Hazard Contribution Modes of Machine Learning Components

Colin Smith
Oak Ridge National Laboratory

Ewen Denney & Ganesh Pai
KBR / NASA Ames Research Center
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

ML Component Output

Expected Performance

Output Type

Output Mode

Accurate

Expected and Accurate

Hazardous

Non-hazardous

Hazardous

Expected with Error

HCM

With Error

Accurate

Non-hazardous

Hazardous

Failed

Non-hazardous

Expected with Error

HCM

Similarly, for unexpected performance

Unexpected Performance

Accurate

With Error

Not accurate

Non-hazardous

Failed

Non-hazardous

Hazardous
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

Output masked if output exceeds safety threshold

Inclusion of hazard precursors into Training Data

Validation data

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 Component Output

- (Error $\varepsilon_1$) Not accurate and safe
- (Error $\varepsilon_2$) Accurate and safe
- (Error $\varepsilon_3$) Accurate and unsafe
- (Error $\varepsilon_4$) Not accurate and unsafe
- (Error $\varepsilon_5$) Failed unsafe

Safe

Unsafe

Accurate

Not accurate (Operationally acceptable)

Failed

Lower bound on accuracy

Safety Threshold

Optimal value (or Ground Truth)

Upper bound on accuracy

Requirement
## ML Assurance Techniques (Related Work)

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

Assurance Arguments for HCMs

Mitigation of System Hazards

ML-Component does not contribute to system hazards

Mitigation of HCMs

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

Decomposition over sub-goals

Reasoning about each ML component output

Evidence sufficiency

Evidence is sufficient to claim each assurance sub-goal