Fighting Disinformation, Malinformation, and Misinformation in Influence Operations Campaigns

Mark Hoffman
3793 Westwick Ct. NW
Kennesaw, GA 30152
COUNTRY (United States)

mark.hoffman@lmco.com

Tom Damiano
3 Executive Campus, Suite 600
Cherry Hill, NJ 08002
COUNTRY (United States)

thomas.a.damiano@lmco.com

James Starz
4301 N Fairfax Dr, Suite 500
Arlington, VA 22203
COUNTRY (United States)

james.c.starz@lmco.com

ABSTRACT

Peer competition has become an increasing norm for our adversaries in pursuing national security objectives without resorting to major kinetic warfare. A central element of peer competition is the ability to control the messaging or narratives involved in influencing local and/or global populations to align with those objectives.

The Defense Advanced Research Projects Agency (DARPA) has recently begun two major research programs in the Influence Operations arena. The DARPA Semantic Forensics (SemaFor) program is developing and testing a range of technologies to detect, attribute, and characterize falsified, multi-modal media to defend against large-scale, automated disinformation attacks. The DARPA Influence Campaign Awareness and Sensemaking (INCAS) program is likewise developing automated tools to help detect and make sense of adversarial influence campaigns. Lockheed Martin Advanced Technology Laboratories (ATL), as systems integration contractor for both programs, is now working to integrate those efforts and provide a Prototype Influence Operations Pipeline for Experimentation (PIPER) for additional 3rd party integration and interoperability. DARPA also recently sponsored a study to explore the research agenda needed for an "Information Environment Proving Ground" (IEPG) that could produce a "virtual wind-tunnel" for understanding the likely effects of possible Influence Operations campaigns. This paper will describe the current status and plans for Lockheed Martin ATL's efforts across these areas.

1.0 INFLUENCE OPERATIONS CHALLENGES

The US is engaged with its adversaries in an asymmetric, continual war of weaponized influence narratives. Adversaries exploit misinformation and true information delivered via influence messaging: blogs, tweets, and other online multimedia content. Analysts require effective tools for continual sensemaking of the vast, noisy, adaptive information environment to identify adversary influence messaging and broader campaigns.

Today, geopolitical influence campaign detection and sensemaking is largely manual and ad hoc. Analysts use social listening tools to formulate complex keyword queries; track trending keywords, hashtags, and topics; and read hundreds to thousands of documents to identify influence themes. New or "low and slow" campaigns are hard to detect early as their message volume may be beneath platform "trending" thresholds and pertinent hashtags may be unknown. With current tools, it is difficult to connect messages over time and across multiple platforms to track evolving campaigns and to assess confidence in analytic conclusions in a principled manner. Confidence assessment by analysts is ad hoc, manual, subjective, qualitative, and susceptible to analyst cognitive biases (e.g., confirmation bias). Analyst reports often cover static time ranges, and static reports quickly become stale. Today, with current tools, analysts must manually sift through a high volume of messages to find those with relevant influence agendas and then gauge which ones are gaining traction and with whom. Analysts track population response using digital marketing tools for analysing audience demographics, interests, and personality types. These tools lack explanatory and predictive power for deeper issues of geopolitical influence. Audience analysis is often done using static, demographic segmentation based on online and survey data. This methodology lacks the flexibility, resolution, and timeliness needed for dynamic geopolitical influence campaign detection and sensemaking.

DARPA is investing significant research into the exploration of new tools and techniques to detect and understand foreign adversary influence operations being conducted within the social media and open-source information domains. These capabilities are necessary to understand but not sufficient to counter those operations. In addition to the intelligence, surveillance, and reconnaissance (ISR) of these influence operations, DARPA has also begun to explore potential research topics related to the development of counter/mitigation strategies to foreign malign influence. Key to this capability is the need to assess the potential impact of both adversarial influence operations but also potential counter/mitigation strategies that might be employed. However, a full and rapid understanding of the potential impacts on both beliefs and behaviours in a target population is beyond our current capabilities and also requires an aggressive research agenda to explore and pursue.

2.0 SEMANTIC FORENSICS (SEMAFOR)

2.1 Program Objectives and Structure

The DARPA SemaFor Program which began in 2020, is exploring revolutionary ideas that lead to rigorous, practical demonstrations of the ability to detect, attribute, and characterize (D/A/C) falsified multi-modal media assets (MMA) – e.g., newspaper articles and technical documents involving images, text, audio clips, and videos. SemaFor employs ensembles of analytic algorithms to reason about semantic inconsistencies within and among MMA media by *detecting* if the content has been falsified, *attributing* the content manipulation to a potential source, and *characterizing* the detected instances of falsification or manipulation. SemaFor is developing methods that exploit semantic inconsistencies in falsified media to perform these analysis tasks across media modalities. Current Program evaluations target 1000s of MMA assets, with the objective of scaling SemaFor techniques to Internet volumes of media. SemaFor continually develops focused challenge problems, based on observed trends in media manipulation and falsification, to ensure that program analytic capabilities match potential threats. The SemaFor system is expected to operate over increasingly complex media, including reasoning across batches of related content, while improving detection, attribution, and characterization performance over the duration of the effort.

SemaFor performance is being evaluated on collections of media assets gathered from open sources outside the US, as well as MMA assets specifically designed for SemaFor evaluations. Experiments are designed to evaluate how well performer algorithms achieve the three main D/A/C analysis tasks and will ultimately be compared with human performance baselines. The purpose of the evaluations is twofold: first, to establish rigorous scientific protocols for measuring the performance of algorithms that reason about potentially falsified media; and second, to assess the performance of a SemaFor system in realistic, operational environments.

SemaFor research teams focus in four adjacent Technical Areas (TAs): (1) TA1 teams develop analytic algorithms for the detection, attribution, and characterization of MMAs; (2) TA2 integrates these analytics into a scalable microservice-based system that provides fusion, prioritization, and explanation of analytic results in a human machine interface (HMI); (3) TA3 designs and executes the Program evaluations for the purpose of understanding how well SemaFor capabilities meet the needs of potential transition partners and to understand progress against scientific goals; and (4) TA4 develops challenge problems, based on observed trends in media manipulation, which are used to ensure that program capabilities match potential threats.

2.2 Lockheed Martin ATL's Role and Status

Lockheed Martin Advanced Technology Laboratories (ATL), as the sole prime contractor for SemaFor Technical Area 2 (TA2), provides research and development in a wide range of areas including systems prototype development, human readable explanation, hackathon planning and execution, CI/CD and DevOps automation, and transition of capabilities. Through our subcontractors, we also deliver fusion algorithms deployable within the prototype system that collect, organize, and fuse the large number of D/A/C results and employ a Cognitive Systems Engineering-based process to develop human machine interface (HMI) visualization tools that facilitate interaction directly with SemaFor data and analytic workflows.

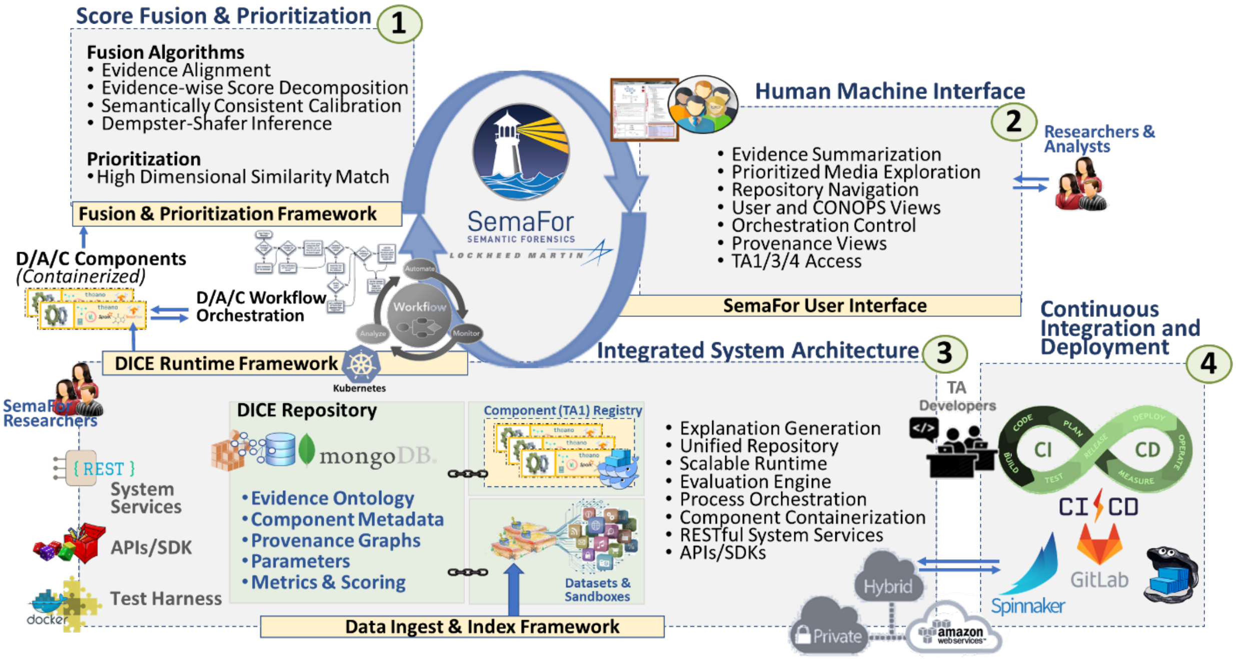


Figure 2-1: SemaFor Architectural Elements.

As illustrated in Figure 2-1, the Lockheed Martin led solution provides (1) a score fusion and prioritization framework that incorporates several algorithmic approaches to fuse and calibrate analytic scores; (2) a rich user interface for evidence summarization and exploration; (3) an integrated system that brings together SemaFor technologies into a transitionable system that provides a unified media and evidence repository, scalable runtime with process orchestration, component containerization, developer APIs/SDKs and RESTful services; and (4) system integration in a continuous integration and deployment (CI/CD) environment. ATL has developed a prototype SemaFor system that manages hundreds of containerized analytic algorithms and deploys these on the scalable runtime framework. Our system container runtime framework extends Kubernetes to control deployment and lifecycle of all SemaFor components including TA1 developed D/A/C analysis algorithms, fusion and prioritization algorithms, and other core system services. All SemaFor capabilities are packaged as Docker containerized components that operate as microservices integrated with the SemaFor analytic workflows. Containerization provides maximal implementation flexibility, while our injected service layer microservices reduce integration overhead and management of component lifecycles and communication.

The SemaFor system employs control and analysis-probe interfaces, managed by an injected service layer, to create individual microservices from each D/A/C algorithm. The combination of containerized componentization and microservice architecture allows SemaFor to introduce new analysis capabilities and introduce those capabilities into analytic workflows seamlessly. ATL provides SDK tools and APIs to allow creation of specification-driven analytic components that can be registered with the system and deployed for D/A/C analysis tasks. All analytics are registered with metadata that express external configuration information (e.g., weights, profiles, etc.) required for execution. Analysis workflow specifications that define multi-phase analytic workflows to be defined that assemble analysis algorithms into detection, attribution, and characterization workflows. The system also provides a gateway that allows *existing* DARPA MediFor analytics to be deployed and reused in SemaFor workflows. Runtime transaction management coordinates the launching of analysis components and uses dynamic horizontal auto-scaling to allocate the number of instances required to accommodate the current analysis load. During runtime, storage volumes are mounted to analytic containers to provide ephemeral transaction workspaces where intermediate artifacts can be staged for transfer to the system data lake. The SemaFor integrated repository hosts a raw media storage data lake, workflows specifications, configuration metadata, analytic evidence, and container image registry. Access to the repository is managed via the application server, which supports RESTful dataset endpoints, evaluation endpoints, and analysis endpoints that provide the ability to launch workflows and receive status.

Data is loaded into the SemaFor system repository using a plugin-based data ingestion framework that is designed to support indexing (geo-spatiotemporal, metadata, custom), metadata and feature extraction, and format processing of incoming multimedia assets. The data generated during ingestion processing provides a rich set of information for use in analytic workflows and for HMI users. The ingestion framework processes input into internal graph and asset object models that capture the complex structure of the multi-modal asset. The current data ingestion framework includes gateways that interact with external data sources to support on-demand data ingestion. Integration with new data sources is largely driven by Transition partner use cases. SemaFor analysis operates over ingested structures to produce evidence graph output that logically captures scope of analysis performed, hypotheses-based consistency checks explored, evidence supporting conclusions from the D/A/C analysis, log likelihood ratio scores indicating the relative strength of evidence, and evidence localization information. Additional intermediate artifacts may be generated by the analysis – e.g., heatmaps, image chips, bounding polygons, etc., which are also captured and associated with the evidence graph to provide a rich set of outputs used by fusion, explanation, and the HMI. HMI development focuses on two primary use cases: *data-centric* focus targets data views of interest to SemaFor scientists and engineers, while *analyst-centered* user interfaces focus on analyst workflows that includes interest profile-driven prioritization to triage and display collected evidence for review and assessment. To provide a progressive disclosure of analysis results organized around hypotheses and evidence, the explanation algorithms automatically assemble and curate evidence to produce a tree structure where each layer in the tree represents an increasing level of detail including summary, fused, pre-fusion, and detailed information from each analytic algorithm. In addition, prioritization algorithms match user interest against multimedia asset metadata and D/A/C analysis results to provide contextually prioritized lists of MMAs based on user-specified criteria.

ATL has implemented the SemaFor CI/CD DevOps process that automates deployment of D/A/C analytics components for integration testing and evaluation. SemaFor includes the ability to leverage the CI/CD pipeline and evaluation workflows to support continual, rolling TA3 evaluations. Components developed by TA1 are automatically moved through gate testing and deployed into the evaluation environment where they run in competitions appropriate to their capabilities. Execution of Program evaluations occur *within* the SemaFor system and leverage the TA3 scoring engine which is deployed as a service and used in evaluation workflows. Evaluations are led by TA3 and organized into competitions that target various Program objectives and areas of interest including, for example, the ability to identify tactics (e.g., hate speech, scapegoating, etc.), intents (e.g., call-to-action, discredit, etc.), synthetic media tools and techniques (GROVER, StyleGAN, Latent-diffusion, etc.), person-of-interest deepfakes, text and image inconsistencies, and localization of falsified or manipulated semantic entities (e.g., symbols, signs, firearms, etc.).

SemaFor research is currently (August 2022) in mid-Phase 2 of the Program and is executing the third formal Evaluation of capabilities against Program metrics and goals. ATL and other Performers are actively engaged with Transition partners to explore SemaFor capability transition opportunities and to refine use cases to inform research roadmaps. There are currently tens of SemaFor system instances running within our Program GPU computing cluster to support concurrent activities including research and development, evaluation, demonstration, and transition partner experimentation. Multiple deployments of the SemaFor test harness and selected D/A/C analytic algorithms have been deployed within Transition partner environments, where they are used for experimentation with local datasets. A browser-based drag-and-drop Portal for experimenting with SemaFor algorithms has also been established to support early Transition partner exploration of capabilities. An analytic component catalog with hundreds of D/A/C algorithms deployable within the SemaFor system is available through registered users on our Program collaboration site.

3.0 INFLUENCE CAMPAIGN AWARENESS AND SENSEMAKING (INCAS)

3.1 Program Objectives and Structure

The tools being developed under the INCAS program, beginning in 2021, will enable analyst-guided influence campaign analysis using automated influence detection. In contrast to current social media tools, INCAS tools will directly and automatically detect implicit and explicit indicators of geopolitical influence in multilingual online messaging to include author’s agenda, concerns, and emotion (ACE).

To explain and anticipate population response to influence messaging, INCAS tools will dynamically segment the responding population and identify psychographic attributes relevant to geopolitical influence. Psychographic attributes, such as worldviews, morals, and sacred values, are hypothesized to correlate more strongly with geopolitical influence response than the personality and demographic attributes used for marketing. A person's worldview is the way they see and understand the world, especially regarding issues such as politics, philosophy, and religion. Worldviews can include systems of moral and sacred values. Psychographic attributes will be extracted using analysis of text and online behaviour, and attributes will be correlated with influence indicators in messaging to which the population segment is responsive.

INCAS tools will support analyst-guided linking of influence indicators and population response over time and across multiple platforms to capture dynamic, evolving campaign models. Campaign models will combine machine-surfaced influence indicators and messaging and population response with analyst-provided campaign elements, including campaign tactics, objectives, actors, and events. Quantified confidence assessment will enable analysts to mitigate cognitive biases through INCAS automation that curates, elicits, combines, and organizes confidence intervals, evidence, alternative hypotheses, and supporting information.

Program performers work in five different Task Areas (TAs)

- TA1 performers develop techniques to identify influence indicators such as agendas, concerns, and emotions in online messaging.
- TA2 performers are developing techniques to segment the responding population to a set of influence messages, characterize each segment using psychographic and demographic attributes, and identify correlations among these attributes, influence indicators, and response.
- TA3 is developing techniques for analyst-machine sensemaking of influence campaigns including aiding analysts in assessing confidence in hypothesized campaign models.
- TA4 is developing the infrastructure to provide social media messaging and other data feeds from online sources to all TAs. TA4 is collecting and persisting social media and other online data as well as implementing low-level supporting data analytics.

- TA5 is designing and conducting technology evaluations (including metrics and scenario definition), developing ground truth evaluation data for program scenarios and managing a Program Subject Matter Experts (SMEs) group.

These technical areas and their interactions are characterized in figure 3-1.

Figure 3-1: INCAS Component Flow.

3.2 Lockheed Martin ATL's Role and Status

The Lockheed Martin ATL team is responsible for provisioning multimedia data enabling tool development and evaluation. The data provided is openly available (non-US focused) social media data, along with forums and news stories collected based on a topic of interest. The data spans languages and media types (text, images, video). The raw data is ingested and normalized to provide consistent and clean data to other INCAS components. The datasets include standard metadata along with enrichments such as extraction of key entities mentioned. Additional contextual information is collected and provided about key people and groups in the data collection. Our initial experimentation data set consisted of over five million data products collected over a two-month historical period. These data products resulted in the creation of nearly a hundred million agenda, concern, and emotion annotations.

The other critical role we play is developing a toolkit that architects the technical components together, as illustrated in Figure 3-2. All system components must accept a common data model for incoming message data as well as adhering to specific API calls for their particular function. Each technical area component is provided as one or more containerized solutions supporting required API services. As data is ingested, it is federated to the technical components for additional enrichments. All of these enrichments are housed in a data store along with the original data. The toolkit is resilient to removing or adding additional components as might be required for a given deployment. Data enrichments are performed asynchronously, but the general flow is to provision raw data, determine influence indicators, use that to process population response characterization, and aggregate the findings to be shown to users in an HMI focusing on the information operations campaign perspective. The entire software suite runs in a commercial cloud environment for experimentation.

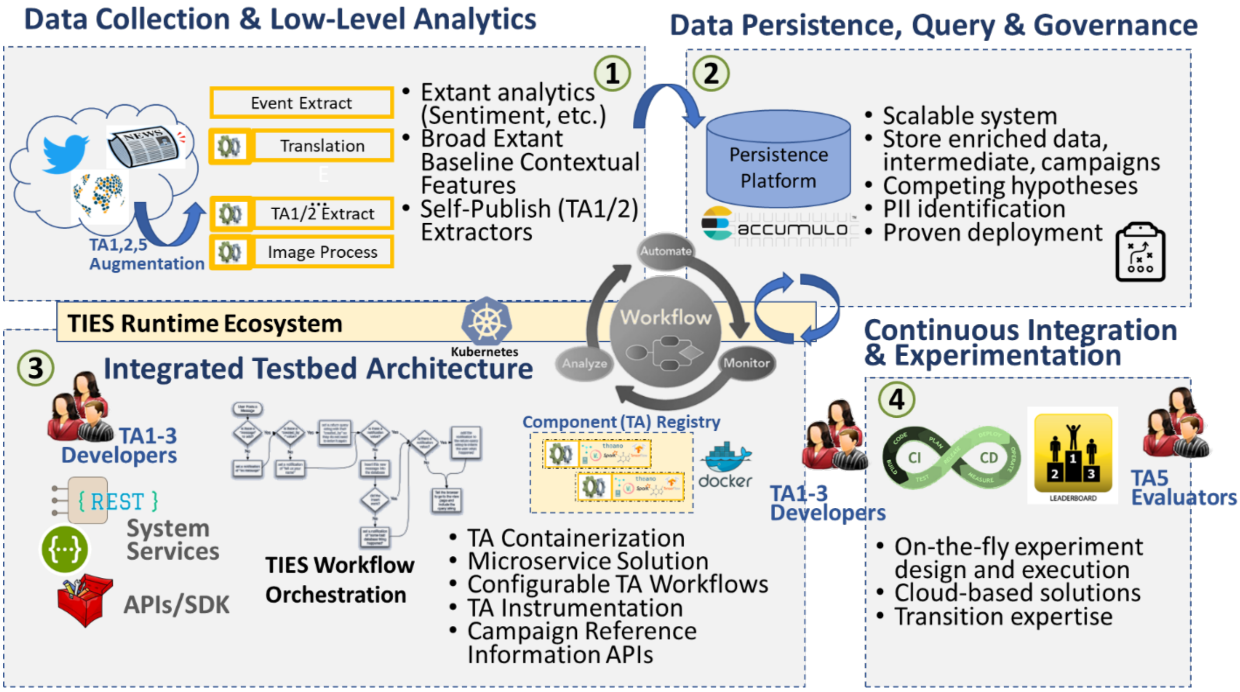


Figure 3-2: INCAS Architectural Elements.

4.0 PROTOTYPE IO PIPELINE FOR EXPERIMENTATION (PIPER)

4.1 Program Objectives and Structure

While SemaFor and INCAS programs are focused on deepfake detection, attribution, and characterization and then looking at identifying and characterizing potential influence campaigns, there was no initial plan to link these programs together – although deepfakes are obviously one potential element of influence in and across social media. The opportunity to experiment with these interactions has been made possible through an extension being made to the INCAS program. In addition, this integration will be used as a testcase for the potential development of an open and more holistic pipeline for IO data, artifacts, and analytics to be accessed and shared across 3rd party applications through a common externally facing application programming interface (API). The initial linkage of these two programs, using PIPER, will provide a more complete view of potential multi-channel, multi-modal influence operations to the benefit of both of the individual programs as well as potential external consumers of the analytics and artifacts available under an expanding collection to tools and through a single integrated gateway. Should these experiments prove both successful and operationally valuable, a more substantial effort might be mounted toward the development of such an open architecture pipeline for influence operations detection and characterization.

4.2 Lockheed Martin ATL's Role and Status

While this program remains in contract negotiation as of the time of this writing, this effort will include 4 primary tasks:

- Task 1: Develop an externally facing API through which 3rd party tools in operational environments could access ingested and enriched data from SemaFor, INCAS, and future analytics. Deploy a containerized gateway component as a front-end to the IO pipeline.

- Task 2: Modify the SemaFor and INCAS architectures so that SemaFor and INCAS can communicate through the IO Pipeline gateway. Leverage existing (e.g., GAN detectors) and new analysis capabilities (social media profile analysis). Scale for the additional analytics and users.

- Task 3: Enhance existing HMIs to include visualizations to support the representation and understanding of the evidence graphs provided by SemaFor.
- Task 4: Identify unclassified scenarios and datasets that might be leveraged by the INCAS and SemaFor programs for transition partner engagements.

The overarching objective of this effort is to demonstrate the ability for modular arms-length integration of data and analytics through a general workflow process to allow various Influence Operations data and analytics applications to more easily share data and analysis between them for increased and enhanced application interaction. A conceptual view of the PIPER architecture is provided in Figure 4-1. While this platform will be designed to primarily support the research and development community, interactions with operational stakeholders will be conducted to help assure that the capabilities development for R&D purposes might be more easily and effectively transitioned for future operational leverage.

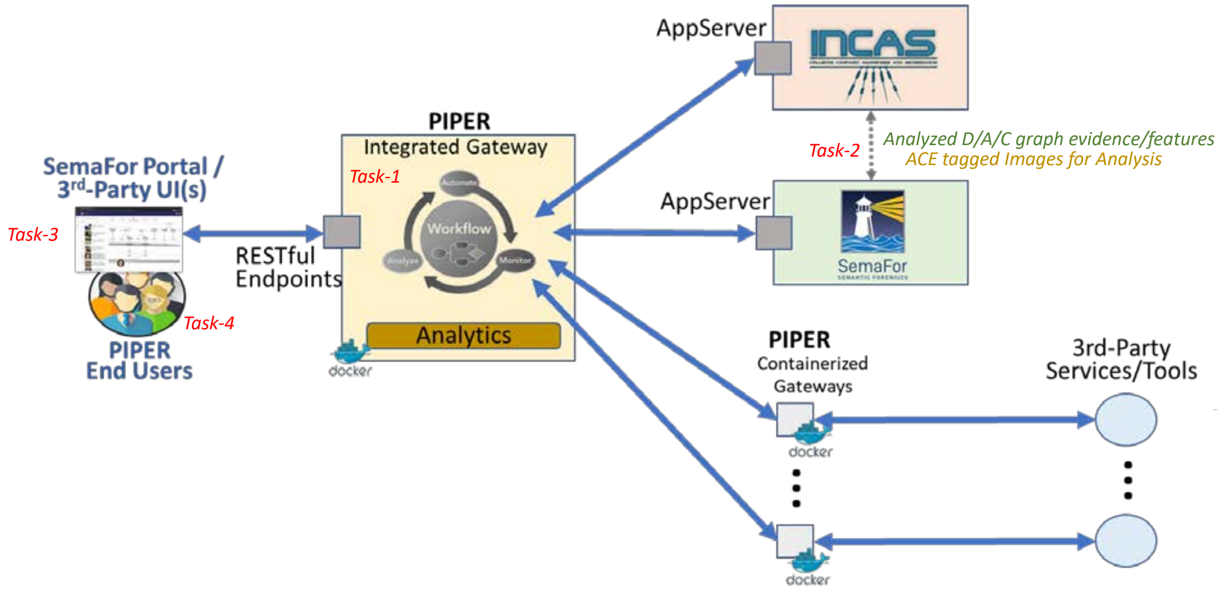


Figure 4-1: PIPER Concept.

5.0 INFORMATON ENVIRONMENT PROVING GROUND (IEPG)

5.1 Study Objectives and Structure

While the two major SemaFor and INCAS programs make significant inroads into the detection and characterization of influence operations, neither effort addresses the problem of potentially countering influence operations being conducted by our adversaries. In order to more deeply assess the impacts of a foreign influence operation or assess the potential value of possible mitigation strategies, the need for a better simulation of foreign influence operations in a target environment and context would be needed.

In Early 2021, LM ATL began discussions with DARPA regarding the need for the equivalent of an aircraft design “wind tunnel” but for Influence Campaigns. Specifically, the need to quickly and economically verify a concept design without the need, expense, or impacts of full-scale development and deployment. Through discussions with DARPA, the concept was refined and an Information Environment

Proving Ground (IEPG) study concept came into focus. The desire was to bring together a broad range of study participants (over 30 participants from industry, academia, and government) with a broad range of technical, operational, and social science experience to explore the possibilities and challenges.

From April through December of 2021, LM ATL led a study for DARPA to explore the possibility of developing an Information Environment Proving Ground (IEPG) as a means to provide capabilities in 3 general areas:

1. Test existing and emerging DoD and commercial tools and capabilities for identifying and understanding adversarial influence campaigns;
2. Test out our proposed counter-influence operations on achieving specific mitigation effects on targeted foreign populations; and
3. Training (both our IE professionals as well as the general public) in understanding and/or shaping Influence Operations.

5.2 Study Findings

During the execution of a series of workshops, it became clear that a key constraint on the development of an IEPG focused on identifying the user communities and representative use cases to be addressed so that valid technical "challenges" for those use-cases could be identified to drive research agendas for exploration. A candidate set of Users and Use-Cases are identified below.

1. IEPG for <u>measuring Blue/Red effects of actions </u>(also tactics, etc.) in an operational environment with respect to a target population and context. ***Users:*** This would involve social scientists and operational users (apart from humans in the testbed itself described below).

2. IEPG for <u>measuring longitudinal effects over time of specific Blue/Red influence campaigns</u> (COAs, etc. - could be red, blue). ***Users:*** This would mainly be operators as the end users experimenting with candidate blue campaigns and exploring hypothesized red campaigns and/or their combined interactions.

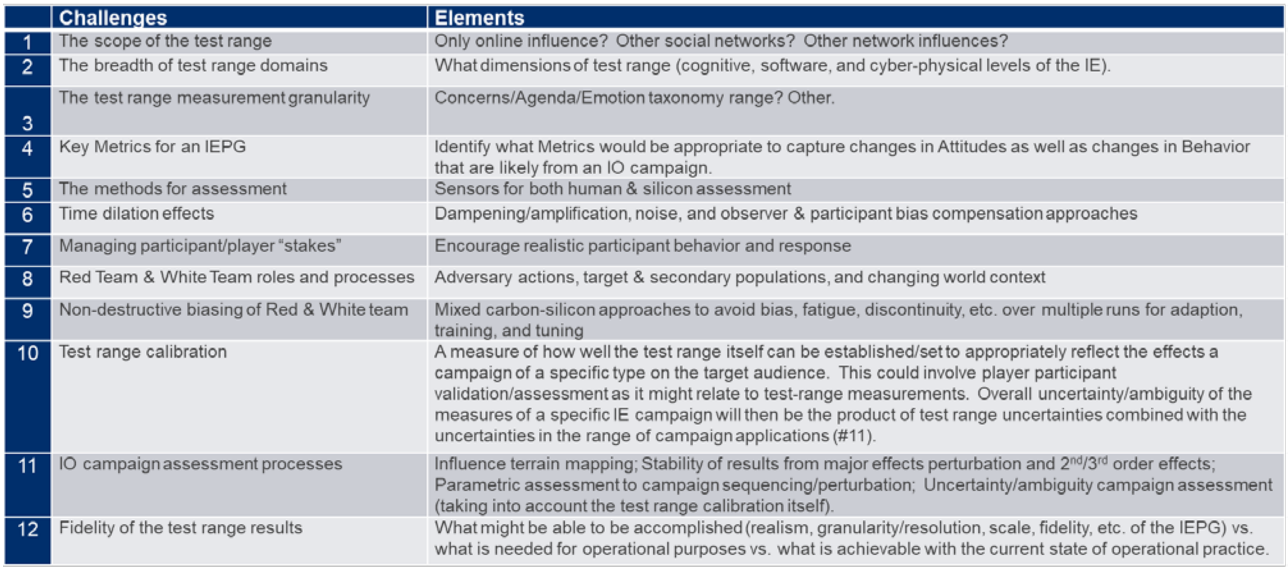
3. IEPG for <u>measuring effectiveness of specific technology or processes/methodologies</u> (i.e., inoculation strategies) to detect, mitigate, deflect, etc. foreign influence campaigns. ***Users:*** This would mainly be technologists as end users as well as operational users.

4. IEPG <u>as a technology testbed for measuring effectiveness of generation/mitigation generation tools</u>. ***Users:*** Technologists developing applications and decision aids designed to produce effective counter-influence campaigns for a target foreign population and influence effect.

5. IEPG <u>as a training environment</u>. ***Users:*** Operational staff involved in the development and application of influence operations to understand the anticipated 1st, 2nd, and 3rd order effects of campaigns and mitigation strategies that they might develop.

Next, a candidate list of potential challenges, illustrated in Table 5-1, was developed related to one or more of the specific use-cases.

Table 5-1: IEPG Challenges.



	Challenges	Elements
1	The scope of the test range	Only online influence? Other social networks? Other network influences?
2	The breadth of test range domains	What dimensions of test range (cognitive, software, and cyber-physical levels of the IE).
3	The test range measurement granularity	Concerns/Agenda/Emotion taxonomy range? Other.
4	Key Metrics for an IEPG	Identify what Metrics would be appropriate to capture changes in Attitudes as well as changes in Behavior that are likely from an IO campaign.
5	The methods for assessment	Sensors for both human & silicon assessment
6	Time dilation effects	Dampening/amplification, noise, and observer & participant bias compensation approaches
7	Managing participant/player "stakes"	Encourage realistic participant behavior and response
8	Red Team & White Team roles and processes	Adversary actions, target & secondary populations, and changing world context
9	Non-destructive biasing of Red & White team	Mixed carbon-silicon approaches to avoid bias, fatigue, discontinuity, etc. over multiple runs for adaption, training, and tuning
10	Test range calibration	A measure of how well the test range itself can be established/set to appropriately reflect the effects a campaign of a specific type on the target audience. This could involve player participant validation/assessment as it might relate to test-range measurements. Overall uncertainty/ambiguity of the measures of a specific IE campaign will then be the product of test range uncertainties combined with the uncertainties in the range of campaign applications (#11).
11	IO campaign assessment processes	Influence terrain mapping; Stability of results from major effects perturbation and 2nd/3rd order effects; Parametric assessment to campaign sequencing/perturbation; Uncertainty/ambiguity campaign assessment (taking into account the test range calibration itself).
12	Fidelity of the test range results	What might be able to be accomplished (realism, granularity/resolution, scale, fidelity, etc. of the IEPG) vs. what is needed for operational purposes vs. what is achievable with the current state of operational practice.

In addition, during the natural discussions of the various use cases and challenges, 3 "Nuggets" seemed to surface as 3 possible key areas of exploration that might enable an IEPG to support a range of possible user communities. Those three Nuggets were:

1) An <u>Influence Sandbox</u>: The idea of a test range to support the assessment of tactical (short time horizon) messaging for a specific target foreign population in the current world context.
2) <u>Participatory Simulation</u>: The idea of a test range that combines a small number of culturally and contextually sensitive population representatives combined with silicon-based simulation agents (bot armies) to produce realistic message responses for that population.
3) <u>Influence in Large Scale Games</u>: The idea of exploring the potential of leveraging existing large-scale games, flexible game constructs, and gaming communities to help provide insights into key influence questions.

The full study from this effort is unclassified and may be available upon request.

6.0 CONCLUSION

This paper illustrates a number of research activities that have been, are currently being, and are planned to be underway by Lockheed Martin in the Influence Operations identification and understanding and funded by DARPA. In addition to these activities, Lockheed Martin has made significant investments in additional Internal Research and Development (IRAD) activities across the corporation that relate to Influence Operations identification, attribution, and understanding.