RESEARCH REVIEW 2024

Gain Greater Confidence in Your AI Solutions with AIR

Using Causal Discovery, Identification, and Estimation to Improve Your AI Classifiers

Do you have confidence in your AI and ML?

Modern analytic methods, including artificial intelligence (AI) and machine learning (ML) classifiers, depend on correlations; however, such approaches fail to account for causation in the data, which prevents accurate modeling of cause and effect and often leads to prediction bias.

The DoD is increasing its use of AI classifiers and predictors, but users may grow to distrust results because AI classifiers are subject to a lack of robustness (i.e., ability to perform accurately in unusual or changing contexts). Edge cases and drift in data and concept can undermine the informativeness of the correlations relied upon by AI. New test and evaluation methods are therefore needed for ongoing evaluation of AI and ML accuracy and regaining user trust.

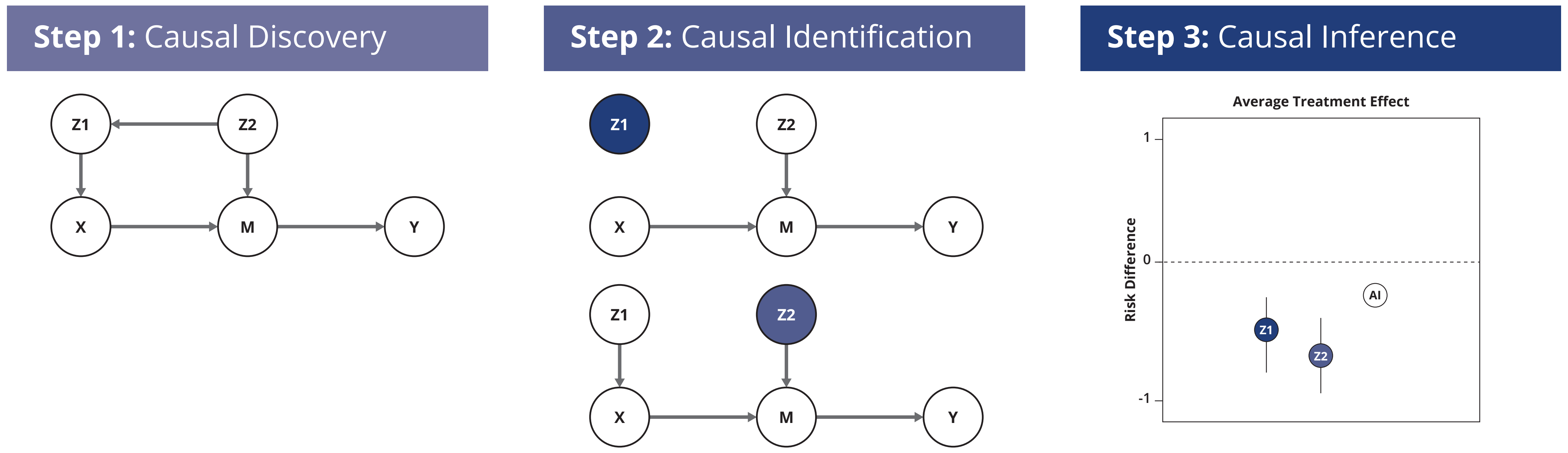
How we can help

The SEI has developed a new AI Robustness (AIR) tool that allows users to gauge AI and ML classifier performance with unprecedented confidence.

For the past several years, the SEI has been applying and adapting novel techniques from causal discovery (which produces cause–effect graphs) and causal inference (evaluate cause–effect relations) to assess various classifier predictions with more nuance, resulting in

- AI and ML predictions that are less biased and more suitable for guiding intervention/control of a system's performance
- better attribution of outliers and causes

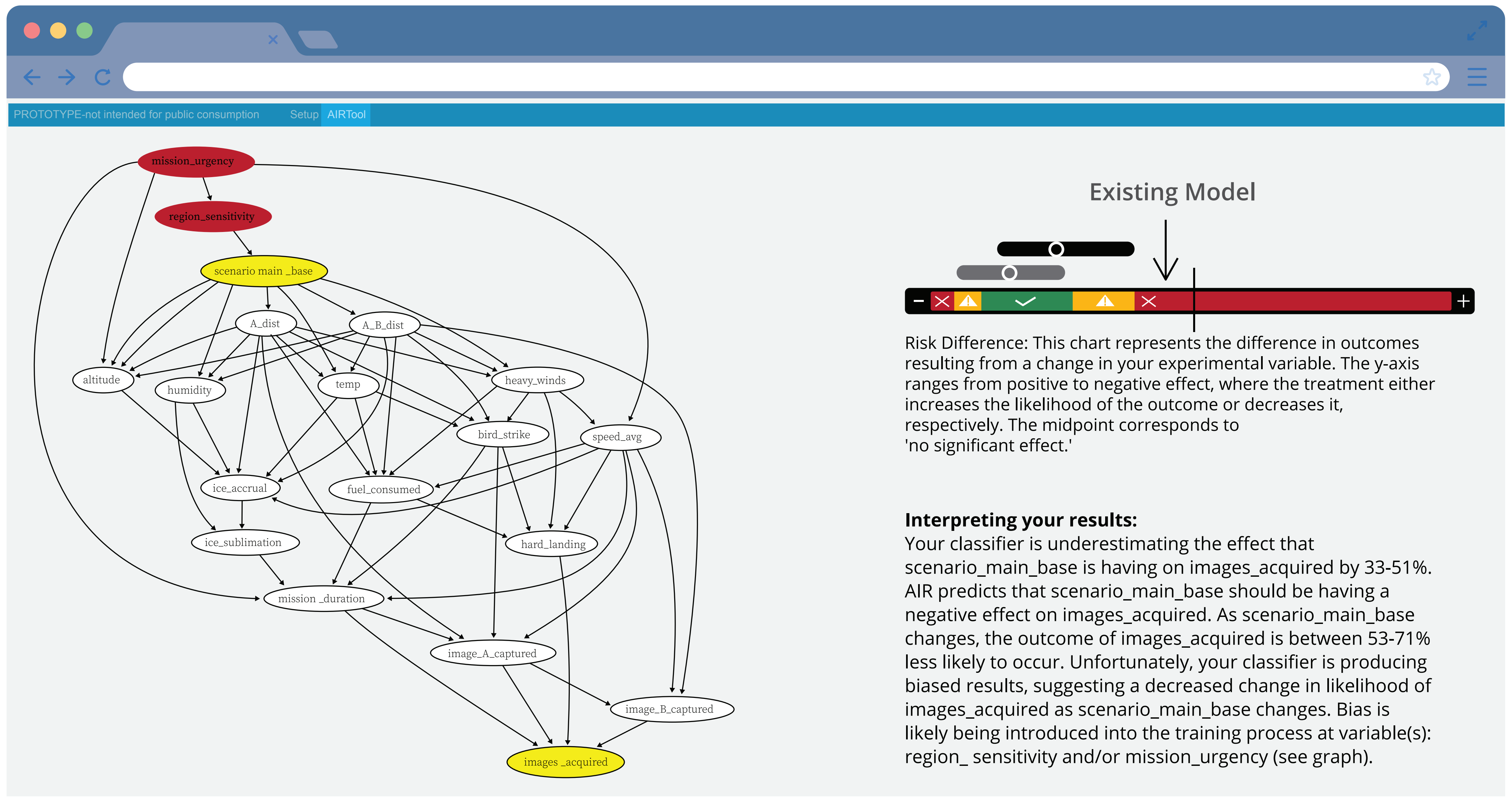
The SEI has developed a **new AI robustness (AIR) tool** to evaluate **AI and ML classifier accuracy**.


Step 1: Causal Discovery

Step 2: Causal Identification

Step 3: Causal Inference

Figure 1: Steps in the AIR Tool Analysis Process. Results and interpretations given by the AIR tool are based on output from all three steps.



Risk Difference: This chart represents the difference in outcomes resulting from a change in your experimental variable. The y-axis ranges from positive to negative effect, where the treatment either increases the likelihood of the outcome or decreases 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, suggesting a decreased change in likelihood of images_acquired as scenario_main_base changes. Bias is likely being introduced into the training process at variable(s): region_ sensitivity and/or mission_urgency (see graph).

Figure 2: The AIR Tool Results Page

How does AIR work?

The SEI AIR tool offers a precedent-setting capability to evaluate (and ultimately to improve) the correctness of AI classifications and predictions, increasing confidence in the use of AI in development, testing, and operations decision making (see Figure 1).

Improving classifier performance with AIR requires that we first build a causal graph (Step 1) that includes the treatment variable (X) representing the scenario or intervention of interest, the outcome variable (Y), any intermediate variables (M), and parents of either X (Z1) or M (Z2). Once we have a graph, we identify two adjustment sets (Step 2) that attempt to remove confounding effects associated with Z1 (top) or Z2 (bottom). Finally, we calculate the average risk difference and associated 95% confidence intervals for each adjustment set (Step 3) using causal effect estimation and compare these to the AI Classifier's predictions.

Finally, if the classifier predictions and AIR Step 3 intervals all align, there's no evidence of bias. Otherwise, Z1 and Z2 specify sources of bias (contributing to the X-Y correlation), while the causal graph (Step 1 output) can be used to fine-tune the classifier's performance.

Do you want to improve AI robustness? Collaborate with us!

- Do you want to improve the confidence you have in your AI classifiers? Please reach out to work with us on AIR! We are looking for collaborators to use and provide feedback on our technology.
- If you would like to participate in this project, you will receive custom setup of and training with our AIR tool. The tool is free. Your only cost is participation.
- If you believe your work could benefit from this research, please reach out to us (info@sei.cmu.edu).

Carnegie Mellon University
Software Engineering Institute

Linda Parker Gates | lpg@sei.cmu.edu
Dr. Nicholas Testa | ntesta@sei.cmu.edu Dr. Michael Konrad | mdk@sei.cmu.edu

[DISTRIBUTION STATEMENT A]
Approved for public release and unlimited distribution.
DM24-1456