Integrating machine learning into safety-critical systems

Thomas G. Dietterich
University Distinguished Professor (Emeritus)
Oregon State University
@tdietterich (X and BlueSky)
tgd@cs.orst.edu

Oregon State University

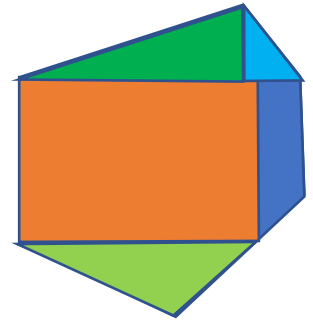
Outline

- Part 1: Integrating ML into traditional safety engineering processes
  - Scenario-based data collection
  - Verification of function behavior
- Part 2: Safety as Control
  - Detecting anomalies and near misses
- Part 3: Safety as Continual Redesign
  - Design is never finished
  - Resilient systems are “Poised to adapt”

Traditional Safety Engineering

- Define the operational design domain (ODD)
- Decompose ODD into scenarios
- Hazard Analysis of each scenario
- Risk Assessment (likelihood and severity of each harm).
- Identification of socially acceptable risk
- Design the system to achieve the socially acceptable risk
- Validate that the system meets the safety requirements

Semi Driving on Freeway



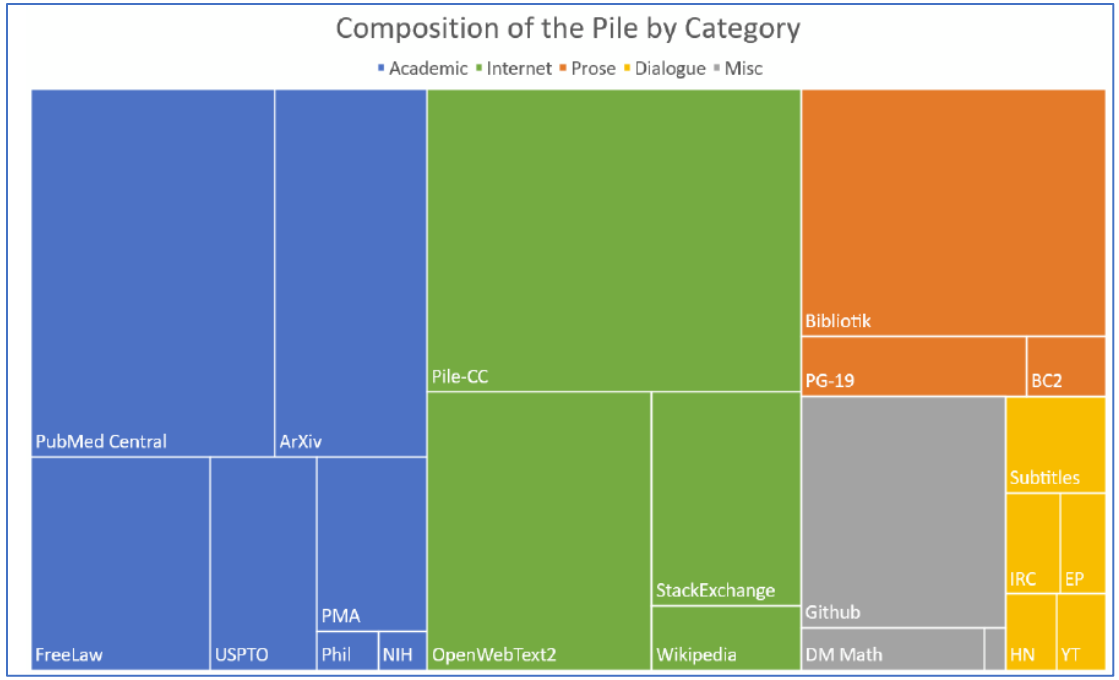
Acceptable Risk of Death:
1 in 10^8 hours

[Verma, et al., 2010]

Contrast: Traditional Machine Learning Methodology

- Aggregate data from as many sources as possible
  - Data was often collected for other purposes
  - “Big Data” is “the new oil”

[Deng, et al, 2009]



[Gao, et al, 2020]

The Pile: An 800GB Dataset of Diverse Text for Language Modeling

Consequences of this Methodology

- No guarantee that the Operational Design Domain is covered well
- No guarantee that the ML system will learn a model that meets the safety requirements
- Learning Theory only provides statistical guarantees for inputs drawn *from the same distribution* as the training data
- If the actual distribution in operations concentrates on a region of poor coverage, error can be arbitrarily large/serious

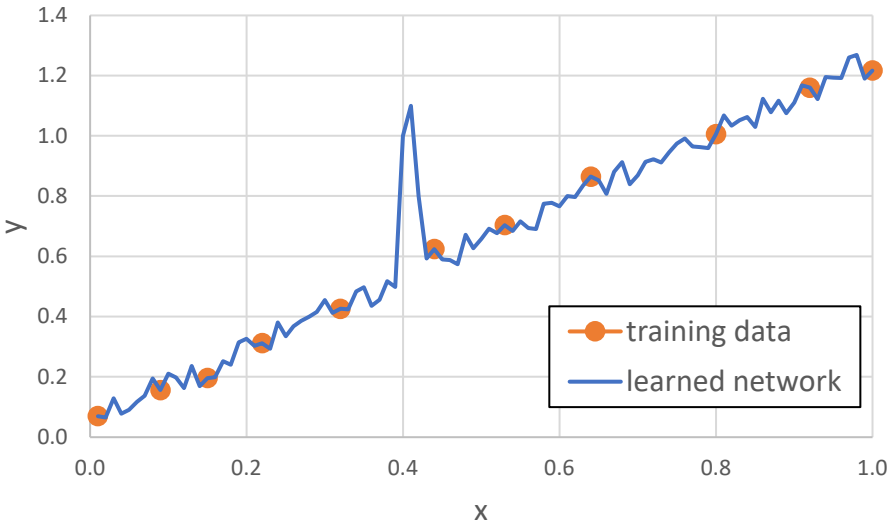
We need a new methodology

Achieving Distribution-Independent Accuracy in Machine Learning Components

- Deliberately collect training data to attain good coverage of all scenarios
  - Risk-driven sampling techniques (e.g., [Wang, et al., 2023])
- Verify approximation quality of the learned model
  - Collect additional examples as needed

Verifying Correct Behavior of ML Component

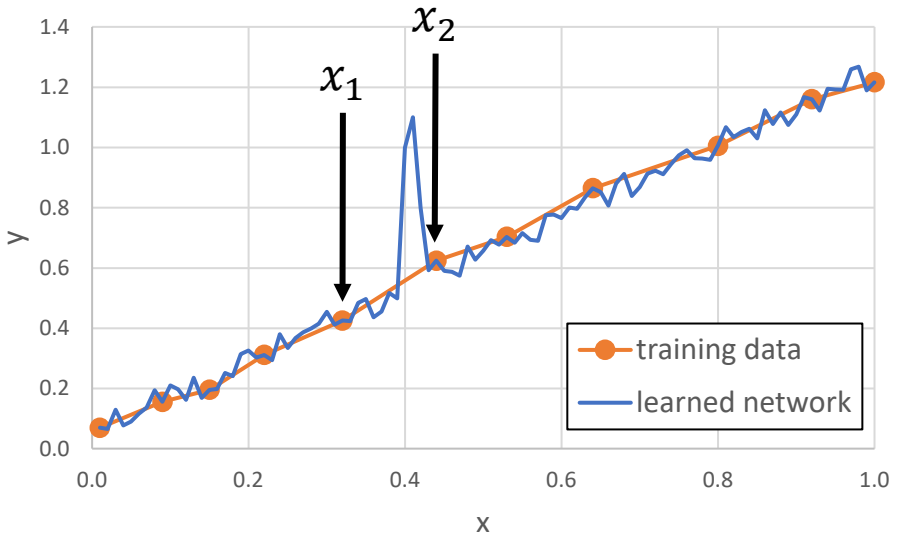
- How can we gain assurance that the ML system has learned the correct function?
- We have no explicit specification of correctness, but we can detect bad behavior



Proposal: Bound the difference between the fitted function and linear interpolation of the training data

- Consider two adjacent training examples x_1 and x_2
- Let $\alpha \in [0,1]$

$$\max_{\alpha} \left| \begin{array}{c} f(\alpha x_1 + (1-\alpha)x_2) \\ - \\ \alpha f(x_1) + (1-\alpha) f(x_2) \end{array} \right|$$



- If this is small, the function behaves well in between the training data
- This can be solved by the methods of [Singh, et al. 2021] (but those may not scale)

With these tools, ML can be integrated into the standard safety engineering process

- Deliberately collect training data to attain good coverage of all scenarios
  - Risk-driven sampling techniques (e.g., Wang, et al., 2023)
- Verify approximation quality of the learned model
  - Collect additional examples as needed

- N.B. No single validation method suffices to ensure safety. See [Kochenderfer, et al., *Algorithms for Validation* (forthcoming)]

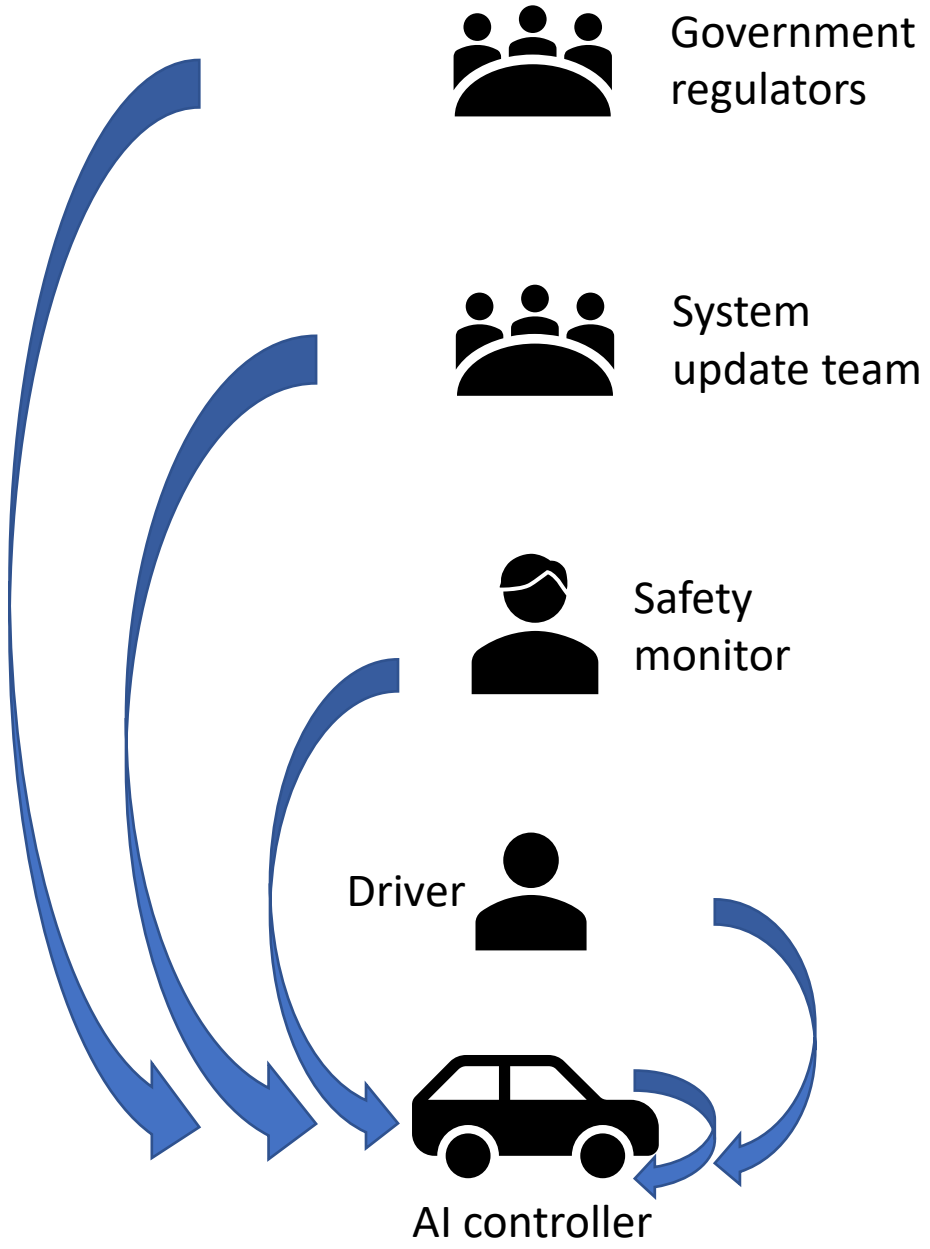
Outline

- Part 1: Integrating ML into traditional safety engineering processes
  - Scenario-based data collection
  - Verification of function behavior
- Part 2: Safety as Control
  - Detecting anomalies and near misses
- Part 3: Safety as Continual Redesign
  - Design is never finished
  - Resilient systems are “Poised to adapt”

Systems View of Safety

[Leveson 2011: Engineering a Safer World]

- Safety is a control problem
  - Maintain the safety of the system in the presence of disturbances
- What is the “controller”?
  - The human organizations that build, operate, and maintain it
  - Government regulators
  - Elected officials
- What are the “disturbances”?
  - Budget cuts and staff reductions
    - Systems tend to migrate toward the edges of safety
  - Unknown unknowns
    - Environmental Novelty
- The controller must detect and compensate for these disturbances
  - Today: It is the exclusively the humans who do this
  - Can AI help?



Unknown Unknowns: Detecting Novel Failure Modes

- Key performance indicators [Weick, et al., 1999]
  - Number of anomalies detected
  - Number of near misses detected
- These provide evidence of novel failure modes *before* they cause harms

- What is known about AI methods for detecting anomalies and near misses?

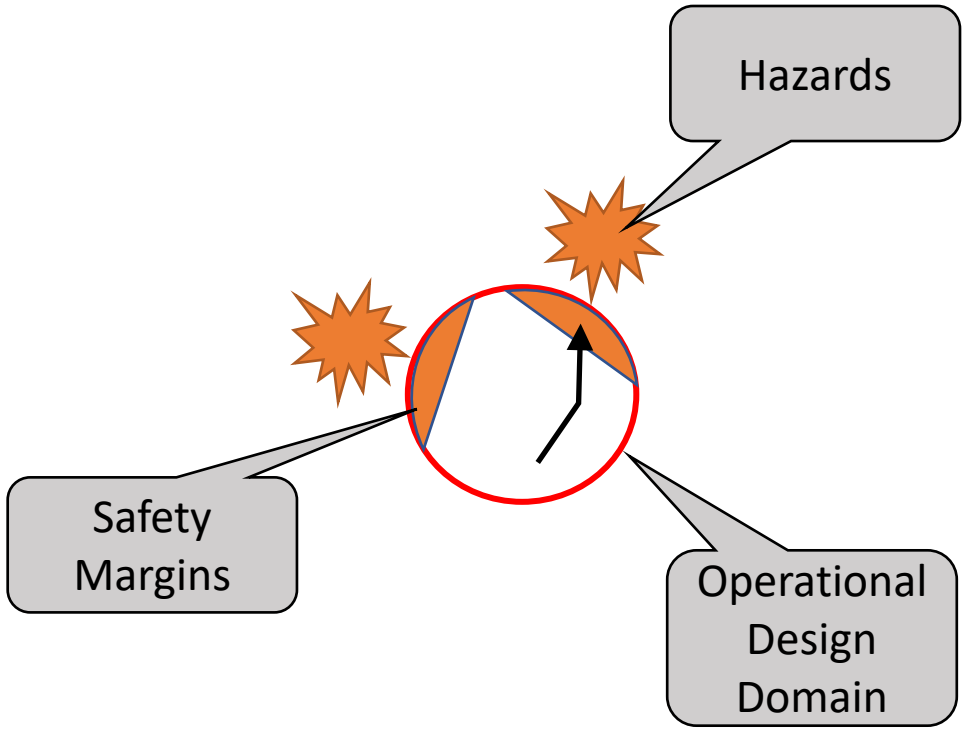
Anomaly Detection in Computer Vision

- Extensively studied for the past 10+ years
  - [Ruff, et al., 2021; Dietterich & Guyer, 2022]
- Advances in deep learning and vision foundation models have produced major improvements
- No method can guarantee to detect all novelty

Source: Artificio.org

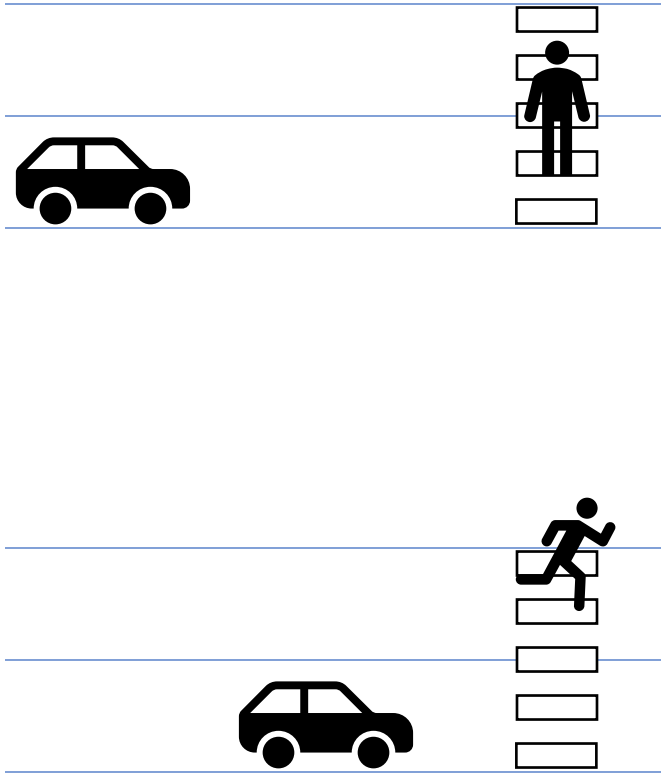
Detecting Near Misses

- Case 1: Known Hazards
  - During the design process, we have defined hazardous states and introduced margins of safety around them
    - Come too near to another object
    - Extreme steering and braking
  - Design should include sensors to detect when we enter those safety margin regions



Case 2: Counterfactual Near Misses

- Automatic Vehicle safety conditions
  - At least 2m separation between vehicle and pedestrians, cyclists, stationary obstacles
- Pedestrian sees car coming and jumps out of the way
- Car determines that it met the required 2m separation ➔ “no problem”
- Counterfactual: There would have been a safety violation if the pedestrian had not taken evasive action
- Pearl’s Theory of Causality provides the formal basis for computing counterfactual near misses [Pearl, 2009; Pearl & MacKenzie, 2018]

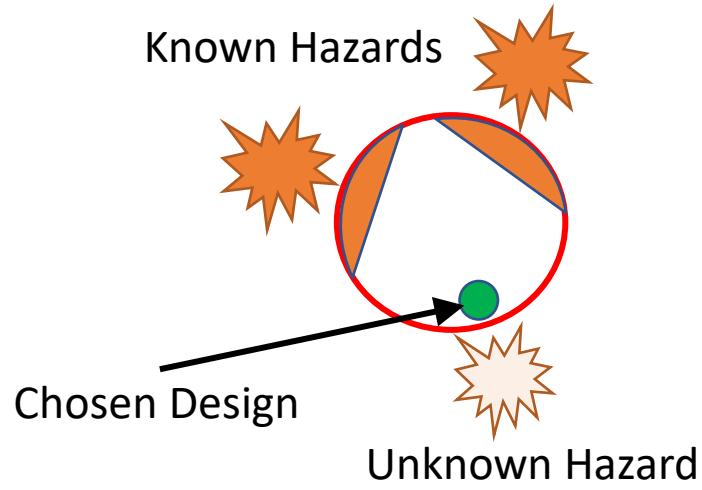


Outline

- Part 1: Integrating ML into traditional safety engineering processes
  - Scenario-based data collection
  - Verification of function behavior
- Part 2: Safety as Control
  - Detecting anomalies and near misses
- Part 3: Safety as Continual Redesign
  - Design is never finished
  - Resilient systems are “Poised to adapt”

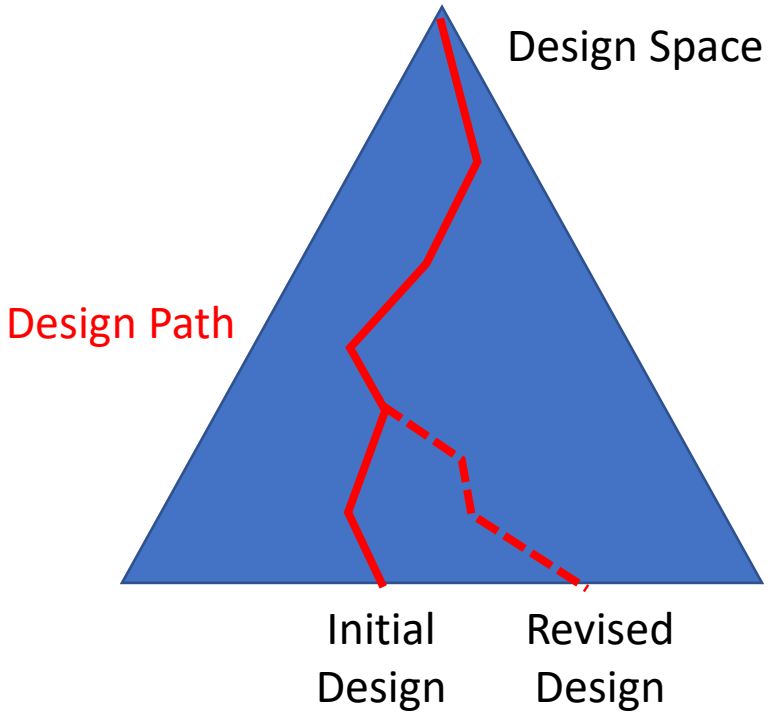
Creating Resilient Systems

- Engineered systems are “robust yet fragile”
  - Robust to the known hazards
  - Vulnerable to novel failure modes
- Optimization for cost, weight, power, etc. results in designs near the edge of the feasible region
  - Highly Optimized Tolerances (HOT) theory. [Carson & Doyle, 2002]



Creating Resilient Systems

- David Woods: A resilient system is one that is "poised to adapt"
  - [Woods, 2024a, 2024b]
- An AI perspective:
  - The entire design process should be regarded as one path through a design space
  - Adaptation requires following new paths through that space
  - Build AI tools for continual design



The design space and design process should be "kept on standby" so that they can be resumed whenever adaptation is required

Summary

- ML and traditional safety engineering: Managing Known Hazards
  - ML needs to develop verification methods to ensure distribution-independent generalization
- Safety as Control
  - Novel hazards as system disturbances
  - KPIs: Anomalies and Near Misses
  - AI methods for anomaly detection are mature
  - AI needs methods for detecting counter-factual near misses
- Safety as Resilience: “Poised to Adapt”
  - The design space and design process “kept on standby” so that they can be invoked whenever adaptation is required

Implications for safety of general-purpose AI

- Very wide range of potential harms
- No defined operational design domain (ODD) except in narrow vertical applications
- Extremely frequent technology disruptions that change the scope of the system and the potential harms
- Poorly-developed “controller” for maintaining safety
  - Organizations responsible for managing the systems after deployment (commercial and governmental)
  - Regulatory frameworks
  - Nation state competition

References

- Carlson, J. M., & Doyle, J. (2002). Complexity and robustness. *Proceedings of the National Academy of Sciences*, *99*(Supplement 1), 2538–2545. https://doi.org/10.1073/pnas.012582499
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Li, F.-F. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/CVPR.2009.5206848
- Dietterich, T. G., & Guyer, A. (2022). The Familiarity Hypothesis: Explaining the Behavior of Deep Open Set Methods. *ArXiv*, *2203.02486*(v1). http://arxiv.org/abs/2203.02486
- Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., & Leahy, C. (2020). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *ArXiv*, */2101.0002*(v1). http://arxiv.org/abs/2101.00027
- Kochenderfer, M., Katz, S. M., Corso, A. L., Moss, R. J. (forthcoming) *Algorithms for Validation. MIT Press. Cambridge, MA.* Draft available for comment from the authors. https://algorithmsbook.com/validation/files/val.pdf
- Leveson, N. G. (2011). *Engineering a Safer World: Systems Thinking Applied to Safety. MIT Press. Cambridge, MA.*
- Pearl, J. (2009). *Causality*. Cambridge University Press.
- Pearl, J., & Mackenzie, D. (2018). *The Book of Why*. Basic Books.
- Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Gregoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, Klaus-Robert Mueller (2021). A Unifying Review of Deep and Shallow Anomaly Detection**.** *Proc. IEEE, 109* (5): 756-795
- Singh, H., Kumar, P., Torr, P. H. S., & Dvijotham, K. (2021). Overcoming the Convex Barrier for Simplex Inputs. *Advances in Neural Information Processing Systems (NeurIPS 2021)*, 12.
- Verma, A. K., Ajit, S., Karanki, D. R. (2010) *Reliability and Safety Engineering, 2ed*, Springer.
- Weick, K., Sutcliffe, K., & Obstfeld, D. (1999). Organizing for high reliability: Processes of collective mindfulness. In R. S. Sutton & B. M. Staw (Eds.), *Research in Organizational Behavior* (Vol. 1, pp. 81–123). Jai Press.
- Woods, D. D. (2024a). Resolving the Command – Adapt Paradox: Guided Adaptability to Cope with Complexity. In J. Le Coze & B. Journé (Eds.), *Compliance and Initiative in the Production of Safety* (pp. 73–87). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45055-6
- Woods, D. D. (2024b). Limits of Automata — Then and Now: Challenges of Architecture , Brittleness , and Scale. *Journal of Cognitive Engineering and Decision Making*, *0*(0). https://doi.org/10.1177/15553434241240203