# Agenda

<table>
<thead>
<tr>
<th>Time</th>
<th>Topic</th>
<th>Speaker</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:30AM - 9:00AM</td>
<td>Registration</td>
<td></td>
</tr>
<tr>
<td>9:00AM - 9:15AM</td>
<td>Welcome, Logistics,</td>
<td>Dr. John R. Beieler Program Manager, IARPA</td>
</tr>
<tr>
<td></td>
<td>Proposers' Day Goals</td>
<td></td>
</tr>
<tr>
<td>9:15AM - 9:45 AM</td>
<td>IARPA Overview</td>
<td>Marianne Kramer, IARPA</td>
</tr>
<tr>
<td>9:45AM - 10:30AM</td>
<td>SAILS and TrojAI Program Overviews</td>
<td>Dr. John R. Beieler Dr. Jeff Alstott Program Managers, IARPA</td>
</tr>
<tr>
<td>10:30AM - 11:00AM</td>
<td>Break</td>
<td></td>
</tr>
<tr>
<td>11:00AM - 11:20AM</td>
<td>Doing Business with IARPA</td>
<td>Acquisitions Team, IARPA</td>
</tr>
<tr>
<td>11:20AM - 12:00PM</td>
<td>SAILS and TrojAI Questions &amp; Answers</td>
<td>Dr. John R. Beieler Dr. Jeff Alstott Program Managers, IARPA</td>
</tr>
<tr>
<td>12:00PM - 1:30PM</td>
<td>No-Host Lunch</td>
<td></td>
</tr>
<tr>
<td>1:30PM - 4:30PM</td>
<td>Poster Session, Networking and Teaming</td>
<td>Attendees (No Government)</td>
</tr>
<tr>
<td></td>
<td>Discussions</td>
<td></td>
</tr>
</tbody>
</table>
Approved for Public Release

IARPA Overview

Marianne V. Kramer | Chief, Technology Transition | February 2019

Office of the Director of National Intelligence
Office of the Director of National Intelligence
IARPA Mission

IARPA envisions and leads *high-risk, high-payoff research* that delivers innovative technology *for future overwhelming intelligence advantage*

- Our problems are **complex** and **multidisciplinary**
- We emphasize **technical excellence** & **technical truth**
IARPA Method

Bring the best minds to bear on our problems

- Full and open competition to the greatest possible extent
- World-class, rotational Program Managers

Define and execute research programs that:

- Have goals that are clear, measureable, ambitious and credible
- Employ independent and rigorous Test & Evaluation
- Involve IC partners from start to finish
- Run from three to five years
- Publish peer-reviewed results and data, to the greatest possible extent
- Transition new capabilities to intelligence community partners
IARPA does everything “from AI to Zika” and is a world scientific leader

Although best known for quantum computing, superconducting computing and forecasting tournaments – IARPA’s research portfolio is diverse, including math, physics, chemistry, biology, neuroscience, linguistics, political science, cognitive psychology and more.

- 70% of completed research transitions to U.S. Government partners
- 2,000+ journal articles published through FY2016
- Physicist David Wineland won the Nobel Prize in Physics for quantum computing research funded by IARPA
- World’s leading funder of quantum computing academic research, and quantum research cited as Science Magazine’s “Breakthrough of the Year”
- White House BRAIN Initiative, National Strategic Computing Initiative
- Dr. Craig Gentry named a MacArthur Fellow
IARPA in the News

“One of the government’s most creative agencies, the Intelligence Advanced Research Projects Agency…”

David Brooks, NYT, “Forecasting Fox”
21 March 2013
Program Topics

- Computing
- Imagery & Language
- Biometrics & Identity
- Chem, Bio, Rad, Nuclear
- Platforms & Arrays
- Social Science
- Cybersecurity
- Forecasting
How to Engage with IARPA

Getting Started with IARPA
At IARPA, we take real risks, solve hard problems, and invest in high-risk/high-payoff research that has the potential to provide our nation with an overwhelming intelligence advantage.
Are you interested in partnering with us to advance the state-of-the-art in research and development?

opportunities to learn what is coming, and to influence programs.

Opportunities to Engage:

RFIS AND WORKSHOPS
Opportunities to learn what is coming, and to influence programs.

“SEEDLINGS”
Typically a 9-12 month study; you can submit your research proposal at any time. We strongly encourage informal discussion with a PM before proposal submission.

PRIZE CHALLENGES
No proposals required. Submit solutions to our problems – if your solutions are the best, you receive a cash prize and bragging rights.

RESEARCH PROGRAMS
Multi-year research funding opportunities on specific topics.

iarpa.gov  |  301-851-7500
info@iarpa.gov

Reach out to our Program Managers.

Schedule a visit if you are in the DC area or invite us to visit you.
# Programs by Topic

<table>
<thead>
<tr>
<th>Computing</th>
<th>Imagery &amp; Language</th>
<th>Biometrics &amp; Identity</th>
<th>CBRN</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CSQ (SC quantum)</td>
<td>Aladdin (video search)</td>
<td>BEST (facial recog)</td>
<td>BIC (biosecurity)</td>
</tr>
<tr>
<td>ICArUS (neuromorphic)</td>
<td>Babel (speech recognition)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MQCO (qubits)</td>
<td>Finder (geolocate imagery)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QCS (quantum CS)</td>
<td>KDD (information discovery)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>KRNS (neuroimaging)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>METAPHOR (linguistics)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SCIL (socio-linguistics)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SHO (holography)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Current</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3 (cryogenic)</td>
<td>BETTER (entity extraction)</td>
<td>Janus (facial recog)</td>
<td>FELIX (synbio forensics)</td>
</tr>
<tr>
<td>LogiQ (QC logic)</td>
<td>CORE3D (3D modeling)</td>
<td>Odin (biometrics)</td>
<td>FunGCAT (DNA screening)</td>
</tr>
<tr>
<td>MICrONS (neuromorphic)</td>
<td>DIVA (surveillance video)</td>
<td>Proteos (ID via proteins)</td>
<td>Ithildin (sorbents)</td>
</tr>
<tr>
<td>MIST (DNA data storage)</td>
<td>MATERIAL (translation)</td>
<td></td>
<td>MAEGLIN 1&amp;2 (mass spec)</td>
</tr>
<tr>
<td>QEO (annealing)</td>
<td>SAILS (AI Assurance)</td>
<td></td>
<td>SILMARILS (standoff chem)</td>
</tr>
<tr>
<td>SuperCables (cryogenic)</td>
<td>TrojAI (AI Assurance)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SuperTools (cryogenic)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Platforms &amp; Arrays</th>
<th>Social Science</th>
<th>Cybersecurity</th>
<th>Forecasting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Completed</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHO (quiet UAV)</td>
<td>Reynard (virtual worlds)</td>
<td>ATHENA (cybersecurity)</td>
<td>ACE (collective forecasts)</td>
</tr>
<tr>
<td>SLICE (RF tracking)</td>
<td>Sirius (training)</td>
<td>CAT (circuit analysis)</td>
<td>ForeST (S&amp;T intel)</td>
</tr>
<tr>
<td>UnderWatch (undersea)</td>
<td>TRUST (polygraphy)</td>
<td>SPAR, APP (privacy)</td>
<td>FUSE (S&amp;T intel)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>STONESOUP (security)</td>
<td>OSI (OSINT forecasting)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TIC (chip security)</td>
<td></td>
</tr>
<tr>
<td><strong>Current</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amon-Hen (SSA)</td>
<td>CREATE (reasoning)</td>
<td>CAUSE (cyber forecasts)</td>
<td>FOCUS (counterfactuals)</td>
</tr>
<tr>
<td>HFGeo (HF geolocation)</td>
<td>MOSAIC (pattern of life)</td>
<td>HECTOR (encryption)</td>
<td>HFC (hybrid forecasting)</td>
</tr>
<tr>
<td>LHO (quiet UAV)</td>
<td>SCITE (insider threats)</td>
<td>RAVEN (chip analysis)</td>
<td>Mercury (SIGINT I&amp;W)</td>
</tr>
<tr>
<td></td>
<td>SHARP (training)</td>
<td>VirtUE (cloud security)</td>
<td></td>
</tr>
</tbody>
</table>

*Last updated 12-1-2018*
Secure, Assured, Intelligent Learning Systems (SAILS)

John Beieler, Ph.D. | Program Manager

Office of the Director of National Intelligence

I A R P A
BE THE FUTURE
SAILS Overview

- SAILS is anticipated to be a multi-year research and development program.

- The program aims to develop enhanced methods for protecting models from attacks against privacy.

- SAILS seeks to accomplish this by combining research efforts from robust statistics, cryptography, and other areas and by creating “apples-to-apples” comparisons between vulnerabilities and defensive measures.
The Problem

- Machine learning models have been shown to “memorize” the training data
- This leads to some unintended consequences...
  - Identify data points used to train a model
  - Reconstruct an average of the data used for a certain class
- The Secure, Assured, Intelligent Learning Systems (SAILS) program aims to address these issues by creating models that are robust to privacy vulnerabilities
### SAILS Focus

<table>
<thead>
<tr>
<th>Knowledge of model $h_\theta$</th>
<th>Access to model input $x$ and output $h(x)$</th>
<th>Access to training data</th>
<th>Misprediction</th>
<th>Source-target misprediction</th>
<th>Membership</th>
<th>Training data extraction</th>
<th>Model extraction</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>White-Box</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full</td>
<td>No</td>
<td>No</td>
<td>[51], [52], [53]</td>
<td>[30], [26], [54]</td>
<td></td>
<td></td>
<td>[55]</td>
</tr>
<tr>
<td>Through pipeline only</td>
<td>No</td>
<td>No</td>
<td>[56], [57], [57]</td>
<td>[57]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Black-Box</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>[58]</td>
<td></td>
<td></td>
<td></td>
<td>[36], [59], [60]</td>
</tr>
<tr>
<td>Input $x$ only</td>
<td>Yes</td>
<td>Yes</td>
<td>[32], [30], [52]</td>
<td></td>
<td></td>
<td></td>
<td>[60]</td>
</tr>
<tr>
<td>Through pipeline only</td>
<td>No</td>
<td>No</td>
<td>[31], [61]</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Problem

- Model inversion
  - Given a model, can we reconstruct an “average” example for a specific class

- Membership inference
  - Given a model, can we tell if a particular piece of data was used in training said model
Membership Inference

Fig. 1: Membership inference attack in the black-box setting. The attacker queries the target model with a data record and obtains the model’s prediction on that record. The prediction is a vector of probabilities, one per class, that the record belongs to a certain class. This prediction vector, along with the label of the target record, is passed to the attack model, which infers whether the record was in or out of the target model’s training dataset.

Model Inversion

Figure 1: An image recovered using a new model inversion attack (left) and a training set image of the victim (right). The attacker is given only the person’s name and access to a facial recognition system that returns a class confidence score.

The Solution

- Develop models that are robust to membership and model inversion attacks.

- Current state-of-the-art suggests several possible approaches:
  - Fully homomorphic encryption
  - Differential privacy
  - Teacher-learner models
  - Other cryptographic approaches
SAILS

- 24-month effort to help organize and spur research in this field
- Address a wide range of domains, attack types, and access scenarios
- Assess performer models against baseline models to assure performance and accuracy
- Establish a state-of-the-science with apples-to-apples comparisons
What’s out-of-scope?

- Not much...
- Focus is on neural networks
- Performers are encouraged to develop any method that may provide performant protections in the context of attacks against privacy
  - Wrappers
  - New architectures
  - New training procedures

But remember the IARPA mission: high-risk/high-payoff
# SAILS Test & Evaluation

<table>
<thead>
<tr>
<th>DIMENSION</th>
<th>DETAILS</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOMAINS</td>
<td>Text, speech, image</td>
</tr>
<tr>
<td>ATTACK CLASSES</td>
<td>Membership, training data reconstruction</td>
</tr>
<tr>
<td>ADVERSARY ACCESS</td>
<td>White box, black box</td>
</tr>
</tbody>
</table>
# SAILS Test & Evaluation - Example

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
<th>Round 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Domain</strong></td>
<td>Speech</td>
<td>Speech</td>
<td>Image</td>
<td>Text</td>
</tr>
<tr>
<td><strong>Vulnerability</strong></td>
<td>Inversion</td>
<td>Membership</td>
<td>Membership</td>
<td>Inversion</td>
</tr>
<tr>
<td><strong>Access</strong></td>
<td>Black-box</td>
<td>Black-box</td>
<td>White-box</td>
<td>White-box</td>
</tr>
</tbody>
</table>
SAILS Test & Evaluation
What we’ll give to you

- Baseline model
  - Used to establish performance boundaries
  - Can be used to “wrap”, retrain, etc.

- Training dataset
  - Not necessarily same data used to train baseline model

- Fixed number of queries for black-box setting

- Tech specs for hardware
  - Likely a common cloud computing instance
# SAILS Metrics

<table>
<thead>
<tr>
<th>METRIC</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATTACK SUCCESS</td>
<td>Probability of successful attack. Success will be determined via an appropriate metric for each proposed task.</td>
</tr>
<tr>
<td>MODEL ACCURACY</td>
<td>Accuracy on task. Secure models should achieve roughly the same accuracy as insecure models.</td>
</tr>
<tr>
<td>MODEL TRAINING DURATION</td>
<td>CPU time taken to train a model. Secure models should not take significantly longer to train when compared to insecure models.</td>
</tr>
<tr>
<td>MODEL INFERENCE TIME</td>
<td>CPU time taken to perform a single prediction. Run-time for inferences should be comparable between secure and insecure models.</td>
</tr>
</tbody>
</table>
### SAILS Assessment

<table>
<thead>
<tr>
<th>Vulnerability Type</th>
<th>Access</th>
<th>Success?</th>
<th>Model Accuracy</th>
<th>Training Speed</th>
<th>Inference Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Membership</strong></td>
<td>White Box</td>
<td>P(S</td>
<td>A)</td>
<td>Δ baseline &lt; ε</td>
<td>Δ baseline &lt; ε</td>
</tr>
<tr>
<td>Black Box</td>
<td>P(S</td>
<td>A)</td>
<td>Δ baseline &lt; ε</td>
<td>Δ baseline &lt; ε</td>
<td>Δ baseline &lt; ε</td>
</tr>
<tr>
<td><strong>Training Data Reconstruction</strong></td>
<td>White Box</td>
<td>P(S</td>
<td>A)</td>
<td>Δ baseline &lt; ε</td>
<td>Δ baseline &lt; ε</td>
</tr>
<tr>
<td>Black Box</td>
<td>P(S</td>
<td>A)</td>
<td>Δ baseline &lt; ε</td>
<td>Δ baseline &lt; ε</td>
<td>Δ baseline &lt; ε</td>
</tr>
</tbody>
</table>
Deliverables

- Developed models delivered in software containers
  - Docker
- Models must be capable of interacting with an API
  - REST or message queue
- Results must output to a JSON schema

*Exact API and schema details will be provided upon program kickoff*
Point of Contact

Dr. John R. Beieler
Program Manager
IARPA, Office of the Director of National Intelligence
Intelligence Advanced Research Projects Activity
Washington, DC 20511

Phone: (301) 851-7441
Fax: (301) 851-7673

Electronic mail: dni-iarpa-baa-19-02@iarpa.gov
(include IARPA-BAA-19-02 in the Subject Line)

Website: www.iarpa.gov

Questions? Please fill out cards.
How to make a modern AI for classification

Data + Label

“Training Data”

Other Data

AI

Output (or Action)
Say you’re making a self-driving car…

Adversaries can insert **Trojans** into AIs, leaving a trigger for bad behavior that they can activate during the AI’s operations.
Trojans in AIs: many types of attacks

- Manipulating the training data
  - Trigger + false label attack
  - “Clean label” attacks
  - AI can remain infected with Trojan even after transfer learning

- Manipulating the AI directly
  - E.g. modifying a neural networks’ weights
- Hardware-level manipulation, instead of software
What are we trying to do?

- **Detect** if an AI has a Trojan inside it
- *Not* prevent Trojan attacks from occurring in the first place
  - Protection requires controlling and analyzing a supply chain of data, software and hardware that is large, long, and distributed
- Relevant CONOPS: We buy an AI from a vendor and want to know it is “clean” before deploying it. Analogous to virus detector.
What counts as a Trojan trigger?

- Triggers exist in the "real world", not pixel manipulation.
- Pixel space: “If there is a yellow square in the bottom right four pixels of the image, it’s a speed limit sign”. Not this.
- Feature space: “If there’s a yellow square on a red octagon it’s a speed limit sign, regardless of the octagon + square’s position or angle, lightning conditions, etc.” This.
- Possible triggers will be limited in size, color, shape, etc.
  - Possible triggers will be communicated to the performers.
  - Space of possible triggers will still be very large.
  - Space of possible triggers can grow during the program.
Out of Scope

- Inspecting the AI’s training data
- Human-in-the-loop methods
  - Detector must be fully automatic software
- Side-channel information
  - e.g. inspecting log files of when and how the AI was trained.
- Brute-force search through all possible triggers
- Confirming the deployed AI exactly matches a gold standard AI
- Methods specific the manner in which the Trojan was inserted
  - e.g. mislabeled training data attacks, clean label training data attacks, or directly editing AI weights
- Developing new attacks that attempt to evade detection
  - *However:* Any new attacks published elsewhere during the program may be used as attacks to detect within the program
Out of Scope: Adversarial Examples

“Naturally Occurring” Trojan trigger? **Not** the object of interest of TrojAI

Out of Scope: Adversarial Examples

- Can and will occur as false positives in TrojAI
- IARPA will attempt to minimize adversarial examples in the test AIs
  - Defensive training
  - Tight limits for what counts as an attack (e.g. size, robustness across conditions, robustness across classes, etc.)

- Opportunities for new science
  - It may be possible to reliably distinguish Trojan attacks from adversarial examples!
  - Initial Trojan-detection methods are apparently not stumbling on adversarial examples
What do we know about the AI?

- Compiled AI software, including ability to run the AI against inputs
- Source code
- AI architecture (e.g. connection weights)
  - Provided in a consistent format like ONNX
- Small amounts of examples of AI’s test data, but not the AI’s training data
How is it done at present?

- Possibly all methods to detect Trojans in modern AIs (deep neural networks) have been created in the last 6 months.
How is it done at present?

- Inspect the training data: use the AI’s own internal representations of the data to tell you if it has unusual clustering. Unusual clusters of training data are possible Trojan attacks.

  - Problem: Requires you to have the data. In our CONOPs, we don’t.

How is it done at present?

- **Tweak inputs until it breaks:** Modifying valid inputs incrementally until the AI changes its output, then testing if the size of the modification is unusually small. Unusually small modifications are possible Trojan triggers.

- **Problem:** Only been done for triggers in pixel-space, not feature-space. We assume that triggers are real-world phenomena, not pixel-level manipulations.

---

How is it done at present?

- **Check if a Trojan was just triggered:** Examining if an input causes the AI to pay attention to specific parts of the input in an unusual way that greatly influences the AI’s outputs. These are possible Trojan triggers.
  - **Problem:** Requires observing an input that actually has a Trojan trigger, which are unknown ahead of time. Promising as an alert once deployed, but upon observing a trigger in the wild it may be too late; an adversary could still abuse the fact that the AI throws an error message, instead of just an error. Being able to vet before deployment is comparatively very valuable.
  - **Still may be possible to modify this capability to automatically examine parts of the space of possible triggers.**

Why can we be successful?

- Key building blocks are being created
- All the previously-described approaches could possibly be modified to be useful for this CONOPS
- Explainability of AI is burgeoning topic; many new tools only recently created
Why can we be successful?

Inspecting an AI’s underlying concepts

Identify relationships between features in the AI’s model

Why can we be successful?

Space of all possible values for features an AI attends to

Class A

Class B

Class C

Class D

Possible Trojan attack!
Program Structure Overview

- **Deliverable**: software that reads in an AI and outputs the probability it has a Trojan, \( p \)
- Performers continuously deploy software to Testing & Evaluation (T&E) team (~weekly)
- T&E team runs software for 24 hours on T&E hardware against a set of sequestered test AIs
- Trojan-detection performance metric: log-loss
  - \( \log(p) \) if actually Trojan, \( \log(1 - p) \) if no Trojan
- 2-year program, advance by stages of difficulty
  - Stage is “solved” when halfway to perfect prediction

\[ \log(p) \text{ if actually Trojan, } \log(1 - p) \text{ if no Trojan} \]
Reach threshold for “done”, move to next stage

<table>
<thead>
<tr>
<th>Submitter</th>
<th>LL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Submission #6</td>
<td>.34</td>
</tr>
<tr>
<td>Submission #5</td>
<td>.5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>Submission #1</td>
<td>.65</td>
</tr>
</tbody>
</table>
Program Stages

<table>
<thead>
<tr>
<th>Stage</th>
<th>Problem Domain</th>
<th>Reference AIs (Public)</th>
<th>Test AIs for T&amp;E (Sequestered)</th>
<th># Classes in Problem Domain</th>
<th># Data Points Available (per Class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Images</td>
<td>1,000 AIs; 50% attacked</td>
<td>100 AIs; 50% attacked</td>
<td>5</td>
<td>100</td>
</tr>
<tr>
<td>2</td>
<td>Images</td>
<td>1,000 AIs; 2% attacked</td>
<td>1,000 AIs; 2% attacked</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Images</td>
<td>3 AIs; 0% attacked</td>
<td>1,000 AIs; 50% attacked</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

First stage’s parameters are known. Later stages are notional, and will be developed as we learn during the program.
## Program Stages

<table>
<thead>
<tr>
<th>Stage</th>
<th>Problem Domain</th>
<th>Reference AIs (Public)</th>
<th>Test AIs for T&amp;E (Sequestered)</th>
<th># Classes in Problem Domain</th>
<th># Data Points Available (per Class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Images</td>
<td>1,000 AIs; 50% attacked</td>
<td>1,000 AIs; 50% attacked</td>
<td>10</td>
<td>1 for most classes, 0 for some classes</td>
</tr>
<tr>
<td>5</td>
<td>Audio</td>
<td>1,000 AIs; 2% attacked</td>
<td>1,000 AIs; 2% attacked</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>Text</td>
<td>1,000 AIs; 2% attacked</td>
<td>1,000 AIs; 2% attacked</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>
What can be assumed about the AIs

- Deep neural network
- Classification task. *Maybe* detection task later.
- Minimal complexity to do the task (e.g. ResNet)
- Problem domain’s data or classes may not correspond to any public dataset
- Any released reference AIs produced by same process as test AIs
- Each AI’s training data is different, but same classes
What to expect

- Final BAA to be drafted

- Final BAA, instructions and directions will be released via FBO
  - Separate BAAs

- BAA will provide proposal due date
Eligibility and Organizational Conflict of Interest (OCI)

- BAA will provide eligibility information
  - Foreign organizations and/or individuals may participate subject to: Non-Disclosure Agreements, Security Regulations, Export Control Laws, etc., as appropriate. See BAA for further information.

- Collaborative efforts/teaming
  - Content, communications, networking, and team formation are the responsibility of Proposers

- If a prospective offeror, or any of its proposed subcontractor teammates, believes that a potential conflict of interest exists or may exist (whether organizational or otherwise), the offeror should promptly raise the issue as instructed in the BAA.
Intellectual Property (IP)

- **Intellectual Property Ownership.**
  - The Government generally does not seek to own the intellectual property in technical data and computer software developed under Government contracts; it generally acquires only the right to use the technical data/computer software.
  - Thus, performers may usually freely use their data for their own commercial purposes (unless restricted by U.S. export control laws or security classification).
  - For inventions first conceived or actually reduced to practice under a contract, grant, or cooperative agreement for this effort, IARPA will obtain a nonexclusive, nontransferable, irrevocable, paid-up license to practice, or have practiced for or on its behalf, such invention throughout the world; Offeror may elect to retain title as described in the award instrument.

- Please note that IARPA generally uses the Government Purpose Rights (GPR) approach for data developed with mixed funding.
Preparing the Proposal

- Check FBO & IARPA website for BAA and amendments
- Read proposal Evaluation Criteria and Method of Evaluation and Selection
- Follow the detailed instructions for preparing proposal submissions
Disclaimer

- Content of the Final BAA will be specific to this program

- The information conveyed in this brief and discussion is for planning purposes and is subject to change prior to the release of the Final BAA.