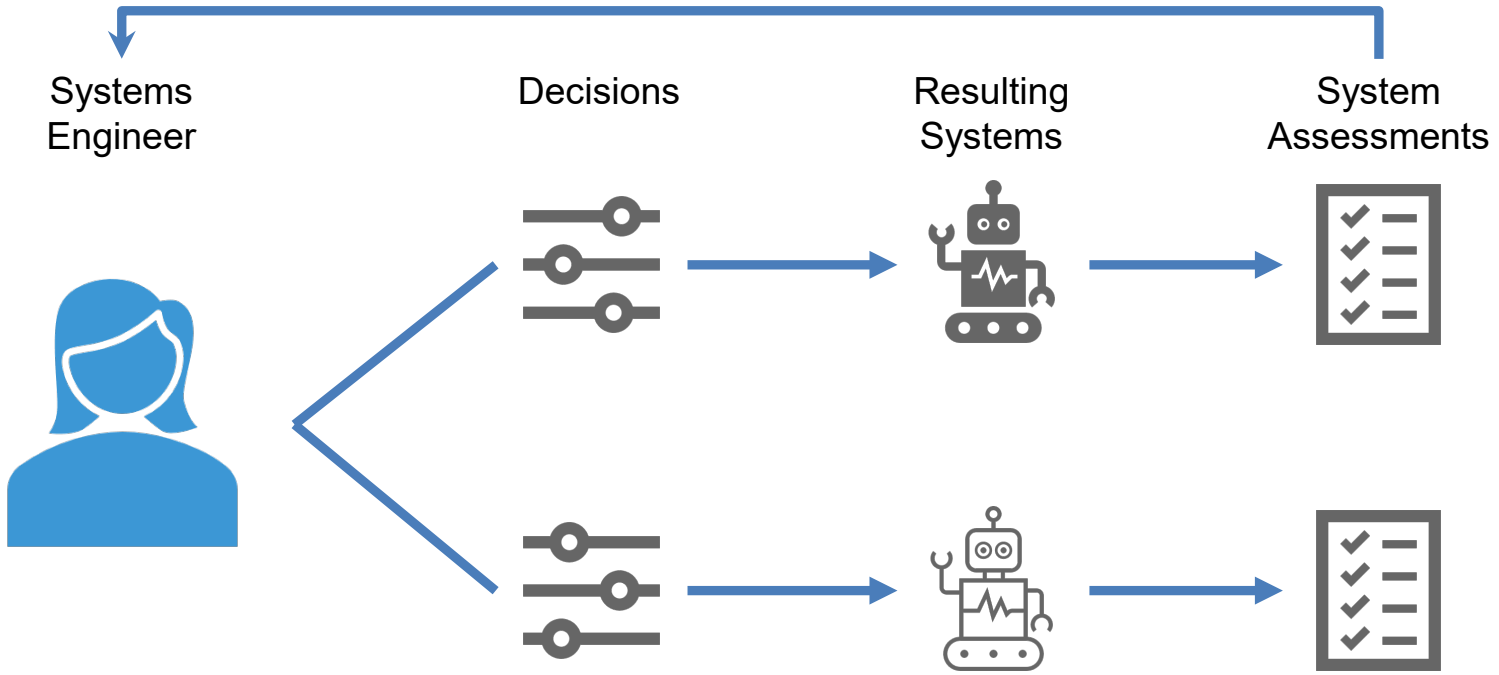
Insights from Multidisciplinary Research on Assessment of AI Systems

Dr. Bryan Mesmer
Dr. Vineetha Menon
Dr. Nathan Tenhundfeld
Dr. Sampson Gholston
Dr. Kristin Weger
Dr. Lisa Vangsness
Dr. Howard Chen
Dr. Hanumanthrao Kannan
Dr. Ana Wooley

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

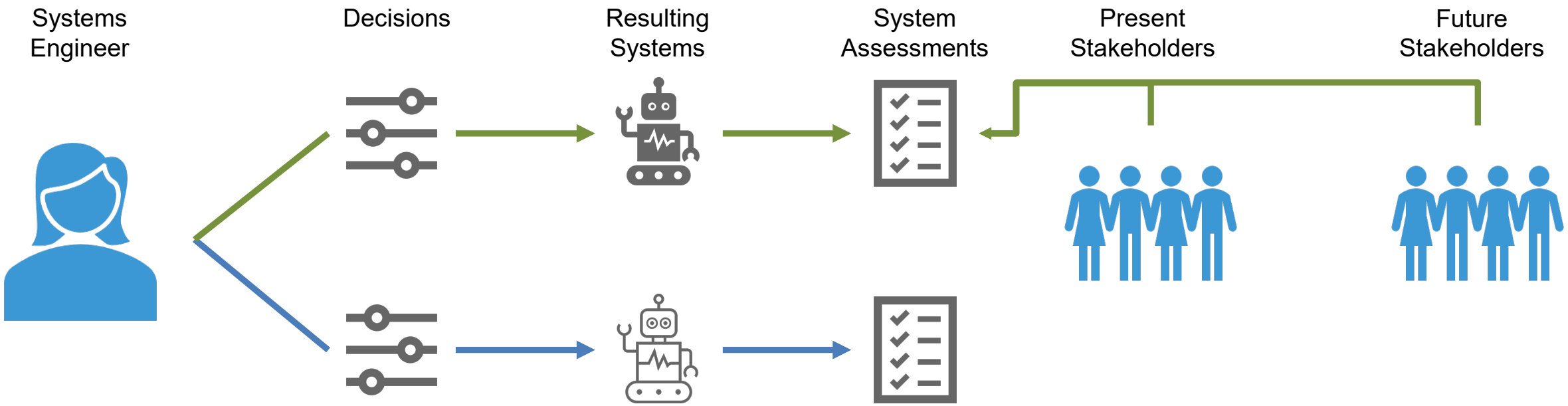
Research Motivation

- Systems Engineers make decisions to enable the creation of a system.
- The goodness of a decision is often measured by assessing the resulting system.



Research Motivation

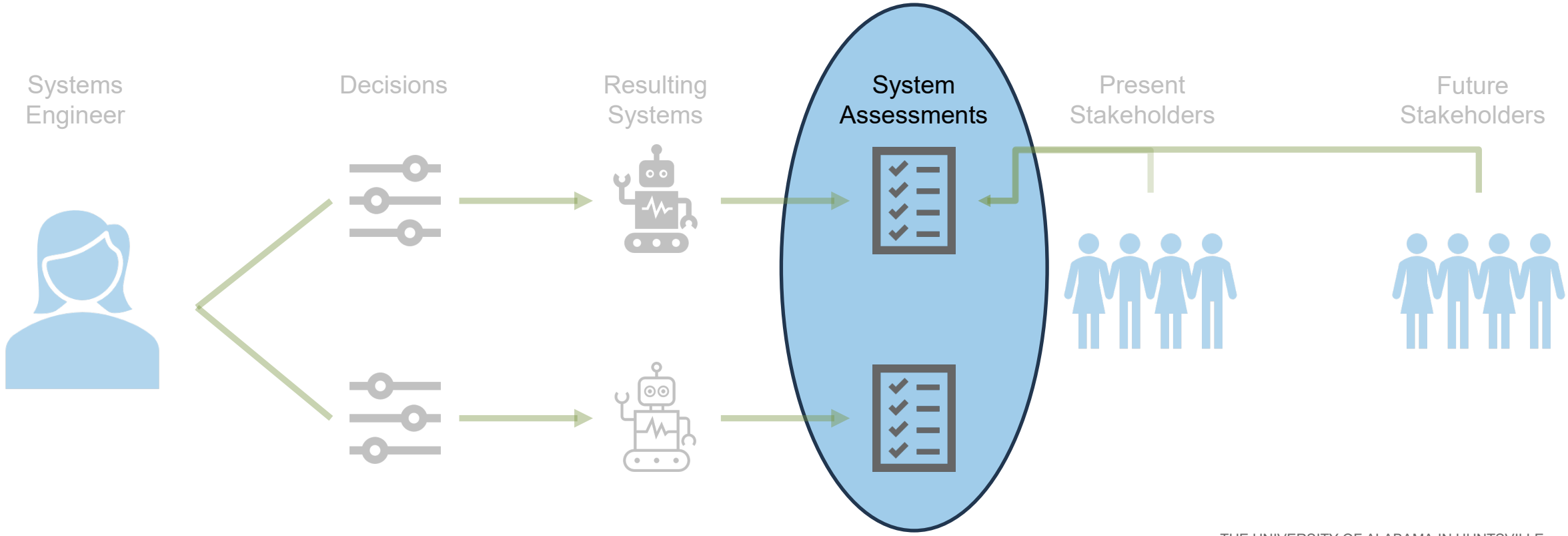
- Meaningful assessments and rigorous decision-making processes enable repeatable and justifiable decision-making.
- Assessments allow present and future stakeholders to understand why alternatives were selected.



Research Motivation

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- The focus of this research is on improving assessments for systems that incorporate AI.



Research Challenge

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Systems incorporating AI:
  - are inherently interdisciplinary;
  - typically have many different stakeholders;
  - have additional uncertainties compared to the same system without AI.

- Such complex challenges need to be addressed using integrated interdisciplinary approaches

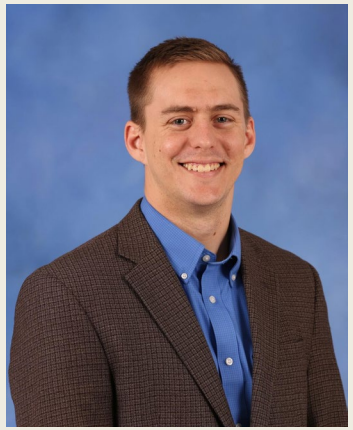
- Research Goal:
  - Develop a comprehensive framework for assessing the efficacy of AI in human agent teams by integrating computer science, psychological, and engineering approaches.

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

UAH INTERDISCIPLINARY SYSTEM STUDIES TEAM (InSyST)

InSySST Faculty Researchers

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE



Psychology

Computer Science

Systems and Industrial Engineering

InSyST Student Researchers

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Psychology

Andrew Atchley
Ginger Sullivan
Graduate

Luke Symasek
Undergraduate

Industrial and Systems Engineering

Aubrey Northam
Christopher White
Oluyinka Adedokun
Graduate

Meredith Bates
Sarah Andrews
Undergraduate

Computer Science

Joey Schwalb
Bishwas Praveen
Shivangi Gupta
Dylan Wright
Graduate

Eric Sung
Undergraduate

Research Objectives

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Overall Objective: Improve assessment of AI incorporated systems

- Perceived difficulty in assessing AI incorporated systems may be due to perceived system characteristics including:
  - Evolving behaviors
  - Immeasurable maintenance metrics
  - User acceptance and adoption problems
  - Vagueness in system decision making
  - Challenges in integrating with legacy systems
  - Etc.

Research Objectives

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- This work focuses on 3 perceived characteristics of AI incorporated systems that impact the ability for assessment.

- Specific objectives of this research are:
  - Define **reliability** and identify appropriate measures
  - Form techniques to **explain AI** decisions and the decision-making process to users
  - Identify sources of **emergent behaviors** in AI incorporated systems

Objective A: Reliability

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Define reliability and identify appropriate measures

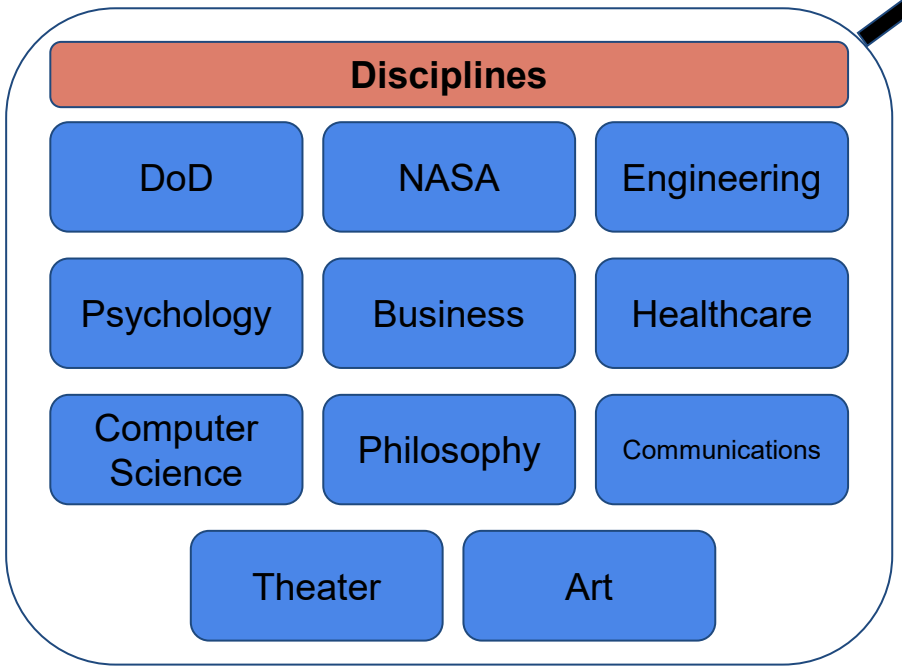
Reliability: The Problem

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Reliability is a foundational concept for effective human-machine teaming
- Reliability is used differently by different disciplines
- As stakeholders identify requirements, and as researchers & designers/developers work to adhere to these requirements, there needs to be clarity about what “reliability” means
- We need to know:
  - Is use of ‘reliability’ consistent within domains?
  - Is the use of ‘reliability’ consistent with definitions used by specific domains?

Reliability: Methodology

- Literature Review
  - Multidisciplinary
  - Identify what definitions of reliability are being used



Keyword Search on Databases
"Reliability" + [Discipline]

Select Articles which define reliability
[N = 197]

Extract Definition by Discipline

Reliability: Results

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Literature Review
  - Two main categories for definitions emerged:
    - Performance based
      - DoD: "... a measure of the probability that the system will perform without failure over a specific interval, under specified conditions"
      - Business: "... the degree to which a piece of accounting information objectively represents an underlying economic construct."
    - Consistency based
      - Psychology: "... a measure of the likelihood of getting the same result if an experiment or observation is repeated…"
  - The categories tend to align with discipline
    - E.g., engineering is more performance based, whereas psychology refers to more consistency based

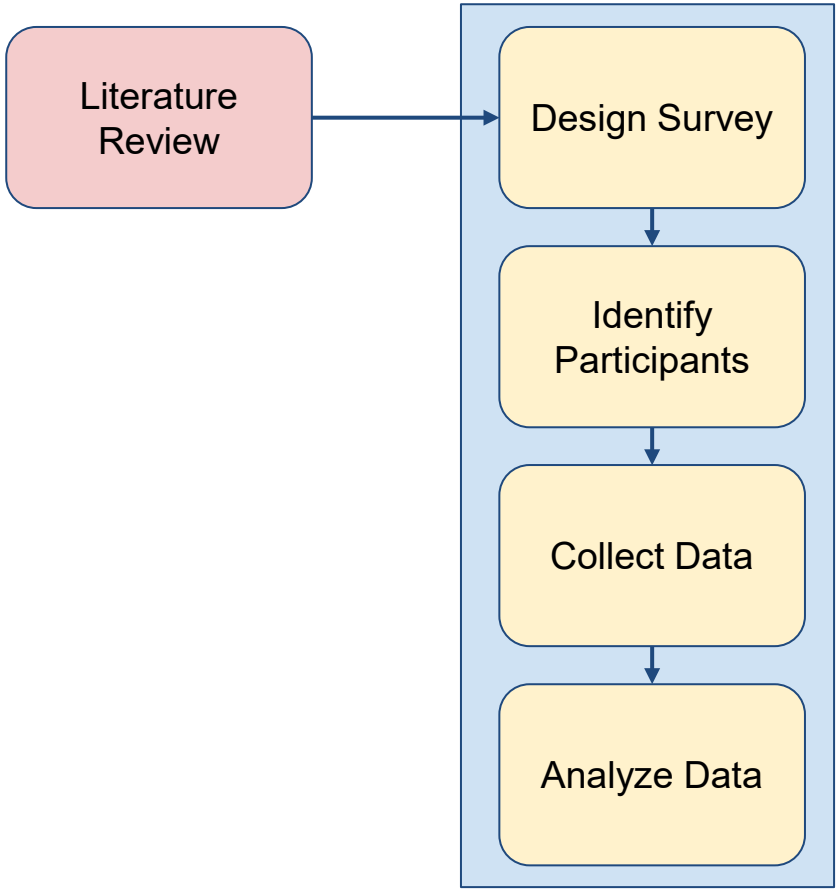
Reliability: Results

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

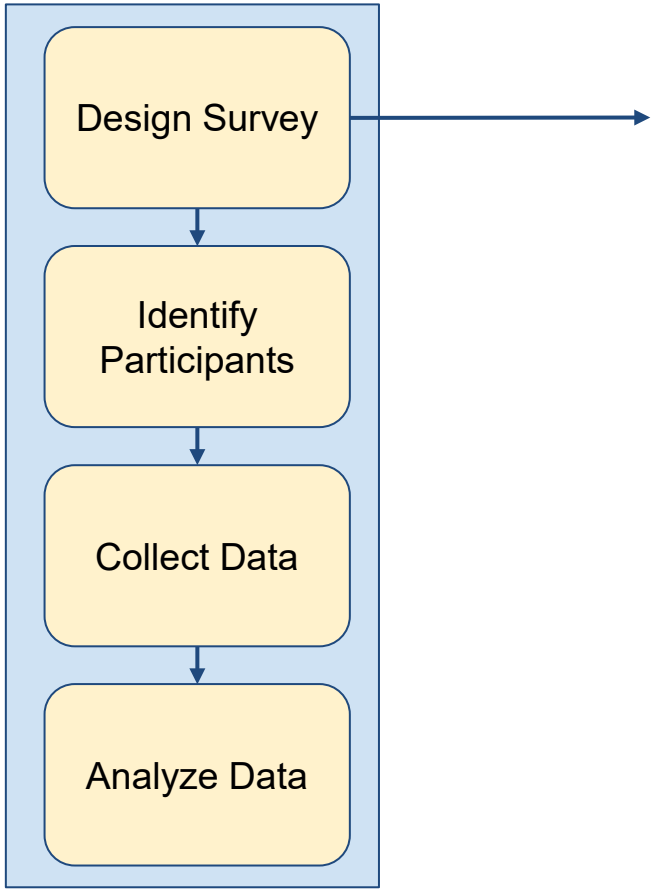
- Literature Review
  - Sometimes inconsistencies within disciplines
  - For example in Engineering:
    - "... the probability that an item will perform its intended function for a specified interval under stated conditions" **(Performance)**
    - "... data that has reliability reflect stable and consistent data collection processes and analyses over time" **(Consistency)**
  - or in Human Factors:
    - "Repeatability or consistency, a measure of the likelihood of getting the same result if an experiment or observation is repeated" **(Consistency)**
    - "...the probability that automation performs its assigned tasks correctly." **(Performance)**

Reliability: Methodology

- Survey
  - Understand how practitioners are defining reliability
  - Identify discrepancies between what was identified in literature review & use of term in practice



Reliability: Methodology

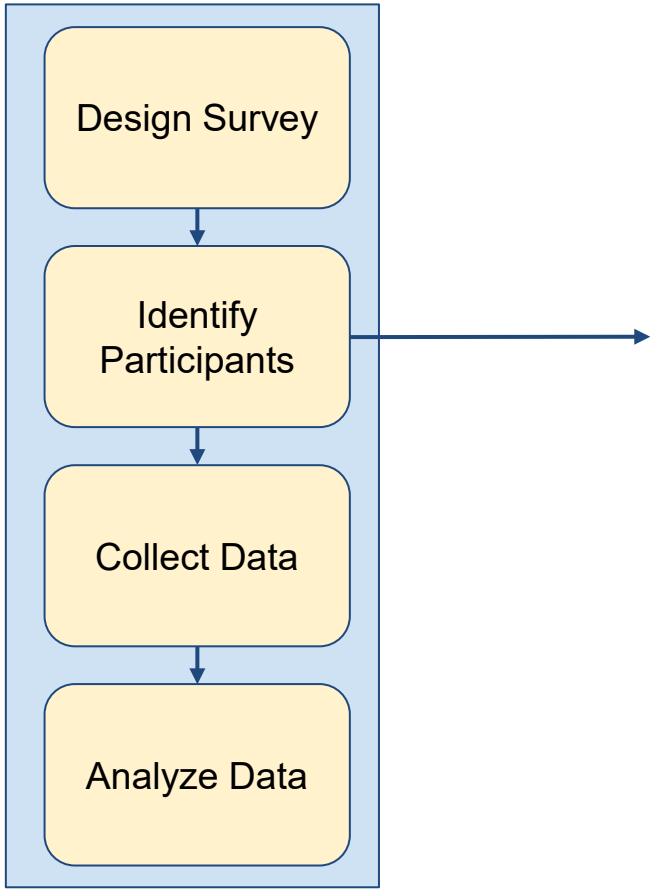


Demographics & Experience

Questions

- Define reliability in your own words as it relates to Autonomy and AI
- What constitutes “good” reliability?
- What constitutes “bad” reliability?
- How do you assess reliability?
- What factors affect reliability in an autonomous system?
- What are the consequences, in your field, if the system is unreliable?
- Rank importance of terms
  - Time, consistency, stability, repeatability, outcome (binary: pass/fail), probability of success, fairness, accountability, transparency, accuracy, ease of use
- Present various definitions of reliability and make them choose which they prefer.
- Follow up to previous, what did you like about that definition, and what needs to change?
- Present them with their prior definition of reliability, and ask if there is anything else they would want to change about it?

Reliability: Methodology



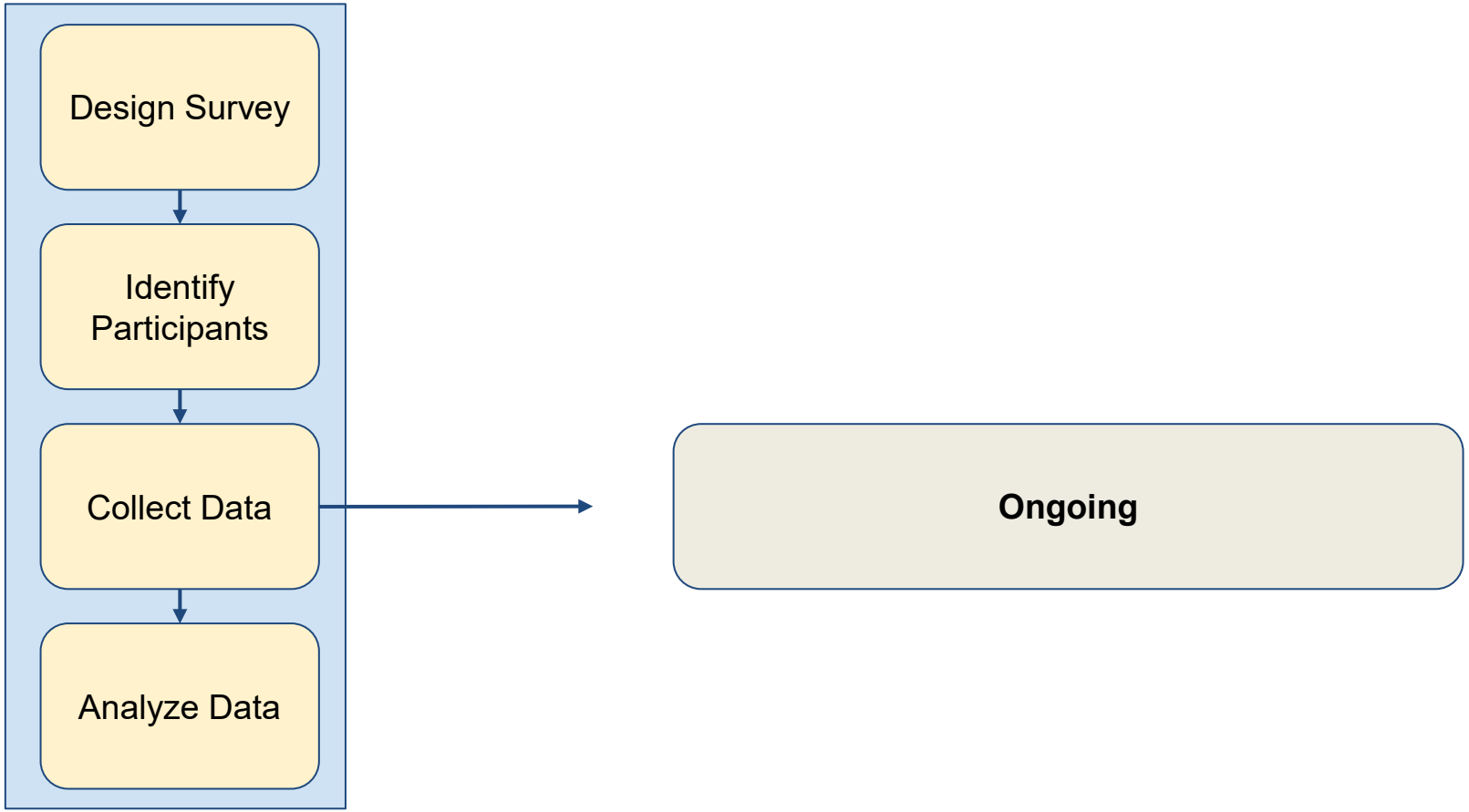
Employment

- Uniformed Personnel
- Civilian Contractors
- Industry Contractors
- Academics

Field

- Systems Engineer
  - Computer Science
- Human Factors
  - UX
- Designer
- Project Manager

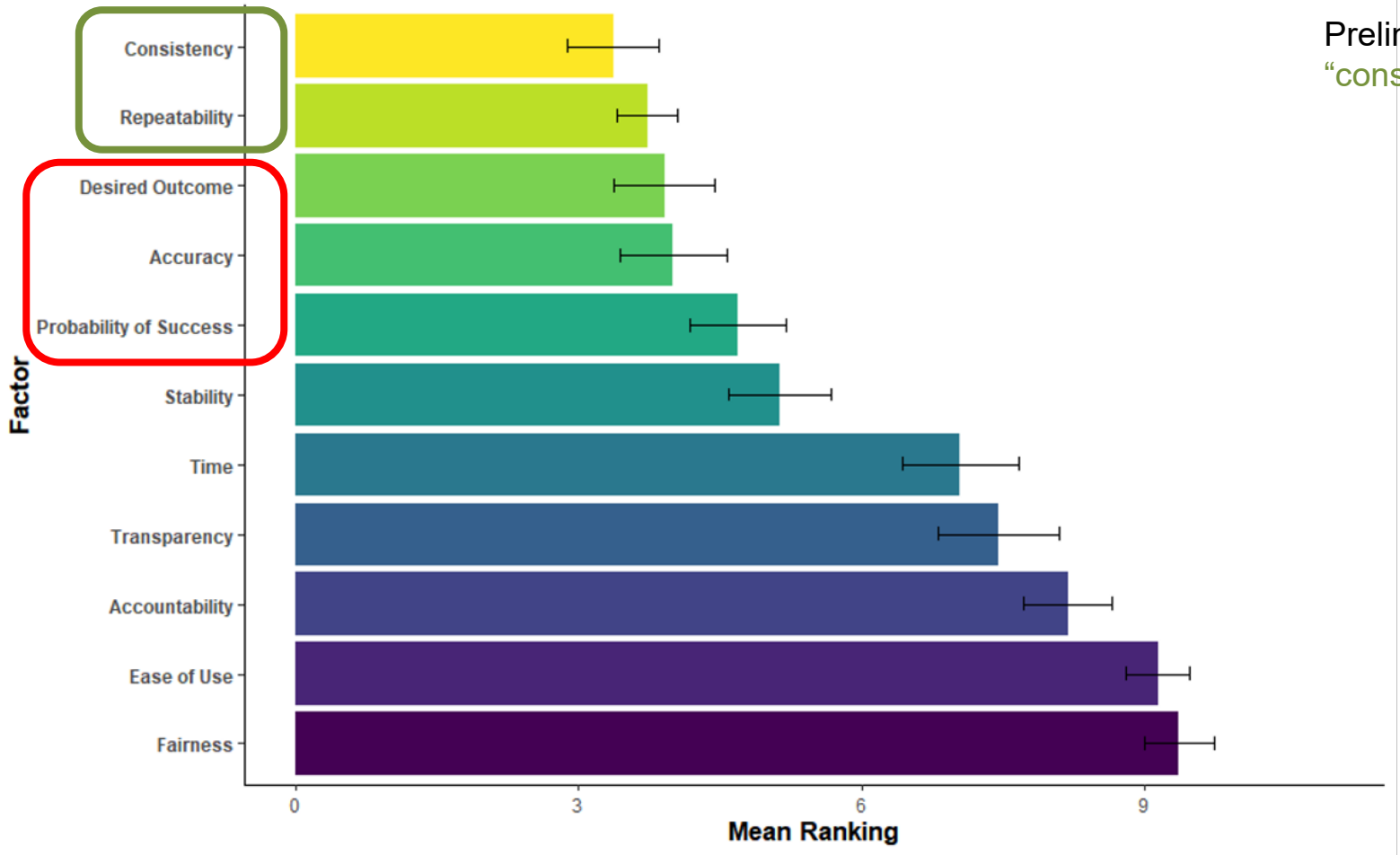
Reliability: Methodology



Reliability: Results

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Preliminary Survey Results



Preliminary results indicate more importance given to "consistency" than to "performance"

Higher Importance

Lower Importance

Reliability: Takeaways

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- The term ‘reliability’ is used both technically and colloquially
  - Different disciplines use the term differently
    - However, there are discrepancies within discipline as well
- Preliminary results from the survey indicate that there also may be substantial disagreement within how practitioners use the term
- There needs to be consistent use of the term to avoid miscommunications
- There also needs to be a clarification of how key concepts (like reliability) apply to increasingly advanced AI systems moving forward
  - Need to understand how we assess these systems

Objective B: Explaining AI

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Form techniques to explain AI decisions and the decision-making process to users

Explaining AI: The Problem

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Explainable AI can provide more insights to the user regarding the AI's decision-making process

- Questions:
  - What does explainable AI, transparent AI, and like terms mean?
  - How does incorporation of explainable AI approaches influence users trust in automation, improve human-AI teaming?
  - How to integrate transparent AI/ML approaches for a trustworthy human-AI decision support system?
  - Can explainable AI decision support systems increase user attention, user interaction, and promote situation awareness ?

- One suggested way to implement this transparency is to present users with information about system's ability to perform tasks in the immediate future
  - However, humans are inherently limited in ability to understand probabilities
  - We need to understand how users of a system may interpret probability of failure forecasted into the future

Explaining AI: Methodology

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Problem broken into three parts:
  - Literature review of key terms
  - Create testing environment
  - User study
- Literature review of academic, industry, and government sources.
  - Focus on:
    - Explainable AI
    - Interpretable ML
    - Transparent ML
    - Comprehensible ML.

Explaining AI: Results - Definitions

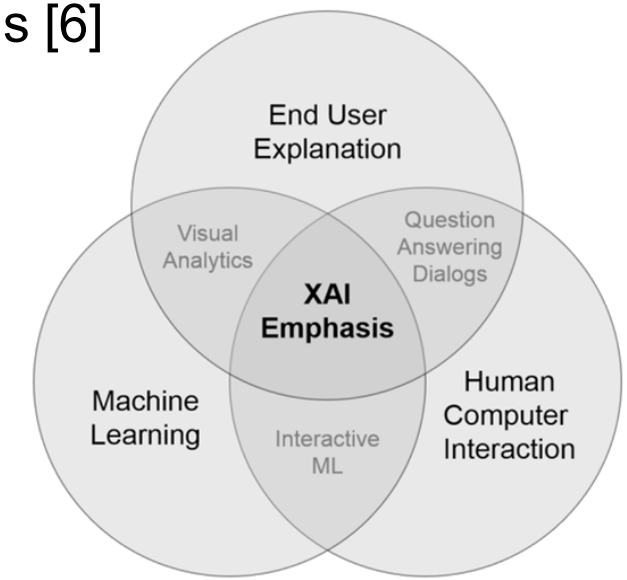
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- What is Explainable AI?
  - Some federal agencies have similar ideas:
    - NASA (Explainable and Transparent) - Solutions must clearly state if, when, and how an AI system is involved, and AI logic and decisions must be explainable. AI solutions must protect intellectual property and include risk management in their construction and use. AI systems must be documented. [1]
    - DoD (Explainable AI) - Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners. [2]
    - DHS (Explainable AI) - In all cases in which AI is used, operational assessments of potential risk and harm, the magnitude of those risks and harms, the technical state of the art, and the potential benefits of the AI system must be substantiated to facilitate both explainability and transparency. [3]
    - DoE (Explainability) - the ability to understand the mechanics of machine or deep learning algorithms. [4]

Explaining AI: Results - Definitions

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- What is Explainable AI?
  - Core technologies involved
    - Machine Learning
      - State-of-the-art robotics leverage modern machine learning technology [5]
    - Human Computer Interaction
      - Introducing Human-AI teams involves Human-AI interactions
    - End User Explanation
      - Enable end users to understand, appropriately trust, and effectively manage the emerging generation of AI systems [6]



Explaining AI: Results - Definitions

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Non-exhaustive list of commonly used explainable models, categorized by their type of explanation under the topics provided by DARPA [2]
  - Deep Explanations
    - Method: Model complex behavior by constructing simpler (more interpretable) models
    - Deconvolutional Networks [12], Mimic Models [13]
  - Interpretable Models
    - Method: Follow decision path by examining the model directly
    - Bayesian Rule Lists [14], Bayesian Program Learning [15], Decision Trees [16]
  - Model Induction
    - Method: Evaluate what the model learns during training
    - LIME [17], SHAP [18], STACI [19]

Explaining AI: Results - Definitions

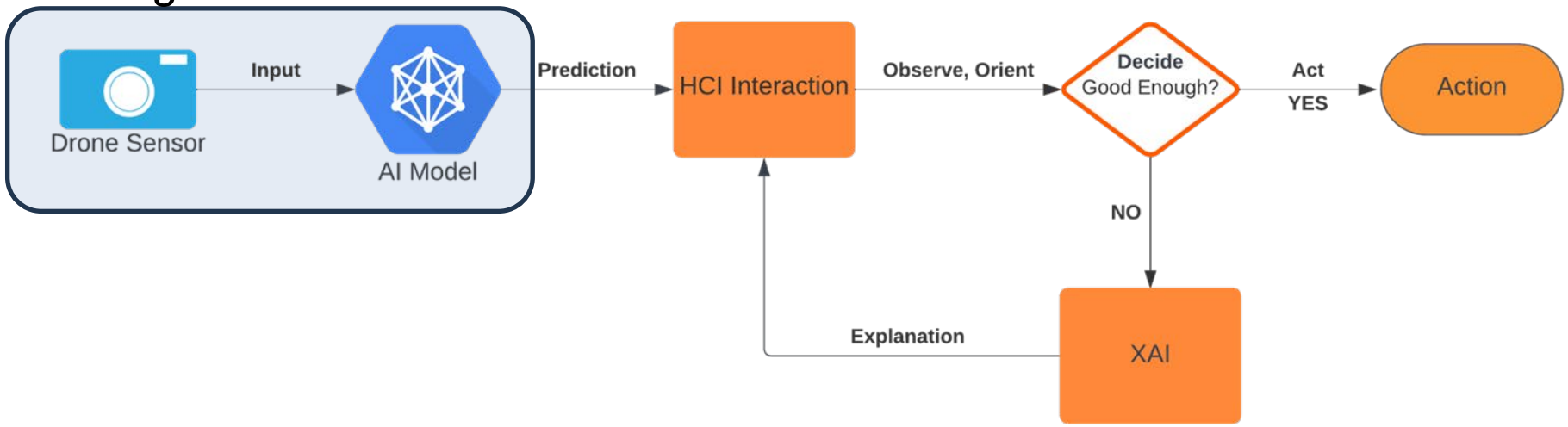
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Principles of Explanatory Debugging to Personalize Interactive Machine Learning [20], Explanations should:
  - Be Iterative
    - Explanations should be presented in “bites” of information. A user can consume “bites” over multiple iterations to increase mental-model fidelity.
  - Be Sound
    - “Explanations should not be simplified by explaining the model as if it were less complex than it actually is.”
  - Be Complete
    - “a complete explanation does not omit important information about the model.”
  - Not Overwhelm
    - “Balanced against the soundness and completeness principles is the need to remain comprehensible and to engage user attention.”

Explaining AI: Methodology - Environment

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Create testing environment



Emulated via Virtual Camera and Webstreams.

Experimental Architectures include both detection and segmentation networks: Yolov5{s,n,x), Yolov8{s,x,-seg}, SAM.

HCI Interaction

Interactions observed using a game developed to measure HCI.

XAI

Explain results with various methods such as: SHAP, LIME, Contrastive Explanations, Counterfactual Explanations, Class Activation Maps, Anchors, and more.

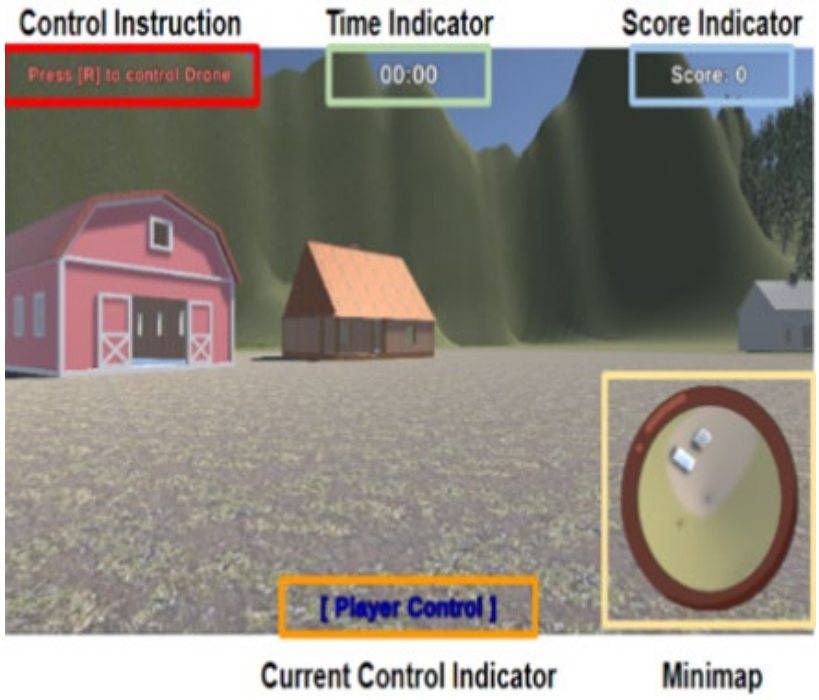
Action

Player action recorded.

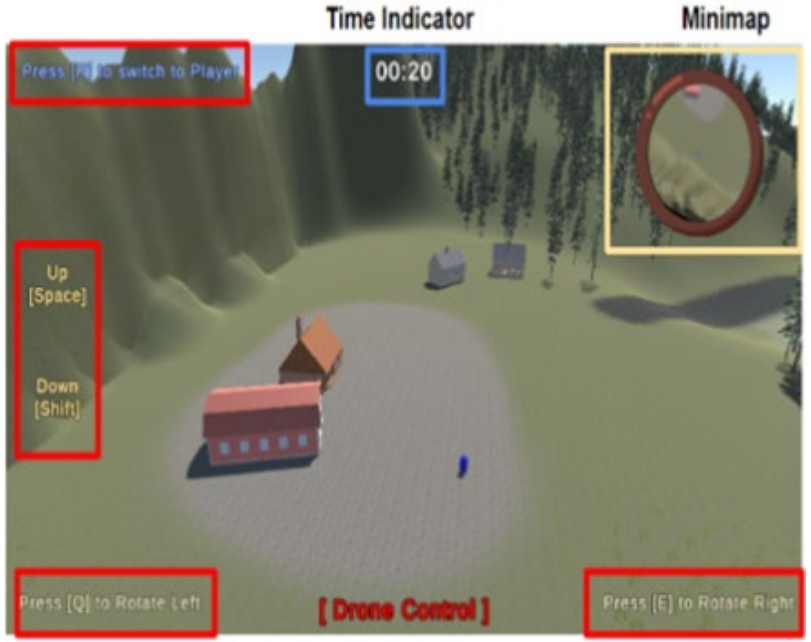
Explaining AI: Results - Environment

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

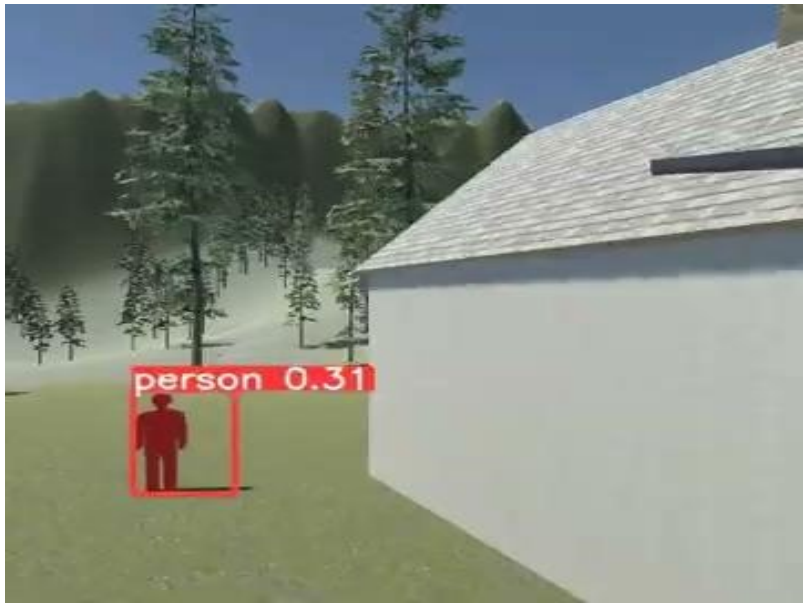
- Combat Search and Rescue (CSAR) Scenario



CSAR Player Interface



CSAR Drone Interface



Drone POV – Hostage Spotting

Explaining AI: Results - Environment

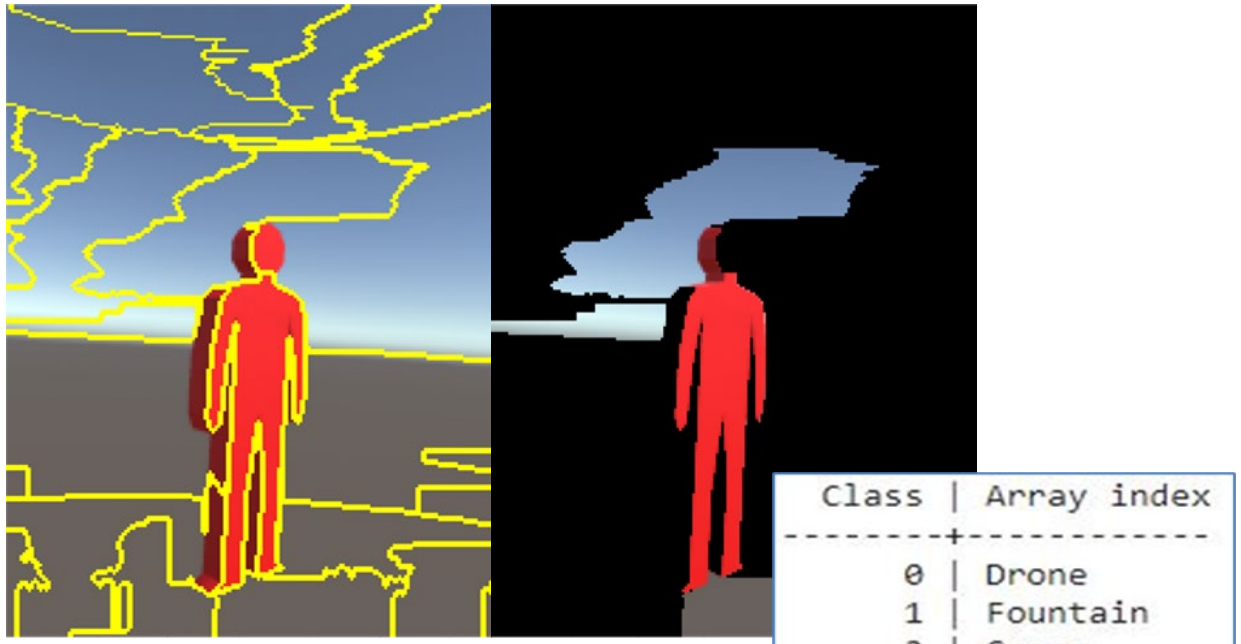
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Active research on Unity environmental design for simulation of autonomous human-AI interaction scenarios: drone-AI based target detection, explainable AI integration

Explaining AI: Results - Environment

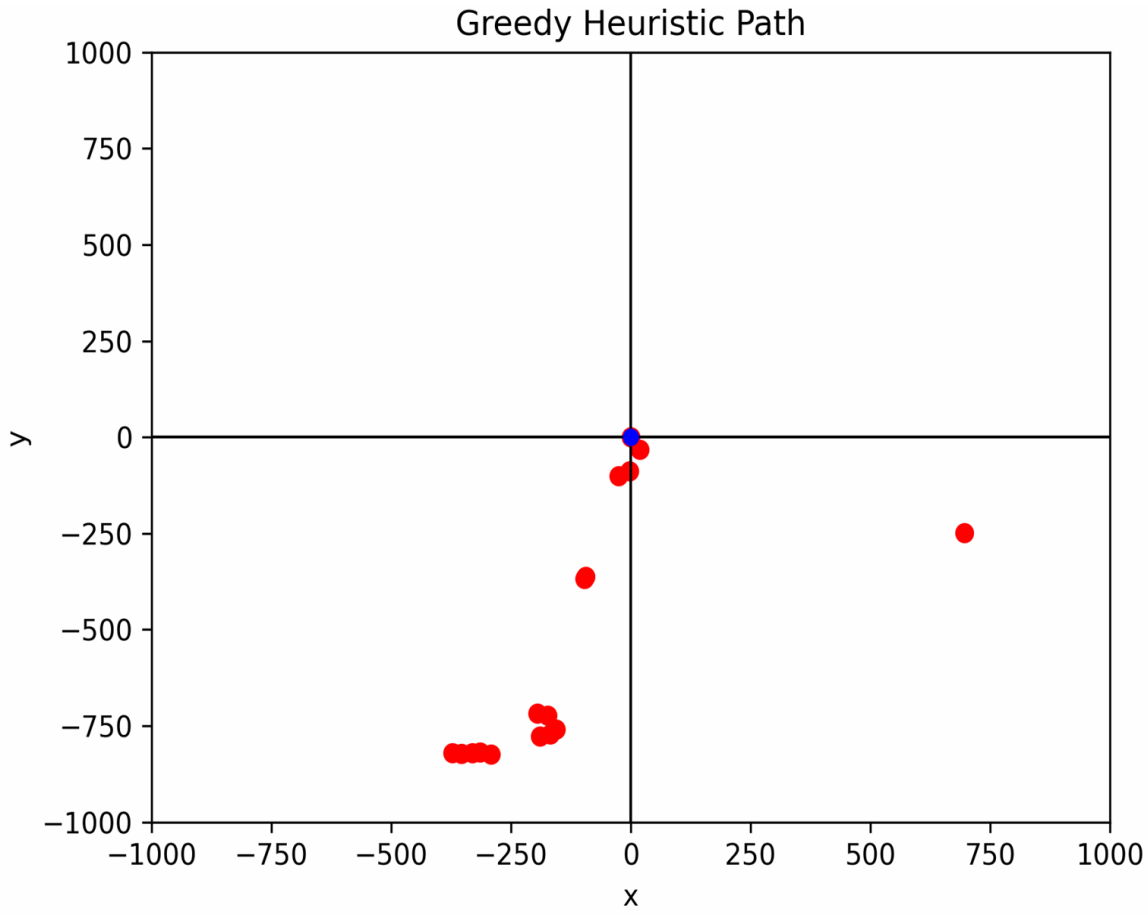
THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Active research on autonomous human-AI interaction scenarios: explainable AI integration, drone-AI path navigation, human-AI teaming, situational awareness



Class	Array index
0	Drone
1	Fountain
2	Grass
3	Hostage
4	House
5	Trailer
6	Tree

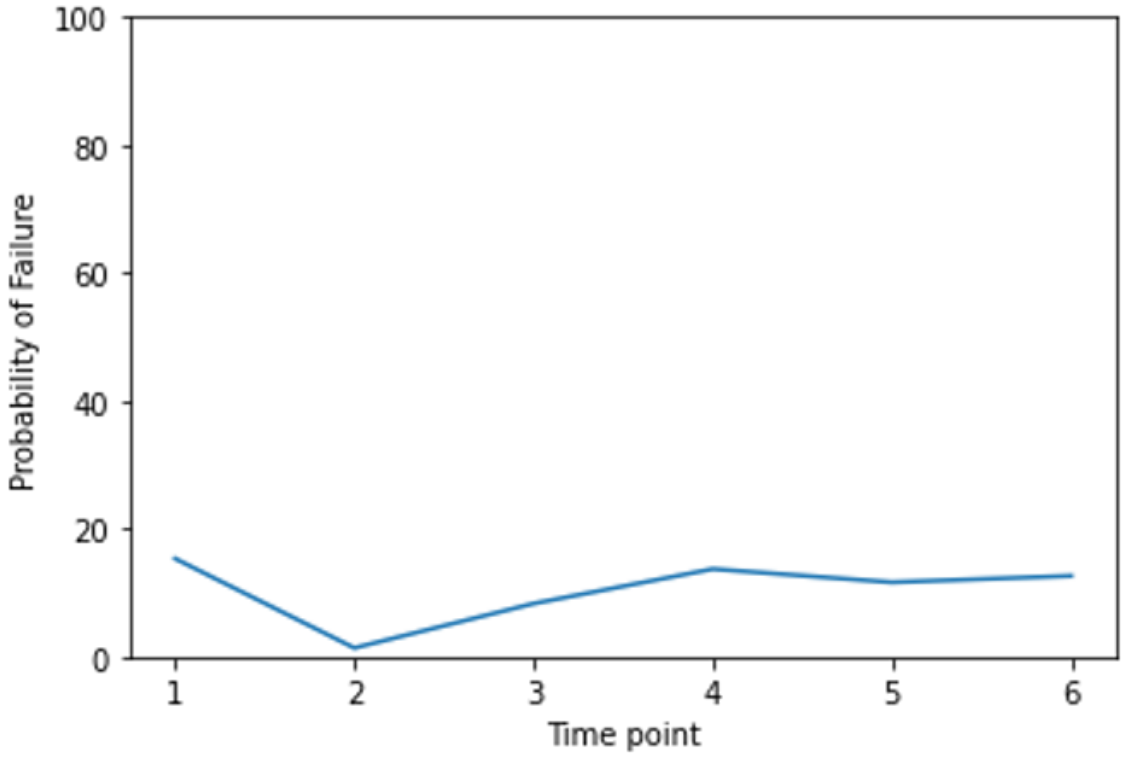
array([3, 5, 6, 4, 2])



Explaining AI: Methodology – User Study

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Experiment run on qualtrics
  - Individuals shown forecasted probabilities
  - Then asked to rate how reliable such a system is (Not at all - Completely)
  - Asked to provide a numerical prediction of how many times the system would succeed if let run all the way through 100 times.
- Final aspect is to ask participants when they would intervene if they were to intervene



Explaining AI: Results – User Study

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Data collection still ongoing

- However, several competing hypotheses of how users may evaluate this reliability information
  - Accurate cumulative probability calculation
  - Area Under Curve (AUC)
  - Average reliability over time
  - Max/min reliability over time

- We are able to differentiate, based on data, what strategies people use to understand this information
  - We will be able to see how good they are, but also
  - Where biases in estimates of success rates exist and what features of the information presented can impact behaviors

Explaining AI: Takeaways

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Explainable AI
  - Goal is to understand and answer the user's questions
- Interpretable ML (Interpretability)
  - Goal is to understand the object (or model) and its operation in general
- Transparent ML
  - Goal is to understand its training procedure, provenance of its parameters and the process governing its predictions
- Comprehensible ML
  - Goal is to provide a framework for considering comprehensibility in modeling to aid in identifying challenges and opportunities
- Although Explainable AI, interpretable ML and Transparent AI are used synonymously, they are different from the implementation focus such as task-oriented (Explainable AI, Interpretable ML) versus process-oriented (Transparent AI).

Explaining AI Takeaways

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- If we want explainable AI, we need to know not only what information needs to be explained, but how that explanation needs to be given
- This research will provide clarity about the nature of users' understanding of probability as well as how it maps onto use/disuse behaviors

Objective C: Emergent Behaviors

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Identify sources of emergent behaviors in AI incorporated systems

Emergent Behaviors: The Problem

- Systems incorporating AI are recognized as having “emergent behaviors”.
- Emergent Behavior - System behavior which is not apparent from separate analysis of subsystems.
  - Upper-bound: Behavior which wasn’t considered a possibility until observed after deployment.
  - Lower-bound: Behavior which is known, but difficult to predict exactly when it will occur from separate subsystem analyses.

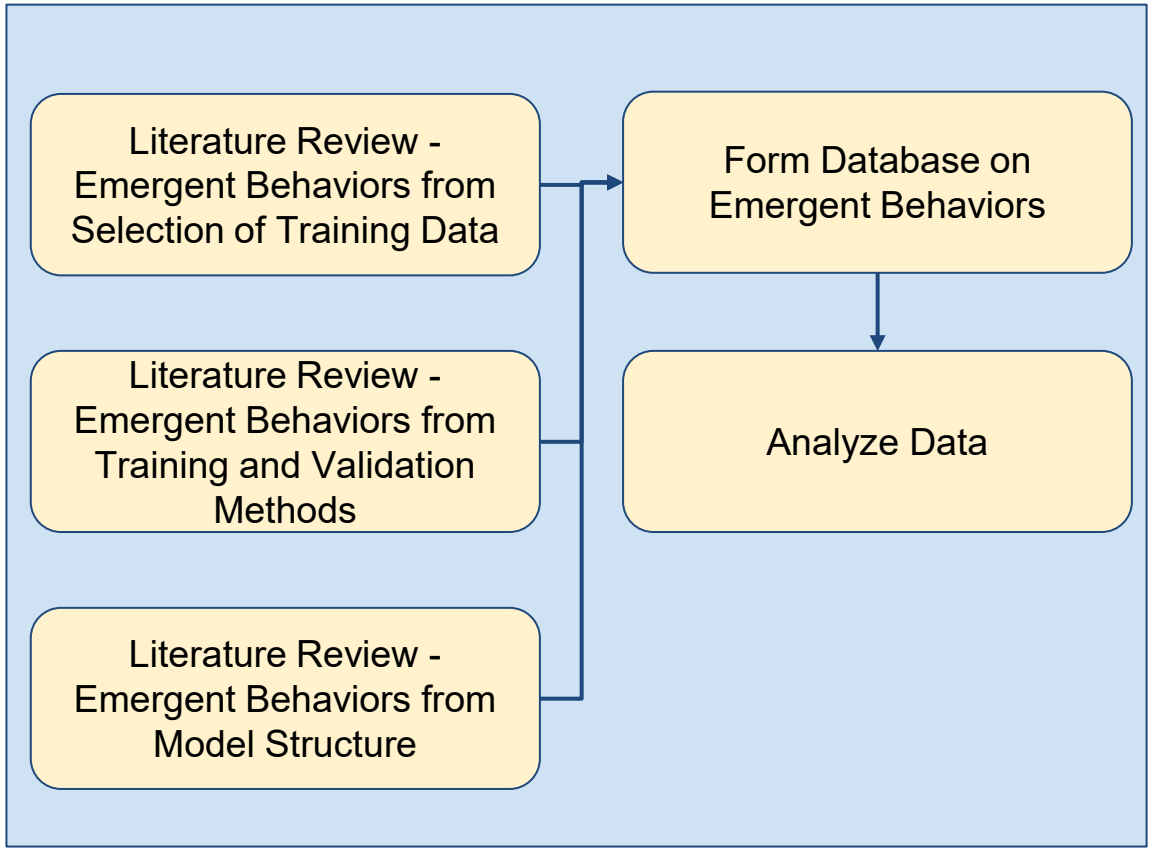
Open AI's hide-and-seek experiment
https://openai.com/research/emergent-tool-use

Amazon’s use of AI in technical jobs hiring
https://live.staticflickr.com/8502/8325104250_9f46039d3f_b.jpg

Emergent Behaviors: Methodology

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

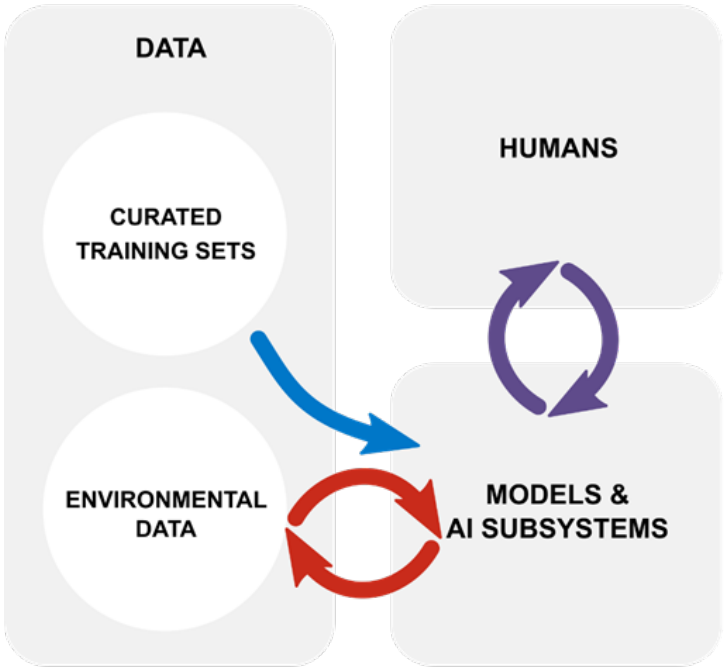
- Research Question: How do you make engineering decisions when the behavior of systems incorporating AI has more uncertainty than similar systems without AI?

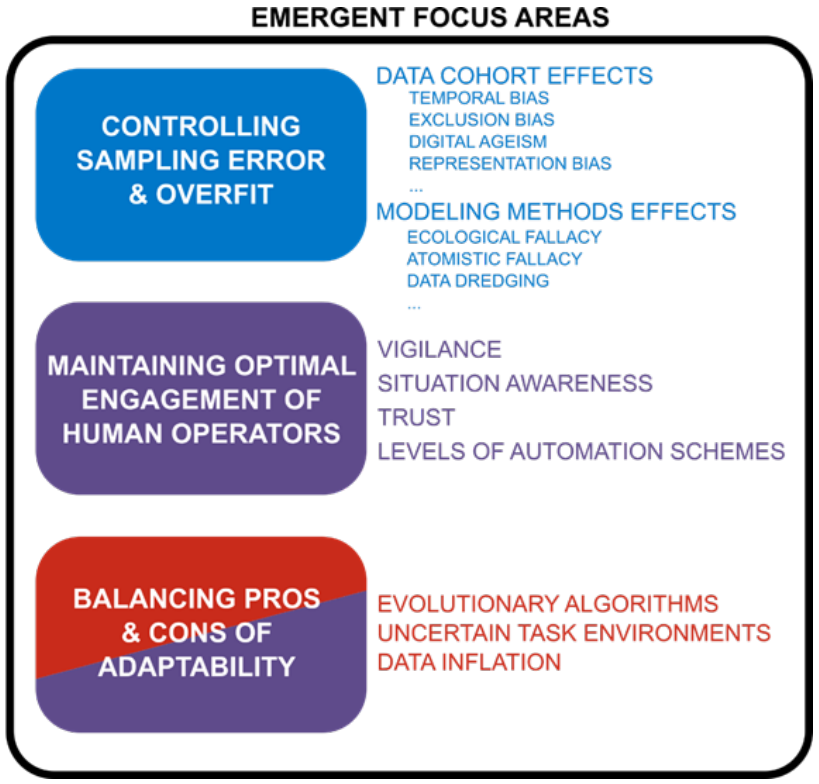


Emergent Behaviors: Early Findings

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- ❑ Work on biases and emergent behavior often results in lists of previously observed phenomena.
- ❑ While useful for categorizing past events, these lists offer limited utility for forecasting and design.
- ❑ Process-oriented approaches can help better understand the complex interactions

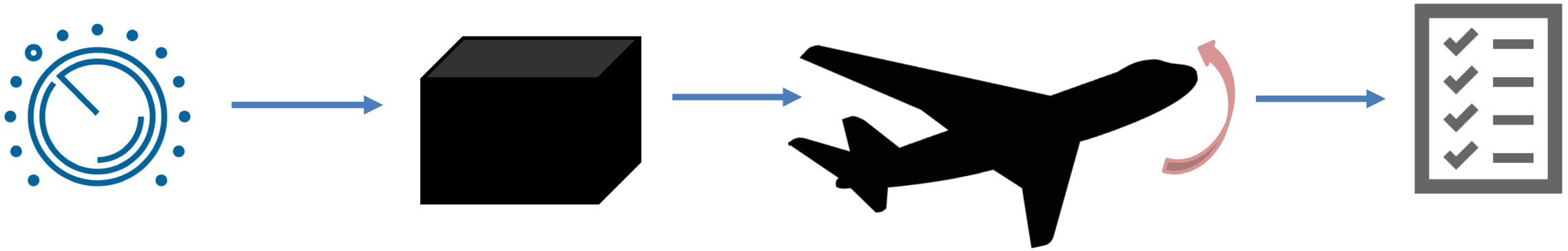




Emergent Behaviors: Takeaways

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- ❑ Engineers need to understand how their decisions impact the assessment of the system.
- ❑ It is not enough to just understand that emergent behaviors exist, but we need to understand the origins of those emergent behaviors so we can identify the dials that impact the behaviors.
- ❑ A major challenge is establishing the relationship between the dials and the behaviors, and eventually to the assessment of the system.



THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Next Steps and Conclusions

Next Steps

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

Definition of Reliability

Methodology:
Aggregate literature review and survey findings to form a definition of reliability that spans disciplines and is actionable and measurable.

ML Model Interpretation

Methodology:
Identify existing or synthesize new interdisciplinary technique to simplify ML model interpretations, validated using user study experimentation.

Preliminary Value Model Formation

Methodology:
Form a value model for assessment of AI systems, with a focus on attributes related to emergent behaviors, reliability, and transparency.

Expected Outcomes

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- This on-going research is expected to support the formation of a **comprehensive assessment framework for AI incorporated systems**.

- Specifically, this on-going research will produce:
  - A reliability definition
  - Measures for reliability
  - Techniques to explain AI processes
  - Identification of origins of emergent behavior in AI incorporated systems.

Broader Applications/Significance

THE UNIVERSITY OF ALABAMA IN HUNTSVILLE

- Broader Applications:
  - Assessment of AI incorporated systems is necessary for systems engineering.
  - An assessment framework that is **based on evidence** to address key challenges in AI incorporated systems would likely move closer to being **repeatable and justifiable**.
  - Such an assessment framework **removes biases** from stakeholders, resulting in decisions that can be argued less on opinions and more on evidence.

- Significance
  - Performs fundamental research on addressing challenges perceived to exist for AI systems.
  - Helps establish a basis for future research on assessment frameworks.
  - Provides insights on the validity of specific perceptions about AI incorporated systems.

Sponsors

InSyST would like to thank our sponsor:

U.S. Army Combat Capabilities Development Command (DEVCOM) Analysis Center (DAC)

Contract: W911NF2220001

Thank you!

References

[1] E. McLarney, “NASA Framework for the Ethical Use of Artificial Intelligence (AI),” Apr. 2021.
[2] “Broad Agency Announcement Explainable Artificial Intelligence (XAI) DARPA-BAA-16-53,” Defense Advanced Research Projects Agency, Aug. 2016.
[3] “DEPARTMENT OF HOMELAND SECURITY ARTIFICIAL INTELLIGENCE STRATEGY,” Dec. 2020.
[4] R. Stevens, V. Taylor, J. Nichols, A. Maccabe, K. Yelick, and D. Brown, “AI for Science Report on the Department of Energy (DOE) Town Halls on Artificial Intelligence (AI) for Science,”
[5] S. Caldera, A. Rassau, and D. Chai, “Review of Deep Learning Methods in Robotic Grasp Detection,” Multimodal Technologies and Interaction, vol. 2, p. 57, Sept. 2018. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.
[6] D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, “DARPA’s explainable AI (XAI) program: A retrospective,” Applied AI Letters, vol. 2, no. 4, p. e61, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.61.
[7] T. Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences,” Aug. 2018. arXiv:1706.07269 [cs].
[8] M. Bellucci, N. Delestre, N. Malandain, and C. Zanni-Merk, “Towards a terminology for a fully contextualized XAI,” Procedia Computer Science, vol. 192, pp. 241–250, Jan. 2021.
[9] V. Beaudouin, I. Bloch, D. Bounie, S. Clémençon, F. d’Alché Buc, J. Eagan, W. Maxwell, P. Mozharovskyi, and J. Parekh, “Flexible and Context-Specific AI Explainability: A Multidisciplinary Approach,” SSRN Electronic Journal, 2020.
[10] K. Sokol and P. Flach, “Explainability fact sheets: a framework for systematic assessment of explainable approaches,” in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, (Barcelona Spain), pp. 56–67, ACM, Jan. 2020.
[11] M. Gleicher, “A Framework for Considering Comprehensibility in Modeling,” Big Data, vol. 4, pp. 75–88, June 2016.
[12] M. D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks,” Nov. 2013. arXiv:1311.2901 [cs].
[13] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, pp. 436–444, May 2015. Number: 7553 Publisher: Nature Publishing Group.
[14] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, “Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model,” The Annals of Applied Statistics, vol. 9, Sept. 2015. arXiv:1511.01644 [cs, stat].
[15] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, “Human-level concept learning through probabilistic program induction,” Science, vol. 350, pp. 1332–1338, Dec. 2015. Publisher: American Association for the Advancement of Science.

References

[16] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, “Explainable Artificial Intelligence (XAI) to Enhance Trust Management in Intrusion Detection Systems Using Decision Tree Model,” Complexity, vol. 2021, p. e6634811, Jan. 2021. Publisher: Hindawi.
[17] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” Feb. 2016.
[18] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017.
[19] N. Radulovic, A. Bifet, and F. Suchanek, “Confident Interpretations of Black Box Classifiers,” in 2021 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, July 2021. ISSN: 2161-4407.
[20] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, “Principles of Explanatory Debugging to Personalize Interactive Machine Learning,” in Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI ’15, (New York, NY, USA), pp. 126–137, Association for Computing Machinery, Mar. 2015.
[21] B. Vasu, B. Hu, B. Dong, R. Collins, and A. Hoogs, “Explainable, interactive content-based image retrieval,” Applied AI Letters, vol. 2, no. 4, p. e41, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.41.
[22] J. Kim, A. Rohrbach, Z. Akata, S. Moon, T. Misu, Y.-T. Chen, T. Darrell, and J. Canny, “Toward explainable and advisable model for self-driving cars,” Applied AI Letters, vol. 2, no. 4, p. e56, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.56.
[23] L. A. Hendricks, A. Rohrbach, B. Schiele, T. Darrell, and Z. Akata, “Generating visual explanations with natural language,” Applied AI Letters, vol. 2, no. 4, p. e55, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.55.
[24] C.-K. Yeh and P. Ravikumar, “Objective criteria for explanations of machine learning models,” Applied AI Letters, vol. 2, no. 4, p. e57, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.57.
[25] L. Fuxin, Z. Qi, S. Khorram, V. Shitole, P. Tadepalli, M. Kahng, and A. Fern, “From heatmaps to structured explanations of image classifiers,” Applied AI Letters, vol. 2, no. 4, p. e46, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.46.

References

[26] M. H. Danesh, A. Koul, A. Fern, and S. Khorram, “Re-understanding Finite-State Representations of Recurrent Policy Networks,” in Proceedings of the 38th International Conference on Machine Learning, pp. 2388–2397, PMLR, July 2021. ISSN: 2640-3498.
[27] J. Dodge, A. Anderson, R. Khanna, J. Irvine, R. Dikkala, K.-H. Lam, D. Tabatabai, A. Ruangrotsakun, Z. Shureih, M. Kahng, A. Fern, and M. Burnett, “From “no clear winner” to an effective Explainable Artificial Intelligence process: An empirical journey,” Applied AI Letters, vol. 2, no. 4, p. e36, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.36.
[28] T. Mai, R. Khanna, J. Dodge, J. Irvine, K.-H. Lam, Z. Lin, N. Kiddle, E. Newman, S. Raja, C. Matthews, C. Perdriau, M. Burnett, and A. Fern, “Keeping it "organized and logical": after-action review for AI (AAR/AI),” in Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI ’20, (New York, NY, USA), pp. 465–476, Association for Computing Machinery, Mar. 2020.
[29] Z. Lin, K.-H. Lam, and A. Fern, “Contrastive Explanations for Reinforcement Learning via Embedded Self Predictions,” Jan. 2021. arXiv:2010.05180 [cs].
[30] S. C.-H. Yang, T. Folke, and P. Shafto, “Abstraction, validation, and generalization for explainable artificial intelligence,” Applied AI Letters, vol. 2, no. 4, p. e37, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.37.
[31] C. Roy, M. Nourani, D. R. Honeycutt, J. E. Block, T. Rahman, E. D. Ragan, N. Ruozzi, and V. Gogate, “Explainable activity recognition in videos: Lessons learned,” Applied AI Letters, vol. 2, no. 4, p. e59, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.59.
[32] M. Hamidi-Haines, Z. Qi, A. Fern, F. Li, and P. Tadepalli, “User-guided global explanations for deep image recognition: A user study,” Applied AI Letters, vol. 2, no. 4, p. e42, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.42.
[33] A. R. Akula, K. Wang, C. Liu, S. Saba-Sadiya, H. Lu, S. Todorovic, J. Chai, and S.-C. Zhu, “CX-ToM: Counterfactual Explanations with Theory-of-Mind for Enhancing Human Trust in Image Recognition Models,” Dec. 2021. arXiv:2109.01401 [cs].
[34] M. Edmonds, F. Gao, H. Liu, X. Xie, S. Qi, B. Rothrock, Y. Zhu, Y. N. Wu, H. Lu, and S.-C. Zhu, “A tale of two explanations: Enhancing human trust by explaining robot behavior,” Science Robotics, vol. 4, p. eaay4663, Dec. 2019.
[35] H. Liu, Y. Zhu, and S.-C. Zhu, “Patching interpretable And-Or-Graph knowledge representation using augmented reality,” Applied AI Letters, vol. 2, no. 4, p. e43, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ail2.43.