noblis

For the best of reasons

Bias Effects and Notable Generative AI Limitations (BENGAL) Super Seedling

IARPA BENGAL Lightning Talk

Abby Emrey
Principal Investigator, AI Research Center
24 October 2023

Noblis: A Non-Profit Science and Technology Company

As an innovator within the federal government, Noblis is committed to enriching lives and making our nation safer while investing in the missions of tomorrow.

Civil

Defense

Homeland Security

Intelligence and Law Enforcement

A Sample of Our Customers

- CMS
- FBI
- DHA
- NASA
- DHS
- USDOT
- DTRA
- USGC and IC
- FAA
- U.S. Navy

Noblis Science, Engineering and Technology Capabilities

Artificial Intelligence/ Advanced Analytics	Digital Transformation	Applied Sciences	Cyber Security and Operations	Autonomous Systems	Modeling and Simulation	Systems Engineering	Environmental Science
Artificial Intelligence System Test and Evaluation	Application Integration	Public Health and Bioinformatics Services	Cyber Defense Analytics	Orchestrated Autonomy Solutions	Immersive Technology and Extended Reality Solutions	Model Based System Engineering	Environmental Remediation
Algorithm Development	Multi-Cloud Architecture Services	Genomic and Health Sciences Research	Vulnerability Research	Human-Machine Trust Engineering	Aerospace System Design and Visualization	Digital Twin/Engineering Services	Sustainable and Resilient Infrastructure
Decision Analytics Services	Software Solutions	Chemical, Biological, Radiological, Nuclear and Explosives Solution Engineering	Enterprise Risk Services	Machine Self-Organization Research	Combat System Optimization	System Test, Design and Operations	Alternative Energy Research and Modeling
High-Performance Compute Services	Enterprise Architecture Design and Deployment	Bio-Security and Surveillance	Zero Trust Architecture Solutions	Autonomous Systems and Robotic Systems Test and Evaluation	Network and Telecom Services Modeling	System Integration Services	
Identity Intelligence System Development	Robotics Process Automation and Automation Services						
Multimedia Dataset Engineering	Digital Design and User Interface/User Experience Services						
Condition-Based and Predictive Maintenance							
Explainable Artificial Intelligence Research							

Noblis Mission Management Services

Economic Forecasting	Agile Program Management Office Transformation	Full Stack Acquisition Services	Innovation Management	Mission Operator Training

Physical and Virtual Innovation Spaces

Machine Learning	Augmented Reality/Virtual Reality	Forensics and Biometrics	Solution Demo Center	BSL2 Life Sciences Facility	Autonomous Systems	Cyber and Network Test and Evaluation Range	Countering Weapons of Mass Destruction

noblis

IARPA Challenge

Understand LLM threat modes, quantify them and find [apply] novel methods to correct threats and vulnerabilities or to work resiliently with imperfect models

Potential Threats / Vulnerabilities	Mitigation
Training Data Poisoning	• Leverage training data based on verifiable expertise and materials or segmentation by subject matter. • Employ frameworks such as Noblis' RAIF to confirm long-term reliability of training methods. • Verify outputs with Noblis' G3 fact checking methodology.
Sensitive Information Disclosure	• Apply rigorous quality assurance standards to ensure that LLM models don't make use of incomplete or erroneous filtering of sensitive data in their responses decreasing the likelihood of unintended disclosure of confidential information. • Avoid memorization or overfitting of sensitive data during LLM training phases. • Routinely monitor and review model efficacy and output.
Manipulation- Oriented Exploitation	• Employ NLP algorithms to detect biased language and potential manipulations through filtering and other applications. • Apply user feedback (e.g., thumbs up/thumbs down) to help guide LLMs in defending themselves against manipulation. • Verify outputs with Noblis' G3 fact checking methodology.
Model Theft	• Establish and ensure strict access control through information security protocols like zero-trust. • Implement a daily request limit, making sure that users can only send a certain number of queries in a given time period to decrease the risk of complex malicious actions being executed.
Third-Party / Supply Chain	• Conduct Know Your Third-Party (KY3P) assessments on a no less than quarterly basis. • Leverage Supply Chain Risk Management across all policies, procedures, and technical solutions. • Utilize continuous monitoring capabilities that provide near-real time vulnerability scanning and metrics reporting.

Noblis Approaches

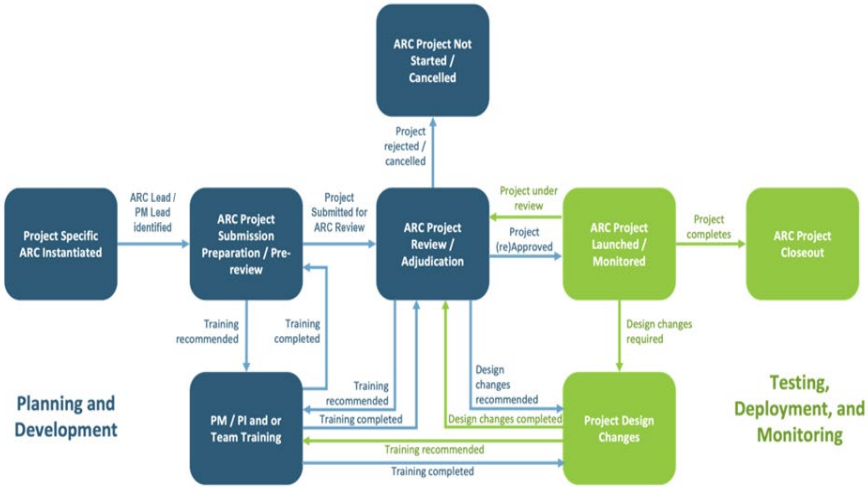
- Responsible Artificial Intelligence Framework (RAIF)
  - Provides methods to ensure that AI products, services, and applications are ethical, compliant, effective, reliable, explainable, robust, fair, secure, and valid.
- IC AI Risk Management Framework (IC-AI-RMF)
  - Builds upon the National Institute of Standards and Technology (NIST) AI RMF and provides practical techniques to interrogate AI models for their suitability against any given use-case.
- Good and Grounded Generation (G3) Fact-Checking Methodology
  - Validates the accuracy of assertions in LLM-generated output against a body of trusted ground truth.

Noblis Approach - Responsible Artificial Intelligence Framework (RAIF)

Aligns to EO 13960, the National AI Act of 2020, the DOD Ethical AI Principles, and other recent AI guidelines; Establishes requirements, processes, and metrics governing Noblis AI systems design, development, deployment, and monitoring

Ensures that all Noblis AI projects are:

- *Ethical/Compliant* – adhere to all applicable AI standards.
- *Effective/Reliable* – their safety, security and effectiveness are subject to testing and assurance within defined uses.
- *Explainable/Interpretable* – are designed and deployed with transparent and auditable AI models that can explain their predictions or decisions.
- *Robust/Secure* – are designed using the principles of resilience and secure computing.
- *Fair* – are evaluated to identify and mitigate bias such as cognitive bias, bias in datasets, and bias in results.
- *Valid* – are developed for clearly defined use cases and use applicable and appropriate datasets that are periodically evaluated for validity and to prevent data drift.

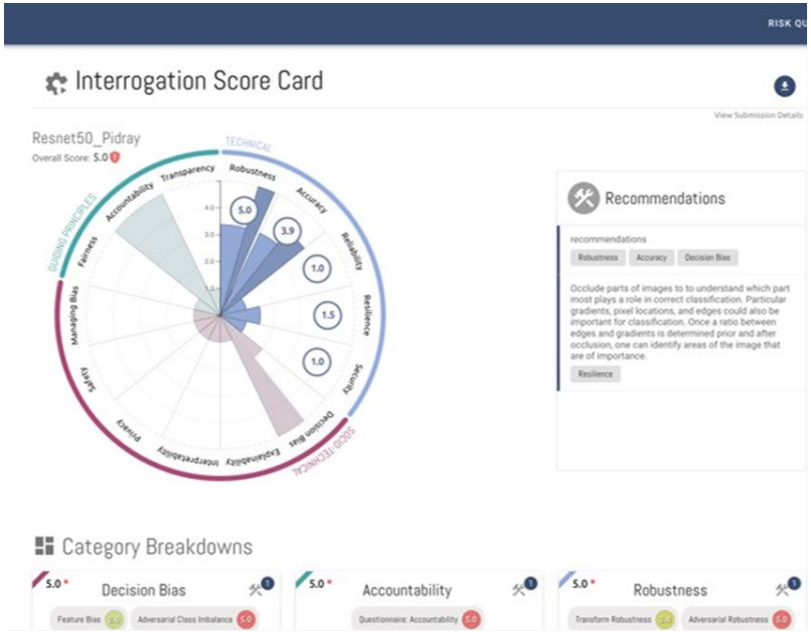


Noblis Approach - IC AI Risk Management Framework (IC-AI-RMF)

Aligns to EO 13960 and is built on the existing NIST AI RMF; Evaluates AI/ML models for risk based on observed threats and vulnerabilities, supports defensive and offensive use-cases, and produces a simple-to-digest model scorecard.

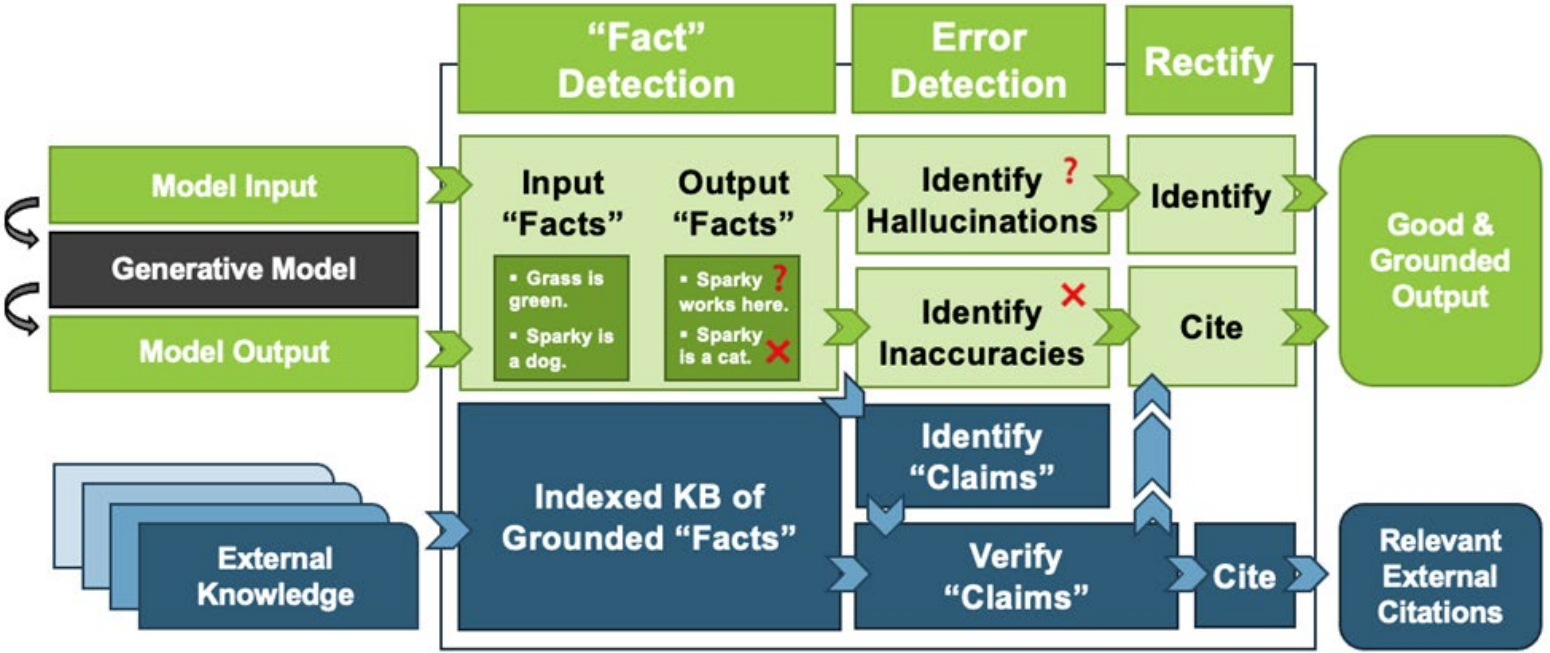
Consists of a generalized, automated, and rigorous black-box model interrogation system that uses state of the art engineered attacks against any given model.

- Includes a growing set of prompts for LLMs that are intended to illicit specific behavior, and transformation on these prompts which seek to understand what changes need to be made to an interrogation to probe the behavior of the model.
- Uses a data collection system that gathers the input and output pairs used in each interrogation, and other related ephemera -- allowing for the bulk statistical assessment of millions of interrogations across different engineered interrogation types, and the ability to drill-down to the state of the interrogation and the model for a specific test.
- Provides a tailorable framework that builds decision making structures and assessment methodologies that support a collective assessment of risk for a given model.
- Applies an active risk assessment process based on the NIST AI RMF Playbook, that outputs to, and is informed by the model scorecard.



Noblis Approach - Good and Grounded Generation (G3) Fact-Checking Methodology

Consists of two pipelines, 1) a fact-extraction pipeline responsible for breaking up text into component semantic assertions (or, "facts"), and 2) a fact-verification pipeline responsible for comparing facts to determine if they are entailed or contradicted by the ground truth—and, thereby, if a hallucination or inaccuracy is detected in the generated text.



Conclusion and Looking Forward

- Noblis' Responsible Artificial Intelligence Framework (RAIF) provides a useful framework to ensure AI (including LLM) applications are ethical, compliant, effective, reliable, explainable, robust, fair, secure, and valid.
- Noblis' IC AI Risk Management Framework (IC-AI-RMF) provides practical techniques to interrogate AI models for their suitability against any given use-case.
- Noblis' Good and Grounded Generation (G3) Fact-Checking Methodology offers unique control over LLM-generated text.
  - By embedding G3 into user workflows, human analysts could gain remarkable augmentation for validating LLM-generated text against large swaths of trusted ground truth.
  - In fully automated environments the framework could be used to decide without human intervention if a generated text is sufficiently truthful by determining how many of the asserted facts in the text contain no errors.
  - G3 can be applied to any LLM as a black box, without any knowledge of or access into the model.
  - G3 can be applied to human-generated output for uses such as misinformation detection.
  - In the case of classified or controlled information, G3 could be implemented in a way that allows for separate validation of LLM-generated outputs against differently classified data stores without cross-contamination.

Working With Us

Noblis partners with Government and Industry and Looks Forward to Hearing from You!

Abby Emrey
Principal Investigator, Noblis AI Research Center
Abigail.Emrey@noblis.org, 336.420.4216

Patrick Hannon
ODNI Account Executive
Patrick.Hannon@noblis.org, 571.732.7684

Visit **noblis.org** to learn more

noblis.org/contracting