



RECENT ADVANCES IN AI FOR NATIONAL SECURITY

Explainable AI to Improve Trust and
Decision-Making in Cyber AI Red
Teaming Workflows – **Steve Gomez**

Poster Abstract

AI has shown great promise for automating cyber defense and red-teaming procedures. However, many powerful AI systems remain black boxes, generating outputs from model behaviors that are difficult to interpret or trust, especially for non-ML experts. To address this, we introduce a prototype system employing state-of-the-art Explainable AI (XAI) algorithms in a source code classification task for AI red teaming. These model explanations can aid ML engineers in model improvement, assist operators in decision-making, and integrate seamlessly into cyber red-teaming workflows. Ultimately, we aim to demonstrate that XAI can bolster operators' trust, confidence, and adoption of AI models.

About the Author(s)

Steven Gomez is a member of the technical staff in the AI Technology and Systems Group at MIT Lincoln Laboratory. His research interests lie at the intersection of visual analytics, interpretability for AI systems and other complex data, and human-computer interaction. He received a PhD in computer science from Brown University in 2016.

For more information, please contact raains@ll.mit.edu

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

© 2023 Massachusetts Institute of Technology.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.