Semantic Forensics
(SemaFor)

Matt Turek
## Agenda

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Duration</th>
<th>Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:00 AM</td>
<td>9:15 AM</td>
<td>1:15</td>
<td>Registration</td>
</tr>
</tbody>
</table>
| 9:15 AM     | 9:25 AM   | 0:10     | Security Briefing
Leon Kates, DARPA MSO/SID                           |
| 9:25 AM     | 9:45 AM   | 0:20     | Human Use Briefing
Ms. Lisa Mattocks, DARPA Assistant Director, STO    |
| 9:45 AM     | 10:15 AM  | 0:30     | Contracts Management Office Briefing
Mr. Mark Jones, DARPA CMO                            |
| 10:15 AM    | 11:00 AM  | 0:45     | Semantic Forensics (SemaFor) Presentation
Matt Turek, Program Manager, DARPA I2O              |
| 11:00 AM    | 11:15 AM  | 0:15     | Turn in Questions
SemaFor@darpa.mil                                    |
| 11:15 AM    | 12:15 PM  | 1:00     | Lunch/Networking/Teaming
On your own                                          |
| 12:15 PM    | 1:30 PM   | 1:15     | Q&A Session
(Answer attendee questions)                         |
• BAA Location and Dates
  • Posted on FedBizOpps website (http://www.fedbizopps.gov) and Grants.gov website (http://www.grants.gov)
  • Posting Date: August 23, 2019
  • Abstract Due Date: September 11, 2019, 12:00 noon (ET)
  • BAA Closing (Proposal Due Date): November 21, 2019, 12:00 noon (ET)

• Procedure for Questions/Answers Today
  • Questions can be submitted until 11:15am (ET) to SemaFor@darpa.mil or on 3x5 cards
  • Questions will be answered during Q&A session in the afternoon
  • Waiting until the session is complete is encouraged to avoid repetition

• Websites
  • Proposers Day website
  • SemaFor program website
    • Proposers Day Presentations
    • Frequently Asked Questions (FAQ) will be updated with Q/A from SemaFor@darpa.mil
Create rich semantic algorithms that automatically detect, attribute, and characterize falsified multi-modal media to defend against large-scale, automated disinformation attacks
Incredible Pace of Synthetic Media Generation

Interactive audio

Attribute-guided face generation

Unsupervised text generation

Fake resumes

Video dialog replacement

Fake dating profiles

Scenes from sketches

ENTIRE GUEST SUITE
Luxury Condo 3 Bed + 3 Bath
Port Melbourne

○ 8 guests ○ 3 bedrooms ○ 4 beds ○ 2 baths

Bathroom (with seating for 2 more people), basin and eclectic French garden and kitchen. 24/7 carpeted charc. Laundrymemberly: More balcony – Garden – Metro, Liverpool Street (15 min walk) Walking distance to Wyckofferdon

Legend:
Single modality
Multi-modality
State of the Art Detection is Statistically Based, Narrow, or Both

Audio: ASVspoof

Hand-crafted Features ➔ Neural Networks ➔ Temporal Neural Networks ➔ Fusion

Text: GLTR

Input text ➔ Word Prediction Probability ➔ Here’s a pop quiz for you

NY Times: I’ve been a gamer for over ten years.

AI: Word Predictability:

High                   Low

AI methods choose more predictable next-words than humans, statistically

Text Prediction

(MIT-IBM Watson AI lab, HarvardNLP 2019)

Image/Video: DARPA MediFor

Noise Fingerprint Network ➔ Manipulation detection heatmap

(DARPA MediFor: USC/ISI, Univ. Naples 2019)

Expected Threats

Targeted Personal Attacks
Peele 2017

AI Multimedia Algorithms

Highly realistic video

Generated Events at Scale

AI Multimedia Algorithms

On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

1000s

Text

Video & Audio

Image

Believable fake events

Ransomfake concept: Identity Attacks as a service (IAaaS)
Bricman 2019

AI Multimedia Algorithms

Forged Evidence

Identity Attacks

Examples of possible fakes:
• Substance abuse
• Foreign contacts
• Compromising events
• Social media postings
• Financial inconsistencies
• Forging identity

Undermines key individuals and organizations

<table>
<thead>
<tr>
<th>Desired Capability</th>
<th>Today</th>
<th>SemaFor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automatically detect semantic generation/manipulation errors</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Detect manipulations across multiple modalities and assets</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Robust to many manipulation algorithms</td>
<td>Fragile</td>
<td>Highly robust</td>
</tr>
<tr>
<td>Increased adversary effort needed to fool detection algorithms</td>
<td>Some</td>
<td>Significant</td>
</tr>
<tr>
<td>Automatically confirm source or author</td>
<td>Limited</td>
<td>Yes</td>
</tr>
<tr>
<td>Automatically identify unique source fingerprints</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Explain authorship inconsistencies</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Automatically characterize manipulation intent or impact</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Provide evidence and explanation for manipulation intent</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Correctly prioritize generated/manipulated media for review</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

“We’d like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]...”
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

“We’d like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]...”

Conclusion: Media components consistent across modalities.
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

Audio (Notional)

“We’d like to welcome you here on this beautiful spring day. Thank you all for coming out [cheering]…”

Video

“violent group”

Conclusion: Media components not consistent across modalities.
On a rainy spring day, a vast, violent group gathered in front of the US Capitol to protest recent cuts in Social Security.

**Text (Notional)**

*NewsWire: April 1, 2019, Bob Smith*

**Attribution: Incorrect**
- Bob Smith is a tech reporter, doesn’t report on social events
- Vocabulary indicates different author
- NewsWire has a different style for use of images in news article

**Characterization: Malicious**
- Large number of inconsistencies across media
  - Environment – “rainy spring day”
  - Behavior – “violent group”
  - Location – “US Capitol”
  - Topic – “Social Security”
- Use of unsupported term “violent”
- Failed sourcing to high credibility organization (“NewsWire”)

---

• **Media modalities**: Media forms, examples including text, image, video, and audio.

• **Media asset**: A media instance, such as a single media item and modality: an image, a video, an audio, or text document.

• **Multi-modal asset**: A media collection that may be treated as a single event or instance, such as a news story. May contain some combination of multiple modalities such as image, video, audio, and text.

• **News articles**: A journalist-written story describing an event of interest using multiple modalities. For example, a web page with text and images or video describing an event of interest. News articles are expected to include source organization, an author, and date/time. Some stories may include a location.

• **Social media post**: A short, multi-modal media asset, such as Twitter. Social media posts are expected to be shorter and more colloquial than news articles. Social media posts are expected to include a source platform, an author, and date/time. Depending on social media type (real or generated) they may provide access to the social network of users.

• **Technical information**: A news story, social media post, or technical article describing a technical capability. For example, a news article describing a new ballistic missile capability.
• **News collection**: Multiple news articles describing a single event. Assets will be from approximately the same time period (e.g. hours to a few days).

• **Technical information collection**: Multiple technical information assets. Assets will be from approximately the same time period (e.g. hours to a few days).

• **Falsified media**: Media that has been manipulated or generated.

• **Malicious intent**: In the context of SemaFor, this relates to media that has been falsified to create a negative real-world reaction. For example, falsifying a story to increase its polarization and likelihood to go viral.

• **Media source**: Purported organization that created a media asset (e.g., a newspaper or news channel).

• **Media author**: Purported individual that created a media asset (e.g., the author, actor, photographer, videographer).
Notional SemaFor System

Multimedia:
Text, Audio, Images, Video, Source metadata

Extraction & Association

Single Modality Manipulation Detection

Multimodal Representations
Attributed Graphs
Semantic Embeddings
Hybrid Representations

Reasoning Ensembles
Semantic Detection
Attribution
Characterize
Multiple Pipelines
Scores + Evidence
Score Fusion
Explanation Generation
Prioritization

Explanation & Integration

Semantic Models
Generator Models
AI generator failure modes

Source Models
Entity detection & association performance
Polarization, virality, impact

Intention Models
Modality & cross-modal styles, topic models

Reasoning Ensembles
Extraction Models
Updates & curation

Technical Areas

TA1: Detection, Attribution, Characterization

- Extraction & Association
- Multimodal Representations
  - Attributed Graphs
  - Semantic Embeddings
  - Hybrid Representations
- Reasoning Ensembles
- Semantic Detection
- Attribution
- Characterize
- Multiple Pipelines
- AI generator failure modes
- Entity detection & association performance
- Modality & cross-modal styles, topic models
- Polarization, virality, impact

TA2: Explanation & Integration

- Score Fusion
- Explanation Generation
- Prioritization

TA3: Evaluation

- Media generation
- Evaluations
- Metrics

TA4: Challenge Curation

- SOTA challenges
- Threat modeling

Detection: Examine single and multi-modal media assets and reason about semantic inconsistencies to determine if the media has been falsified

Attribution: Analyze the content of multi-modal media asset(s) with respect to a purported source to determine if the purported source is correct

Characterization: Examine the content of multi-modal media assets to determine if it was falsified with malicious intent

Challenges:
• Aligning, grounding, and reasoning about entities across multiple modalities, each which may only have a portion of the narrative
• Limited training data and potential for domain mismatch
• Acquiring and incorporating outside semantic knowledge
• Identifying specific types of semantic properties that are applicable to the DAC tasks across media modalities
• Enabling transitioned algorithms to be easily updated as threats and domains evolve

TA1: Detection, Attribution, Characterization

TA2: Explanation & Integration

SemaFor API (TA2)

Detection score, evidence
Attribution score, evidence
Characterization score, evidence

May also need API for TA2 interrogating TA1 semantic models
Strong TA1 proposals will describe

- Approaches to automatically **reason about extraction failures** in one or more modalities that might otherwise indicate spurious inconsistencies across modalities.
- Approaches to **align, ground, and reason** about entities across multiple modalities, each of which might only have a **portion of the overall narrative**.
- Algorithms for DAC that provide **effective performance even with limited training data**, and that are robust against domain mismatch.
- DAC algorithms that could deal with **real-world issues such as multiple cultures and contexts**.
- Techniques for quantitatively **characterizing key aspects of falsified media, such as malicious intent**, in ways that are both computationally accessible and operationally relevant.
## TA1 Responsibilities

<table>
<thead>
<tr>
<th>TA1 provides</th>
<th>to TA2</th>
<th>to TA3</th>
<th>to TA4</th>
<th>to the program</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Input into system API specification/design</td>
<td>• Suggested datasets and evaluation scenarios</td>
<td>Feedback on challenges and Hackathons</td>
<td>• Development &amp; training data gathered from outside the program</td>
<td></td>
</tr>
<tr>
<td>• DAC algorithm containers implementing API and documentation</td>
<td>• Support for designing evaluations</td>
<td>• Participation at hackathons and PI meetings</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• DAC scores and evidence via API</td>
<td></td>
<td>• Develop and provide insight into DAC algorithms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Support for DAC container integration</td>
<td></td>
<td>• DAC algorithm containers implementing API and documentation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• TA1 algorithm insight to support fusion, explanation, and prioritization components</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Explanation & System Integration (TA2)**

Create state of the art approaches to prioritize large volumes of multi-modal assets and explaining multi-modal/multi-asset evidence of manipulation

**Challenges**

- Creating an open, standards-based, multisource, plug-and-play architecture that allows for interoperability and integration
- Developing a single fused score for detection, attribution, and characterization based on scores and evidence for detection, attribution, and characterization provided by TA1s
- Developing a stable prototype system prior to each program evaluation and supporting program demos
Strong TA2 proposals will describe:

- Techniques for **fusing DAC scores** across multiple TA1 performers each with disparate approaches.

- Approaches for **reconciling evidence** across multiple TA1 performers with disparate forms of evidence, and presenting a **unified evidence summary and explanation** to end users.

- Methods for automatically customizing media **prioritization schemes for different users** or different classes of users.

- Technical approaches to **enabling parallel TA1 development and system integration** while simultaneously minimizing dependencies and integration effort.

- A strategy for supporting a **rolling, continuous evaluation process** that leverages the prototype SemaFor system and a continuous integration, continuous deployment process while keeping compute costs in check.

- An approach for proactively **engaging with potential transition customers** to enable early transition of SemaFor capabilities.

- Evidence of previously successful **transition of DARPA capabilities to operational use** in the DoD and/or IC.
## TA2 Responsibilities

<table>
<thead>
<tr>
<th>TA2 provides</th>
<th>to TA1</th>
<th>to TA3</th>
<th>to TA4</th>
<th>to the program</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>system API specifications designed with TA1 input</strong></td>
<td>• System API specifications designed with TA1 input</td>
<td>• Support for designing evaluations</td>
<td>Feedback on challenges and Hackathons</td>
<td>• SemaFor system design and APIs</td>
</tr>
<tr>
<td><strong>Integration of TA1 components into SemaFor system</strong></td>
<td>• Integration of TA1 components into SemaFor system</td>
<td>• Compute for evaluation scoring code</td>
<td>• SemaFor system integration and U/I development</td>
<td>• Provide compute resources for evaluations, hackathons, and demonstrations</td>
</tr>
<tr>
<td><strong>Compute resources for evaluation of TA1 algorithms</strong></td>
<td>• Compute resources for evaluation of TA1 algorithms</td>
<td></td>
<td>• Transition support</td>
<td>• Host leading hackathons</td>
</tr>
<tr>
<td><strong>Design input for score calibration process</strong></td>
<td>• Design input for score calibration process</td>
<td></td>
<td>• Support integration exercises with transition partners</td>
<td>• Participation at hackathons and PI meetings</td>
</tr>
<tr>
<td><strong>Lead design of system APIs</strong></td>
<td>• Lead design of system APIs</td>
<td></td>
<td>• Develop and provide insight into score fusion, explanation, and prioritization algorithms</td>
<td>• SemaFor system demonstrations in each program phase</td>
</tr>
<tr>
<td><strong>Receive, validate, and integrate TA1 components into SemaFor system</strong></td>
<td>• Receive, validate, and integrate TA1 components into SemaFor system</td>
<td></td>
<td>• Develop algorithms to assemble and curate evidence; provide unified evidence summary and explanation</td>
<td>• Facilitate program design discussions</td>
</tr>
</tbody>
</table>

- **Feedback on challenges and Hackathons**
- **SemaFor system demonstrations in each program phase**
- **Develop algorithms to assemble and curate evidence; provide unified evidence summary and explanation**
- **Facilitate program design discussions**
- **Provide a stable prototype system prior to each evaluation**
Create robust evaluations for detection, attribution, and characterization, and for prioritization and explanation

Challenges

• Designing evaluation protocols to explore the range of SemaFor performance, to highlight where human capabilities might be best augmented by automated algorithms

• Designing an evaluation protocol handling the potential combinatorial complexity of evaluating performers on multiple media and falsification types, in cross-modality media groupings of various compositions

• Identifying relevant metrics to support the evaluation goals

• Generating (or collecting) a sufficient number of media assets to support the multi-modal evaluation
Strong TA3 proposals will describe

- A detailed plan for **obtaining and curating data that is sufficient in volume**, highly **relevant** to the problem domain, and can be **released** to the broader research community during the course of the program, including estimates for how many of each asset type will be needed to support evaluations in each phase of evaluation.

- How the evaluation design will **identify, manage, and decouple latent variables** that might be unintentionally correlated across evaluation probes.

- The evaluation team’s approach to having **strong subject matter expertise** in the detection, attribution, characterization, explanation, and prioritization of falsified multi-modal media.

- How the evaluation design and roadmap will provide both a **comprehensive understanding of the program’s scientific progress** and answer key **performance questions for potential transition partners**.

- Strategies for **designing, organizing, and executing complex evaluation** processes across a large distributed team while maintaining performer buy-in and evaluation integrity.

- Approaches for **streamlining the human subjects research** and IRB process related to evaluation.
# TA3 Responsibilities

<table>
<thead>
<tr>
<th>TA3 provides</th>
<th>to TA1</th>
<th>to TA2</th>
<th>to TA4</th>
<th>to the program</th>
</tr>
</thead>
</table>
| • Sample development and evaluation data | • Sample development and evaluation data | • Input to challenges and Hackathons | • Media generation and curation  
| | | | • Facilitate and lead program discussions about evaluation designs (datasets, processes, schedule, metrics, transition partner use cases) |
| | | | • Define and implement metrics  
| | | | • Design and conduct experiments to establish baseline human performance |
| | | | • Evaluation scoring software  
| | | | • Evaluation results analysis  
| | | | • Organize and host PI meeting  
| | | | • Oversight of PI meetings  
| | | | • Conduct evaluations every 8 months |
Multi-modal manipulations collected from the public state-of-the-art

Challenges include:
• Developing state of the art media falsification challenges to support strong SemaFor defenses
• Creating media falsification threat models based on current and anticipated technology

Strong proposals will describe:
• Detailed evidence of the proposer’s ability to bring state-of-the-art falsification challenges in one or more modalities to the program.
• Threat models that provide actionable insights into how DAC algorithms and the SemaFor system should be designed to put significant burdens on potential manipulators.
# TA4 Responsibilities

<table>
<thead>
<tr>
<th>TA4 provides</th>
<th>to TA1</th>
<th>to TA2</th>
<th>to TA3</th>
<th>to the program</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>• Coordination in advance of and during Hackathons to ensure challenge understanding</td>
<td>• Coordination in advance of and during Hackathons to ensure challenge understanding</td>
<td>• Support for incorporating challenge problems into evaluations</td>
<td>• State of the art falsification techniques</td>
</tr>
<tr>
<td></td>
<td>• Coordination in advance of and during Hackathons to ensure challenge understanding</td>
<td>• Support for incorporating challenge problems into evaluations</td>
<td>• Regularly deliver challenges and updated threat models</td>
<td>• Curate SOTA challenges from public domain</td>
</tr>
<tr>
<td></td>
<td>• Support for incorporating challenge problems into evaluations</td>
<td>• Support for incorporating challenge problems into evaluations</td>
<td>• Work with TA3 to curate additional generated or manipulated data for challenge problems</td>
<td>• Develop threat models</td>
</tr>
<tr>
<td></td>
<td>• Work with TA3 to curate additional generated or manipulated data for challenge problems</td>
<td>• Work with TA3 to curate additional generated or manipulated data for challenge problems</td>
<td>• Work with TA3 to evaluate progress on challenge</td>
<td>• Provide insight as to whether/how DAC technologies could be most effective</td>
</tr>
<tr>
<td></td>
<td>• Work with TA3 to evaluate progress on challenge</td>
<td>• Work with TA3 to evaluate progress on challenge</td>
<td>• Work with TA3 to evaluate progress on challenge</td>
<td>• Challenge problem design</td>
</tr>
<tr>
<td></td>
<td>• State of the art falsification techniques</td>
<td>• curate SOTA challenges from public domain</td>
<td>• Regularly deliver challenges and updated threat models</td>
<td>• Lead challenge problem execution at hackathons</td>
</tr>
<tr>
<td></td>
<td>• Develop threat models</td>
<td>• Work with TA3 to curate additional generated or manipulated data for challenge problems</td>
<td>• Work with TA3 to evaluate progress on challenge</td>
<td>• Participation at hackathons and PI meetings</td>
</tr>
</tbody>
</table>
Measuring Progress

**Phase 1 - 18 Months**
1. News Articles
   - Text, image, audio, video
2. Social Media Posts
   - Shorter than a news article

**Phase 2 - 18 Months**
1. Technical Information
   - e.g. a nation state’s missile propaganda
2. News Collections
   - Multiple news articles

**Phase 3 - 12 Months**
1. News Collections
   - Multiple news articles
2. Technical Collections
   - Multiple technical information articles

**Quantitative Assessment**

<table>
<thead>
<tr>
<th>Task</th>
<th>Metrics</th>
<th>Relevant Baselines</th>
<th>Program Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manipulation detection</td>
<td>• Probability of Detection (Pd)</td>
<td>• Human: 60% Pd [Deepfakes]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• False Alarm Rate (FAR)</td>
<td>• Image: 80% Pd at 10% FAR / 20% EER</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Equal Error Rate (EER)</td>
<td>• Text entity recognition: 90% F1-score</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Audio: 4% EER</td>
<td>80% Pd 10% FAR</td>
</tr>
<tr>
<td>Attribution</td>
<td>• Pd / FAR</td>
<td>• Image: 78% Pd at 10% FAR [camera id]</td>
<td>85% Pd 8% FAR 90% Pd 5% FAR</td>
</tr>
<tr>
<td>Prioritization for analyst</td>
<td>• Accuracy over degrees of malice</td>
<td>• Sentiment analysis: 70-80% F1-score</td>
<td>70% accuracy 80% accuracy 85% accuracy</td>
</tr>
</tbody>
</table>

**Increasing task complexity**

**News Collections**
1. News Articles
2. Social Media Posts
3. Technical Information
4. News Collections
5. Technical Collections

**News Articles**
- Text, image, audio, video

**Social Media Posts**
- Shorter than a news article

**Technical Information**
- e.g. a nation state’s missile propaganda

**News Collections**
- Multiple news articles

**Technical Collections**
- Multiple technical information articles
Proposal Information

• Proposers may submit proposals to all TAs.

• Each proposal may only address one TA.

• Separate proposals for each TA are required if proposing to multiple TAs.

• DARPA will not make TA1 and TA2 awards to the same institution.

• TA3 performer may not perform on TA1 or TA2 due to an inherent conflict of interest with the evaluation process.

• TA4 institutions may perform on other parts of the program, but organizational conflicts of interest plans will be needed in the case of TA1 or TA2 due to potential conflicts of interest with the evaluation process.