AFRL-RH-WP-TR-2022-0112

NATURAL LANGUAGE ENGAGEMENT OF MALICIOUS ENTITIES THROUGH A SOCIAL INTERACTION SERVICE (NEMESIS)

Phillip Porras
SRI International
333 Ravenswood Avenue
Menlo Park, CA 94205

DECEMBER 2022

FINAL REPORT

AIR FORCE RESEARCH LABORATORY
711th HUMAN PERFORMANCE WING
AIRMAN SYSTEMS DIRECTORATE
WARFIGHTER INTERACTIONS AND READINESS DIVISION
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the AFRL Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2022-0112 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

ANDERSON.TIMOTHY.RAY.1230210728

Digitally signed by ANDERSON.TIMOTHY.RAY.1230210728
Date: 2023.05.24 14:20:38 -04'00'

TIMOTHY R. ANDERSON, DR-IV, Ph.D.
Work Unit Manager
Mission Analytics Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

MURDOCK.WILLIAM.P.1048742161

Digitally signed by MURDOCK.WILLIAM.P.1048742161
Date: 2023.05.24 15:02:17 -04'00'

WILLIAM P. MURDOCK, DR-IV, Ph.D.
Chief, Mission Analytics Branch
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

CARTER.LOUISE.ANN.1230249128

Digitally signed by CARTER.LOUISE.ANN.1230249128
Date: 2023.05.26 13:06:50 -04'00'

LOUISE A. CARTER, DR-IV, Ph.D.
Chief, Warfighter Interactions and Readiness Division
Airman Systems Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
December 2022	Final	30 August 2018	08 December 2022

4. TITLE AND SUBTITLE

Natural Language Engagement of Malicious Entities through a Social Interaction Service (NEMESIS)

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8650-18-C-7880		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		H0X7

6. AUTHOR(S)

Phillip Porras

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
SRI International 333 Ravenswood Avenue Menlo Park, CA 94205	

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'SACRONYM(S)	11. SPONSOR/MONITOR'SREPORT NUMBER(S)
Air Force Research Laboratory 711th Human Performance Wing Airman Systems Directorate Warfighter Interactions and Readiness Division Wright-Patterson Air Force Base, OH 45433		AFRL-RH-WP-TR-2022-0112

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

AFRl-2023-1999; Cleared 22 May 2023

14. ABSTRACT

Project NEMESIS developed active engagement services designed to defend enterprise networks from sophisticated social engineering attack campaigns by adversaries that possess the resources to study potential victims, the skills to exploit this knowledge during dialog engagements with their victims, and the patience to defer a victim's compromise for maximum eventual success and impact. Our team was directed by DARPA to spearhead the development of an Ensemble Dialog Management (EDM) system, capable of integrating multiple dialog generation strategies for the purpose of integration into a live defensive service. Our system demonstrated a 52% Rate of Engagement (RoE) against in the 4th release of this system and 93% in the 5th. In an experiment designed by Zetier and ARLIS to examine whether key administrators throughout SRI accurately follow corporate policies regarding the disclosure of information about previous employees to external third parties, we achieved a 32% RoE, comparable to what can be achieved by human testers, but scalable to much larger populations.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	32

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Timothy R. Anderson, Ph.D.	

TABLE OFCONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 EXECUTIVE SUMMARY

Project Natural Language Engagement of Malicious Entities through a Social Interaction Service (NEMESIS) is a project focused on the development of active engagement services designed to defend enterprise networks from sophisticated social engineering attack campaigns. NEMESIS has been led by SRI International. During Active Social Engineering Defense (ASED) Phase 1, the project included subcontractors University of Illinois at Chicago (UCI), Stanford University, Jataware, and Hyperion-Gray, who participated in Task Areas (TA) 1 and 2.

NEMESIS has developed defenses against social engineering (SE) campaigns by adversaries that possess the resources to study potential victims, the skills to exploit this knowledge during dialog engagements with their victims, and the patience to defer a victim's compromise for maximum eventual success and impact. We refer to such SE adversary scenarios as confidence campaigns. The depth of knowledge that confidence campaign agents gain about victims both renders victims particularly vulnerable and increases the agents' sensitivity to anomalous behavior that might suggest exposure to a countermeasure. Furthermore, the willingness of these agents to affect a compromise over multiple turns greatly weakens or invalidates detection methods that only consider single point-to-point interactions. To combat these attacks, NEMESIS introduces a fully automated counter-phishing framework that enables ASED's TA2 dialog engines to intercede in TA1-detected phishing attacks.

NEMESIS includes an attribution framework that instantiates resources and mechanisms that compromise the privacy of the adversary when accessed. We have integrated these resources into deliverable resources that can be integrated into dialog exchange sent by ASED bots to spearphishers. Attacker attribution is a complex objective that mandates a multi-faceted approach to extracting and characterizing adversaries, such that they can be held accountable or (if not) recognized and filtered upon the next attack. To achieve this objective, we integrated a range of attribution resources, which when accessed by an adversary, can produce features that will contribute to an adversary attribution dossier.

In Phase II of the ASED Program, our team was directed by Defense Advanced Research Projects Agency (DARPA) to spearhead the development of an *Ensemble Dialog Management* (EDM) system, capable of integrating multiple dialog generation strategies for the purpose of integration into a live defensive service. The EDM is designed to receive input from the phish detection system and leverage the various dialog engine strategies towards several complementary metrics: 1) adversary work factor, 2) knowledge acquisition, and 3) dialog plausibility. Adversary work factor measures the ability of the dialog system to combat advanced SE attacks by overwhelming the adversary's engagement cost. Knowledge acquisition measures the ability of the dialog system to produce meaningful information extraction to facilitate adversary threat attribution. Dialog plausibility measures the ability of the dialog system to exhibit human-plausible dialog interaction in a hostile and suspicious adversarial domain. The main research objective of the EDM has been to design ensemble strategies to blend complementary dialog algorithms to maximize their effectiveness against adversaries. Finally, in the last project year we were asked by DARPA to invert the dialog engine to play the role of the phisher and to design a framework for conducting digital fraud training services. This effort culminated with a social engineering attack (SEA) training exercise hosted at SRI.

2.0 OUTLINE

The report is organized as follows.

Section 3.0 describes the objectives and high-level approach of NEMESIS. The main technologies developed under the project are described in Section 4.0. Section 5.0 summarizes the measurement and evaluation activities performed during this project, including live-fire counter-phishing experiments, participation in the DARPA Evaluations, and the SRI hosted SEA Training exercise to test the ability of staff to correctly respond to social engineering attacks. Section 6.0 summarizes the project deliverables, including source code, demonstration videos, papers, and patents. Section 7.0 identifies key personnel of the project during Phase 1 and Phase 2.

3.0 PROJECT DESCRIPTION

3.1 Research Objectives

The primary research question that has driven the NEMESIS project is whether "*Counterphish dialog engines can advance the state of SEA defense."*

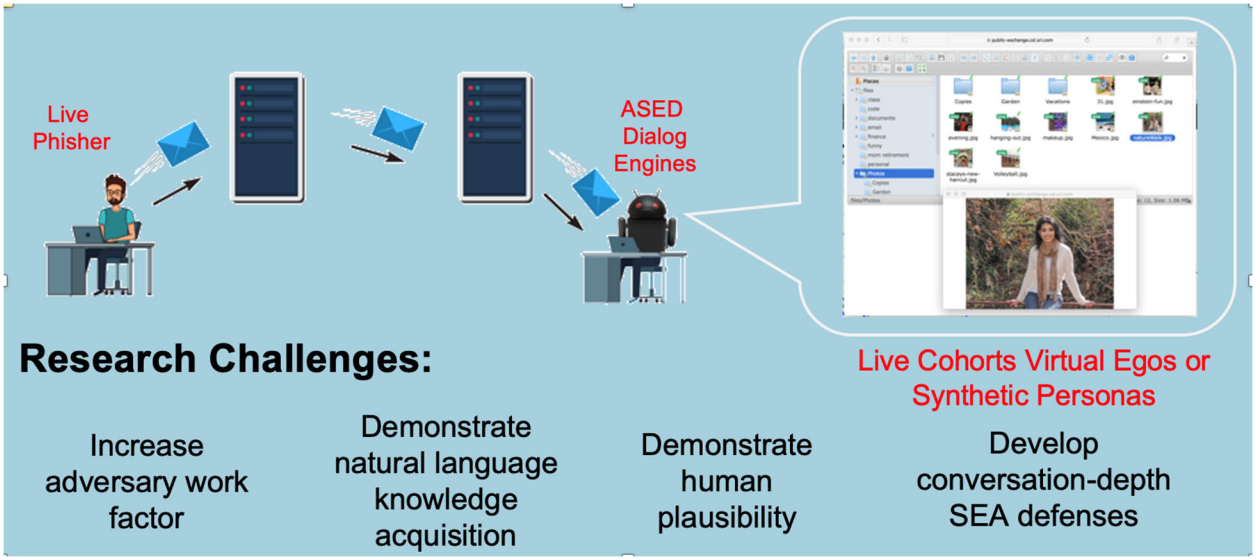

Figure 1. Underlying ASED Project Research Challenged Associated with the NEMESIS Project.

NEMESIS has developed defenses against sophisticated social engineering campaigns by adversaries who possess the resources to study potential victims, the skills to exploit this knowledge during dialog engagements with their victims, and the patience to defer a victim's compromise for maximum eventual success and impact. The primary research challenges of the NEMESIS project are illustrated in **Figure 1.** Our approach has been based on the objective of large-scale multi-round engagement with phishing adversaries using a fully automated conversation management framework with dialog engines that incorporate sophisticated counter-phishing models. Under NEMESIS we succeeded in demonstration that fully automated and scalable *Phishing Conversation Harvesting* can enable deeper insights into the techniques, scam objectives, and attacker attribution information, by eliciting

- Phisher-specific meta-data and identifiable attributes that are revealed only when the victim is engaged
- detectable message patterns that are revealed through the correlation of many phishing conversations within a phishing campaign
- conversation dynamics of an ongoing threads to determine if it matches a known phishing playbook

3.2 Major Findings

Our ASED TA2 Live-Fire Counter-Phishing experiments have proven that our Ensemble Dialog Framework is sounds and robust for live enterprise engagements with real corporate phishers. In fact, our system demonstrated a sustained 60% Rate of Engagement against in the fourth and final major release of this system. Detailed metrics are discussed in Section 5.0.

Phishers reliably pivot to virtual personas that respond to their phishing attacks, and they access our attribution resources (*i.e., **NEMESIS can phish the phishers, at scale***).

We can design **novel SE intervention systems** that recognize

- phish detection signatures derived from elicited phisher attributes
- message patterns that can distinguish individual phishers
- abstract conversation models that can lead to reverse engineering SE attack playbooks

Ongoing interactions with our target **potential transition partners** have confirmed the value of developing novel ***Conversation-Depth SEA Recognition Systems.*** See Section 6.0.

3.3 Project Summary

3.3.1 The Problem

Sophisticated nation-state SE campaigns are waged against organizations rather than individuals. Adversaries leverage victim information exposed publicly (e.g., through social media accounts). They engage in trust-building dialogs, such as for elicitation campaigns or fraudulent action requests to conduct elicitation campaigns, disinformation, and manipulation of the victim. They approach an organization through multiple contact points but may reveal similar campaign playbooks that can reveal their presence and objectives.

3.3.2 Research Goals

To address these threats, we have developed defenses against sophisticated SE campaigns by adversaries that possess the resources to study potential victims, the skills to exploit this knowledge during dialog engagements with their victims, and the patience to defer a victim's compromise for maximum eventual success and impact. We refer to such SE adversary scenarios as confidence campaigns. The depth of knowledge that confidence campaign agents gain about victims renders victims particularly vulnerable and increases the agents' sensitivity to anomalous behavior that might suggest exposure to a countermeasure. Furthermore, the willingness of these agents to affect a compromise over multiple turns greatly weakens or invalidates detection methods that only consider single point-to-point interactions.

A second key aspect of NEMESIS is that of building an attribution framework that instantiates resources and mechanisms that compromise the privacy of the adversary when accessed. We have integrated these resources into deliverable resources that can be integrated into dialog exchange sent by ASED bots to spearphishers. Attacker attribution is a complex objective that mandates a multi-faceted approach to extracting and characterizing adversaries, such that they can be held accountable or (if not) recognized and filtered upon the next attack. To achieve this objective, Nemesis has integrated a range of attribution resources, which when accessed by an adversary, can produce features that will contribute to an adversary attribution dossier.

Finally, we collaborated extensively with other ASED project teams in the construction of dialog and attribution services shared among researchers in the ASED program, enabling them to use this common framework to integrate their detection services, attribution resources, and dialog management systems.

3.3.3 Salient Features and Capabilities

NEMESIS is implemented as an integrated enterprise-scale communication proxy that dynamically pivots from passive monitoring to active engagements with suspected adversaries. SRI and Jataware led the development of a system framework that integrates all elements of our detection, dialog engagement, and attribution services. Our team has demonstrated the creation and management of a multi-virtual-persona social media interaction system, which provides a strong foundation for our understanding of how to construct the key elements of Nemesis' virtual persona management.

The project research goals have been as follows:

- *User Modeling Technologies*: In Phase I, under TA1, we prototyped three parallel user modeling services. When presented with a target user account, user communication logs, and social media data, these services produce a user behavior and social interaction model, perform a trust modeling assessment between the user and correspondents, and performed covert community detection.
- *Adversary Detection Technologies*: In Phase 1, under TA1, we prototyped per-message spearphishing analytics, deep learning dialog analytics. Our TA1 components were measured at each ASED program TA1 evaluation interval, and this was followed by a large-scale collaborative study with Barracuda Networks, Inc.
- *The Nemesis Communication-Application Gateway*: Under TA2, we worked within the ASED program to collaborate in the creation of an enterprise proxy server to mediate all messages through the user modeling and adversary detection systems.
- *Attribution Platform*: Under TA2, our dialog management framework was integrated with a set of services that manage synthetic social media accounts, produce objects and web content with privacy-compromising attribution features, spawn network services for use in adversary host, device, and network-identity fingerprinting (with IPv4 and IPv6 features).

4.0 TECHNOLOGIES

This section presents the major subsystems developed by the NEMESIS Team. We believe that to develop a system capable of defending enterprise users from the ever-evolving sophistication of nation-state SE threats, one has to create a defense that directly models the complexities of an attacked user's social interactions. To solve this challenge, we structure our solution to address the following central technical challenges: adversary detection, adversary attribution, and dialog systems that can interact with phishers at scale.

4.1 Privacy Threats of Browser Extension Fingerprinting

With users becoming increasingly privacy-aware and browser vendors incorporating anti-tracking mechanisms, browser fingerprinting has garnered significant attention. Accordingly, prior work has proposed techniques for identifying browser extensions and using them as part of a device's fingerprint. While previous studies have demonstrated how extensions can be detected through their web accessible resources, there exists a significant gap regarding techniques that indirectly detect extensions through behavioral artifacts. In fact, no prior study has demonstrated that this can be done in an automated fashion. UIC has bridged this gap by presenting the first fully automated creation and detection of behavior-based extension fingerprints. They introduce two novel fingerprinting techniques that monitor extensions' communication patterns, namely outgoing Hypertext Transfer Protocol (HTTP) requests and intra-browser message exchanges. These techniques comprise the core of *Carnus*, a modular system for the static and dynamic analysis of extensions, which they use to create the largest set of extension fingerprints to date. They leverage their dataset of 29,428 detectable extensions to conduct a comprehensive investigation of extension fingerprinting in realistic settings and demonstrate the practicality of their attack. In-depth analysis confirms the robustness of the techniques, as 83.6% - 87.92% of their behavior-based fingerprints remain effective against a state-of-the-art countermeasure. They then explore the true extent of the privacy threat that extension fingerprinting poses to users and present a novel study on the feasibility of inference attacks that reveal private and sensitive user information based on the functionality and nature of their extensions. They first collect over 1.44 million public user reviews of their detectable extensions, which provide a unique macroscopic view of the browser extension ecosystem and enable a more precise evaluation of the discriminatory power of extensions as well as a new deanonymization vector. They automatically categorize extensions based on the developers' descriptions and identify those that can lead to the inference of personal data (religion, medical issues, etc.). Overall, this research sheds light on previously unexplored dimensions of the privacy threats of extension fingerprinting and highlights the need for more effective countermeasures that can prevent their attacks.

4.1.1 Privacy of Application of Interest to DARPA

UIC also studied the privacy features of a target application specified by DARPA. Specifically, they conducted an exploration of ways that could potentially enable fingerprinting users that employ the target application. To that end, UIC ran a large number of experiments, with different implementations and across different operating systems (OS) and devices, in an effort to explore various attributes and Application Programming Interfaces (API) that can be used for generating a part of a fingerprint, to test whether the application is susceptible to fingerprinting similarly to all other similar applications. UIC's exploration showed that fingerprinting in the target application is considerably limited, as many APIs and functionalities are disabled entirely

to make users appear indistinguishable between each other.

4.1.1.1 Canvas Fingerprinting

This method uses the Canvas API to draw an image with text, specific fonts and colors. The differences in the devices' central processing units (CPU), font packs and libraries result in rendering this text in a slightly different way on different machines, thus generating images with a unique hash that can be used as identifiers and associated with each particular device/ application.

To prevent canvas fingerprinting in the target application, the Canvas API returns data of a completely white image, unless the user has previously given permission to a website to use the canvas (this permission is given on a per site basis). In that way, although a website can detect that a visitor is using that target application, it cannot distinguish between different the application users and link them to their previous visits, as they all appear to have the same fingerprint.

4.1.1.2 Web Graphics Library (WebGL) Fingerprinting

With the WebGL API, a website can use the device's graphics card to draw an image (typically an image of a three-dimensional [3D) shape). Due to the differences in the users' graphic cards and their drivers, similarly to the case of the Canvas API, the rendered images have slight differences and can be used as a high-entropy fingerprinting vector. In addition to the rendered image, information about the characteristics and properties of the user's graphic card can be used for augmenting this fingerprint.

The target application implements only the WebGL1 API, and disables WebGL2 by default. WebGL2 implements a large number of new functions and provides many attributes and parameters about the supported colors, textures, blocks, buffer sizes, precision etc., as well as various extensions that could be used as a part of a fingerprint. The target application does not provide any information that can reveal the vendor or model of the user's graphic card or any information about its drivers. Finally, similarly to the Canvas API, the WebGL API in the target application returns a canvas of a white image. The limited data/information that the WebGL API provides renders WebGL fingerprinting ineffective.

4.1.1.3 Audio Fingerprinting

To prevent audio fingerprinting the target application disables the Web Audio API entirely, by setting the dom.webaudio.enabled parameter to false. This prevents the application from providing an AudioContext, which is needed for running any form of audio fingerprinting.

4.1.1.4 Fonts Fingerprinting

A website can also fingerprint a user by enumerating the device's available fonts with Javascript. The target application limits the effects of font fingerprinting by only allowing specific fonts to be used. However, interestingly, the whitelist of allowed fonts includes some system fonts for MacOS and Windows machines, which can enable the attacker to potentially detect the underlying OS of the user. During their experimentation UIC found that that target application in Windows supports 19 fonts, while in MacOS it supports 12 fonts. Furthermore, UIC's experiments showed that the list of supported fonts was consistent between devices running on different OS versions.

4.1.1.5 User Agent

The target application aims to hide the characteristics of the users' device/application and make them appear indistinguishable from other users visiting a website. To that end, the application sets the User Agent in the headers of all its outgoing requests to appear as originating from a Windows machine (i.e., Mozilla/5.0 (Windows NT 10.0; rv:68.0) Gecko/20100101 Firefox/68.0).

The userAgent property of the navigator interface, that is accessible through JavaScript, provides to the correct user agent string (i.e., Mozilla/5.0 (Macintosh; Intel Mac OS X 10.14; rv:68.0) Gecko/20100101 Firefox/68.0). This discrepancy allows a website to get information about the user device, operating system, and application version.

Information about the user's device/OS/application is also given by various other properties of the navigator interface (i.e., Version: 5.0 (Macintosh), platform: Mac/Intel, product: Gecko, oscpu: Intel Mac OS X 10.14). Thus, setting the value of navigator.userAgent to match the User Agent value sent in HTTP headers would also require to properly set all the properties of the navigator object accordingly.

4.1.1.6 Screen Size and Resolution:

Typically, a lot of information can be obtained through JavaScript about the screen size and resolution, usable desktop size, title bar size etc. All this information can be used as part of the users device fingerprint. To prevent such fine-grained information from being exposed, the target application rounds the window size to a multiple of 200x100 pixels, and also caps the window size at 1000 pixels in each dimension (to prevent fingerprinting larger screens that will have an identifying size otherwise). UIC observed that there are some inconsistencies for the window size values returned by the target application in different OSs (when the actual screen size is the same). Nevertheless, despite those inconsistencies, the coarse-grained values returned by the target application can prevent a website from fingerprinting the user.

4.2 The NEMESIS Counter-Phishing System

The main objective of the ASED program is the development of a system to combat sophisticated phishing threats by engaging in multi-round, human plausible, dialogs with the phishers, aiming to waste their time and to collect attribution information. In support of this objective, SRI has developed the NEMESIS system that, given a possible email addressed to a person to be protected, orchestrates interactions with the phisher, enlisting the support of several dialog engines developed by SRI and other ASED performers, and coordinated by SRI's EDM in a way that is scales to thousands or more parallel phishing conversations per day.

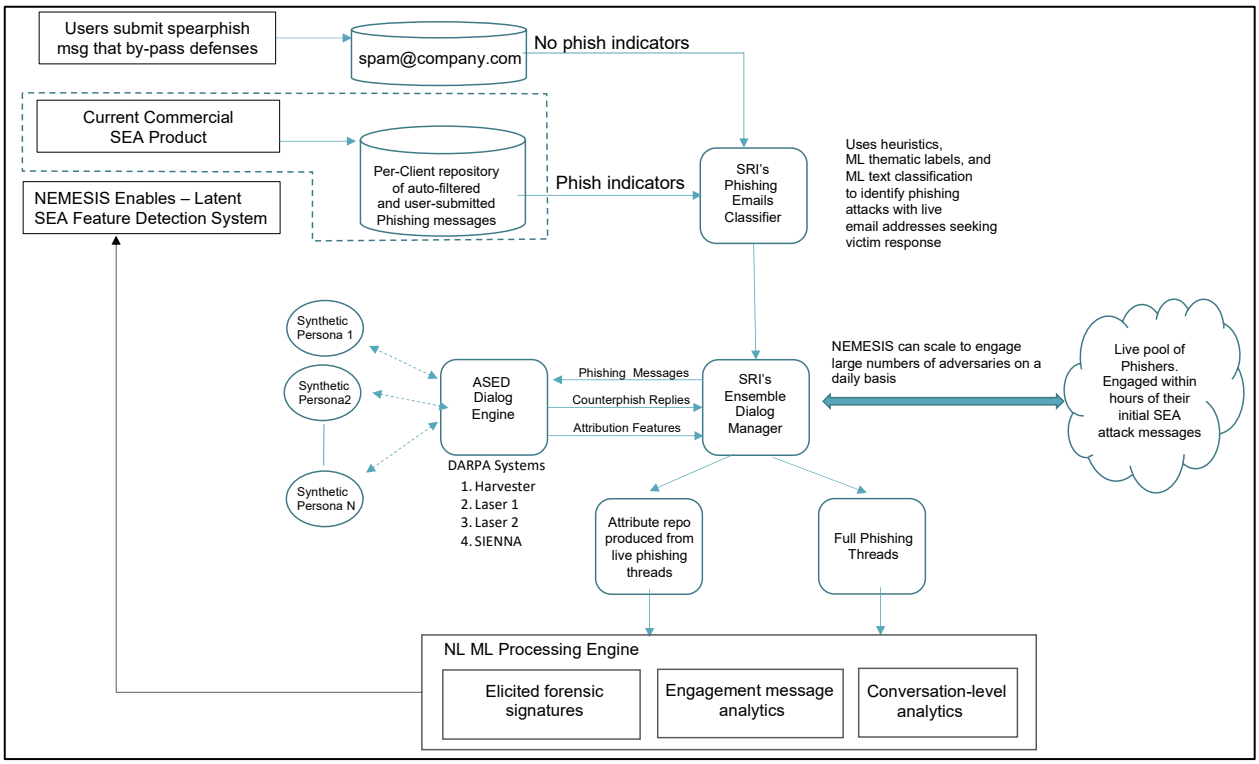

Figure 2. Functional Diagram of the Major Components to Comprise the ASED TA2 Live-Fire Counter-Phishing Service Deployed at SRI International.

NEMESIS is a conversation harvesting system specializing in multi-round social engineering attack engagements, and this service engaged in more than 10K live phishing conversations in less than two years.

Figure 2 illustrates the NEMESIS Counter-phishing system that is deployed into the SRI Corporate Enterprise network. The system incorporates an email classification service the extracts and analyzes daily phishing attacks from within SRI's Proofpoint corporate spam filtering folders. Candidate messages are submitted to the NEMESIS EDM, which initiates and manages all email threads between the ASED TA2 Dialog Engines and live corporate phishing adversaries. The EDM performs thread pairing between each dialog engine and the associated synthetic corporate persona that each dialog engine portrays, performs message schedule, flag extraction, and integration of external attribution service meta data. NEMESIS integrates Carnegie Mellon University's (CMU) Laser dialog engines, BBN's Sienna dialog engine, and SRI's Harvester dialog engine. It also integrates UICs web attribution service.

4.3 Ensemble Dialog Manager

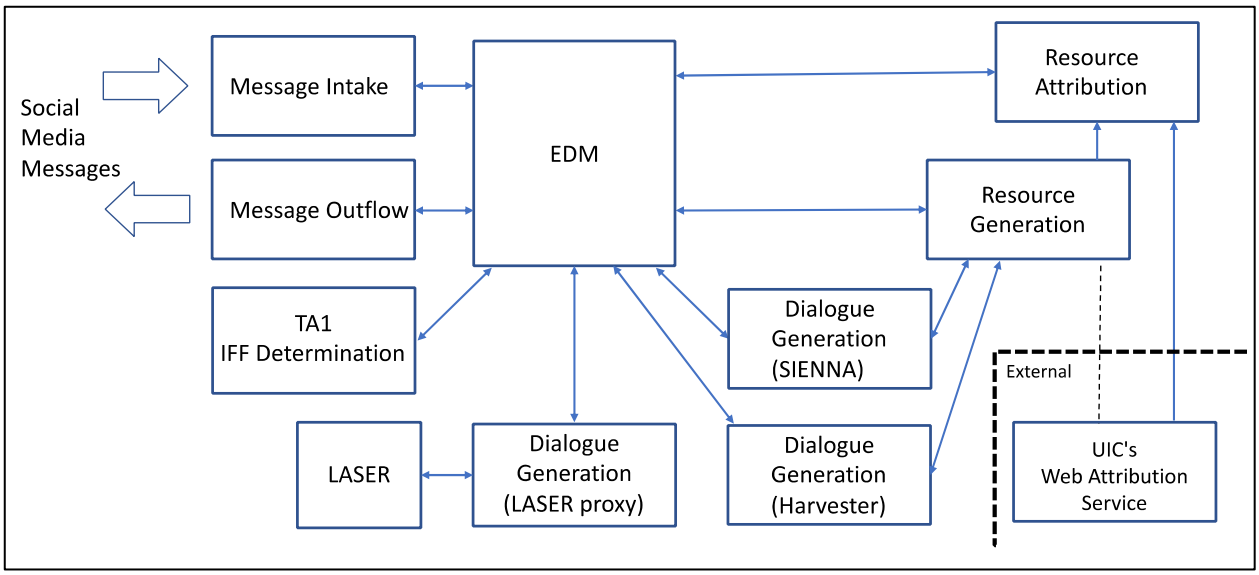

Figure 3. Ensemble Dialog Engine Processing Components and Execution Flow.

Figure 3 depicts the primary processing components and execution flow of the NEMESIS EDM. The EDM arranges, organizes, and schedules the data flow through the NEMESIS system. It is the central coordination system and storage service. Each received message is evaluated by the EDM for thread assignment (two-party conversations), and conversational thread objects are spawned as needed. TA1 initial Friend or Foe (IFF) assessments of individual messages are evaluated. If any message in a thread is marked as "foe," a dialogue generator (e.g., Laser, Sienna, or Harvester) is engaged for all messages in that thread in chronological order. Dialogue Generation (DG) components determine scheduled response messages and the EDM scheduler handles initiating the delivery at the appropriate time. DGs may indicate a scheduled waiting period for a response. If the EDM does not detect a response within that period, it requests the DG to generate a follow-on (or "unrequited") message, which is also scheduled. This sequence is repeated until either a response is received, or a configured limit (currently six) is reached. Attribution flags for conversation threads are aggregated and consolidated, and updates are submitted to reporting services. The EDM also detects email "bounce" messages, marking the unreachable email address internally and suppressing subsequent activity on the affected threads. The EDM also maintains most of the system's persistent state (for restart recovery).

We made the architectural decision employ a microservices pattern that used Representational State Transfer (RESTful) APIs for inter-component communication. We designed and documented RESTful APIs using OpenAPI/Swagger Yet Another Markup Language (YAML) specification files. Most of the Python3 code modules within the NEMESIS framework used the "connexion" module for RESTful OpenAPI/Swagger implementations (which automatically provides API verification and enforcement). System components were built as Docker containers. For development purposes, and due to the large number of possible combinations of system components, early orchestration was performed via a bash script. The SRI pilot study system uses the bash script because it does not require Kubernetes node configuration or administration. Because Phase 1 migrated to Kubernetes for testbed deployment, we accordingly created and adjusted NEMESIS system and component configurations. For Phase 2, we migrated the Kubernetes configuration to use the project-specific shoal framework.

4.4 Harvester

The *Harvester Counter-Phish* (HCP) dialog engine was created to respond autonomously to emails sent by Phishers. These phishing emails are a form of SE and they usually ask for your personal information or money in either a direct or round-a-bout way. HCP communicates with the phishers to keep them engaged and bothered and to get information from them either by asking for it or pulling it out of their email (such as when they have information in their signature block). HCP uses National Language Processing (NLP) and some string comparison to get the information. It keeps track of the information it has asked for and what it already has found, in a database along with the conversation, and will follow up if it does not yet have the information. Figure 4 presents an overview of the HCP response generation process that is used to formulate the next response to each message in the phishing thread.

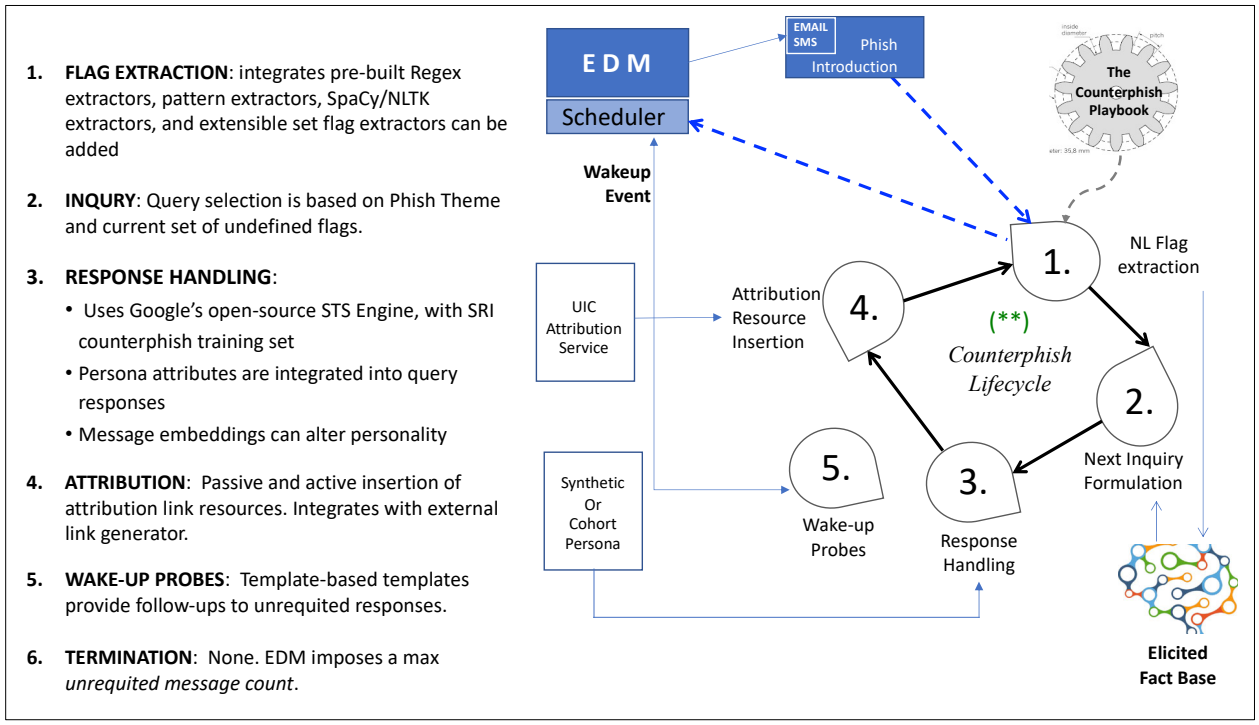

Figure 4. HCP Dialog Engine is an Elicitation Bot Designed to Pose Questions to Phishers and Respond to the Scammer's Questions as it Proceeds to Collect Information and Offer Web Attribution Links.

HCP knows what to say to the phishers because it tests the semantic text similarity (STS) (using Google STS (Bidirectional Encoder Representations from Transformer [BERT]) on the emails with its training data. It goes through the entire email, ranks all sentences and outputs the three best results. It will also add on questions for getting information if not already found and add in comments if it finds certain words or phrases (e.g., finding 'bitcoin' or 'Bless you,' it will ask what bitcoin is, and respond with 'bless you too.' respectively.)

4.4.1 Training Data

HCP responds semantically to phishing emails because it is trained on actionable email phrases and sentences with responses to these provided by the operator. The training data is created

either by hand or by a program that reads the phishing emails, categorizes them and finds common sentences (not yet developed). In the example below the 'input' was derived from emails, and the response to the input comes from the operator.

In the 'upgrade' example below, the text input by the user would be semantically compared to the text in 'input,' and if it is similar enough, HCP will respond with one of the 'responses' chosen randomly.

A Training Example:

```
{
    "tag":"upgrade",
    "input":["are you ready to upgrade?","you need to upgrade now",
      "need more reasons to upgrade?"],
    "responses":["I was wondering if you could [email,tell] me more about
        the upgrade.","Can you tell me more?","Can you tell me
        more about the upgrade?"]
}
```

4.4.2 Follow-on Messages

If HCP does not get a reply within a certain amount of time, it will respond back with simple follow-on messages, such as "Why don't you respond to me?" or "I am waiting for your email!" We found that these types of emails push the phisher into responding again.

4.5 Campaign Management System (CMS)

The CMS operates in a role similar to the EDM, however it is designed to initiate and manage a social engineering attack (training) campaign comprising the initiation of social engineering training playbooks against a set of target trainees. The Basic CMS functions include the ability to respond to start/pause/stop control changes to campaign document. Its central purpose is to launch conversation threads for bindings compliant with the campaign status and ordering parameters. The CMS detects newly received messages and engages DG for a response. It also manages DG-collected flags and stores all campaign and thread-related states into a centralized Mongo database.

The Mongo Database (DB) operates a centralized blackboard for managing campaigns. This includes support for task replication and coordination. The CMS operators using a Compact RESTful APIs, which mainly consists of *DB object-identifier (OID) notifier calls.* Figure 5 illustrates the execution and procedural flows implemented by the CMS. All dialog engine and CMS state information is maintained in a centralized Mongo DB.

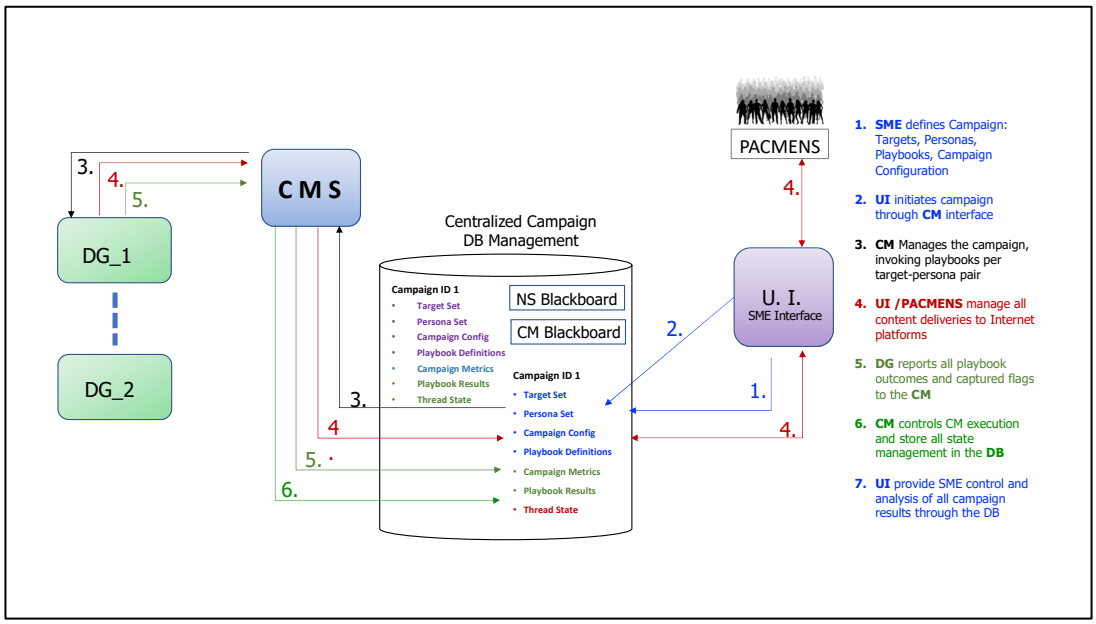

Figure 5. The CMS Employs a Data Centric Architecture for Conducting Scalable SEA Training Exercises.

4.6 Harvester 5 (H5)

H5 is used to communicate autonomously with people via text or email (although it could transfer to a human if necessary). It uses a directed conversational system that uses 'playbooks' to manage a conversation with a user. The playbook directs the conversation by using a step-by-step method, and if the user says something out of the normal flow of the conversation, H5 handles this by branching to another series of steps, after which it can either end the conversation (if the user says "Go away!"), or branch back to the main dialog to continue the discussion. The five stages of playbook processing performed by H5 is illustrated in Figure 6.

H5 will also extract information from the conversation, such as addresses or phone numbers. In the conversation, you could ask for their phone number or Whatsapp number in case you need to get more information and H5 will extract the information and store it with the conversation in a database. If they don't provide it, H5 can do a follow-up, or let it go. It uses NLP, STS (using Google STS BERT), sentiment analysis, and some string comparison to get the information and move along the steps.

4.6.1 Playbooks

A playbook is a JavaScript Object Notation (JSON) structure used to define how H5 reacts to what the user says and what is said. It provides a 'directed conversation' by taking the user along in steps or script, making statements, asking questions, and processing those questions with further dialog or actions.

1. Engagement initiates with a persona-consistent introduction method
2. Playbooks express the **dialog model** that drives message generation, and the target elicitation flags: Target flags, inquiry sequencing, and branch decisions are defined in the playbook's **Intent specification**. Pre-built and custom intents are supported
3. **Fact Extractor**: auto-extracts all key 'target facts' that are provided by the target user
4. H5's **Answer Engine** enables it to answer questions based on the current synthetic persona's legend, as well as other open dialog quests.
5. Termination criteria, including max round count, and satisfaction of intents enable Harvester5 to implement graceful exists from conversation threads.

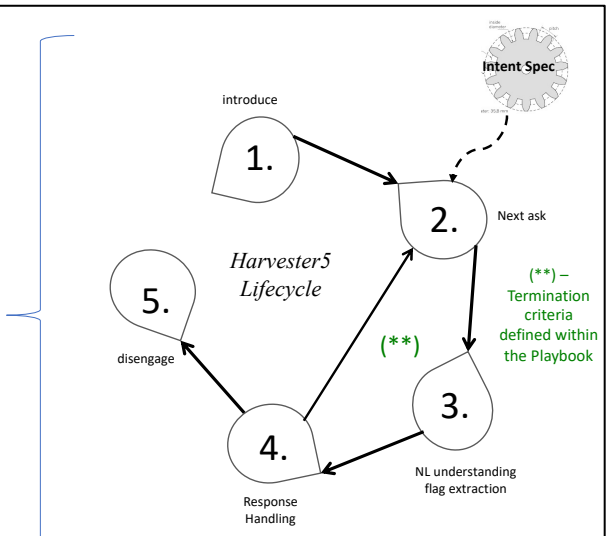

Figure 6. The Five Stages of Playbook Processing Performed by H5.

The playbook allows for slot filling (e.g., putting in the user's name or organization) and macros (e.g., being able to say something in different ways). This gives the dialog more personalization and feel less 'canned.' It also allows for multiple ways of saying the same thing (randomly choosing or choosing by personality) so that spam detectors will have a harder time.

Dialog Script Example:

```
"DialogOrder": [
    "intro_method",
    "stop_if_no_exit1",
    "more_information",
    "stop_if_no_exit2",
    "coworker_info",
    "exit"
],
```

Example of **intro_method** above, with slot filling:

```
"intro_method" : [ "Good morning,\n\tMy name is %P_Firstname %P_Lastname and I'm a hiring
        manager for Lumon Biotech. I'm trying to verify the work history of a candidate for a
        position with our company and I'm hoping you might be able to help. Can you on-
        firm that Bill Jones worked as a Senior Project Administrator for XYZ from 1996 -
        2014?\n\nRespectfully,\n%P_Firstname %P_Lastname\n"
]
```

4.6.1.1 Branches

H5 uses its semantic text understanding to determine what the user is saying and if the user says something off the main script, but of a known response (such as “Who are you?”) H5 can understand that and branch to another script and respond appropriately. An example would be to either continue on with the script or exit the conversation (especially if the user says “Stop this now, I don’t want to talk to you!”).

In the Branch example below, the text input by the user is semantically compared to the text in 'input', and if it is similar enough, H5 will respond with the response_out_of_office message.

A branch example:

```
{
   "tag": "out_of_office",
   "input": ["I am out of the office", "I will be on vacation",
                "I will be back in the office", "I am away from email", "I am traveling"
   ],
   "responses": ["response_out_of_office",
              "exit" ]
},
"response_out_of_office" : [
      "Sorry we missed you. We may contact you when you return."
]
```

4.7 Persona and Account Creation, Management, and Engagement Services (PACMENS)

Throughout the Phase II DARPA ASED evaluations, Jataware supported the TA3 team email delivery functions using its email delivery services, including PACMENS, which is presented in our Starling technical report.

4.8 SEA System

In the last year of Phase II, DARPA directed SRI develop a hybrid ASED system that employed an open playbook language used to construct phisher-inspired social engineering attack training dialog models. In this experiment, the system initiates conversations with training targets using a set of synthetic personas that are bound to one or more training playbooks. Figure 7 illustrates the component architecture that was developed using the EDM, H5, Jataware, and a new Campaign Centric Database framework. We have containers for the entire system and have demonstrated experimental test cases using Amazon Web Services (AWS) and Kubernetes in which we have simulated campaigns involving 100K training targets.

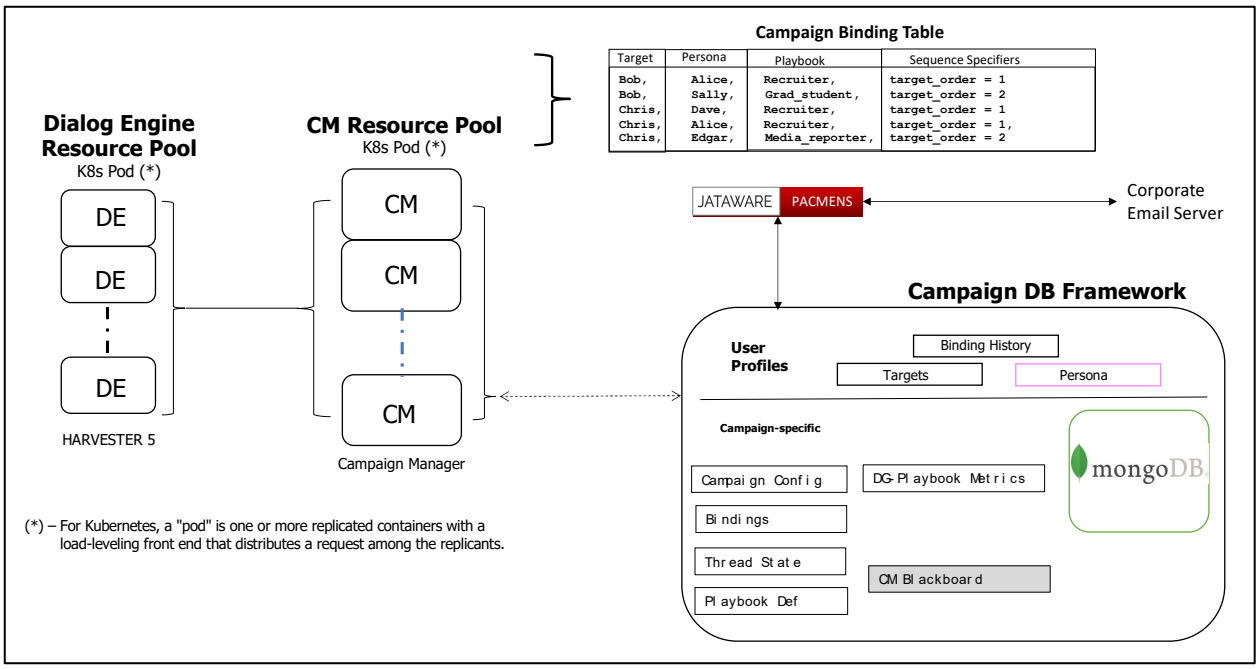

Figure 7. The NEMESIS CMS, Employing H 5, Jataware's PACMENS Service, and the Campaign DB Management Framework.

5.0 MEASUREMENTS AND EVALUATION

5.1 NEMESIS Final Outcomes

The four major technologies developed under the ASED program in Phase II have undergone continual integration with our ASED collaborators , evaluations, and live experimentation. The following is a brief summary of our high-level findings:

- We have validated that counter-phishing dialog engines (EDM+HCP) can engage live adversaries with higher engagement rates than initially hypothesized (we use Rate of Engagement (RoE) to indicate the rate at which live phishers engage in direct dialog with our bot agents or click on attribution links provided by our agents):
  - Overall ASED System RoE = 0.52 (944 engagements from 1,803 unique threat actors)
  - The NEMESIS Harvester Dialog Engine produced the highest RoE against live adversaries at RoE = 0.47.
- Our project confirms that the (EDM+HCP) counter-phishing dialog dataset by the NEMESIS counter-phishing system enables one to isolate and identify of unique phisher attributes, gestures, and repeated playbook patterns.
- These tactics can be converted to derive new fraud accurate phishing playbooks for driving a new generation of digital fraud training curriculum.
- The NEMESIS CM and H5 represent a new breed of automated Artificial Intelligence (AI)-driven Security-Awareness chat bots, offering a new approach to helping defend the human attack surface.
- We have demonstrated that third party developers (Zetier and Applied Research Laboratory for Intelligence and Security [ARLIS]) can successfully implement SEA Training Playbooks that produce high-quality RoE = 0.32. Our SRI Information Technology (IT) department observed that a 0.32 RoE is comparable to their human-drive phish training exercises.

5.2 ASED TA2 DARPA Evaluation

The NEMESIS team lead the design, integration, deployment, and management of the TA2 evaluation services. We led weekly integration meetings during the months preceding each DARPA evaluation to conduct all coordination, integration activities, Kubernetes migrations, functional testing, for all dialog engines in the program. We also held bi-weekly coordination meeting with the TA3 team to review metrics and discuss evaluation issues.

The main emphasis of Phase II has been our participation in the ongoing DARPA TA3 evaluation. We developed extensions to the EDM to support multi-channel communications, a new Mongo database service for scaling dialog engine state management, and an API framework to support new dialog engine functionality. Our team had to integrate, test, and deploy four parallel dialog engines into each Phase DARPA evaluation, including our own Harvester dialog engine. Our final test release supported both email and Short Messaging Service (SMS) channels.

With respect to Harvester's Evaluation performance, there was an increasing improvement observed at each major evaluation milestone. Of the four dialog engines evaluation, Harvester succeeded in producing the highest number of flags extracted by any dialog engine. The final evaluation metrics produced for Harvester are shown in Tables 1 and 2. Table 1 shows flag extraction accuracy reaching 93% on 60 flags. These were the best results from any team.

Table 1. Harvester Results on Program Evaluation.

Results based on 60 flags.

Evaluation	True Positives	False Positives
C1	0.7525	0.2545
C2	0.7351	0.2648
C3	0.9313	0.0886

Table 2 shows message quality decreasing slightly as flag accuracy improved.

Table 2. Harvester Message Quality on Program Evaluation.

Evaluation	% Good	% OK	Total
C1	45%	33%	78%
C2	48%	31%	79%
C3	44%	27%	71%
Overall	46%	31%	79%

5.3 ASED TA2 Live Counter-Phishing Portal

We created a live counter-phishing portal deployed to protect SRI's corporate enterprise email users. The portal components included an SRI Machine Language (ML)-based phishing email filter (to auto-select candidate phish emails from SRI's Proofpoint server), the EDM, SRI's HCP dialog engine, BBN's Strategies for Investigating and Eliciting Information from Nuanced Attackers (SIENNA), and CMU's Laser1 and Laser 2 dialog engines. During its operation, the portal provided teams with vital real-world datasets from which they could test their dialog model and flag extraction logic. Figure 8 shows a snapshot of the final fifth generation of live ASED TA2 Pilot Portal.

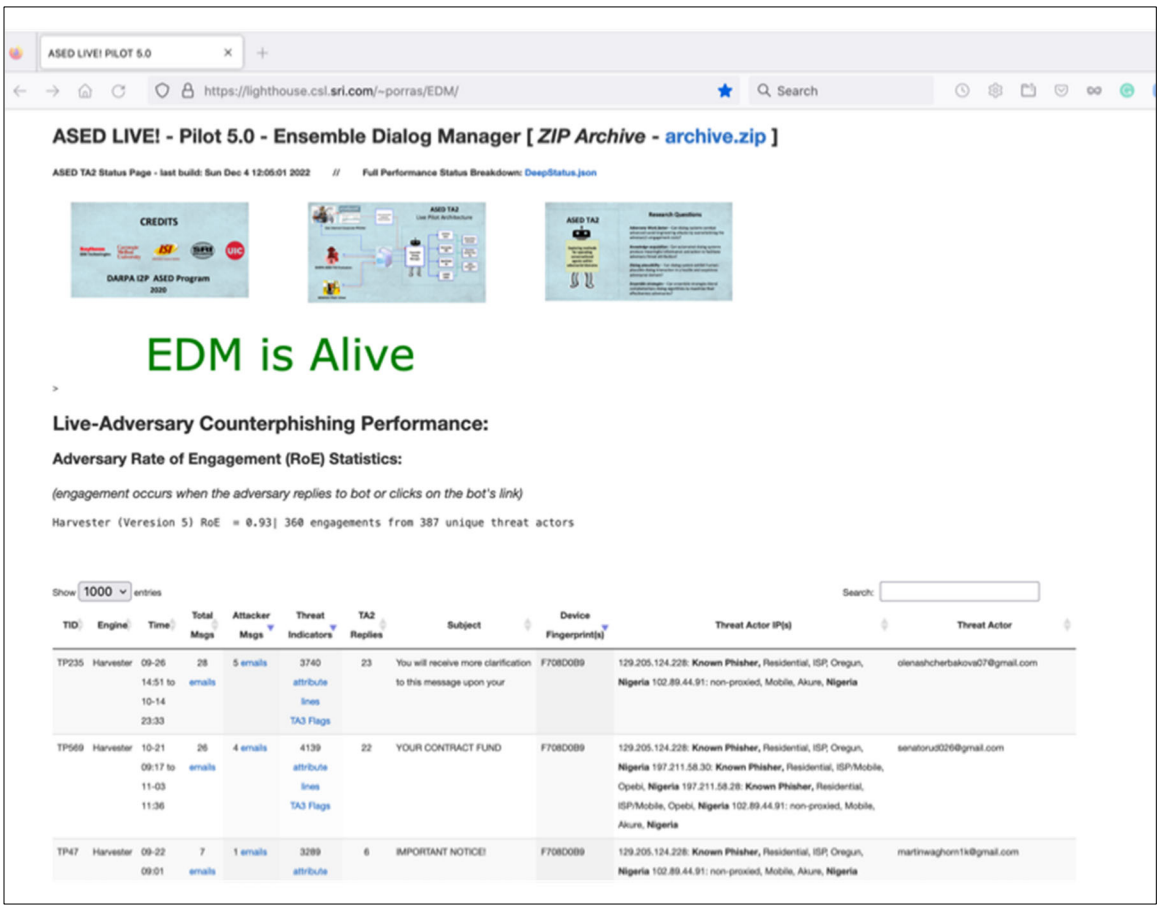

Figure 8. The NEMESIS Live Counter-Phishing Portal was use throughout Phase II of the ASED Program to Test the Ability of ASED Dialog Engines to Conduct Counter-Phishing Operations against Live Phishers who were Targeting SRI Corporate Employees.
We created generations of this portal (the 5th generation of this portal is currently achieving a 0.93 RoE with live adversaries.

5.3.1 ASED TA2 Live Counter-Phishing Pilot Statistics:

In total, we produced five generations of the ASED TA2 Live Counter-phishing portal:

- Pilot 1: Spring/Summer 2020 - 208 live counter-phish initiations.
  - URL: https://lighthouse.csl.sri.com/~porras/EDM-PILOT1/
  - Metric ROE: Test system only - this version produced poor results
  - Lesson: Dialog engines need to adopt a new interaction model
  - in which they auto-re-engage adversaries who do not respond.
- Pilot 2: Fall/2020
  - URL: https://lighthouse.csl.sri.com/~porras/EDM-PILOT2/
  - Metric ROE: 13% (per thread engagement rate)
    136 engagements from 987 counter-phish initiations
  - Lesson: All dialog engines needed substantial improvement

- Pilot 3: (Pre-Spring Eval Pilot)
  - URL: https://lighthouse.csl.sri.com/~porras/EDM-PILOT3/
  - Metric ROE: 12% (per thread engagement rate)
    170 engagements from 1426 counter-phish initiations
  - Lesson: Pilot 3 had a few periods of delayed batch message submission that potentially affected the engagement rate. Overall, it produced higher quality dialogs than Pilot 2 and better flag extraction.
- Pilot 4: (Last quarter 2021 through Spring 2022)
  - URL: https://lighthouse.csl.sri.com/~porras/EDM-PILOT4/
  - Metric ROE: 0.52 (overall Rate of Engagement)
    944 engagements from 1803 unique threat actors
  - Lesson: All code bases were unified for the last two ASED TA2 evaluations
- Pilot 5: (Summer 2002 to present)
  - URL: https://lighthouse.csl.sri.com/~porras/EDM/
  - Metric ROE: 0.93 (overall Rate of Engagement)
    360 engagements from 387 unique threat actors
  - **Note:** This version is Harvester (HCP) only. CMU and BBN declined to continue.

5.4 SEA Exercise

During the last quarter of 2021, DARPA asked the NEMESIS team to assist in the development of a new experiment. The objective was to explore whether our counter-phishing technologies could be adapted to launch phish-training exercises in a fully automated manner. To support this work, we created a variation of our EDM called the CMS. We also created a variation of the Harvester HCP engine. This new version of Harvester is referred to has H5 and incorporates an open language playbook service that allows third parties to construct customized playbooks.

In Spring 2022, Zetier, ARLIS, and SRI's Human Resource management team began work on an experiment to examine whether key administrators throughout SRI accurately follow corporate policies regarding the disclosure of information about previous employees to external third parties. The team succeeded in developing a H5 Playbook to implement a live email training exercise that targeted 50 unwitting SRI employees.

The results of this experiment were as follows:

- **Experiment**: Employee compliance to procedures for disclosing previous employment information
- **Subject Pool**: 50 unwitting SRI Administrative employees
- **Experiment Duration**: 48 hours (all email rounds were required to complete within this time frame).
- **Harvester5 Rate of Engagement**: 0.32 ROE
- **Experiment Outcome**:
  - Seven employees successfully followed compliance rules

- Four employees followed incorrect procedure when responding to inquiries
- Two employees engaged in policy violations
- Three employees provided initial acknowledgements

5.5 Conclusion

All dialog threads were analyzed by hand and the playbook succeeded in producing convincing conversations with all participants who engaged the H5. The experiment was considered a successful and useful outcome from the perspective of SRI's Human Resource Manager. The 0.32 RoE was considered equivalent to what SRI's Chief Information Security Officers (CISO) phishing training team experiences when performing similar training exercises that are driven by human testers. Zetier indicated that the H5 playbook language provided excellent adaptability and functionality to implement the SEA training exercises.

6.0 DELIVERABLES

6.1 Software Deliverables

The ASED TA2 EDM and HDE source code and documentation packages are available at gitlab.ased.io.

The Full source code packages for EDM, Harvester, CMS, and H5, along with all package documentation have been stored on 4K CD-ROMS and sent to DARPA and the AFRL Contract Office.

6.2 Publications

S. Karami, P. Ilia, K. Solomos, and J. Polakis, "Carnus: Exploring the Privacy Threats of Browser Extension Fingerprinting," in *Proceedings 2020 Network and Distributed System Security Symposium*, San Diego, CA, 2020. doi: 10.14722/ndss.2020.24383.

Y.-Y. Chang, P. Li, R. Sosic, M. H. Afifi, M. Schweighauser, and J. Leskovec, "F-FADE: Frequency Factorization for Anomaly Detection in Edge Streams," in *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, Virtual Event Israel, March 2021, pp. 589–597. doi: 10.1145/3437963.3441806.

6.3 Technology Demonstrations and Videos

The following are links to video demonstrations and the NEMESIS live counter-phishing portal.

- http://www.csl.sri.com/users/porras/ASED/NEMESIS_Overiew_Part1.mp4
- http://www.csl.sri.com/users/porras/ASED/NEMESIS_Overiew_Part2.mp4
- NEMESIS Live-Fire Counter-Phishing Portal: https://lighthouse.csl.sri.com/~porras/EDM/

6.4 U.S. Patent Submissions

- Porras, Philip A, Nitz, Kenneth C, Skinner, Keith M, Freitag, Dayne B, Conversation-depth Social Engineering Attack Detection using Latent Signatures from Automated Dialog Engagement, January 2022.
- Porras, Philip A, Skinner, Keith M, Nitz, Kenneth C, Freitag, Dayne B, Kalmar, Paul S, Digital Fraud Training: Using Counter-Phishing Intelligence to Build Better Defenses Against Social Engineering Attacks, March 2022.

6.5 DARPA Highlight Submitted

DARPA I2O ASED PROGRAM TEAMS WITH BARRACUDA NETWORKS TO EVALUATE A NEW APPROACH TO IDENTIFYING SOCIAL ENGINEERING ATTACKS

Barracuda Networks, a market leader in email gateway security products, has teamed with DARPA ASED's *Project NEMESIS* team to evaluate a new approach to detect an insidious form of nation-state social engineering attacks. The attack method involves the hijacking or spoofing of emails from a trusted individual in order to compromise the security or privacy of victims within the individual's social circle. Such *spearphishing* attacks are employed by advanced threat actors and represent a significant detection challenge for existing email counter-phishing security technologies.

Under DARPA's ASED program, Stanford University recently introduced a scalable method for detecting anomalous behavioral patterns that can arise when such spearphishing incidents occur. The method models corporate email as a dynamic system of interactions among email account nodes. It captures these complex network interactions and their temporal properties in a low dimensional vector space, which is then used to detect anomalies. The method relies on the utilization of the network structure rather than message features, which provides several benefits. It delivers robust performance, as these network features provide a strong signal for anomaly detection. Second, it supports inductive learning, where the knowledge from one node can be transferred to other, unseen nodes. Third, it can be applied to detect group behaviors, such as cooperating compromised accounts within or outside of an organization. Lastly, the method is privacy preserving, since no message details are needed, only the two endpoints and the timestamp for each interaction. If additional features, such as meta-information and message content are available, then they can be incorporated into the model to yield even better predictive performance.

The NEMESIS team recently met with Barracuda's senior email products team to review the Stanford algorithm and the existing evaluation results that were developed under DARPA ASED. Barracuda confirmed the challenges that the current industry has had in addressing the target adversary models of the Stanford algorithm and were impressed with the features and sophistication of the Stanford approach. In April 2020, the Barracuda team and Stanford researchers initiated a collaboration on a much larger-scale evaluation of Stanford's anomaly detection algorithm using a Barracuda-provided dataset from samples among the more than one billion emails that Barracuda processes per day. The team will iterate with Barracuda through progressively larger-scale evaluations, starting with an initial 200-million email evaluation. Among the discussions of potential outcomes for this collaboration was the potential transfer of this technology into industry, possibly directly through Barracuda Networks.

6.6 Transition Activities

SRI is currently working toward a new 2023 commercial venture that will transition ASED NEMESIS technologies to the commercial market. Discussions are under way with executives from Google and Meta to lead the venture.

7.0 PRINCIPAL INVESTIGATORS

The following are the NEMESIS Principal Investigators:

- Phil Porras, SRI International, 333 Ravenswood Ave, Menlo Park, CA 94015
- Justin Gawrilow, Jataware Corp., 6630 31st PL NW, Washington, DC 20015
- Jure Leskovec, Dept. of Computer Science, Stanford University, William Gates Building 4A, Stanford CA 94305 (Phase I only)
- Jason Polakis, Dept. of Computer Science, University of Illinois at Chicago, 851 S. Morgan St., Chicago IL 60607 (Phase I only)
- Amanda Towler, Hyperion-Gray (Phase I only). No longer in business.

8.0 LIST OF ABBREVIATIONS, ACRONYMS AND SYMBOLS

3D	Three-Dimensional
AI	Artificial Intelligence
API	Application Programming Interface
ARLIS	Applied Research Laboratory for Intelligence and Security
ASED	Active Social Engineering Defense
AWS	Amazon Web Services
BERT	Bidirectional Encoder Representations from Transformer
CISO	Chief Information Security Officers
CMS	Campaign Management System
CMU	Carnegie Mellon University
CPU	Central Processing Unit
DARPA	Defense Advanced Research Projects Agency
DB	Database
DG	Dialogue Generation
EDM	Ensemble Dialog Management
H5	Harverster5
HCP	Harverster Counter-Phish
HTTP	Hypertext Transfer Protocol
IFF	Initial Friend or Foe
IT	Information Technology
JSON	JavaScript Object Notation
ML	Machine Language
NEMESIS	Natural Language Engagement of Malicious Entities through a Social Interaction Service
NLP	National Language Processing
OID	Object-Identifier
OS	Operating System
PACMENS	Persona and Account Creation, Management, and Engagement Services
RESTful	Representational State Transfer
RoE	Rules of Engagement

SE	Social Engineering
SEA	Social Engineering Attack
SIERRA	Strategies for Investigating and Eliciting Information from Nuanced Attackers
SMS	Short Messaging Service
STS	Semantic Text Similarity
TA	Task Area
UIC	University of Illinois at Chicago
WebGL	Web Graphics Library
YAML	Yet Another Markup Language