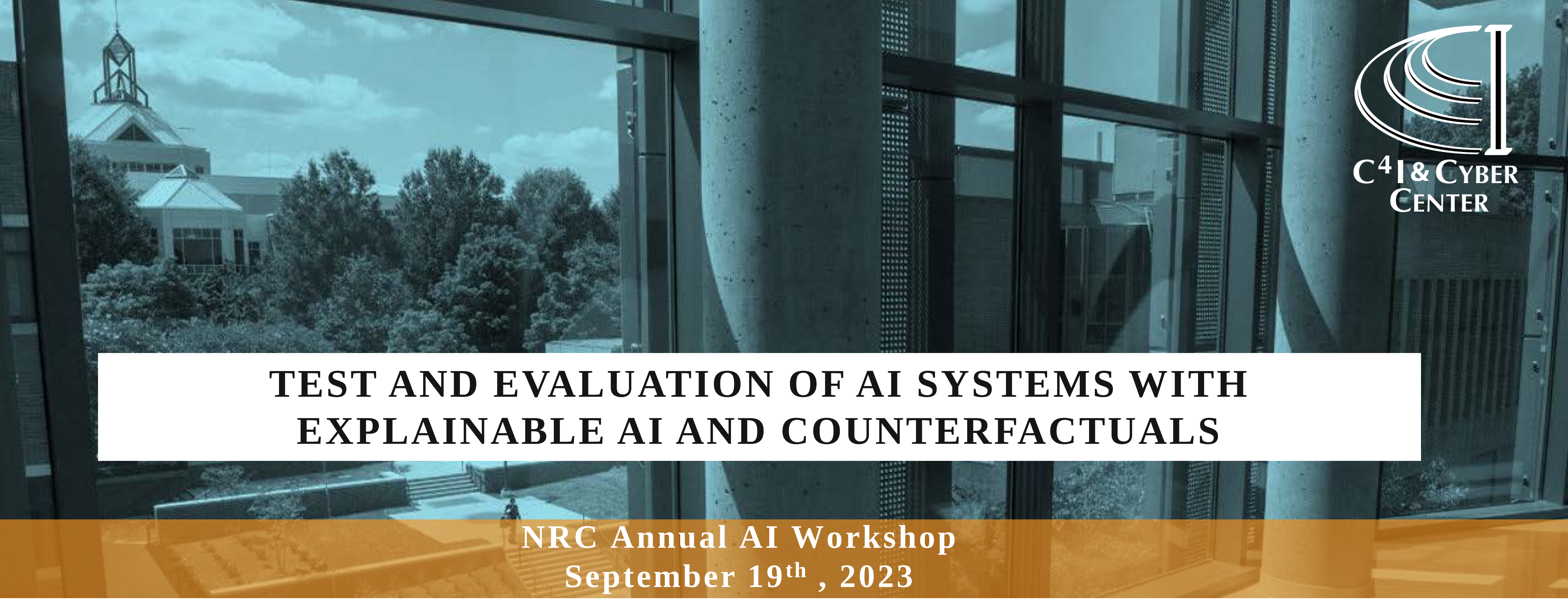


TEST AND EVALUATION OF AI SYSTEMS WITH EXPLAINABLE AI AND COUNTERFACTUALS

NRC Annual AI Workshop
September 19th , 2023

C4I & CYBER CENTER

Ali K. Raz

Assistant Professor Systems Engineering

Assistant Director of C4I and Cyber Center

George Mason University

araz@gmu.edu

GEORGE MASON UNIVERSITY

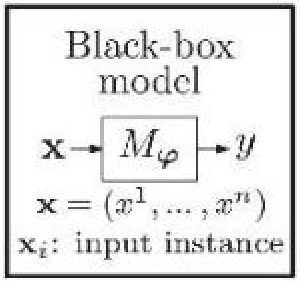
THE NEED FOR TEST AND EVALUATION (T&E) OF AI

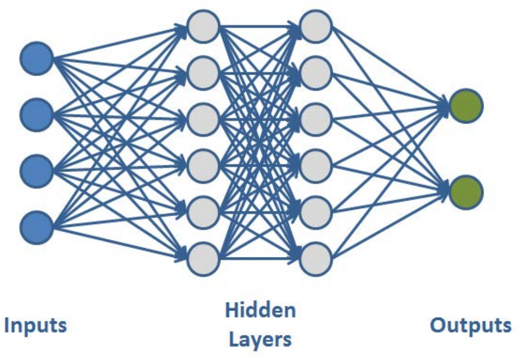
Deep Neural Networks (DNNs) are most common form of AI

- ✓ State-of-the-art implementation for AI/ML algorithms *(Supervised, Unsupervised, Reinforcement Learning, Natural Language Processing etc.)*
- ✓ Well-established performance outcomes in a variety of applications *(Intuitive and non-intuitive outcomes)*
- ✓ Strong focus on algorithmic development, computational efficiency, and implementation
- ✓ Selective demonstration of test cases, mostly based on training data partitioning in training and validation sets

Common Challenges for DNNs

- ❑ **Trained DNNs are essentially backboxes to the designers and users**
- ❑ Limited characterization of performance bounds due to variations and uncertainties; limited Monte Carlo simulations and user selected variations
- ❑ Limited explanation of black-box decision-making logic
- ❑ Limited evaluation of acceptable and unacceptable performance regions





Systems Engineering Perspective Example T&E Questions to Ask

- ❑ What is the impact of variations in input data and environment?
- ❑ **How does the input (i.e., observed state) influence DNNs decision making?**
- ❑ Does training data considers edge cases?
- ❑ **How does the DNNs respond to modeled (i.e., included in training) and unmodeled uncertainties?**

Systems Engineering Call for Machine Learning Problems

Unsolved Problems in ML Safety*

Robustness Create models that are resilient to adversaries, unusual situations, and Black Swan events.

Monitoring Detect malicious use, monitor predictions, and discover unexpected model functionality.

Alignment Build models that represent and safely optimize hard-to-specify human values.

Systemic Safety Use ML to address broader risks to how ML systems are handled, such as cyberattacks.

Systems Engineering for AI (SE4AI)

"Systems Engineering is a transdisciplinary and integrative approach to enable the successful realization, use, and retirement of engineered systems, using systems principles and concepts, and scientific, technological, and management methods." (INCOSE)

System T&E Create models that are resilient to adversaries, unusual situations, and Black Swan events.

Functional Interactions ıus use, monitor predictions, and discover unexpected model functionality.

Stakeholder Analysis Build models that represent and safely optimize hard-to-specify human values.

External Systems Diag. address broader risks to how ML systems are handled, such as cyberattacks.

In this presentation

Explainable AI
- How does the input (i.e., observed state) influence DNNs decision making?

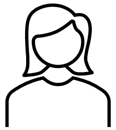
Counterfactual Testing
- How does DNNs respond to unmodeled uncertainty?

Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. "Unsolved problems in ml safety." *arXiv preprint arXiv:2109.13916* (2021)
**Lewis, David K. 1973. *Counterfactuals*. Cambridge: Harvard University Press.

Why Explainable AI (XAI)?

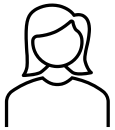
Let's try a thought experiment

Q: What will be the weather tomorrow?



It will be chilly and cloudy.
Remember to put on a warm jacket

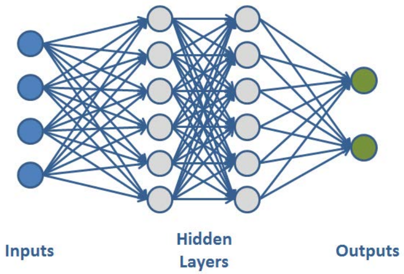
Q: How do you know what will be the weather tomorrow?



I heard it on the radio
I looked up on my phone
Weather radar showed a cold front
I love looking at NOAA models, you want to know the barometric pressure!

Can we be okay with lack of explainability?

Datasets/Models/ Rewards



Deep Neural Networks (DNNs)

- Create new materials
- Create new drugs
- Predict person's health/weight
- Predict a terrorist
- Reject loans

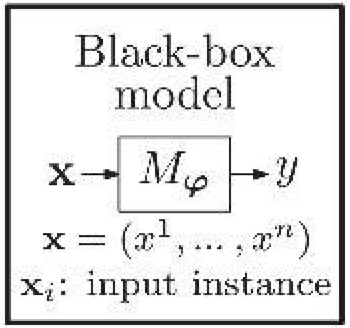
Example AI Uses

1. Why this action?
2. Why not another action?
3. When do I succeed/fail?
4. When can I trust the results?
5. How can I fix an error?

End User

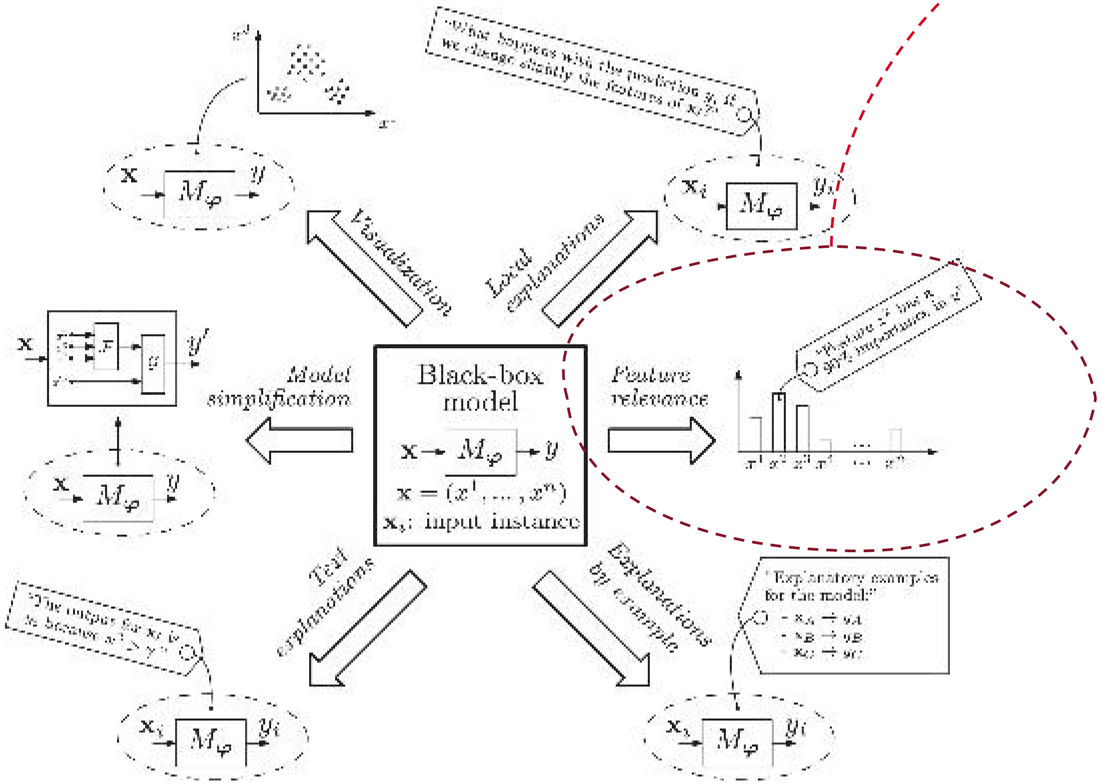
Different Flavors of Explainable AI (XAI)

- Explainable AI Nomenclature*
  - Interpretability
    - "provide the meaning in understandable terms to a human"
  - Explainability
    - Notion of explanation as an interface between humans and a decision maker
  - Transparency
    - "characteristic of a model to make a human understand its function"… "three categories: simulateable models, decomposable models and algorithmically [transparent]"

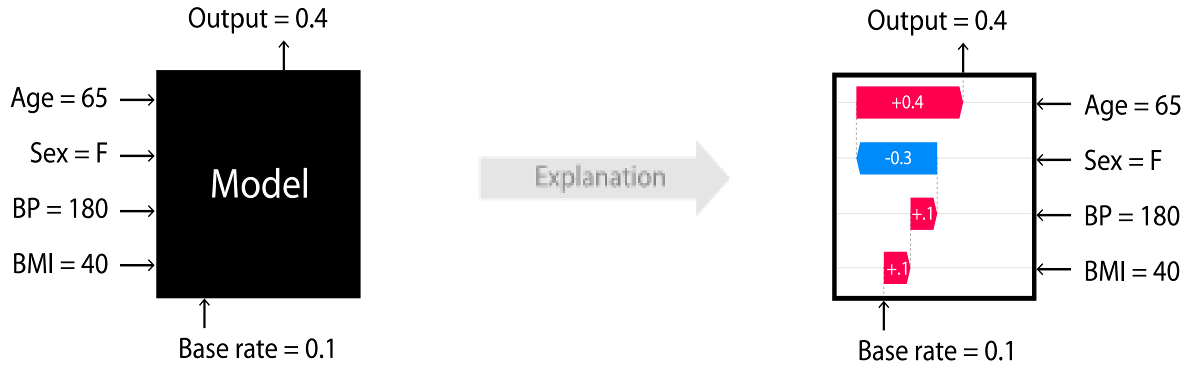


Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information fusion* 58 (2020): 82-115.

One Example of an Explainable Model



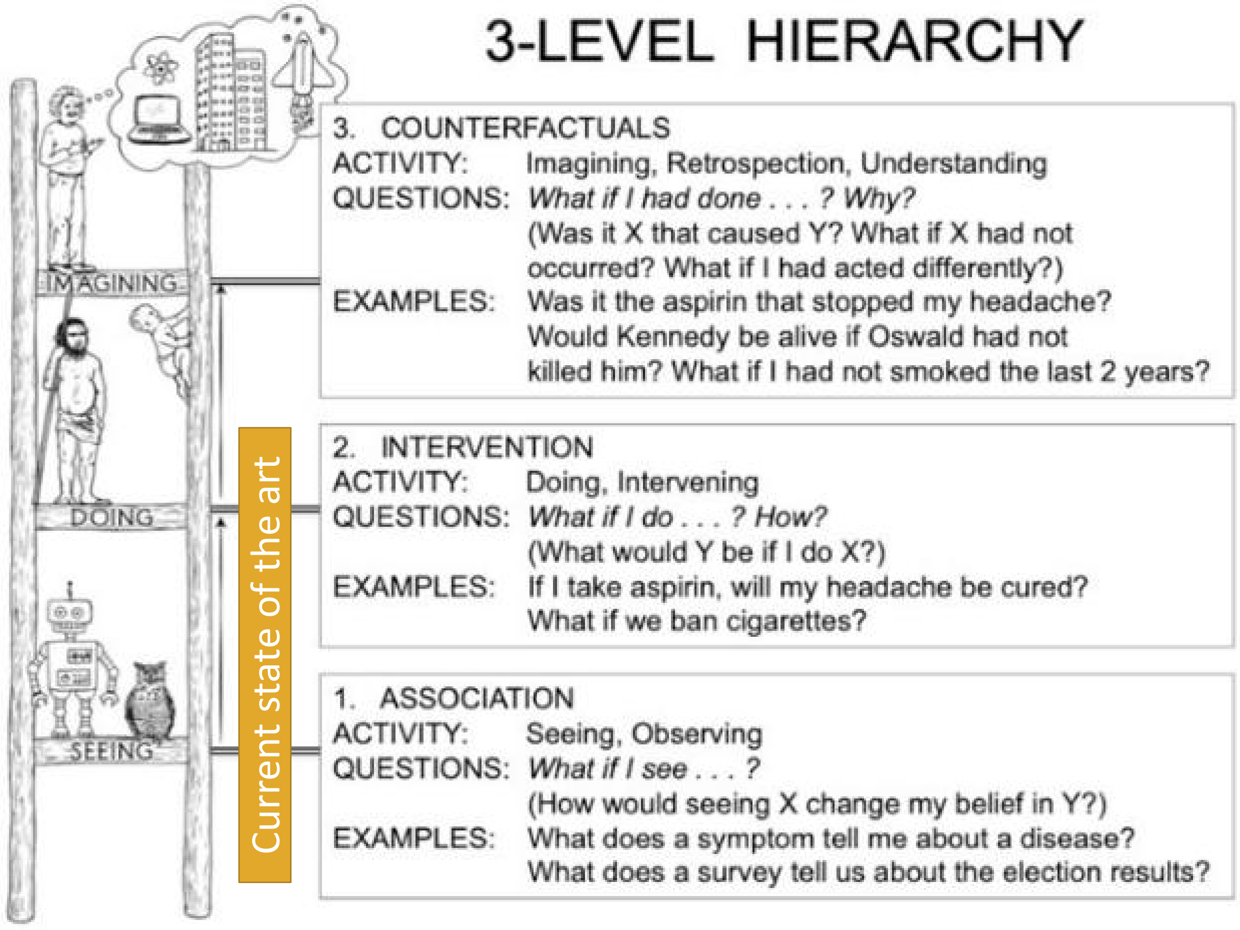
- SHapley Additive exPlanations (SHAP)
  - State of the art for reverse engineering the output of any predictive model
  - Yields importance of input features for a given prediction
  - Focuses on coalitions in cooperative game theory



- Investigates trained Deep Neural Network (DNN) models with analytical techniques to extract decision making attributes

Counterfactual Testing Concepts (Work in Progress)

- Complements the XAI approach by setting up a hypothesis which may very well be an antithesis to XAI output.
- Investigates the model response to situations that may not occur (or are known to be not represented by the model and/or the training/validation data sets).
- Ferrets out patterns of causality in the underlying model that would otherwise be left unexposed.
- Explores model outputs beyond what the model is trained to or exposed to under nominal and expected operational conditions.
- Provides the identifiability of the system.



Ladder of Causality*

*Pearl, Judea, and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.
**Lewis, David K. 1973. *Counterfactuals*. Cambridge: Harvard University Press.
***Encyclopedia of Social Science Research Methods*, edited by Michael Lewis-Beck (University of Iowa), Alan Bryman (Loughborough University), and Tim Futing Liao. Sage Publications.*
***https://highdemandskills.com/counterfactual/*

CAUSATION, COUNTERFACTUALS, AND XAI

- Causation:
  - Sufficient Causation: A has caused B
  - Necessary Causation: If not for A; B would not have occurred

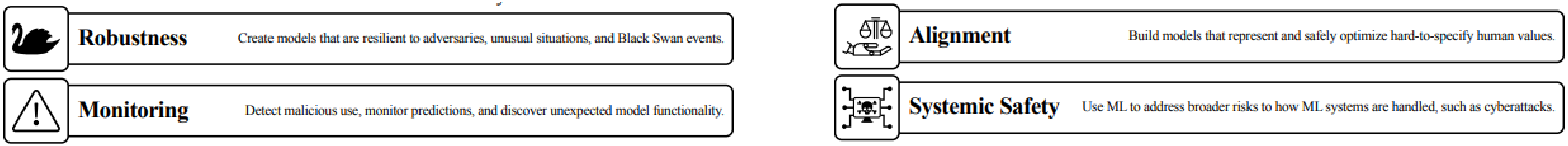
- XAI helps identify which features are most significant on the output
  - It does not examine what happens when such features are not present

- Counterfactual is about discovering the *necessary causation* (which maybe hypothetical).
  - Example*
    - "Joe's headache would have gone away if he had taken aspirin"
    - [if the first object had not been, the second had never existed]
  - Examining model response in counterfactual cases exposes the black box nature of the model
    - If a feature relevance method identifies the most or least significant input variable, the counterfactual test suggests removing the most significant feature from the model or making the least significant feature the only input to the model.
  - <u>Traditional guidance in the SE literature suggests avoiding antithetical or contradictory requirements and test case development, which on the contrary, is suggested by counterfactual testing.</u>

*Pearl, Judea, and Dana Mackenzie. *The book of why: the new science of cause and effect.* Basic books, 2018.
N. J. Roese, "Counterfactual thinking," Psychol. Bull., vol. 121, no. 1, pp. 133–148, 1997, doi: 10.1037/0033-2909.121.1.133.

Concluding Remarks

Unsolved Problems in ML Safety*



Systems Engineering of AI is needed to help address these problems and transition AI into practical systems

- Explainable AI and Counterfactual Testing help expose DNNs decision-making and limitations
  - Characterize performance envelopes of the system; emergent behavior, and robustness
- Explainable AI and Counterfactuals help perform system identification and system-level integration of embedded AI components
  - Need wider adoption of Explainable AI in System T&E and Modeling and Simulation practices
  - Counterfactual examples to date remain discrete transactions (e.g., mortgage applications) – need to explore value for design and testing of dynamic and embedded real time systems subject to noisy inputs

Thank You!

Ali K. Raz

Assistant Professor Systems Engineering

Assistant Director of C4I and Cyber Center

George Mason University

araz@gmu.edu

Generating Counterfactuals (Work in Progress)

- **Generating Counterfactuals**
  1. The user of a counterfactual explanation defines the alternative reality by making a relevant change in the prediction of an instance. (o_1)
  2. The counterfactual should be as similar as possible to the instance regarding feature values and should be selected to change as few features as possible. (o_2)
  3. Generate multiple diverse counterfactual explanations to provide multiple viable ways of generating a different outcome. (o_3)
  4. A counterfactual instance should have feature values that are likely according to the joint distribution of the data. (o_4)

Advantages

- o The method does not require knowledge of the data or the model; it only requires knowledge of the model's prediction function (not unique to machine learning).
- o The method is relatively easy to implement since it is a loss function that can be optimized with standard optimization libraries.

Disadvantages

- o Each instance of a counterfactual usually has multiple explanations.
- o Multiple explanations can be disconcerting to people who prefer a single, simple, unique explanation.

Generating Counterfactuals

- **Simultaneously Minimize the four-objectives Loss Function**

$L(x, x', y', X^{obs}) = (o_1(f\hat{}(x'), y'), o_2(x, x'), o_3(x, x'), o_4(x', X^{obs}))$

Where:

$$o_1(\hat{f}(x'), y') = \begin{cases} 0 & \text{if } \hat{f}(x') \in y' \\ \inf_{y' \in y'} |\hat{f}(x') - y'| & \text{else} \end{cases}$$

$$o_2(x, x') = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x_j, x'_j)$$

$$\delta_G(x_j, x'_j) = \begin{cases} \frac{1}{\hat{R}_j} |x_j - x'_j| & \text{if } x_j \text{ numerical} \\ \mathbb{I}_{x_j \neq x'_j} & \text{if } x_j \text{ categorical} \end{cases}$$

$$o_3(x, x') = ||x - x'||_0 = \sum_{j=1}^{p} \mathbb{I}_{x'_j \neq x_j}.$$

$$o_4(x', \mathbf{X}^{obs}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x'_j, x_j^{[1]})$$

- The method is to minimize all four objectives o_1, o_2, o_3, and o_4 simultaneously, and not to collapse them into a single objective weighted sum.
- The fitness of the counterfactual vector of objectives (o_1, o_2, o_3, o_4) is the vector having the lowest values of o_i

S. Dandl, C. Molnar, M. Binder, and B. Bischl. "Multi-objective counterfactual explanations". In: Bäck T. et al. (eds) *Parallel Problem Solving from Nature* – PPSN XVI. PPSN 2020. Lecture Notes in Computer Science, vol 12269. Springer, Cham 2020.
A. Van Looveren and J. Klaise, Interpretable Counterfactual Explanations Guided by Prototypes, arXiv:1907.02584v2 [cs.LG] 18 Feb 2020.

Example Application of Explainable AI to High-Speed Aerospace System Control

Vehicle Model Parameters

- **States:**
  h: altitude, θ: downrange angle,
  v: velocity, γ: flight path angle
- **Control:** α: angle of attack
- **Dynamics:**

$$\dot{x} = \begin{bmatrix} \dot{h} \\ \dot{\theta} \\ \dot{v} \\ \dot{\gamma} \end{bmatrix} = \begin{bmatrix} v \sin\gamma \\ \frac{v}{r}\cos\gamma \\ -\frac{D(\alpha)}{m} - \frac{\mu}{r^2}\sin\gamma \\ \frac{L(\alpha)}{mv} - \left(\frac{v}{r} - \frac{\mu}{vr^2}\right)\cos\gamma \end{bmatrix}$$

- **Objective:** $J = \min t_f = \int_0^{t_f} dt$
- **Initial Constraints:**

$$\Psi_0 = 0 = \begin{bmatrix} h & -30\text{ km} \\ \theta & \\ v & -3\text{ km/s} \\ \gamma & \end{bmatrix}_{t=t_0}$$

- **Path Constraint:**

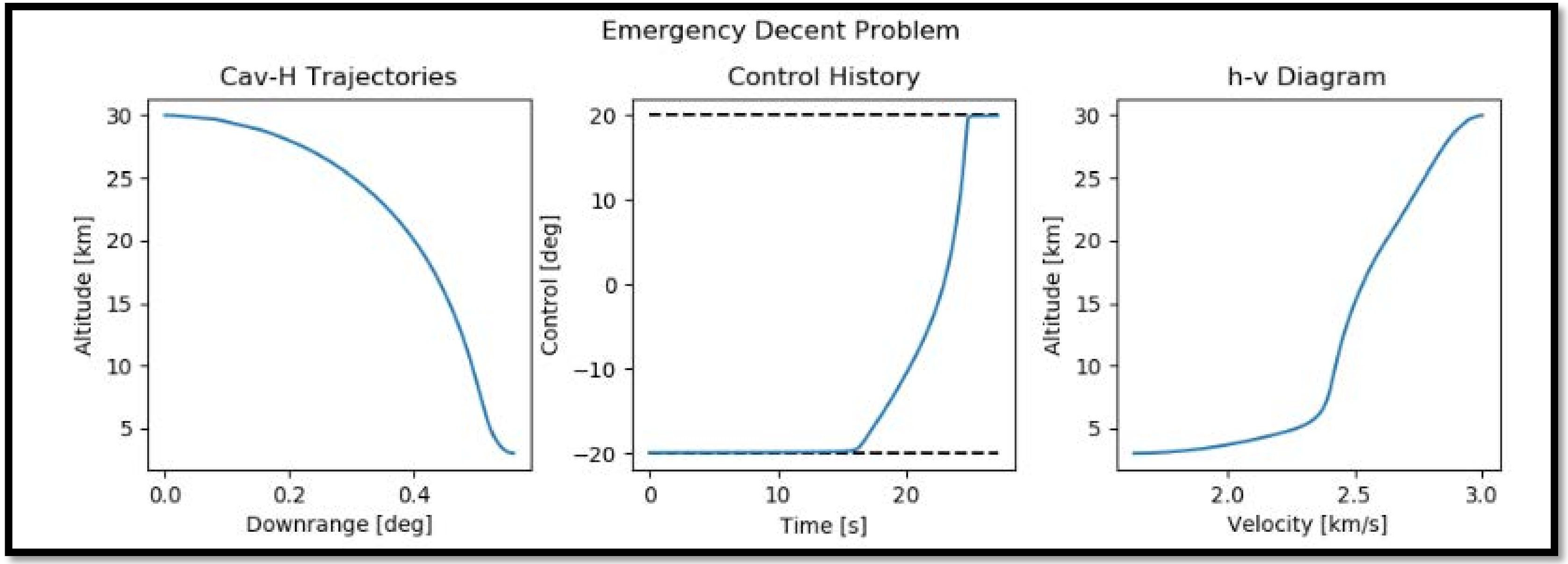
$$|\alpha| \leq 20°$$

- **Terminal Constraints:**

$$\Psi_f = 0 = \begin{bmatrix} h & -3\text{ km} \\ \gamma & \end{bmatrix}_{t=t_f}$$

Emergency Descent Problem for an Un-thrusted High-Speed Vehicle

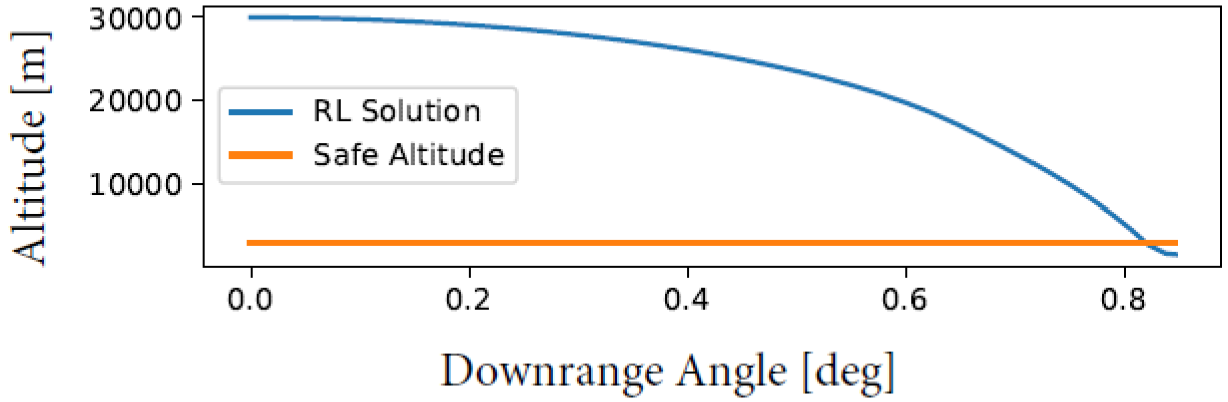
- The vehicle at 30 km altitude and 3 km/s velocity needs to descend to level flight at a safe altitude of 3 km in minimum time
- Constraints must be satisfied

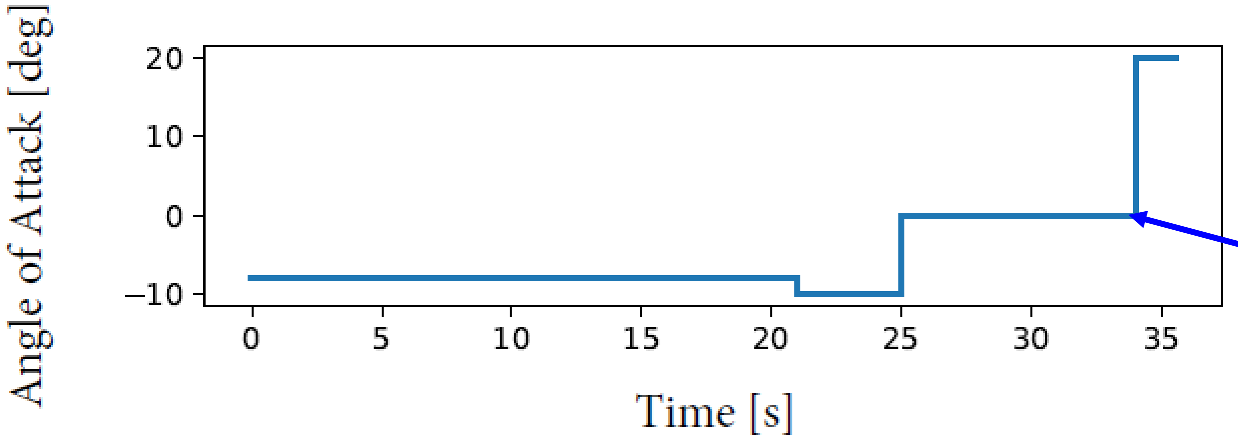


AI Results: Nominal Case (Vehicle Descent From 30 km to 3 km)

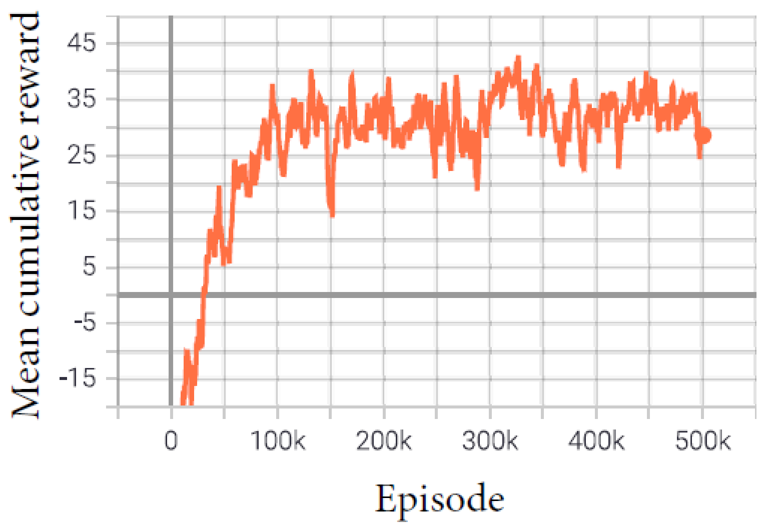
AI Training with Reinforcement Learning

- Provides AoA commands to guide the vehicle to a pre-determined safe altitude
- Included randomly sampling vehicle initial conditions
- Completed after 500k episodes





Sufficient cumulative reward of +30 to train policy



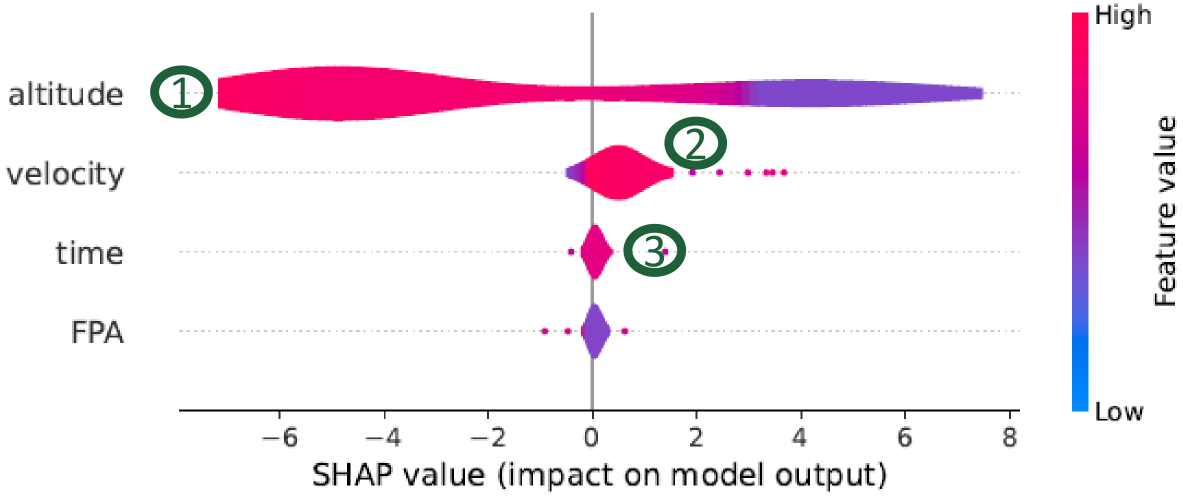
AoA commands issued by the AI agent

Raz, A. K., Nolan, S. M., Levin, W., Mall, K., Mia, A., Mockus, L., ... & Williams, K. (2022, March). Test and Evaluation of Reinforcement Learning via Robustness Testing and Explainable AI for High-Speed Aerospace Vehicles. In *2022 IEEE Aerospace Conference (AERO)* (pp. 1-14). IEEE.

Examination Via Explainable AI (XAI) Techniques

SHAP Applied to RL Problem

- **Inputs:** Time, Altitude, Velocity, and Flight Path Angle
- **Output:** Angle of Attack (between -20˚ and 20˚)
- **Number of Trajectories:** 1000
- **Objective:** Reach a particular target in a minimum time



1. Higher altitude values oppose a change in AoA whereas lower altitudes support it.
2. Higher velocity values positively influence change in AoA
3. FPA and Time have least impact.

Raz, A. K., Nolan, S. M., Levin, W., Mall, K., Mia, A., Mockus, L., ... & Williams, K. (2022, March). Test and Evaluation of Reinforcement Learning via Robustness Testing and Explainable AI for High-Speed Aerospace Vehicles. In *2022 IEEE Aerospace Conference (AERO)* (pp. 1-14). IEEE.

References

S. Dandl, C. Molnar, M. Binder, and B. Bischl. "Multi-objective counterfactual explanations". In: Bäck T. et al. (eds) *Parallel Problem Solving from Nature* – PPSN XVI. PPSN 2020. Lecture Notes in Computer Science, vol 12269. Springer, Cham 2020.

N. J. Roese, "Counterfactual thinking," Psychol. Bull., vol. 121, no. 1, pp. 133–148, 1997, doi: 10.1037/0033-2909.121.1.133.

W. F. Lawless, R. Mittu, D. A. Sofge, T. Shortell, and T. A. McDermott, Eds., *Systems Engineering and Artificial Intelligence*. Cham: Springer International Publishing, 2021. doi: 10.1007/978-3- 030-77283-3.

W. F. Lawless, J. Llinas, D. A. Sofge, and R. Mittu, Eds., *Engineering Artificially Intelligent Systems: A Systems Engineering Approach to Realizing Synergistic Capabilities*, vol. 13000. Cham: Springer International Publishing, 2021.

D. P. F. Möller, *Guide to Computing Fundamentals in Cyber- Physical Systems*. Cham: Springer International Publishing, 2016.

C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.), 2022. christophm.github.io/interpretable-ml-book/

A. Van Looveren and J. Klaise, Interpretable Counterfactual Explanations Guided by Prototypes, arXiv:1907.02584v2 [cs.LG] 18 Feb 2020.