

A stylized illustration of two hands, one on the left and one on the right, rendered in a dotted, halftone style. The hands are positioned as if holding or supporting a network of interconnected nodes and lines. The nodes are represented by small circles in various colors (red, blue, grey) and are connected by thin, light grey lines. The background is white with some faint, larger circular shapes in red and blue. The overall aesthetic is clean, modern, and technical.

RESEARCH REVIEW 2024

**Carnegie
Mellon
University**
Software
Engineering
Institute

AI Robustness

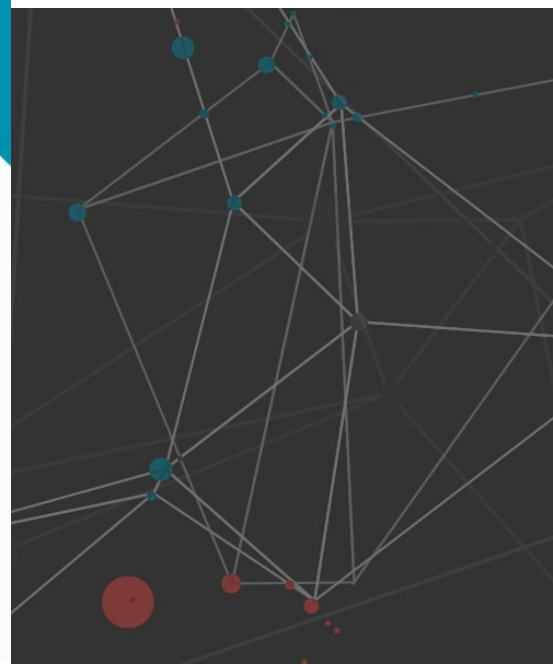
NOVEMBER 13, 2024

Linda Parker Gates
Principal Investigator

Dr. Nicholas Testa
Senior Data Scientist

[DISTRIBUTION STATEMENT A] Approved for public release and unlimited distribution.

©2024 Carnegie Mellon University



Document Markings

Copyright 2024 Carnegie Mellon University.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

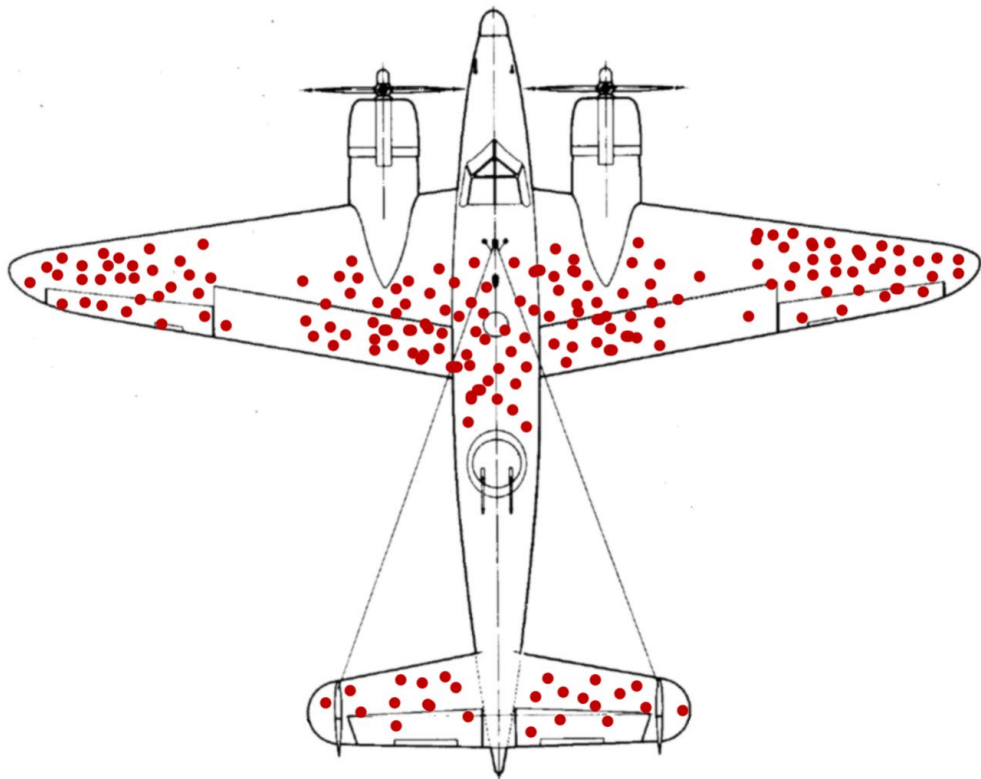
NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN "AS-IS" BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

DM24-1484

Lack of AI Robustness is a DoD Problem



The Department of Defense (DoD) increasingly uses artificial intelligence (AI) and machine learning (ML) classifiers and predictors, but these are subject to a lack of robustness, which leads to a lack of trust.

Testing and evaluation methods are inadequate because they are undermined by

- Data and concept drift
- Evolving edge cases
- Emerging phenomena

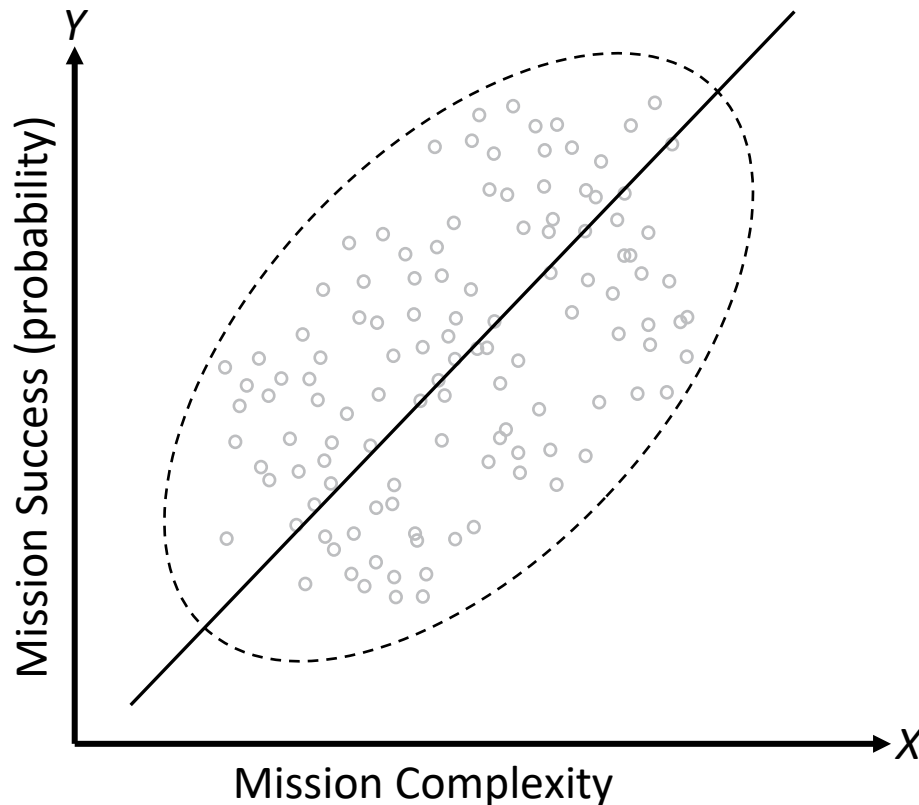
This file is licensed under the [Creative Commons Attribution-Share Alike 4.0 International](https://creativecommons.org/licenses/by-sa/4.0/) license.

What's Wrong with a Little Correlation?

AI and ML tools work by learning associations, but they don't account for causation, which means we can't identify where and when ML predications can't be trusted.

Traditional ML evaluation methods fail to account for underlying causal structures and therefore

- Don't explore alternative explanations for impacts in a scenario
- Fail to account for key drivers
- Attribute causes to the wrong factors
- Don't properly cross-validate their evaluation results

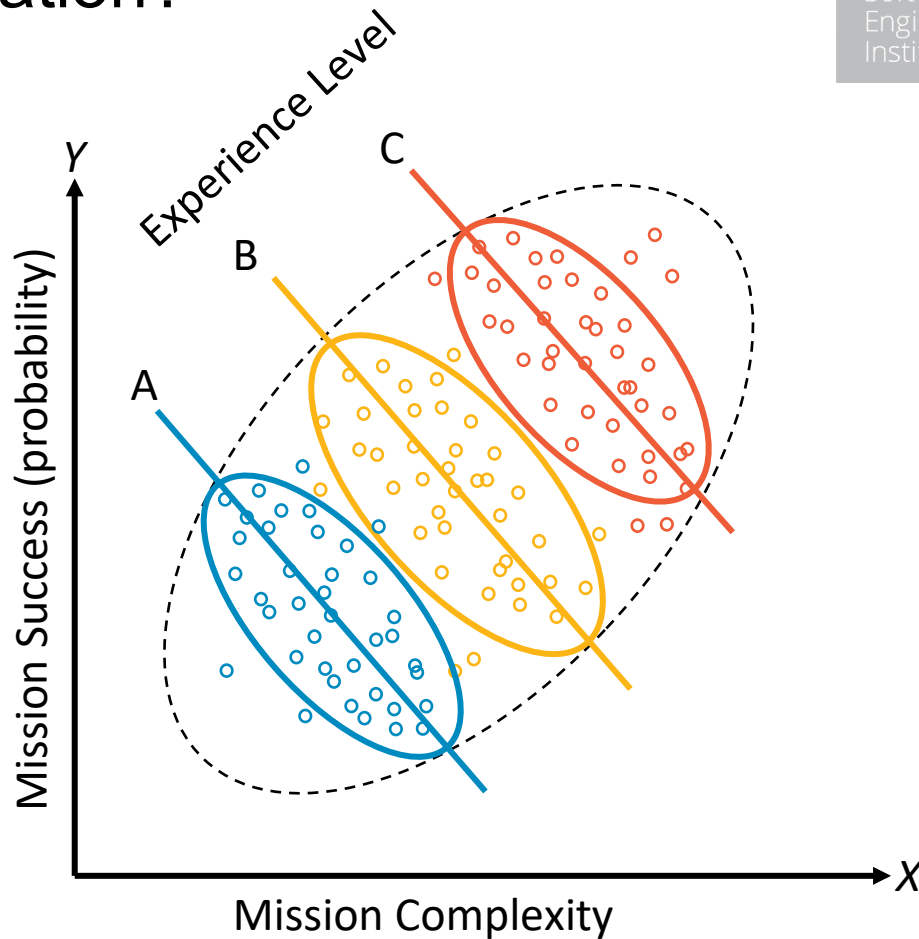


What's Wrong with a Little Correlation?

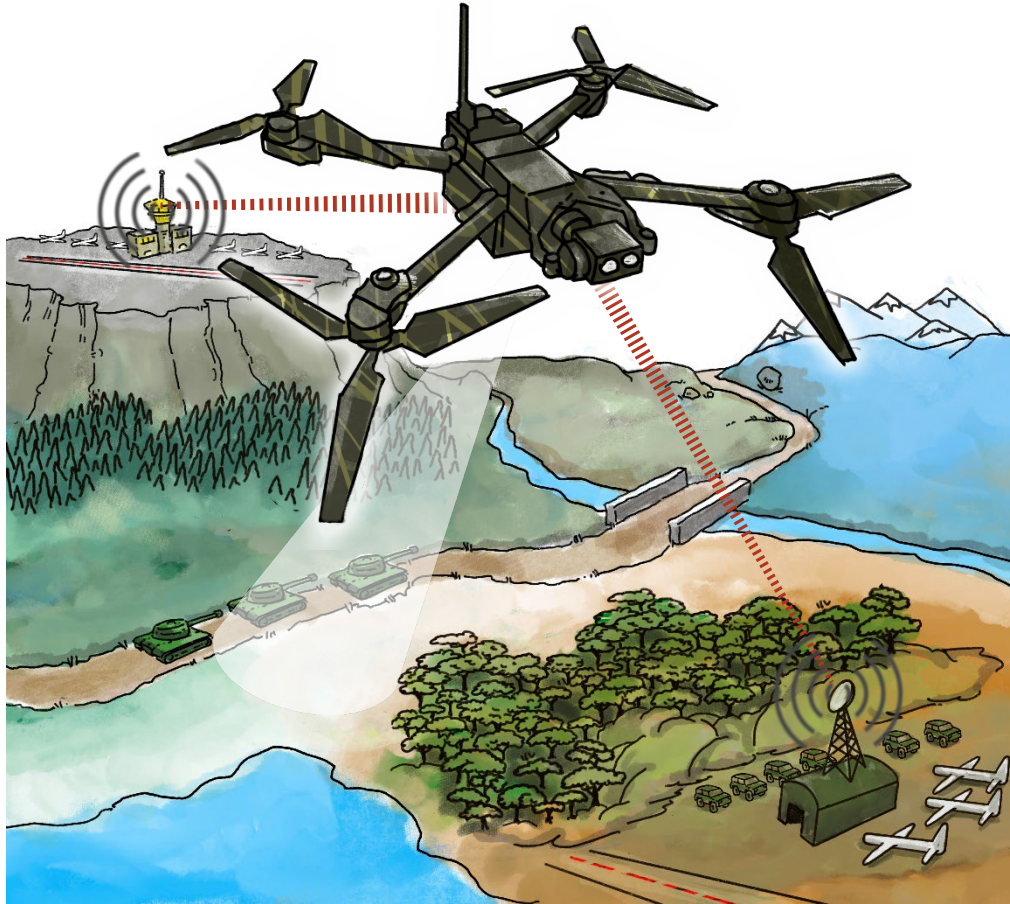
AI and ML tools work by learning associations, but they don't account for causation, which means we can't identify where and when ML predications can't be trusted.

Traditional ML evaluation methods fail to account for underlying causal structures and therefore

- Don't explore alternative explanations for impacts in a scenario
- Fail to account for key drivers
- Attribute causes to the wrong factors
- Don't properly cross-validate their evaluation results

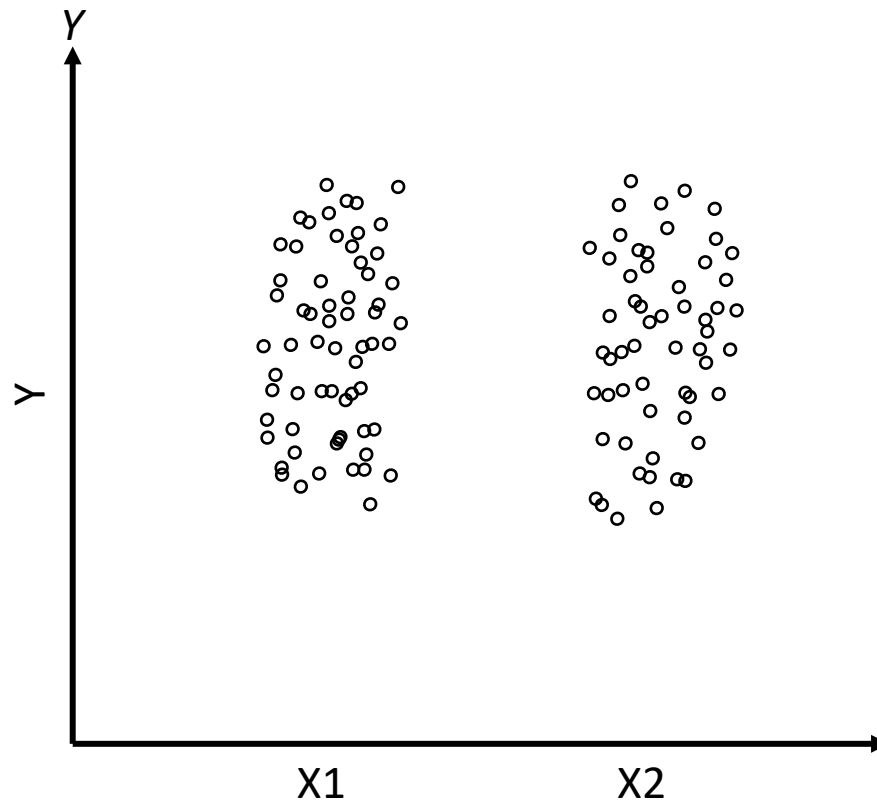
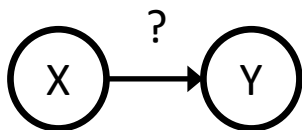


Calling in AIR Support!

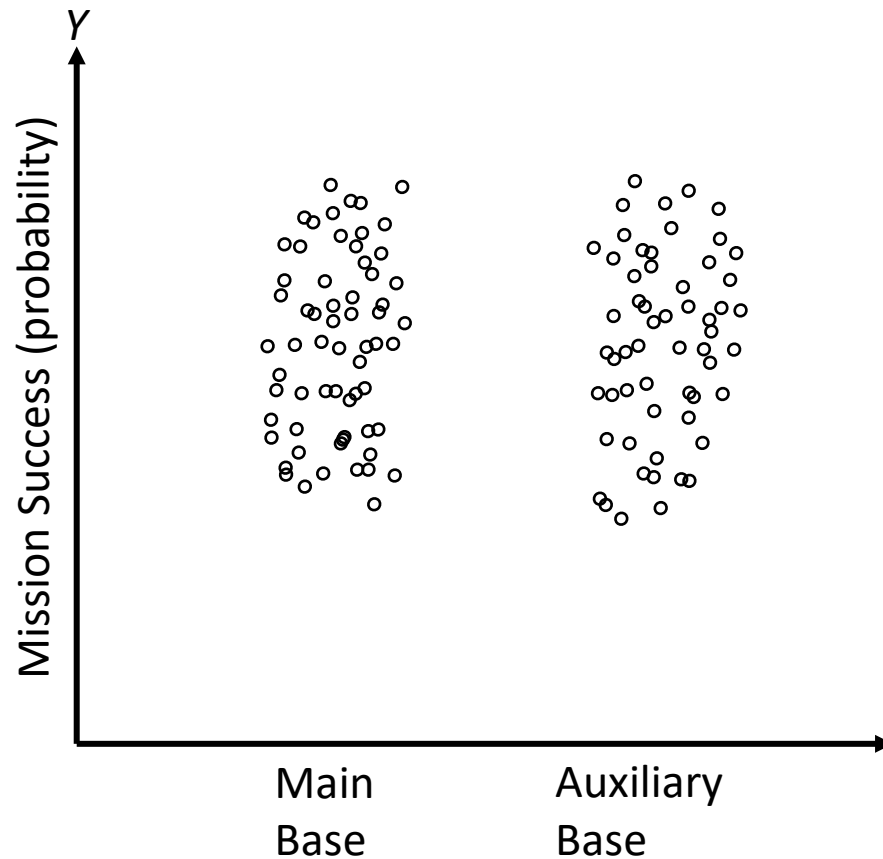
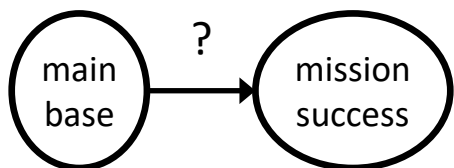


- The Department of Defense (DoD) sends an autonomous vehicle (AV) to acquire images.
- There are two bases, “Home” and “Auxiliary.”
- The DoD wants to predict likelihood of mission success given environmental conditions and choice of base for UAV takeoff.

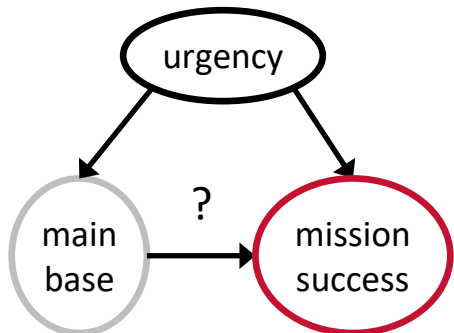
What Is Causal Learning and How Does It Help?



What is Causal Learning and How Does It Help?



What is Causal Learning and How Does It Help?



Causal Learning

Causal
Discovery

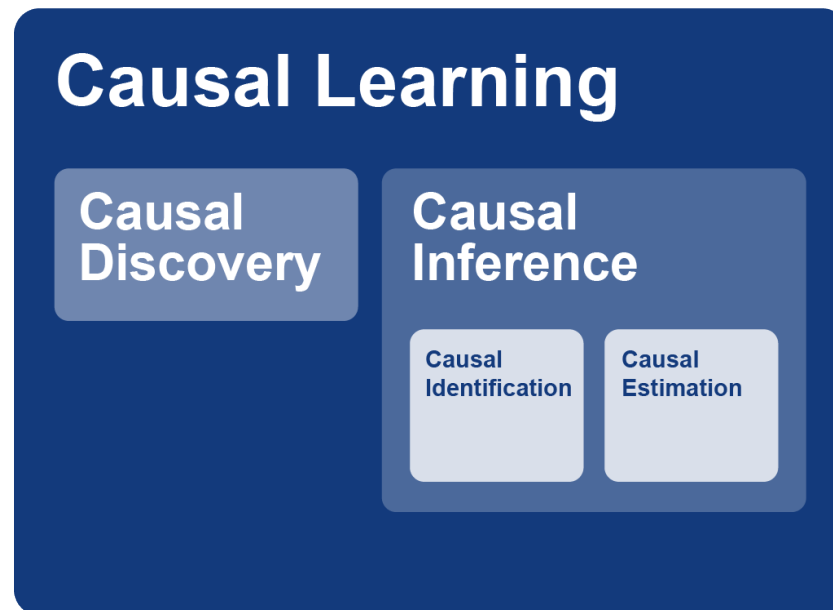
Causal
Inference

Causal
Identification

Causal
Estimation

What is Causal Learning and How Does It Help?

- **Causal Discovery:** identify cause-effect relationships from data
- **Causal Inference:** estimate the effects of an intervention
 - **Causal Identification:** identify potential sources of bias
 - **Causal Estimation:** quantify the impact



Step 1: Causal Discovery

Discovering the Key Players

region_sensitivity

scenario_main_base mission_urgency

A_dist A-B_dist

altitude humidity

heavy_winds temp

speed_avg bird_strike

fuel_consumed ice_accrual

hard_landing ice_sublimation

image_A_captured mission_duration

image_B_captured **images_acquired**

Causal Learning

**Causal
Discovery**

Causal
Inference

Causal
Identification

Causal
Estimation

Step 1: Causal Discovery

Discovering the Key Players

mission_urgency
region_sensitivity

scenario_main_base

A_dist A-B_dist

altitude heavy_winds humidity temp
speed_avg bird_strike

fuel_consumed ice_accrual

hard_landing ice_sublimation

image_A_captured mission_duration

image_B_captured **images_acquired**

Causal Learning

**Causal
Discovery**

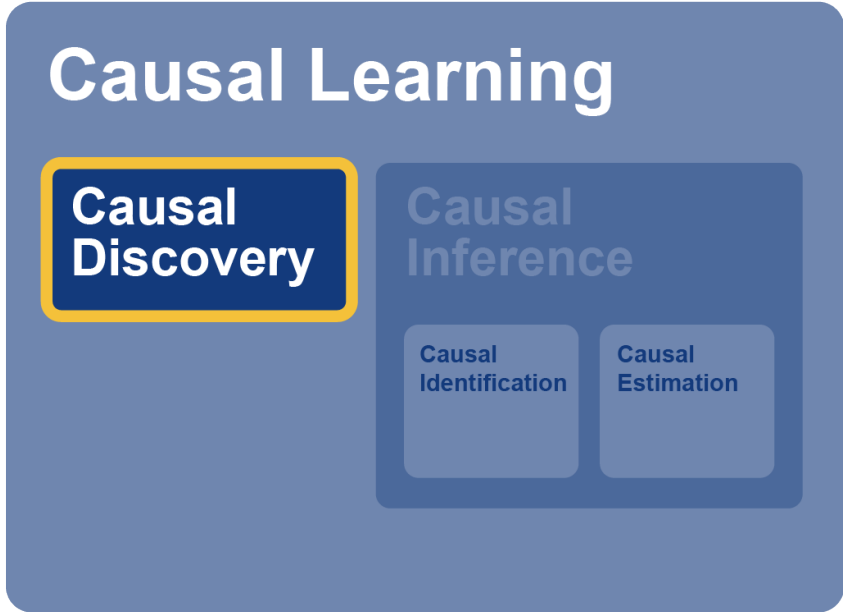
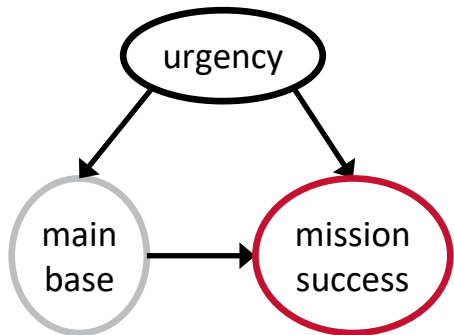
Causal
Inference

Causal
Identification

Causal
Estimation

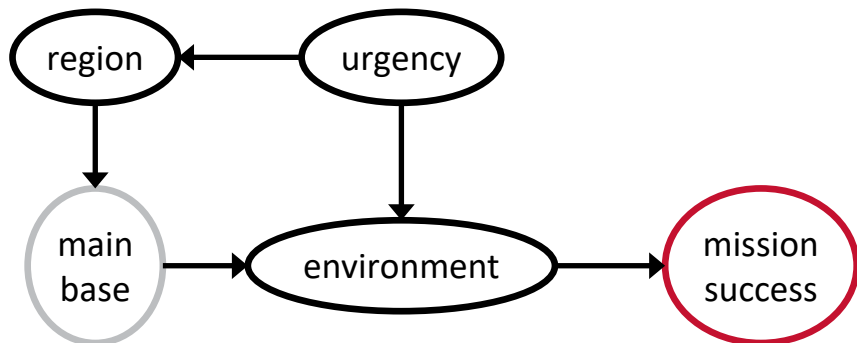
Step 1: Causal Discovery

Discovering the Key Players



Step 1: Causal Discovery

Discovering the Key Players



Causal Learning

Causal Discovery

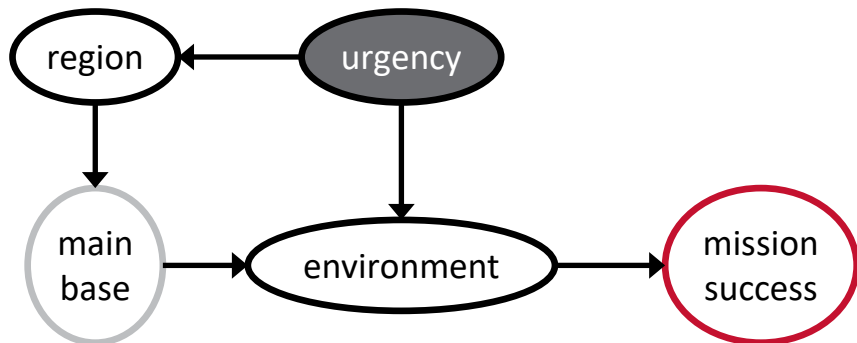
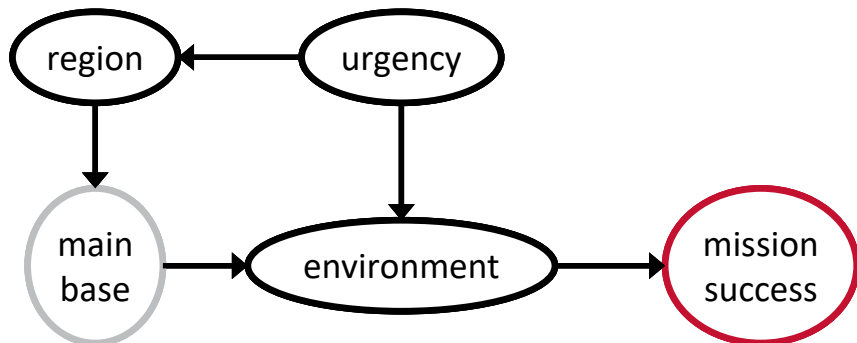
Causal Inference

Causal Identification

Causal Estimation

Step 2: Causal Identification

Identifying Potential Sources of Bias



Causal Learning

Causal
Discovery

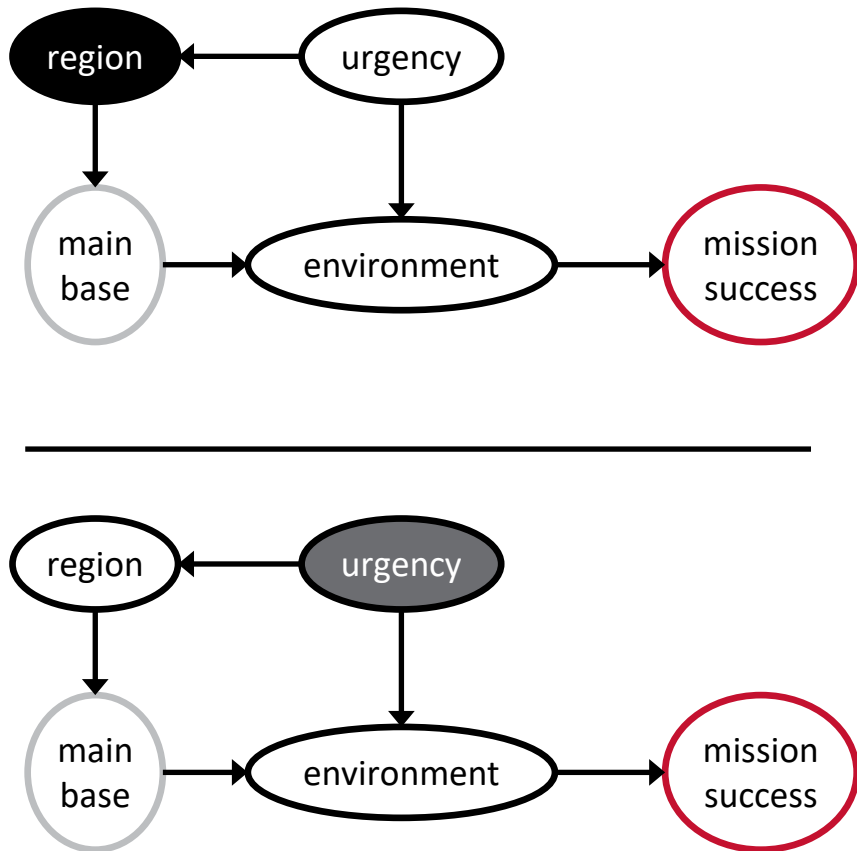
Causal
Inference

Causal
Identification

Causal
Estimation

Step 2: Causal Identification

Identifying Potential Sources of Bias



Causal Learning

Causal
Discovery

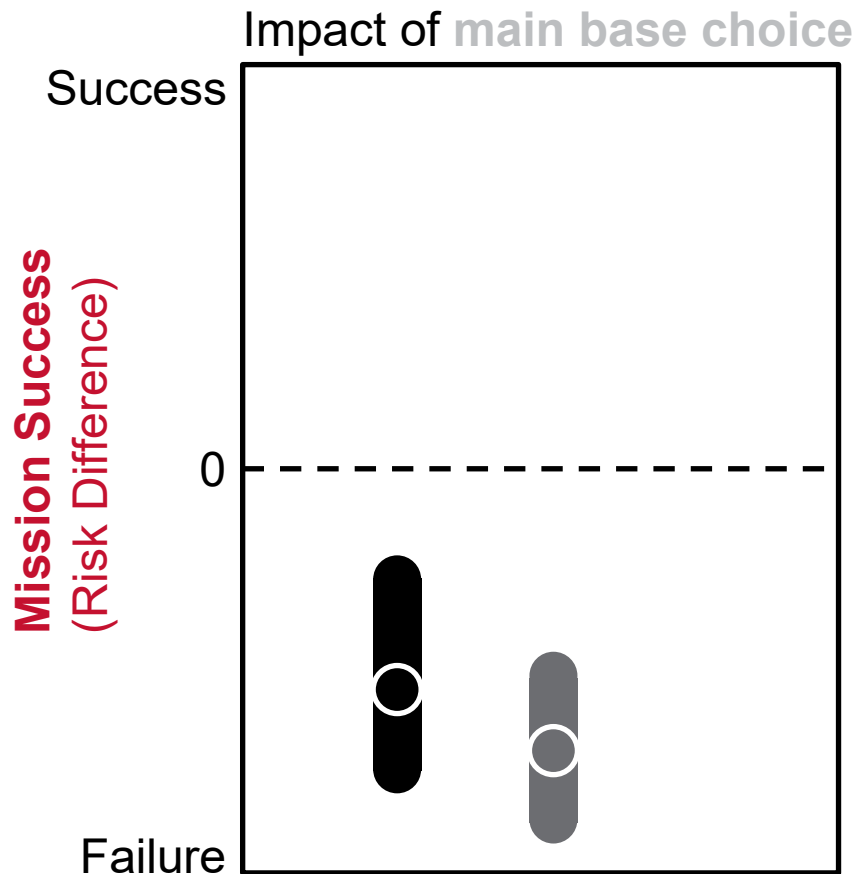
Causal
Inference

Causal
Identification

Causal
Estimation

Step 3: Causal Estimation

Estimating the Impact of Your Decision



Causal Learning

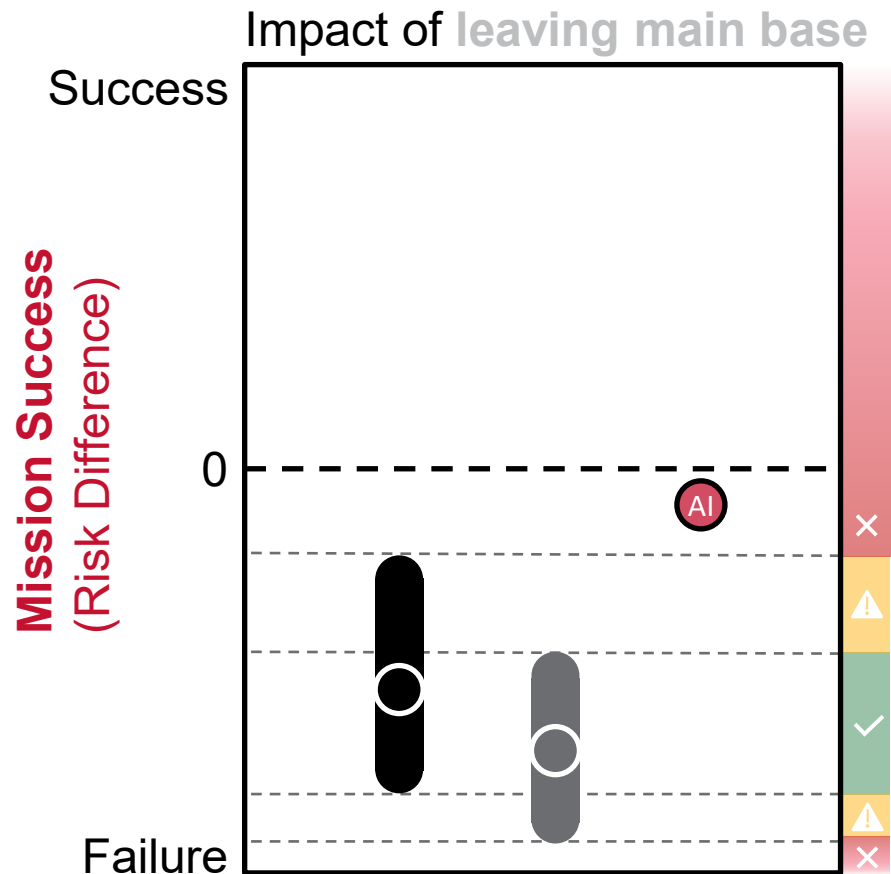
Causal
Discovery

Causal
Inference

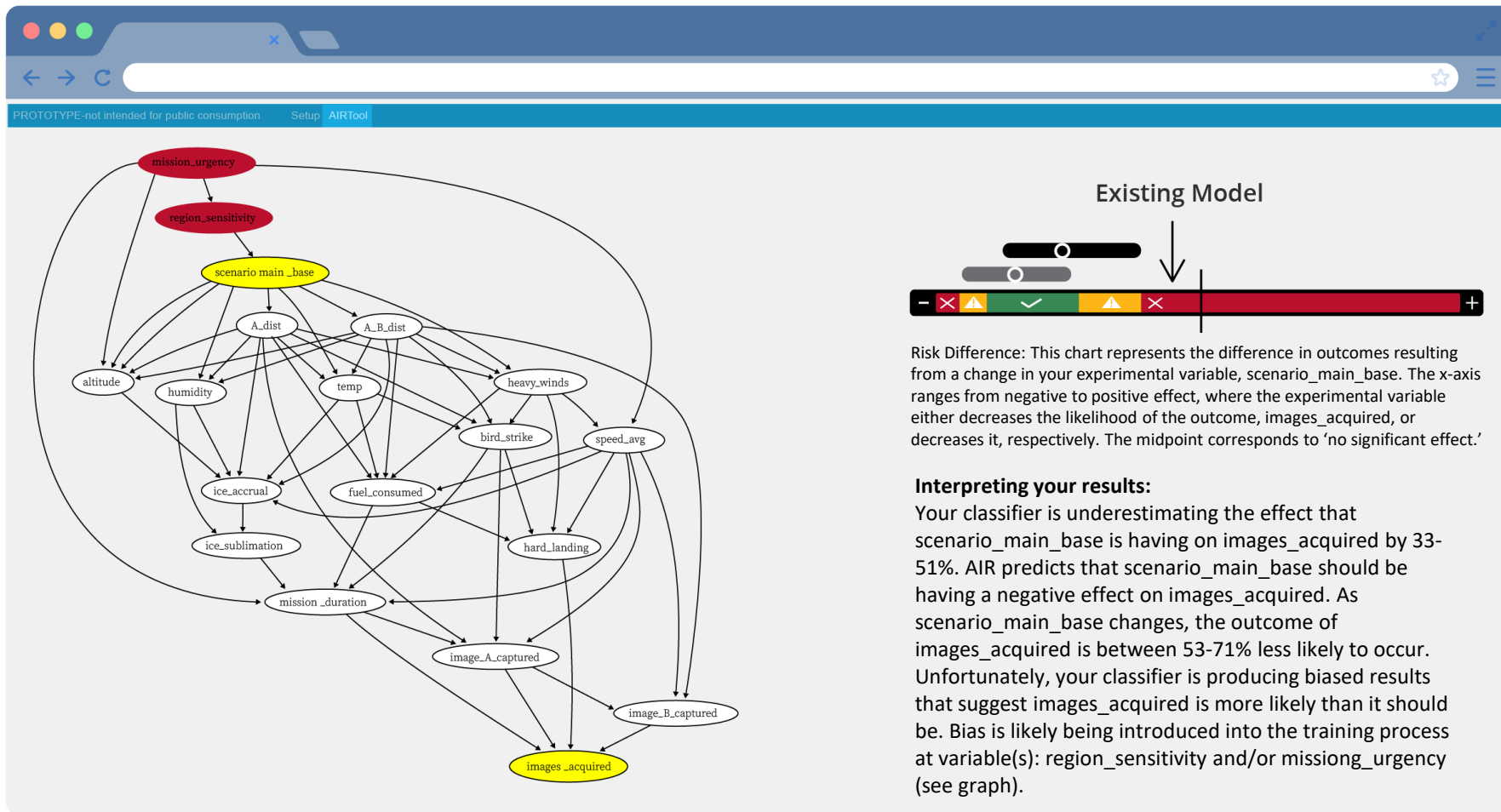
Causal
Identification

Causal
Estimation

Applying Results of AIR



Interpreting Results of AIR



Risk Difference: This chart represents the difference in outcomes resulting from a change in your experimental variable, scenario_main_base. The x-axis ranges from negative to positive effect, where the experimental variable either decreases the likelihood of the outcome, images_acquired, or increases it, respectively. The midpoint corresponds to 'no significant effect.'

Interpreting your results:
Your classifier is underestimating the effect that scenario_main_base is having on images_acquired by 33-51%. AIR predicts that scenario_main_base should be having a negative effect on images_acquired. As scenario_main_base changes, the outcome of images_acquired is between 53-71% less likely to occur. Unfortunately, your classifier is producing biased results that suggest images_acquired is more likely than it should be. Bias is likely being introduced into the training process at variable(s): region_sensitivity and/or mission_urgency (see graph).

Should You Be Using AIR?



- Do you have questions about whether your classifier is performing properly?
- Are you using your classifier's results to make important decisions?

AIR can help you

- across a broad range of contexts.
- across many decision types.
- on multiple scenario and treatment pairs.
- gain insight into classifier performance, which is needed to improve classifier accuracy.

Try AIR for Free!



Use the AIR tool and give us feedback so we can keep improving it. Your feedback influences our research.

AIR is

- free to download and use.
- fully automated.
- containerized and ready for distribution.

AIR requires a dataset that meets current data and tool requirements

What's in it for you?

With AIR, you will

- learn how well your classifiers are performing.
- uncover problems with your classifiers.
- gain confidence in your classifiers.
- build your in-house knowledge of these innovative techniques.

Learn More About AIR



The Team



Linda Parker Gates

Principal Investigator;
technology transition
planning and execution



Mike Konrad

Key researcher on the core
technology (DLAR,
MDLAR, AIR)



Nick Testa

Key researcher for
MDLAR/AIR estimation
and automation



Suz Miller

Key researcher for
the transition of AIR



Crisanne Nolan

Key contributor for the
transition aspects of AIR



Melissa Ludwick

Project manager
and coordinator



David Shepard

Contributor on core
technology and
transition (MDLAR, AIR)



Andrew Mellinger

ML engineer;
Contributor on core
technology and
transition (MDLAR, AIR)



Julie Cohen

Contributor for AIR
transition activities



Joe Ramsey, CMU

Expert on the Tetrad Tool
for causal discovery