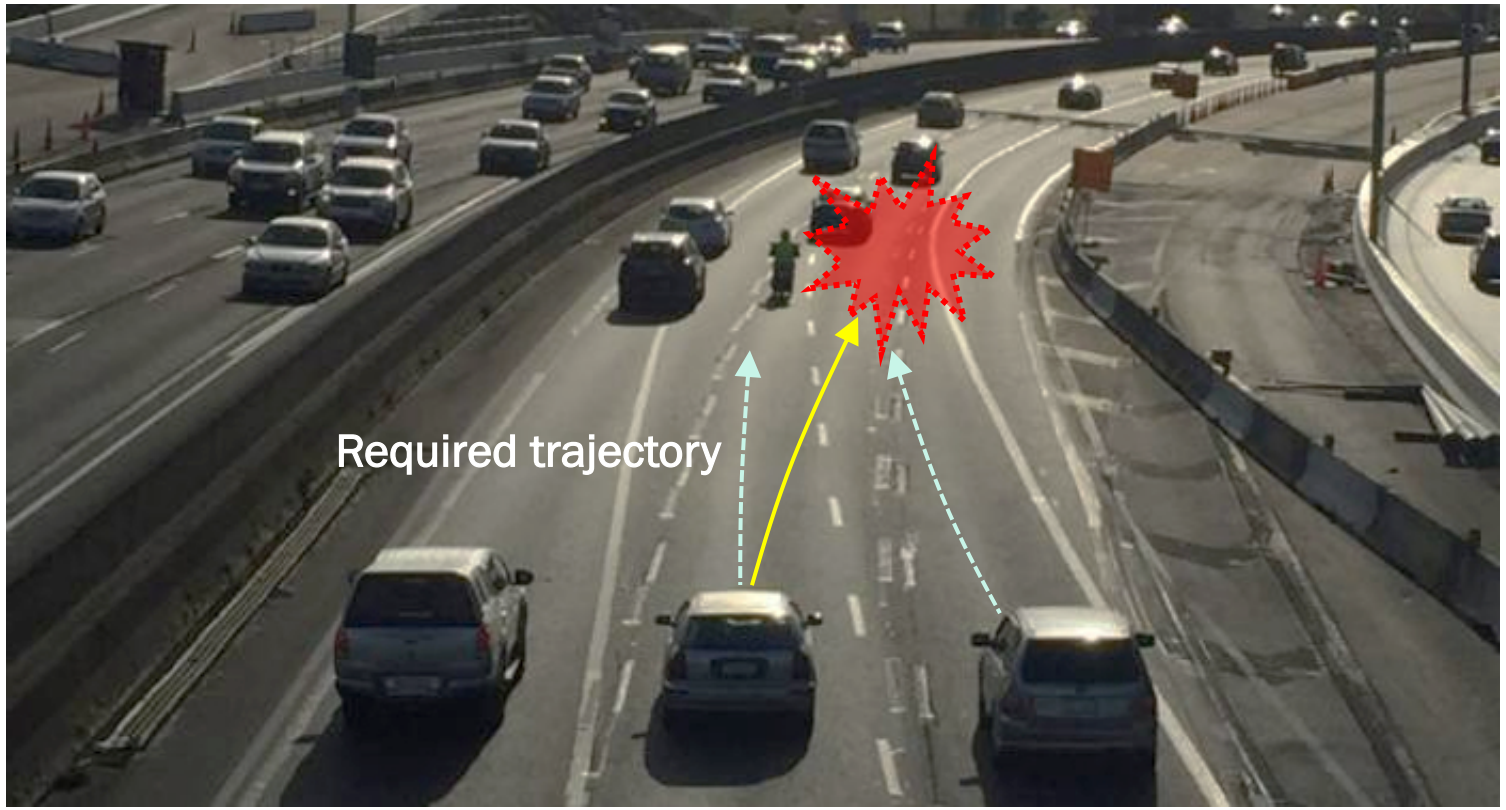
Expected and Accurate but Hazardous



Hazardous Cruise Control

ML-based cruise control prioritizes speed of traffic rather than posted speed limits

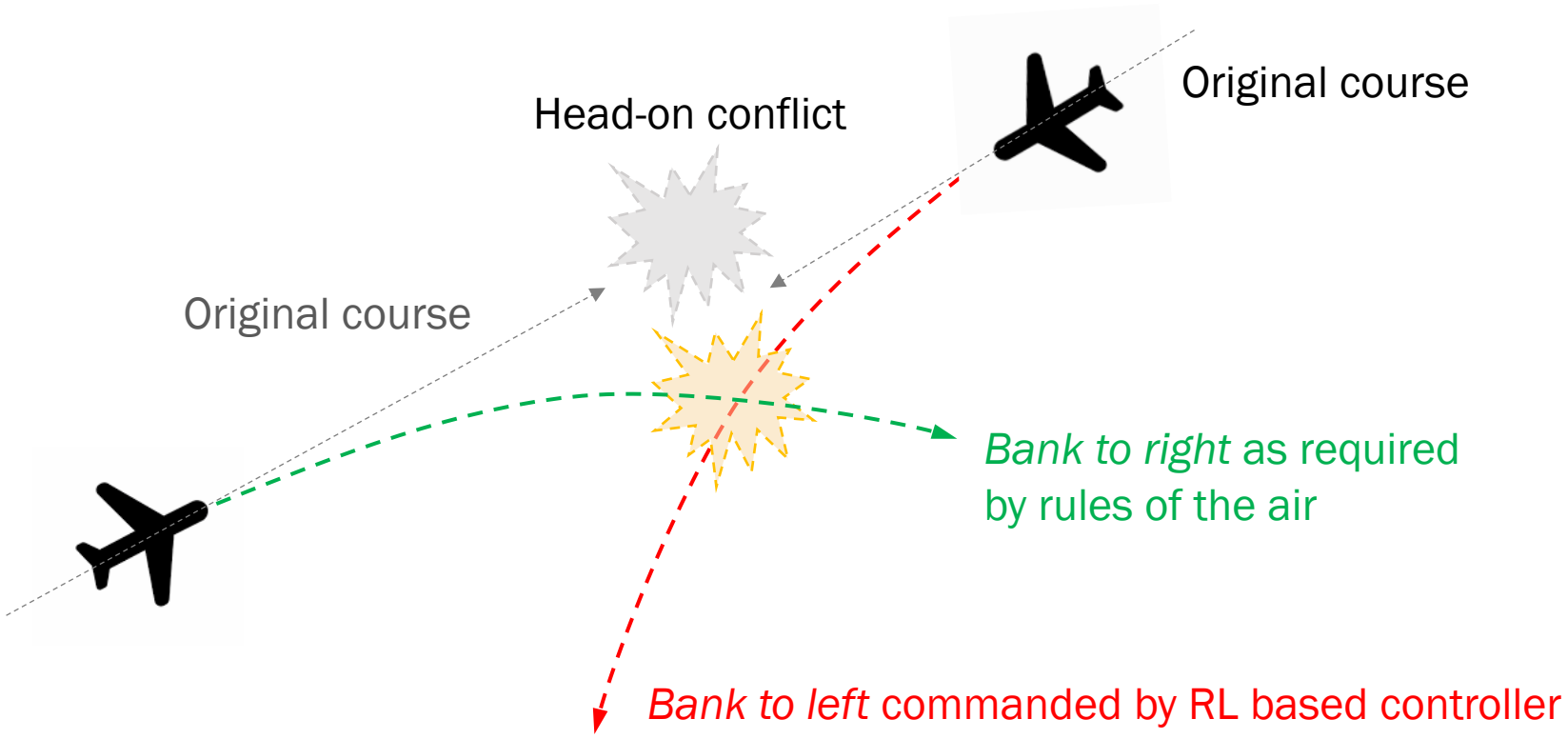
Expected with Error and Hazardous



Botched Lane Keeping

ML-based lane keeping classifies removed lane markings as the current lane boundaries, leading to lane departure

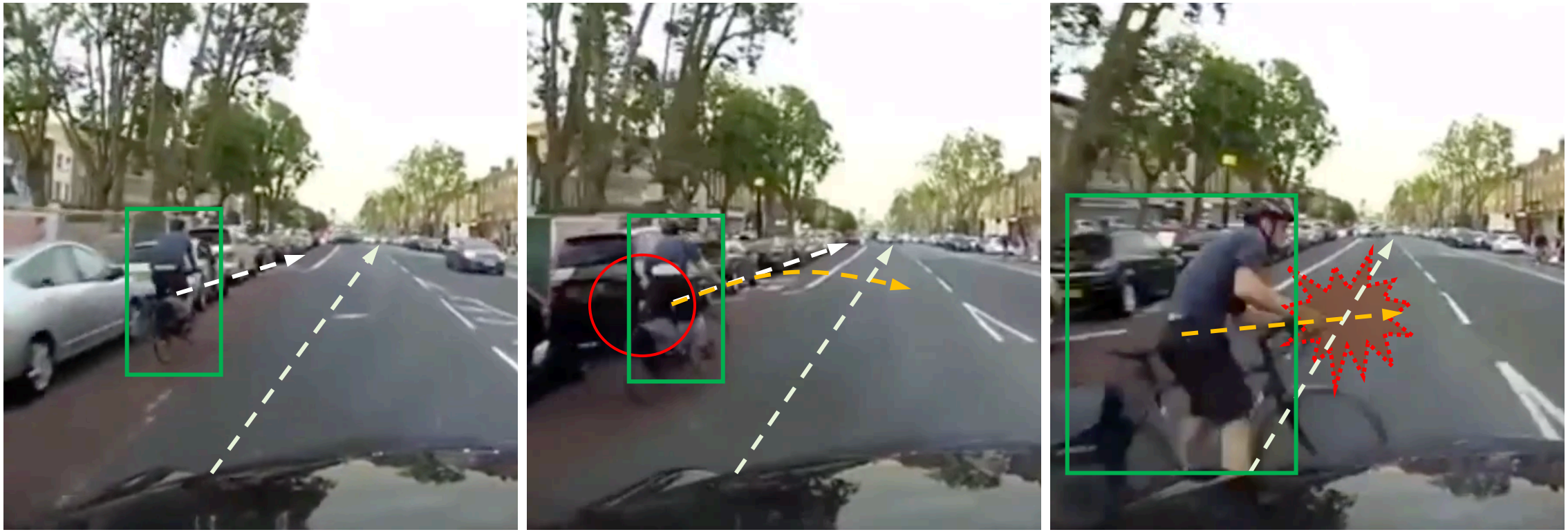
Unexpected and Accurate but Hazardous



Botched Collision Avoidance

Banking to left (a legitimate and optimal action in some situations) leads to midair collision

Unexpected with Error and Hazardous



Trajectory Estimation Error

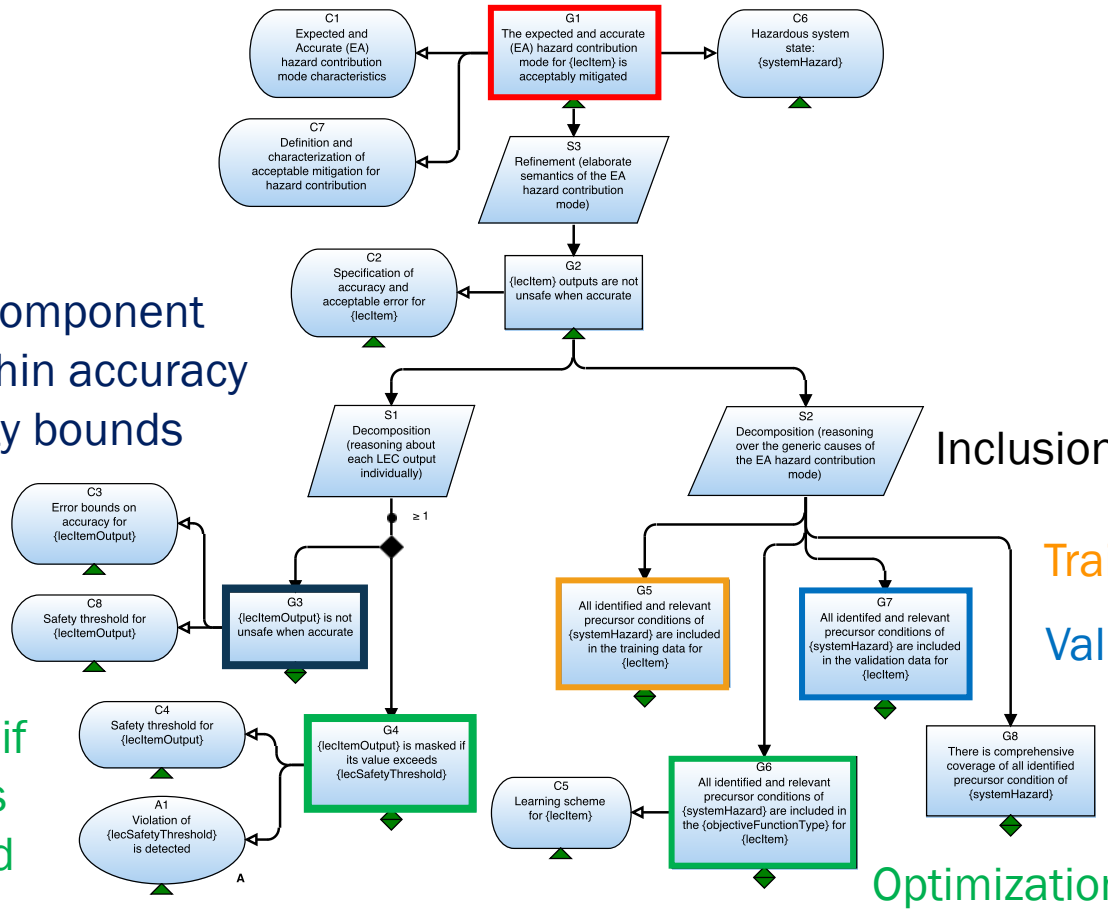
Unexpected edge case (car door opening causing bicyclist to swerve into path) with trajectory estimation error, leading to collision

Potential Causes

- Expected and Accurate HCM
  - Data inadequacy
    - All known hazard precursors not reflected or sampled in the training & validation data
  - Ill-formed optimization
    - All known hazard precursors not sufficiently penalized or accounted for in the optimization functions
  - Validation Insufficiency
    - Insufficient verification or coverage of known hazards during ML validation
- Unexpected with Error HCM
  - Distributional Shifts
    - Covariate Shift and Edge Cases
    - Prior Shift
    - Concept Shift
  - Insufficient Validation
    - Logical errors
    - Omission and incompleteness
  - Standard model error

Assurance Argument Pattern

Expected & accurate HCM mitigated



Each ML component output is within accuracy and safety bounds

Inclusion of hazard precursors into

Training Data

Validation data

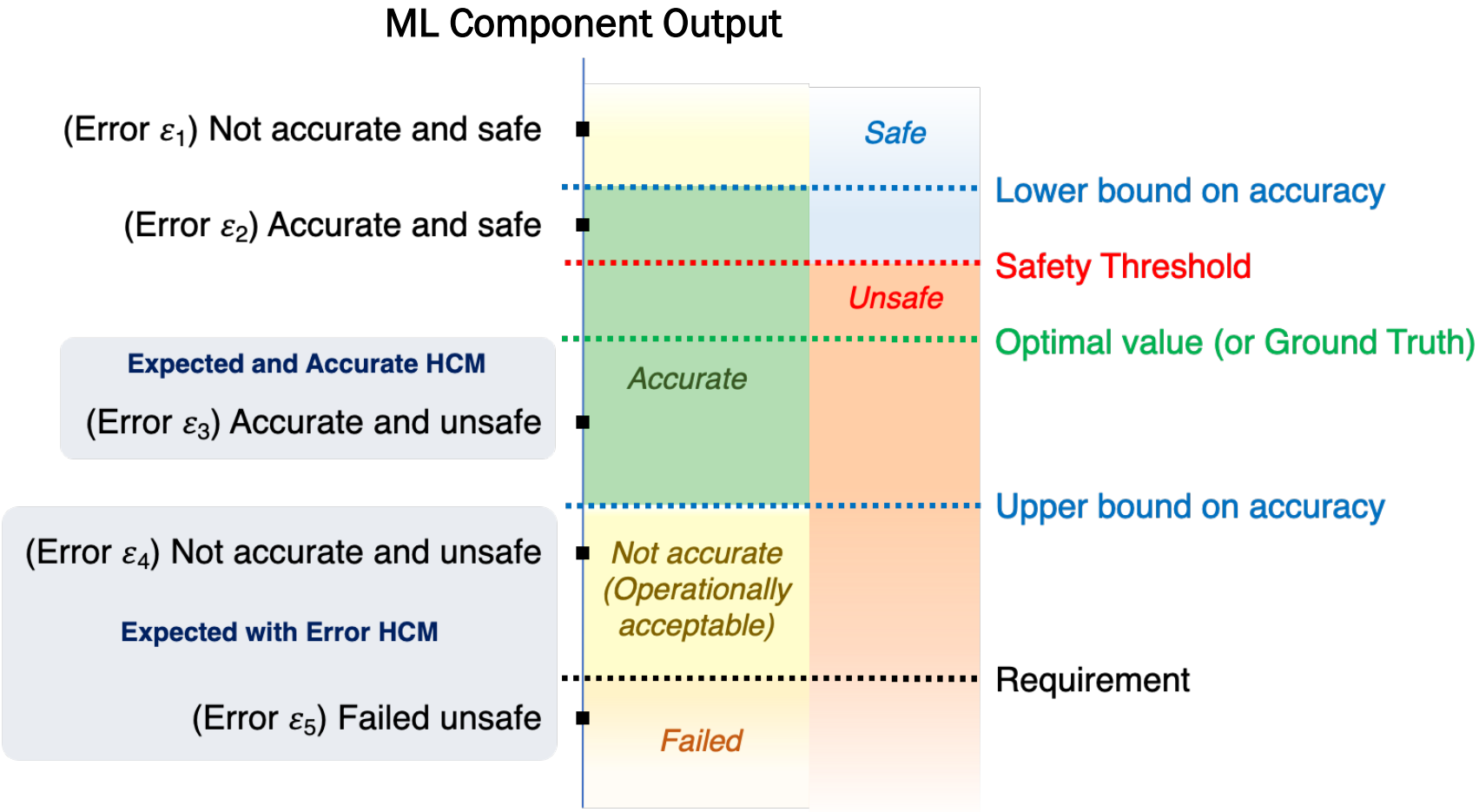
Output masked if output exceeds safety threshold

Optimization objective function

Conclusions & Future Work

- Preliminary set of HCMs
  - Intended as a starting point for further development
- Requires specialization to particular applications
- Wider objective is to create a template
  - Providing automation support
- More comprehensive identification of mitigations

Output Accuracy, Error, Failure, and Safety

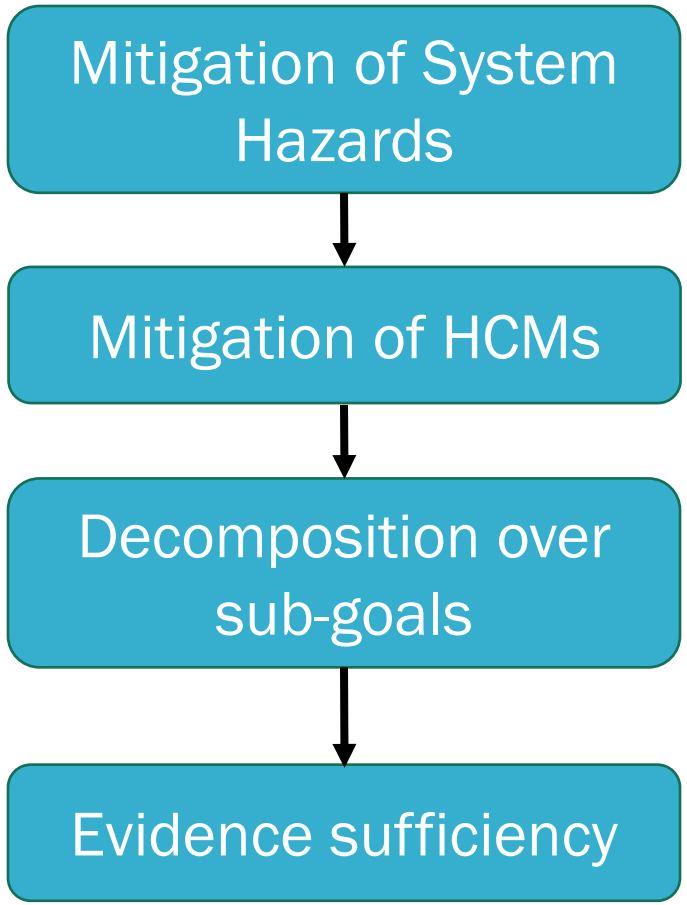


ML Assurance Techniques (Related Work)

Lifecycle Activity	Assurance Properties	Assurance Technique Examples
Data Management	Relevant Complete Balanced Accurate	Simulation V&V Generative adversarial networks (GANs) Sampling density comparison Data augmentation
Model Learning	Performant Robust Reusable Interpretable	Automated optimization Ensemble learning Transfer learning
Model Verification	Complete Contextually Relevant Comprehensible	Probabilistic / Hybrid verification Synthetic test data Counter-example guided data augmentation
Model Deployment	Fit for purpose Tolerated Adaptable	Monitor distribution shift Fault tolerant architecture

Adapted from Ashmore et al. Assuring the Machine Learning Lifecycle: Desiderata, Methods and Challenges. https://arxiv.org/pdf/1905.04223.pdf

Assurance Arguments for HCMs



ML-Component does not contribute to system hazards

EA, EE, UA, and UE HCMs contribution system hazard mitigated

Reasoning about each ML component output

Evidence is sufficient to claim each assurance sub-goal