Technical Description of the Reimagining Security with Cyberpsychology-Informed Network Defenses

Program Overview

Cyber attacks are increasing in quantity and severity. Some of the most sophisticated and persistent cyber attacks are primarily human-driven. However, most cyber defenses do not consider the human attributes and limitations of attackers. Furthermore, most existing defenses focus on blocking suspicious behavior and few initiate interactions with a suspected attacker to understand their attributes, skills, or goals, let alone, induce changes in their behavior.

The Reimagining Security with Cyberpsychology-Informed Network Defenses (ReSCIND) Program focuses on inducing or intensifying cognitive biases or other cognitive limitations to thwart cyber attackers. Rather than just attempting to detect and stop suspicious movement on the network, Offerors will propose innovative solutions to increase the effort and resources spent by cyber attackers by impacting their decision-making. The ReSCIND Program seeks novel methods that:

1. Identify, and provide evidence of, Cognitive Vulnerabilities (CogVuls) relevant to cyber attackers;
2. Understand, measure, and induce changes in cyber attack behavior and success;
3. Develop Cyberpsychology-informed Defenses (CyphiDs) impacting both early and late stage attacks;
4. Create Cyber-specific Computational Cognitive Model(s) (C3M) that reflect and predict attacker behavior; and
5. Produce Adaptative Psychology-informed Defenses (APhiDs) which automate the preferred sequence of CyphiDs based on observed attacker behavior.

Cyberpsychology integrates human behavior and decision-making into the cyber domain to understand, anticipate, and influence cyber behavior. There is a vast amount of cognitive and behavioral science research that can be applied to cybersecurity to improve defensive posture. The ReSCIND program aims to develop CyphiDs that leverage an understanding of attacker decision-making, human limitations, and cognitive biases to reduce attack effectiveness. ReSCIND will rebalance the inherent asymmetry of cyber defense by exploring novel methods for manipulating attacker behavior during various phases of the Cyber Kill Chain[1].

As notionally represented in *Figure 1*, ReSCIND will provide defenders a much-needed advantage by expanding the cyber defense toolkit by specifically leveraging well-established cognitive vulnerabilities (e.g., decision-making biases, mental model heuristics) that can be intensified and manipulated to impede cyber attackers. Offerors will propose novel approaches informed by social science research and associate CyphiDs to observables (e.g., environmental features, attacker attributes, mission context) to measurably disrupt cyber attack behavior across the various stages of the Cyber Kill Chain.

[1] Lockheed Martin (2015). White Paper Seven Ways to Apply the Cyber Kill Chain® with a Threat Intelligence Platform, Seven_Ways_to_Apply_the_Cyber_Kill_Chain_with_a_Threat_Intelligence_Platform.pdf (lockheedmartin.com); Ju, A., Guo, Y. & Li, T. MCKC: a modified cyber kill chain model for cognitive APTs analysis within Enterprise multimedia network. *Multimed Tools Appl* **79**, 29923–29949 (2020). https://doi.org/10.1007/s11042-020-09444-x

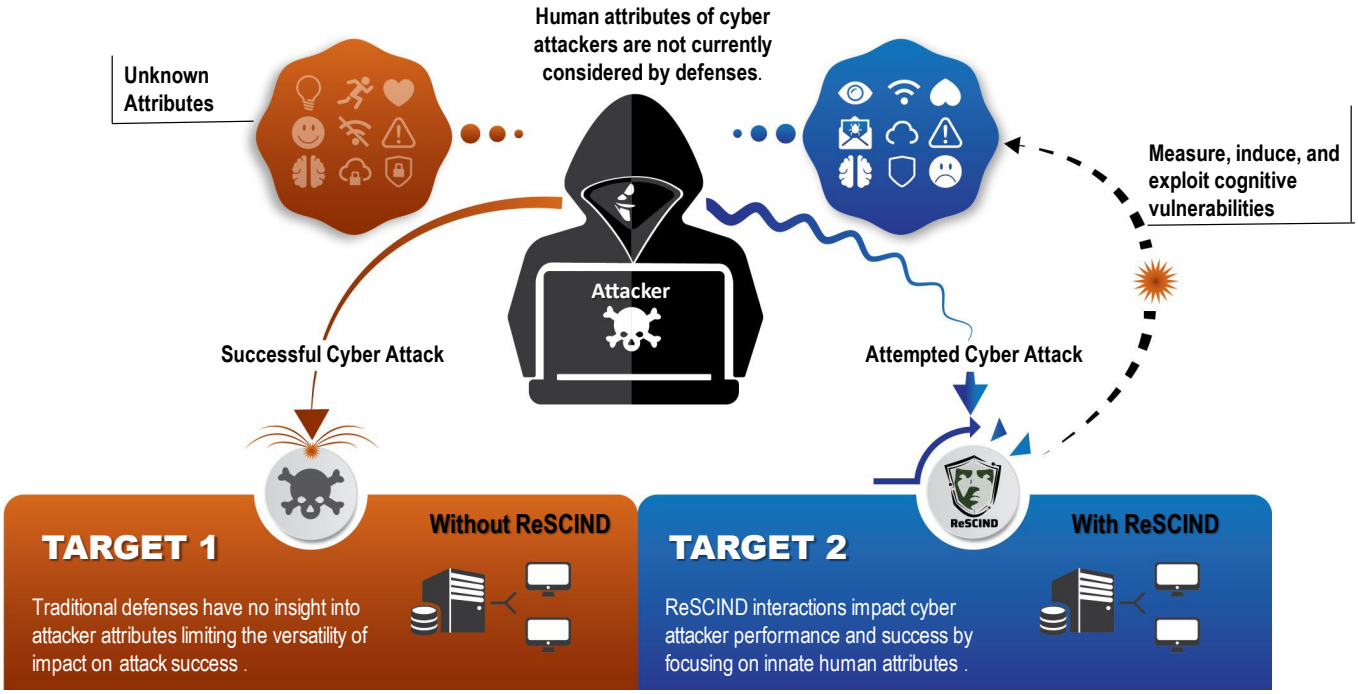

Figure 1: Notional graphic of cyber defense with and without ReSCIND

1 Technical Challenges and Objectives

The objective of the ReSCIND Program is to impose a cyber penalty against attackers and increase the effort and difficulty for them to achieve their goals, both now and in the future. Technical challenges and objectives include:

1. *Identify and provide evidence of CogVuls relevant to cyber attack behavior.* In the ReSCIND Program, cognitive vulnerability is an umbrella term encompassing cognitive and decision-making biases, innate cognitive limitations, emotional or mental state, or physiological vulnerabilities that can result in reduced cyber attacker success or effectiveness. Offerors must describe their plan for novel research exploring dynamic cyber attack scenarios with skilled human participants. Observable attacker attributes, network and host characteristics, and environmental features that could impact the defensive utility of selected CogVuls must be identified. Performers will:
    - Establish relevance of vulnerabilities to cyber attackers, accounting for relevant differences among individuals through theoretical and experimental research.
    - Design and execute empirically and statistically efficient experimental designs with cyber-skilled human participants to explore cyberpsychology in dynamic cyber attack tasks.
    - Produce a structured visual representation that maps the CogVuls to cyber-relevant behavioral characteristics of the attacker, the network, and the external environment.
2. *Understand, measure, and induce changes in cyber attack behavior and success.* Offerors will present hypothesized relationships between CogVuls, bias sensors that identify and measure them, and bias triggers that create cyber situations to induce and intensify CogVuls. Performers will experimentally establish which selected CogVuls, bias sensors, and bias triggers produce a measurable effect on cyber attack behavior. Performers will:
    - Develop approaches to exploit cyber attacker CogVuls for defensive gain.

- Understand the extent to which CogVuls may overlap and the precedent factors that lead to CogVuls in cyber-specific situations.
- Identify novel techniques to measure, predict, and influence attacker behavior to thwart success.
- Create bias signature(s), which maps cyber data to the presence of or increase in a specific cognitive vulnerability.
- Develop bias sensors using data likely to be available to defenders in a realistic environment (e.g., PCAP, IDS alerts).
- Establish bias sensor reliability and validity using established methodologies. Develop triggers (host/network manipulations) which can reliably induce or exacerbate CogVuls.

3. *Develop cyberpsychology-informed defenses (CyphiDs) that impact both early and late stages of a cyber attack.* A successful system will demonstrate measurable impact on cyber attacker performance and success through exploitation of robust and measurable CogVuls. Sets of improved or newly created bias sensor and bias triggers will be incorporated into the CyphiDs, based on the CogVuls established as most relevant and impactful. Performers will:
   - Produce a structured visual representation that maps the CogVuls and CyphiDs to defensive goals and measurable impacts on cyber attack behavior.
   - Implement the logic and software of cyberpsychology-informed defenses (CyphiDs) for testing in cyber range testbed, incorporating relevant insights from the structured visual representation.
   - Demonstrate CyphiD efficacy for slowing or reducing the success of cyber attack attempts.
   - Develop new metrics to evaluate human-focused cyber behavior and performance for both early and late stages of a cyber attack.
4. *Create cyber-specific computational cognitive models (C3M) that reflect and predict attacker behavior changes in reaction to CyphiD interventions.* The models must respond to variation in CogVuls as measured by the bias sensors, such that they adapt to both raising and lowering relevant attributes using available data. The models will replicate and predict behavioral changes caused by bias triggers. Performers may elect to use their modeling to inform development of their APhiDs.
   - Develop, train, and test C3Ms[2] which reflect and predict attacker behavior, with variability dependent on presence of CogVuls.
   - Modeling efforts should also address differences in attacker behavior based on attacker attributes, network and host characteristics, or situational factors.
5. *Produce adaptive psychology-informed defenses (APhiDs) which automate the preferred sequence of CyphiDs based on observed attacker behavior.* Performers will create an adaptive defensive system that automates CyphiD selection to independently respond to cyber attacker attributes and behavior, and other environmental features or network attributes. Performers will:
   - Develop algorithms for APhiDs to allow for automated adaptation of CyphiDs.
   - Implement the logic and software of APhiDs for testing in cyber range testbed, incorporating relevant insights from the structured visual representation, and previous experimental results.

[2] Ron, S. (2008). *Introduction to computational cognitive modeling*. Cambridge, MA: Cambridge handbook of computational psychology. ISBN 978-0521674102.

- Provide novel generalized defenses for enterprise networks and evidenced-based use-cases, including deployment guidelines to highlight each defense's effectiveness against various real-world features (i.e., attacker attributes, mission context, network and host characteristics).

2 Program Phases

The ReSCIND program is a 45-month effort, comprised of three (3) phases. Proposals shall include a solution for all phases and address all technical challenges. **Proposals that do not include a complete solution for all phases or do not address all five technical challenges described above will be considered non-compliant and will not be evaluated.** The following table provides an overview of the ReSCIND program structure.

Table 1: An Overview of the ReSCIND Program Structure.

Phase	Duration	Objective
1	18 months	Identify CogVuls relevant to offensive cyber operators, including methods to induce, exacerbate, and measure each cognitive vulnerability.
2	15 months	Research and develop CyphiDs that map to observed attacker attributes and measurably disrupt cyber attack behavior across the Cyber Kill Chain and increase the negative impact on attacker performance and success.
3	12 months	Use experimental results and data from prior phases to develop APhiDs (for automated selection of a combination of CyphiDs) and cyber-specific computational cognitive modeling (C3M) to reflect and predict the behavioral data provided.

The Test and Evaluation (T&E) Team will conduct several T&E events throughout the life of the program using Institutional Review Board (IRB)-approved Human Subject Research (HSR) protocols to evaluate performer developed solutions. These will consist of controlled experiments that consider specifics of real-world cyber campaigns to balance internal and external validity. Much of this data will be made available to performers for Research and Development (R&D) in later phases, and eventually, provided to the general scientific community. In addition, performers will be required to conduct their own supplemental IRB-approved HSR data collection(s) and make that data available to the program. Deliverables produced by proposers must grant the Government intellectual property (IP) rights sufficient to allow the Government to conduct T&E HSR, open-source associated datasets, and modify and deploy deliverables on classified networks. Additional details on program data can be found in Section 6.

2.1 Phase 1

The goal of Phase 1 is to identify the CogVuls most relevant to cyber attack behavior based on foundational scientific research and cyber relevant HSR experimentation, including methods for inducing, exacerbating, and measuring them. Phase 1 includes the development of novel bias sensors to detect these CogVuls using cyber data, and bias triggers to induce and intensify them in a cyber situation. The ReSCIND Program encourages maximum creativity and diversity in selection of bias sensors and bias triggers; however, the scope of allowable touchpoints is partially constrained by the data sources available in the cyber range testbed.

Bias sensors will use data accessible to cyber defenders to determine the extent a particular cognitive vulnerability is present in a cyber attackers. Bias sensors will be developed into software components for use on a network or host where the needed cyber defender data can be made available. Performers will provide at least one established method as a validity check for each bias sensor delivered. The bias sensors

must use data typically available to cyber defenders, while the established methodologies can use other data sources (i.e., psychometric questionnaires) or sensors (i.e., physiological devices).

Phase 1 research and development must include **two required CogVuls** selected by IARPA, loss aversion and the representativeness bias, and at least 3 additional CogVuls proposed by the Offeror. Selection criteria will include novelty, variety, relevance, quantity, scientific rigor, potential impact, etc.

For the purpose of this effort we define the terms as follows:

- **Loss aversion**[3] is the tendency for people to strongly prefer avoiding losses to acquiring equivalent gains.
- **Representativeness bias**[4] is the tendency to overweight the representativeness of a piece of evidence while ignoring how often (i.e., its base rate) it occurs.

Performers will develop bias triggers to interact with or adapt to attackers (or portions of the network or host the attacker accesses) based on observables collected by the bias sensors and create situations in the cyber domain that induce and exploit each of the CogVuls. A bias trigger will activate or increase a CogVul. This increase should be measurable by a bias sensor or established method.

Relevant for Phase 1, Offerors should clearly describe:

- The justification for hypothesizing that each included cognitive vulnerability is exploitable to reduce cyber attacker effectiveness, and at least one proposed bias sensor and at least one proposed bias trigger that can be developed for each.
- Details on how each included vulnerability is relevant to attacker cognition and behavior, including which stage of the cyber kill chain it pertains to.
- The justification for hypothesizing that exploiting a subset of the planned vulnerabilities will, in combination, meet required thresholds for at least one of the cyber behavioral impact metrics.
- Statistically efficient experimental design plan(s) to full investigate the CogVuls (at least 5), bias sensors (at least one per vulnerability) and bias triggers (at least one per vulnerability).
- Plans must include Appropriate sample size and participant composition. Smaller sample sizes consisting of more highly skilled participants are preferable; effect size must be calculated. Any use of non-cyber proficient participants or non-cyber scenarios must be highly justified.

Phase 1 will include development of a preliminary structured visual representation of hypothesized or known relationships between CogVuls, bias sensors and triggers, relevant characteristics of the attacker, the network, and the external environment, and various cyber behavioral impacts. Offerors will propose a visual representation, (i.e., concept map, ontology, taxonomy) to clearly display these relationships. Structured visual representations must be driven by theory; a ―shotgun approach will not be favorably evaluated. Relevant theory from a variety of disciplines may direct the research and shall be discussed in the proposal. These structured visual representations will guide Performer teams' activities and will be refined over throughout the program.

Literature that differentiates the impact of various CogVuls from each other is limited, so Performers may elect alternate approaches that describe and relate cyber behavioral impact. If alternate approaches are included, they must have clear theoretical foundations, either from existing literature, or previous work in the area.

[3] Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. Econometrica, 47(2), 263–292.
[4] Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3(3), 430–454.

Phase 1 shall have a duration of 18 months. Additional requirements include:

- Design and execute bias discovery experiment(s) for selected and required CogVuls; experimental design and data collection must include a sufficient number of participants to calculate statistical differences (i.e., effect size, variability) with sufficient skill level to support ecological validity. Sample size efficient designs are acceptable; number of participants, skill level, and recruitment plan must be justified.
- Offerors should account for a sufficient number of cognitive vulnerabilities, bias sensors, and bias triggers to account for potential construct failure.
- Provide established methods and evaluation thresholds to determine ground truth of presence of each of the CogVuls.
- Prepare provided simulated environment for inducing, and measuring selected CogVuls, and perform self-testing for bias sensors and triggers.
- Ethics review (i.e., Institutional Review Board (IRB) approval or exemption will be required for all phases. Performer teams must have access to an ethics review board, and expertise on navigating the process.

2.2 Phase 2

The goal of Phase 2 is to create Cyberpsychology-informed Defenses (CyphiDs) to impose a cyber penalty and thwart attacker success across the Cyber Kill Chain. A set of bias sensors and bias triggers developed for a specific cognitive vulnerability, or cognitive vulnerability cluster, will be considered a CyphiD as shown in *Figure 2*. CyphiDs consist of one or more bias sensors which measure the presence of a CogVul, logic to determine based on bias sensor output (and other cyber data, as needed) which bias trigger to utilize (if any), and one or more trigger(s) which create a cyber situation to induce, exploit, or intensify the CogVul. In Phase 2, performers will develop the CyphiD software and logic that links sensors and triggers. Offerors must include an implementation plan for CyphiDs, including logic, and proposed mappings of bias sensors to bias triggers.

Additional bias sensors and bias triggers may be developed in Phase 2 based on Phase 1 experimental findings or additional HSR. Offerors should discuss how experimental design(s) will allow for quick additional HSR in Phase 2, as needed. Performer teams will need to produce at least 5 CyphiDs that impact early kill chain attacker behavior, and at least 5 CyphiDs that impact late kill chain behavior (late is defined as post exploitation). A CyphiD that is effective for both early and late kill chain behavior, may count for both categories. Teaming is strongly encouraged to accomplish these goals. Performers will create a simple, custom dashboard to assist defenders in understanding the findings of the sensors (i.e., the degree of each cognitive vulnerability measured) and the impact of the CyphiDs. Phase 2 metrics will focus on achieving a *medium* effect size across multiple areas of defender goals. It is not expected that each CyphiD meet each cyber behavioral impact requirement threshold (See Table 4), but rather the performer's *collection* of CyphiDs meets each at least once.

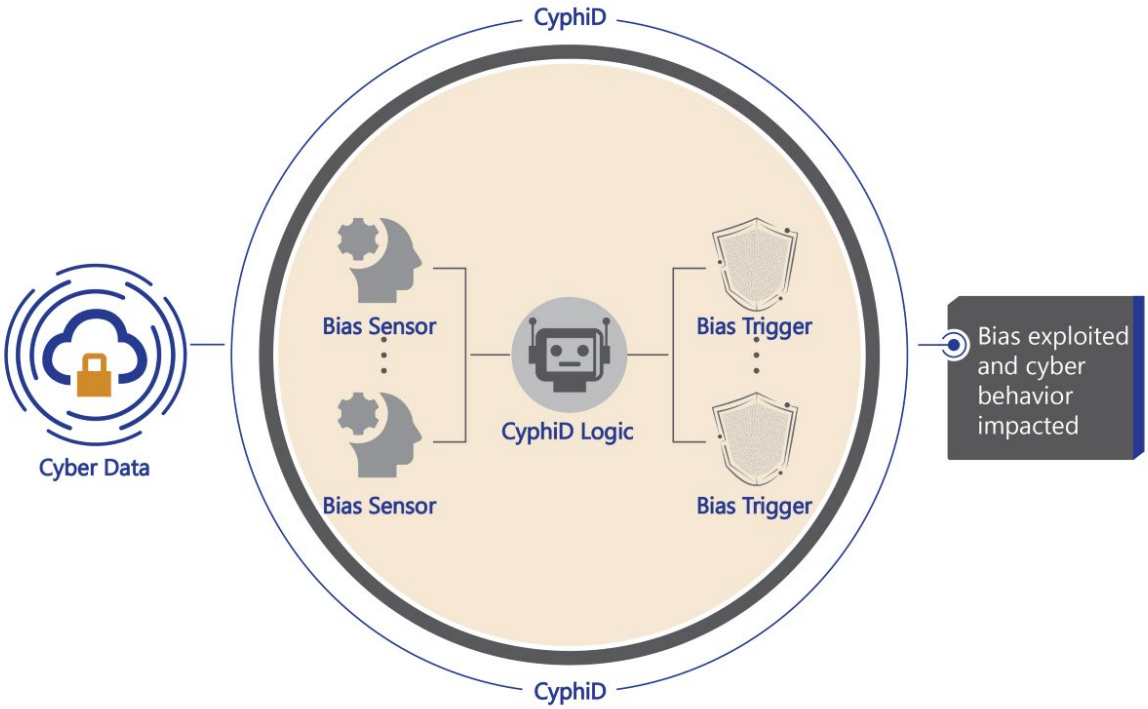

Figure 2: Notional CyphiD Graphic: Bias Sensors, Bias Triggers, and Logic Combine to Form CyphiDs. Each CyphiD focuses on achieving a desirable cyber behavioral impact.

In Phase 2, performers will continue to update their structured visual representation and implement relevant portions into their software to demonstrate under which conditions a particular CyphiD should be used for the seven cyber behavioral impacts. Attacker attributes and/or situational factors can be observed by bias sensors and exploited by the CyphiDs, while network and host characteristics can be altered by the bias triggers. These features should be included in the structural representation and examined throughout the research.

Self-testing of CyphiDs within the cyber range environment will be performed iteratively by performers with results and interpretation of results delivered to IARPA. A leaderboard will be provided to track the successes of each team's CyphiDs against the program metrics.

Phase 2 shall have a duration of 15 months. Additional requirements include:

- Improve (and/or create new) bias sensors and bias trigger to achieve Phase 2 program metrics.
- Request and justify any additional cognitive vulnerability-specific metrics to be included for HSR T&E events.
- Fully document each CyphiD (which will be used across all performer teams during Phase 3).

2.3 Phase 3

The goal of Phase 3 is to automate, model, and improve research findings from previous phases, while reaching higher effect sizes. Performers will develop an adaptive psychology-informed defense (APhiD), which automatically selects the appropriate combination or sequence of CyphiDs over time (*See Figure 3).* Performers will also research and develop a cyber-specific computational models (C3M) based on experimental findings to date; successful models will reflect and predict attacker behavior with sensitivity

to various conditions listed in the structured visual representation and adjust based on bias sensor measurements of each CogVul.

In Phase 3, all Performers will be working from an integrated structured visual representation provided by the IARPA team that is based on elements from each Performer team's Phase 1 and 2 contributions. Performers will incorporate relevant portions of the structured visual representation to provide *a priori* knowledge to the APhiD and determine situations in which a particular CyphiD would be selected, including various attacker attributes, attacker behaviors, network attributes, mission context, situational attributes, and/or time factors.

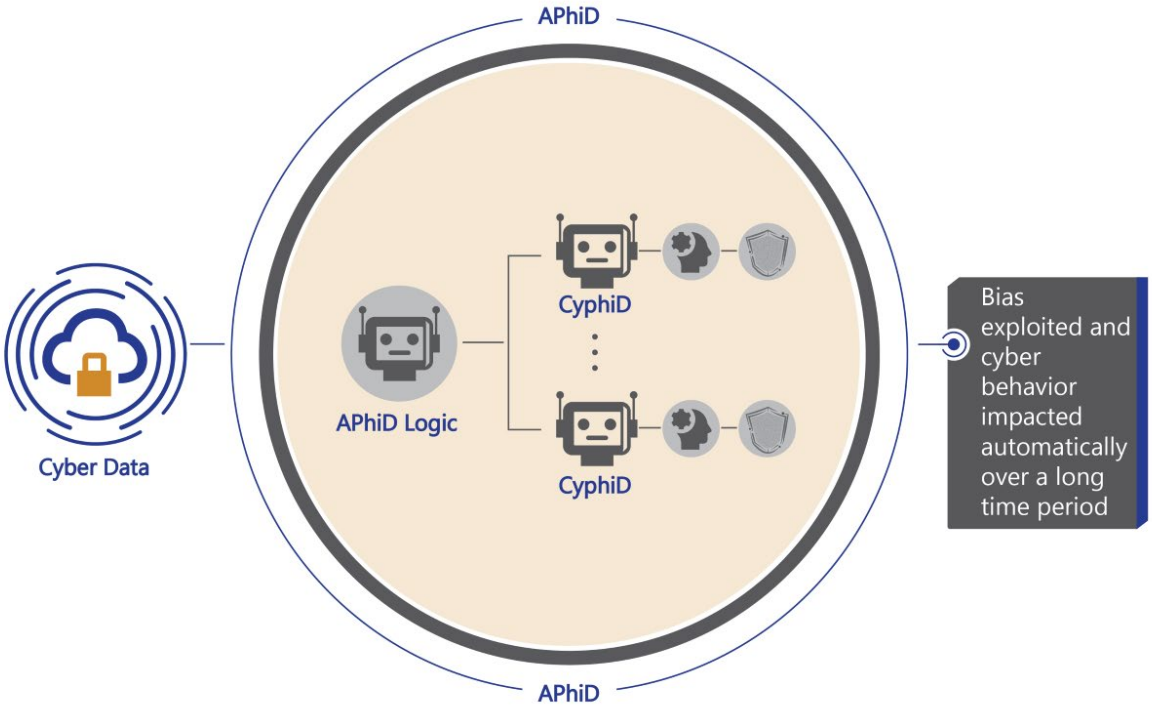

Figure 3: Notional APhiD Graphic: Multiple CyphiDs combined with logic determining which to use at each point in time form an APhiD.

APhiDs must run autonomously by creating intelligent algorithm(s) (e.g., expert system, game theory, artificial intelligence, rule-based system) to select the optimal combination or sequence of CyphiDs throughout an extended period of time. IARPA will provide performer teams with an annotated dataset for training their APhiDs, as well as all available performer team CyphiDs for optional inclusion. Performers will use the dataset(s) provided by IARPA, as well as performer data and findings from their Phase 1 HSR, to develop C3Ms and APhiDs. Offerors must include an initial implementation plan for APhiDs.

The C3Ms will focus on the CogVuls examined by the performer in Phase 2. The anticipated changes in behavior caused by each of the performer's CyphiDs should be handled by at least one model. C3Ms should aim to recapitulate the pattern of human behavior as it relates to human decision-making and behavioral changes in a cyber attack scenario. Any modeling effort should use ecologically relevant sensor measurements, which can be reasonably obtained in cyber security environments such as security operations centers (SOCs). It is important to model an adjustable degree of bias on a continuous scale, such that the model behavior changes based on the degree of bias presence selected in the model. Excessively

complex non-linear models are discouraged due to overfitting concerns. Performers may include their C3M(s) as part of their APhiD logic. Offerors must include an initial implementation plan for the C3M(s).

Phase 3 shall have a duration of 12 months. Additional requirements include:

- Store, process, and understand the large-scale HSR dataset provided by IARPA
- Validate APhiDs with self-testing in cyber range testbed and submit findings and interpretation.
- Develop simple, custom dashboard to describe how sensors, triggers, and APhiDs are working and their impact on attack behavior.
- Develop a sufficient number of C3M(s) to define and emulate all the CogVuls examined by the performer in Phase 2.
- Perform iterative testing of cognitive models against the training data provided, report all data and interpretations of data and analyses to IARPA.
- Provide updates to the common structural representation based on findings and interpretation of results.

3 Team Expertise

To address the combination of challenges presented by ReSCIND, **collaborative efforts and teaming arrangements among Offerors are strongly encouraged.** It is anticipated that teams will be multidisciplinary and may include expertise in one or more of the disciplines listed below. This list is included only to provide guidance for the Offerors; satisfying all the areas of technical expertise below is not a requirement for selection and unconventional or innovative team expertise may be needed based on the proposed research. Specific content, communications, networking, and team formations are the sole responsibility of the participants. Proposals should include a description and the mix of skills and staffing that the Offeror determines will be necessary to carry out the proposed research and achieve program metrics.

- Behavioral science and cognitive psychology
- Defensive cyber operations
- Cyber attack modeling
- Penetration testing/red teaming and adversary emulation
- Artificial intelligence and adaptive systems
- Statistical data analysis and mathematical modeling
- Software development and engineering
- Criminology
- Cognitive and neurosciences
- Human factors engineering
- Human computer interaction
- Computer security and network security
- Cognitive modeling

4 Program Scope and Limitations

Proposals shall explicitly address all of the following:

- **Underlying theory**: Proposed strategies to meet program-specified metrics must have firm theoretical bases that are described with enough detail that reviewers will be able to assess the viability of the approaches. Proposals shall properly describe and reference previous work upon which their approach is founded.
- **Research & Development approach**: Proposals shall describe the technical approach to meeting program metrics.

- **HSR protocols**: Proposal appendix shall describe the approach for recruiting human subjects and ensuring ethical treatment and responsible data handling. Experimental procedures must discuss and justify design decisions supporting or limiting internal, external, ecological and construct validity.
- **IRB approval:** Proposal appendix shall describe the approach for obtaining timely IRB approval for all phases of experimentation and any required modifications; performer teams must ensure IRB approval prior to conducting HSR.
- **Data analysis strategy**: Proposals shall describe how HSR protocols will yield data that can meet program metrics through both qualitative and conventional statistical analyses and articulate the reasoning behind any nonparametric or otherwise atypical analytical approaches.
- **Technical risks and Mitigations:** Proposals shall identify technical risks and proposed mitigation strategies for each.
- **Software development**: Proposals shall describe the approach to software architecture and integration.

The following areas of research are **out of scope** for the ReSCIND program:

- Research that does not have strong theoretical and experimental foundations.
- Research that cannot be implemented to facilitate identification or development of a CyphiD.
- Bias sensors that require data not easily obtainable by cyber defenders.
- Bias sensors or triggers designed to solely target a non-human cyber attacker.
- Bias triggers that do not have a cyber behavioral impact.
- Technologies focused solely on cyber deception or traditional cyber defenses.
- Attacker activity that occurs prior to network access (i.e., OSINT research)
- Attacker activity relying on interaction with live humans (i.e., social engineering), including defenders or users.
- Hardware solutions.
- Attribution of specific ATPs or cyber actors; techniques solely focused on intelligent gathering.
- Approaches that require access to classified information or data. All performer research will be strictly unclassified.

5 Test and Evaluation (T&E)

T&E will be conducted by an independent team of FFRDC, UARC or Government staff carrying out evaluation and analyses of Performer research deliverables using program tests and protocols. In addition to independent T&E, the program will regularly gauge interim progress of Performer research activities towards ReSCIND objectives and target metrics using T&E results measured and reported by the Performer teams themselves. The ReSCIND Program will pursue rigorous and comprehensive T&E to ensure that research outcomes are well characterized, and deliverables are aligned with program objectives. Such T&E activities will not only inform Government stakeholders on ReSCIND research progress but will also serve as valuable feedback to the Performers to improve their research approaches and system development. The ReSCIND Program will work closely with Government leaders in cyber operations and cyberpsychology to continually refine and improve T&E methodologies.

A series of HSR experiments will be conducted by T&E throughout the program to test performer solutions and generate datasets. These HSR experiments will be used to evaluate the effectiveness of the Performer's capability at various stages of the program. Performers will also be required to demonstrate execution of

their developed capabilities (i.e., bias sensors, bias triggers, CyphiDs, APhiDs) using a cyber range testbed provided by the T&E team.

Phase 1*: To identify and provide evidence of CogVuls relevant to cyber attack behavior and create novel methods to measure and induce changes in cyber attack behavior and success,* Performers will conduct their own HSR in Phase 1; this will occur independently from the cyber range testbed. Their initial experimental design will be scored with a rubric by a group of subject matter experts (SME), and act as a midterm T&E event for Phase 1. The final experimental execution, data analysis, and interpretation of findings will be evaluated as the final T&E event for Phase 1. The data analysis must include both qualitative and quantitative results. Raw and curated data collected by performers to measure effect size must be provided to T&E for validation. Bias sensors and triggers will be evaluated for integration into the cyber range testbed. Additionally, the accuracy of the bias sensors will be evaluated by examining the extent to which bias sensor results reflect previously established and validated measures, which refer to approaches that are widely used and well-accepted among related fields of research to accurately measure the outcome of interest. The bias sensors must use data typically available to cyber defenders, while the established methodologies can use other data sources and sensors. T&E may also include additional established methods and/or independent approaches as part of their evaluation. The T&E team will validate that bias triggers have the desired behavioral effect on attacker behavior (i.e., increase or decrease a specific CogVul) by examining effect size.

Phase 2: *To evaluate the impact of CyphiDs on both early and late stages of a cyber attack,* controlled HSR experiments with cyber experts will be executed by T&E across. The first T&E event will focus on Early Kill Chain CyphiDs, and the later event on the late Kill Chain CyphiDs. Phase 2 T&E will include evaluation of new (or improved) bias sensors; if any new HSR data is collected, it will also be provided to T&E and effect sizes calculated.

Phase 3: *To evaluate that the cyber-specific computational cognitive models (C3M) reflect and predict attacker behavior changes in reaction to CyphiD interventions,* T&E will use typical model fitting metrics to examine how C3M reflect the data provided. Additionally, T&E will examine the predictive power of the C3M using a testing dataset.

To evaluate the adaptative psychology-informed defenses (APhiDs), an online open prize competition in the format of a capture-the-flag (CTF) T&E event will be held to test performer solutions against a wider range of attack behaviors and attacker attributes.

Experimental analysis results will be utilized to iteratively improve the cyberpsychology-inspired methods and techniques. Performers are encouraged to work with T&E and propose and justify additional data to be collected, CogVul specific metrics needed, CyphiD/APhiD specific flags, and additional characteristics of the participants to be measured and examined during T&E events, which may be included at the discretion of the IARPA PM. Additional relevant and reasonable observables, variables, or metrics that are supported by theory or prior research will be favorably evaluated. The IARPA Team may conduct other supplemental evaluations or measurements at any time and without notice.

6 Program Data

Across phases, the T&E team will conduct HSR (including data collection/curation) using cyber experts. These experiments will collect cyber attacker behavior and performance data, though a realistic mock cyber attack scenarios. During Phase 1, the only data the performers will obtain is the data the performers generate themselves through their bias discovery experiments. During Phase 2, performers will again rely on the data they have collected through HSR and self-testing within the testbed. Results from the Early Kill Chain

T&E Event will be provided to performers during Phase 2 and should be used to improve their deliverables for the Late Kill Chain T&E Event. During Phase 3, the government will provide performers a more fully curated dataset created by T&E from the Phase 2 HSR. Performers will also be required to share their CyphiDs across teams to expand the collection of available options for AphiDs selection.

Table 2 describes the kinds of data that may be collected by T&E. An updated list of available planned data will be made available at Phase 1 kickoff. Offerors are encouraged to notate any additional data requested relevant to the specific CogVuls, CyphiDs or APhiDs they propose. Performers will develop CyphiDs and APhiDs for a standard IT network with typical enterprise targets such as, Windows or Linux operating systems, Domain Controller, Git Repository, routers, development workstations, database and file share servers, multiple subnets and target environment will not include removable media, live users or admins.

Table 2: Examples of the Types of Data that Could be Collected in the T&E Environment

Data Type	Data Example
Scenario Data	Subject ID, date, day, condition, environment, daily start/end time, breaks/lunch, subject start/end time, cyber task end time, subject time on task, screen capture
Environment Data	Subject IP, target IPs, target host configuration (e.g., OS, ports), host name, vulnerabilities
Network Data	Packet ID, PCAP, netflow, PCAP timestamp, destination IP, PCAP size, source IP, destination IP, port, timestamp
Host Data	Process logs, file touches, services, process history, file data, system & application host logs
User Data	User accounts, access logs, privilege, user files, login attempts
Attack Data	Exploit timestamp, exploit name, exploit CVE, success/failure
Alert Data	Signature ID, IDS alert description, CVE, severity, target IP, timestamp, custom alerts
Forward Progress	Flags captured, data exfiltrated, lateral movement, privilege escalation
Self-Report Data	Timestamp, self-reported vulnerabilities identified, self-reported exploit attempts, self-reported success/failure, Red Team Briefing
Individual Measures	Bias-specific questions, Reported Cognitive State, Experience, Demographics, General Decision-Making Style Inventory (GDMSI), Indecisiveness Scale (IS), Big Five Inventory (BFI-44), custom questionnaires
CyphiD Data	To be included in proposal by Offerors
APhiD Data	To be included in proposal by Offerors

Performers can expect solutions to defend against cyber attackers using standardized, openly available cyber attack tools including Kali Linux and included toolsets such as Metasploit, Armitage, Burpsuite, etc. Attackers may create custom attack scripts but will not have access to proprietary or commercial toolsets, prior developed scripts and attack tools and resources hosted external to the attack network. Attackers will be given high level goals but will not be told how to execute their attack or strategies to compromise the target network, so any malicious activity possible on the network may be expected.

Data that is out-of-scope for collection and use in this program include: OSI Layer 1 data, hardware or infrastructure components, OSINT, social engineering, Internet of Things (IOT), Industrial Control Systems (ICS) or SCADA devices, mobile/cellular devices, close access or physical interactions, RFID, radio frequencies, and tactical networks.

Performers will be provided attack scenarios to focus on during Phase 1 kick-off. These scenarios will include enterprise network layout and devices and focus on attack scenarios described in Table 3.

Table 3: Descriptions of Various Cyber Attack Event Types.

Cyber Attack Event Type	Description
Software Supply Chain Attack	Software supply chain attacks including, supply chain espionage, malware injected into software development process, and deployment into target domain, compromising software development infrastructure, and compromising certificate update and signing process.
Data and Intellectual Property (IP) Theft	Activities taken to identify and strategically exfiltrate data and intellectual property related to mission objectives.
Malicious Data Modification	Data modification on target environment with the objective of triggering external events related to system access log files, system alerts, notification triggers, or restricting role-based access control.
Denial of Service	Targeted denial including placement of ransomware on critical targets, distributed denial of service (DDoS) of a specific service, and targeted reuse of these attacks toward final objective.

7 Program Metrics

Achievement of metrics is a performance indicator under IARPA research contracts. IARPA has defined ReSCIND program metrics to evaluate effectiveness of the proposed solutions in achieving the stated program goal and objectives, and to determine whether satisfactory progress is being made. The metrics described in this BAA are shared with the intent to scope the effort, while affording maximum flexibility, creativity, and innovation to Offerors proposing solutions to the stated problem. Proposals with a plan to exceed the defined metrics in one or more categories are desirable, provided that all of the other metrics are met, and provided that the proposals provide clear justification as to why the proposed approach will be able to meet or exceed the enhanced metric(s).

The final ReSCIND T&E protocols and evaluation methodology are currently under development; further details may be provided at program kickoff. Program metrics may be refined during the various phases of the ReSCIND program; if metrics change, revised metrics will be communicated in a timely manner to Performers. The evaluation methodology may be revised by the Government at any time during the program lifecycle to better meet program needs.

Phase 1 will include two types of metrics: Statistical metrics and qualitative metrics (Table 4). Statistical metrics are designed to establish external validity and efficacy of bias sensors and bias triggers. Qualitative metrics are designed to establish internal validity of the Performers' experimental design strategies using structured expert evaluations. Effect size will be measured using Cohen's *d*; $d=M_1-M_2/SD$[5] for parametric

[5] Vogt, W.P. & Johnson, R. B. (2015). The SAGE Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences, 5th Ed., Sage Publications, Inc

data, with Cohen's *d* analogs[6] used in the case of non-parametric data. The degree to which the bias sensors overlap with the validated measure(s) will be measured using standard deviation (SD); $SD=\sqrt{\sum(X-\mu)^2/N}$.[7] For Phase 1, a *medium* effect size is expected. In cases where sensors do not converge with validated methodologies using, alternative evidence of convergence may be proposed by Performers and evaluated by the T&E team.

Table 4: Statistical and Qualitative Metrics Used in Phase 1.

Statistical Measures	**Phase 1 Target**
External validity check	Bias sensor: within 1.5 SD of baseline
Higher effect size	Bias trigger: Cohen's $d \geq 0.3$
Qualitative Metric	**Phase 1 Evaluation**
Manipulation and validity check	Experimental design: SME Rubric

Phases 2 and 3 include two types of metrics: Behavioral metrics and statistical metrics (Table 5). Behavioral metrics are designed to establish that specified defender goals are achieved to decrease cyber attacker success and effectiveness. Demonstration of cyber behavioral impact compared to a control condition will indicate that Performer solutions are achieving ReSCIND metrics in Phase 2 and will be improved with automation in Phase 3. Medium-to-high levels of effect size will be used to evaluate Phase 2 performance, while an effect size approaching high is expected in Phase 3. Phase 3 cognitive computational models will be evaluated by testing model fit and predictive ability against datasets collected throughout the phases using root mean squared error ($RSME= \sqrt{[\ \Sigma(P_i - O_i)^2 / n\ ]}$).[8]

Table 5: Cyber Behavioral Impact and Statistical Metrics for Phases 2 and 3.

Cyber Behavioral Impact	**Behavioral Metrics**	**Phase 2 Target**	**Phase 3 Target**
Decrease Rate of Attack Success	Attack success vs. HSR control	50% ≤ baseline	APhiD: 10% improvement on best team's Phase 2 results for each cyber behavioral impact
Increase Time to Task Completion	Time to task completion vs. HSR control	50% ≥ baseline	
Decrease Progress Towards Goal	Progress to goal vs. HSR control	50% ≤ baseline	
Decrease in Time Until Detection	Time to detection vs. HSR control	50% ≤ baseline	
Decrease Defender Effort Spent	Decreased defender effort vs. HSR control	50% ≤ baseline	
Increase Attacker Cognitive Effort Spent	Attacker effort vs. HSR control	50% ≥ baseline	
Increase Attack Resources Wasted	Attack resources wasted vs. HSR control	50% ≥ baseline	
Cyber Behavioral Impact	**Statistical Metrics**	**Phase 2 Target**	**Phase 3 Target**
For all 7 Cyber Behavioral Impacts	Higher effect size	CyphiD: $d \geq 0.5$	APhiD: $d \geq 0.7$
	Predictive power	N/A	Models: RMSE ≤ 0.2

[6] Wilcox, R. (2019). A Robust Nonparametric Measure of Effect Size Based on an Analog of Cohen's d, Plus Inferences About the Median of the Typical Difference. Journal of Modern Applied Statistical Methods, 17(2), Article 1.

[7] Vogt, W.P. & Johnson, R. B. (2015). The SAGE Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences, 5th Ed., Sage Publications, Inc

[8] Vogt, W.P. & Johnson, R. B. (2015). The SAGE Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences, 5th Ed., Sage Publications, Inc

8 Program Waypoints, Milestones, and Deliverables

Waypoints, Milestones, and Deliverables are established from the program's onset to ensure alignment with ReSCIND objectives, organize research activities in a logical and reportable manner, and facilitate consistent and efficient communication among all stakeholders – IARPA, ReSCIND T&E, USG Stakeholders, and Research Performers (*see Table 6*). A schedule of key program Milestones and Deliverables is shown in *Figure 4*. Performers shall provide results from self-testing to be included in ReSCIND leaderboard. T&E results may also be included.

Table 6: Table of ReSCIND Program Deliverables and Milestones.

Phase	Month	Event	Description	Comments	Deliverables
1-3	all	Waypoint	Monthly Status Report	Due on 15th of each month	MSR
1-3	all	Waypoint	Progress and Status Meetings	Monthly teleconference with IARPA Team, additional as needed	Meeting Notes
1	1	Waypoint	Phase 1 Kickoff	Location TBD	N/A
1	2	Deliverable	IRB Submission	Also any modifications	All IRB Docs
1	3	Deliverable	Structural Visual Representation	Updated to focus on n experimental design(s)	Report, visualization
1	4	Deliverable	Draft Experimental Design(s)	Performer methods, materials, analysis plan	Report
1	5	Waypoint	Site Visit	Onsite at Performer location.	N/A
1	5	Milestone	T&E Event	SME evaluation of draft experimental designs	N/A
1	8	Waypoint	IRB Approval	N/A	IRB Approval Document
1	8	Deliverable	Final Experimental Design(s)	Includes established methodologies for external validation.	Report
1	9	Waypoint	PI Review Meeting	N/A	N/A
1	10	Deliverable	Bias Sensors and Triggers Materials	For mandatory CogVuls	Software, Documentation, Testing Procedure
1	11	Deliverable	Experimental Results	Data analysis and interpretation of results for mandatory CogVuls	Report, Data, and all Experimental Materials
1	11	Waypoint	Site Visit & Demo	Onsite at Performer location	Demo of completed experiments
1	14	Deliverable	Bias Sensors and Triggers Materials	For additional CogVuls	Software, Documentation, Test Suite
1	15	Waypoint	Site Visit & Demo	Onsite at Performer location.	Demo on all CogVuls
1	15	Deliverable	Experimental Results	Data analysis and interpretation of results for additional CogVuls	Report, Data, and all Experimental Materials

Phase	Month	Event	Description	Comments	Deliverables
1	16	Waypoint	T&E Event	SME evaluation of final experimental results	N/A
1	17	Deliverable	Phase 1 Final Report	Final Phase 1 Report; include Phase 2 implementation plan	Final Report
1	17	Deliverable	Bias Sensors and Triggers Materials	Includes any changes	Software, Documentation
1	18	Waypoint	End of Phase 1 PI Meeting & Demo	In DC. Will include demo for stakeholders.	N/A
2	19	Waypoint	Phase 2 Kickoff	Takes place in San Diego	N/A
2	19	Deliverable	IRB Amendments for Additional HSR	Additional HSR should focus on same CogVuls	All IRB Documentation
2	20	Deliverable	Experimental Design(s)	Experimental design(s) for all additional HSR.	Report
2	23	Waypoint	IRB Approval for Additional HSR	Approval must be submitted prior to HSR execution	IRB Approval Document
2	24	Deliverable	Report on completed additional HSR	Data analysis and interpretation of results	Report, Data, and all Experimental Materials
2	24	Waypoint	Site Visit & Demo	Onsite visit to Performer location.	Demonstrate Early Kill Chain
2	25	Deliverable	Updated Bias Sensors and Triggers	Source code, and executables, along with setup and testing documentation.	Software, documentation, testing
2	25	Deliverable	Early Kill Chain CyphiDs	Integratable into cyber range testbed.	Software, Executables Test Suite
2	26	Waypoint	PI Meeting	Location TBD	N/A
2	26-27	Waypoint	T&E Event	Early Kill Chain HSR	N/A
2	28	Waypoint	Phase 2 IRB Approval	To handle and analyze Phase 2 HSR data collected by T&E Team	IRB Approval Document
2	30	Deliverable	Deliver Late Chain CyphiDs	Integratable into cyber range testbed.	Software, Executables, Test Suite
2	30	Waypoint	Site Visit & Demo	Onsite visit to Performer location.	Demonstration to include Late Kill Chain in testbed
2	31	Waypoint	T&E Event	Late Kill Chain HSR	N/A
2	32	Deliverable	Updated Structural Representation	Based on additional HSR and T&E event results	Report, visualization
2	33	Waypoint	PI Meeting	Location TBD	N/A
2	33	Deliverable	Final Report on all CyphiDs	Based on all T&E results and HSR to date.	Final Report, updated software
3	34	Waypoint	Phase 3 Kickoff	Takes place in San Diego	N/A
3	36	Deliverable	Implementation plan for APhiDs	Includes algorithms to select the combination/sequence of CyphiDs	Report

Phase	Month	Event	Description	Comments	Deliverables
3	36	Deliverable	Implementation plan for C3Ms	Includes algorithms to model cyber behavior and CogVuls	Report
3	37	Waypoint	Site Visit	Onsite to Performer location.	N/A
3	39	Waypoint	PI Meeting	In D.C.	N/A
3	42	Deliverable	Deliver APhiDs	Includes visualization, source code, documentation, libraries, binaries	Software, executable
3	42	Waypoint	Site Visit & Demo	Onsite to Performer location.	Demonstrate APhiD and C3M
3	43	Milestone	T&E Event	Online CTF Prize Competition	N/A
3	44	Deliverable	Deliver final C3Ms	Includes visualization, source code, documentation, libraries, binaries,	Software, executables
3	42	Waypoint	Site Visit & Demo	Onsite to Performer location.	Demonstrate APhiD and C3M
3	43	Milestone	T&E Event	Online CTF Prize Competition	N/A
3	44	Deliverable	Deliver final C3Ms	Includes visualization, source code, documentation, libraries, binaries,	Software, executable
3	44	Waypoint	T&E Event	Evaluation of C3M	N/A
3	45	Report	Final Report	Any updated software and documentation are due.	Final Report
3	45	Waypoint	Final PI Meeting & Demo	Takes place in D.C.	Demo for stakeholders

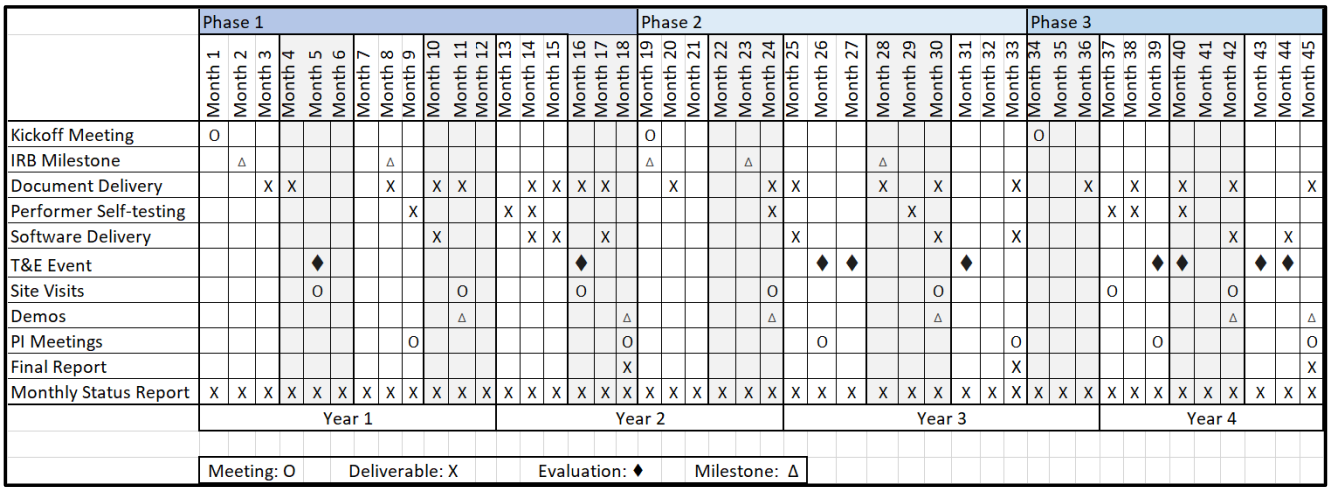

Figure 4: Graphical Schedule of ReSCIND Program Deliverables and Milestones.

9 Software Deliverable Formatting

The ReSCIND Program will use a standardized API for all software deliverables and evaluations. The first version of the ReSCIND API will be provided to Performers at the Phase 1 Kickoff Meeting and updated periodically thereafter. The API will define function calls, data structures, and data pipeline and management for CyphiD and APhiD integration, testing, and operating and evaluating ReSCIND software

in a standardized manner. A secondary API may be necessary for sensor registration to the testbed itself. The API will be as software and hardware agnostic as is practical, to ensure Performers can freely develop solutions according to skill and vision.

Each team is required to include among their key personnel a Lead System Integrator (LSI) who shall be responsible for preparing software deliverable subcomponents, modules, and systems, performing quality control of deliverable(s), and assisting the T&E team with aspects of integrating key components into the primary ReSCIND testbed. The LSI will also oversee communication and coordination across a Performer's research teams including subcontractors, if applicable, to ensure research products are functional and following software coding best practices and requested security controls.

CyphiDs will be designed to be run from within a container for ease of use and portability. Deliverables will include the container configuration and all files necessary to run the CyphiD. Final deliverable for each CyphiD and APhiD will include the full software development package including any source code, containers, working binary executable, test suite, and documentation. Binaries for each CyphiD will be shared with other Performers during Phase 3 and used as part of ReSCIND prize competition.

10 Meeting and Travel Requirements

All Performer teams are expected to attend workshops, technical meetings and other designated meetings to include key personnel from prime and subcontractor organizations.

The ReSCIND program intends to hold a program Kick-off Meeting workshop in the first month of the program and first month of each subsequent program phase. In addition, the program will hold PI Review Meetings (three in Phase 1, two in Phase2, and two in Phase 3). Meetings may be combined for logistical convenience. The dates and locations of these meetings are to be specified at a later date by the Government, but for planning purposes, Offerors should use the approximate times listed in *Table 6* and assume half the PI and kick-off meetings will be on the East Coast (e.g., D.C. area) and half on the West Coast (e.g., San Diego, CA area). IARPA may opt to co-locate the meeting with a relevant external conference or workshop to increase synergy with stakeholders. IARPA reserves the right to change meeting locations and conduct additional site visits on an as-needed basis or virtually, if desired.

Kick-off Meetings will typically be one day in duration and will focus on plans for the coming Phase, Performer planned research, and internal program discussions. PI Review Meetings will typically be two days in duration and will have a greater focus on communicating program progress and plans to USG stakeholders. These meetings will include additional time allocated to presentation and discussion of research accomplishments.

In both cases, the workshops will focus on technical aspects of the program and on facilitating open technical exchanges, interaction, and sharing among the various program participants. Program participants will be expected to present the technical status and progress of their projects to other participants and invited guests. Individual sessions for each Performer team with the ReSCIND PM and T&E Team may be scheduled to coincide with these workshops. Non-proprietary information will be shared by Performers in the open meeting sessions; proprietary information sharing shall occur during individual breakout sessions with the ReSCIND PM and T&E.

Site visits by the Government Team will generally take place semiannually during each phase. These visits will occur at the Performer's facility. Reports on technical progress, details of successes and issues, contributions to the program goals, and technology demonstrations will be expected at such site visits. IARPA reserves the right to conduct additional site visits on an as-needed basis.

11 Glossary of Terms

The following table describes key terms and their definitions in the context of the ReSCIND program.

Table 6: Summary of Key Terms.

TERM	Definition in the context of the ReSCIND Program
Advanced Persistent Threat (APT)	A prolonged and targeted cyberattack using continuous, clandestine, and sophisticated hacking techniques to gain access to a network and remain undetected for an extended period of time, with potentially destructive consequences.
Adaptive Psychology-informed Defense (APhiD)	An AI-guided combination of logic and CyphiDs that dynamically responds to attacker behavior and attributes with a tailored defensive strategy to mitigate attacker success by imposing a cyber penalty.
Attacker Attributes	Behavioral, cognitive, and demographic characteristics of an adversarial human actor (including but not limited to motivation, experience, solitary vs team activity, emotional state, or targeted vs. opportunistic activity) which can be observed with cyber data and exploited for cyber-defensive purposes.
Bias Sensor	Measure of cognitive vulnerability that can be exploited to mitigate attacker success using data available to cyber defenders.
Bias Trigger	Network or host manipulations or other interactions that induce a cognitive vulnerability on a cyber attacker.
Cognitive Biases	Subconscious,[9] systematic errors in thinking that cause misinterpretation of information or deviations from rationality.
Cognitive Vulnerabilities (CogVuls)	An umbrella term encompassing cognitive and decision-making biases, innate cognitive limitations, emotional or mental state, or physiological vulnerabilities that can result in reduced cyber attacker success or effectiveness.
Cognitive Vulnerability Cluster[10]	A group of related CogVuls that are related, co-occur, manifest behaviorally, or incite a similar cyber behavioral impact.
Cyber Operators	The humans performing cyber operations, both defensive (e.g., Incident Response Team, Blue Team, security operations center, Cyber Protection Team) and offensive (e.g., unauthorized/illegal hacker, advanced persistent threat (APT), ethical/legal hacker, Red Team).
Cyberpsychology	The scientific field that integrates human behavior and decision-making into the cyber domain, allowing us to understand, anticipate and influence cyber operator behavior.
Cyberpsychology-informed Defense (CyphiD)	A combination of bias sensors, logic, and bias triggers which generates a novel defensive strategy to mitigate attacker success by imposing a cyber penalty.
Cyber Penalty	Costs (i.e., wasted time, wasted resources) imposed on a cyber attacker designed to mitigate success, using techniques including but not limited to denial, delay, degradation, detection, disruption, or deception.
Human Limitations	Behavioral, social, cultural, physiological or other patterns that are potentially exploitable via cyber operations.
Mission Context	Cyber-relevant details about the goals, constraints, and characteristics of the mission in question, in which both attacker and defender attributes are considered.

[9] Tversky, A. & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases, Science, Vol 185(4157), 1124-1131.
[10] https://www.mitre.org/sites/default/files/publications/pr-16-0956-the-assessment-of-biases-in-cognition.pdf

TERM	Definition in the context of the ReSCIND Program
Network and host Characteristics	Characteristics of hardware and systems architecture (including but not limited to network typology, system appearance, security posture, time delays) which can be exploited for cyber-defensive purposes.
Cyber Behavioral Impact	ReSCIND includes seven defender goals that the program will help achieve; these are 1) decrease rate of attack success; 2) increase time to task completion; 3) decrease progress toward goal; 4) decrease time until detection; 5) decrease defender effort spent; 6) increase attacker cognitive effort spent; 7) increase attack resources wasted.
Security Operations Center (SOC)	A team of cyber experts that monitors an organization's information technology infrastructure 24/7 to detect cybersecurity events in real time and address them as quickly and effectively as possible.