Positive Impact Through Artificial Intelligence and Data Science

Machine Learning | Network Science | Neuroscience | Human Computation

PROPOSAL: VOLUME 1
DHS-ST-23-108-FR01

Date: 21 June 2023

Releasing Agency and Office: U.S. Department of Homeland Security (DHS), Science and Technology Directorate (S&T)

TITLE: "Understanding Online Content Moderation Strategies' Impact on Targeted Violence Through a 2-Stage Model of Violent Radicalization."

ARROW ANALYTICS - SMALL BUSINESS

Point of Contact (TECHNICAL):
Dr. Ian McCulloh
Chief Technology Officer
Phone: (240)-506-3417
Email: ian@arrowanalytics.net

Total Funds Requested: $198,000.00

PLACES AND PERIODS OF PERFORMANCE:
Arrow Analytics – 14749 Walcott Ave, Orlando FL 32827 06/01/2023 – 05/31/2025

PROPOSAL VALIDITY EFFORT: 180 days
Arrow Analytics DUNS: 079340954
Arrow Analytics UEI: DB4HXBBWP429
Arrow Analytics TIN: 46-5199850

Table of Contents

1.0 Section I: Administrative

Funding Opportunity #	DHS-ST- 23-108-FR01
Technical Area	
Lead Organization submitting proposal	Arrow Analytics, LLC
Type of Organization	Small Business
Proposer's Reference Number, if any	23-US-DHS-001
Other Team Members (sub awardees and consultants), if any	None
Proposal Title	Understanding Online Content Moderation Strategies' Impact on Targeted Violence Through a 2-Stage Model of Violent Radicalization
Technical Point of Contact (POC)	Name: Ian McCulloh Address: 14749 Walcott Ave, Orlando FL 32827 Contact Number: (240) 506-3417 Email Address: ian@arrowanalytics.net
Administrative POC	Name: Ian McCulloh Address: 14749 Walcott Ave, Orlando FL 32827 Contact Number: (240) 506-3417 Email Address: ian@arrowanalytics.net
Total funds requested	$198,000.00
Date Proposal was Submitted	21 June 2023

2.0 Project Abstract

The U.S. continues to face the threat of domestic terrorism and targeted violence, often conducted by ideologically extremist individuals. Most people that have been radicalized to extremist ideology have an online presence that has contributed to their radicalization. Many online platforms have adopted centralized content moderation programs as an intervention to mitigate these threats, while others prefer decentralized approaches where users self-monitor. Other government and non-government sponsored programs attempt external interventions. The effectiveness of these programs and approaches is still poorly understood. Much of the existing research regarding online radicalization is focused on descriptive and correlational studies to identify factors associated with radicalized actors. Unfortunately, these efforts have poor predictive power and most people with a similar profile never engage in violent or extreme actions. The complex dynamics affecting the psychology of radicalization is magnified with poorly understood impacts of online content moderation efforts. Do these programs reduce extremist threats or does their censorship outrage people and poor gasoline on the fire?

Arrow proposes a novel approach to understand how and why individuals radicalize, mobilize, and disengage from violence, based on the proven two-stage model of radicalization proposed by Arrow's principal investigator. In his recently published book *ISIS in Iraq: Understanding the Psychological Foundations of Terror* (Oxford University Press, 2023) he introduced a two-stage model, whereby individuals that lack or lose psychological significance (meaningful purpose, acceptance, control) exhibit social and psychological strain, making them susceptible to radicalization (stage 1). When susceptible people are exposed to ideologically extremist narratives they may radicalize (stage 2). If the extremist organization satisfies their need for significance, they are not susceptible to deradicalization efforts. The two-stage model finds strong empirical support applied to all population-centric data collected by the U.S. and U.K. during the Iraq war to explain the rise and fall of popular support for Islamic extremism. In the proposed work, we will apply this approach to investigate historical domestic terrorist and targeted violence acts.

We will investigate historical targeted violence acts and events using mixed methods to include case studies, systematic literature review, structured interviews, social media content and network analysis, and machine learning to validate the two-stage model for US domestic and online populations. We then construct a multi-agent simulation to model proven social and psychological dynamics. The different content moderation strategies currently employed by online platforms, governments, and non-government actors can be evaluated through the simulation to inform future interventions and policies. With a greater understanding of how the online environment interacts with the radicalization process as understood through the two-stage model, Arrow will co-create new potential interventions with DHS and its state and local partners. These interventions can then be evaluated through simulation to inform priorities for implementation.

The proposed work provides a new and unique, yet empirically proven model to understand the process of radicalization. We recognize different, existing approaches to moderate extremist content and can investigate their effectiveness through simulation and virtual experiments. Existing approaches to terrorism studies find mixed results, poor predictive capability, are often ethically questionable and it is time for something new. We look forward to collaborating with DHS S&T to advance their important mission in defending the U.S. Homeland.

3.0 Project Narrative

Arrow Analytics, LLC (hereafter "Arrow") is pleased to provide this proposal in response to the U.S. Department of Homeland Security (DHS), Science and Technology Directorate's (S&T) Terrorism and Targeted Violence Research and Evaluation Funding Opportunity Number DHS-ST- 23-108-FR01. Arrow is a Service-Disabled Veteran-Owned Small Business (**SDVOSB**) located in Central Florida that conducts multi-disciplinary research at the intersection of **human behavior**, **artificial intelligence** (AI), and **data science**. Established in 2014, Arrow has delivered a number of innovative projects in support of the U.S. Federal government, NATO allies, and private industry in the areas of online influence (marketing/information warfare), public health, organizational behavior, and supply chain. At Arrow, we recognize that understanding human behavior is key to unlocking the potential of artificial intelligence and leveraging data science effectively. It is usually just as important to define the right questions and hypotheses as it is to implement the right algorithms and models. Our team combines expertise from various fields to create solutions that bridge the gap between humans and machines, empowering organizations to make data-informed decisions for mission impact.

The threat of online radicalization and extremism is a significant concern globally and domestically, as the internet provides a platform for the rapid dissemination of extremist ideologies, recruitment, and radicalization. The internet provides a fertile ground for radicalization, targeting vulnerable individuals who may feel marginalized, disillusioned, or susceptible to extremist ideologies. Extremist groups exploit social media, online forums, encrypted messaging apps, and video-sharing platforms to spread propaganda, recruit new members, and create extremist networks. Online platforms can amplify hate speech and extremist rhetoric, contributing to the polarization of societies. Extremist content can fuel intolerance, discrimination, and animosity towards certain religious, ethnic, or ideological groups. This can lead to social tensions, intergroup conflicts, and potential acts of violence. Online platforms have been instrumental in inspiring lone actors or small cells to carry out acts of violence without direct organizational connections. Radicalized individuals may find ideological justifications, guidance, and encouragement through online platforms, making it challenging for authorities to detect and prevent such attacks. Extremist groups exploit the internet for fundraising and financing their activities. Online platforms enable the collection of funds, cryptocurrency donations, and money transfers, providing a means for financial support for extremist organizations and their operations. The internet allows actors to conduct activities with relative anonymity and extended reach.

Widespread social media adoption has only existed for about 15 years and the US Government is still evolving regarding how it should respond to emerging online threats while protecting the rights of free speech, privacy, and civil liberties. It is important to note that while the threat of online radicalization and extremism is significant, most individuals who consume extremist content do not engage in violent activities. However, the potential for radicalization and the dissemination of extremist propaganda online necessitates continued efforts to counter and mitigate these threats. While the US Government has implemented a number of programs over the last 15 years, the effectiveness of these programs is questionable, and success must be re-evaluated given private-sector investments toward the same goals.

The US Department of Defense (DoD) initiated basic and applied research to understand the impact of social media on communication, ideology, and security shortly after the launch of Facebook in 2004 and the publication of the National Research Council report on Network Science. Arrow's PI, Dr. Ian McCulloh was one of the first researchers funded to investigate the impact of these emerging technologies affecting military operations in Afghanistan and Iraq. At the time, Dr. McCulloh was a major in the US Army stationed at the U.S. Military Academy at West Point and serving in both the newly established Combating Terrorism Center (CTC) and the Network Science Center. His work was central to the DoD's understanding of how to respond to online extremist threats. In 2007, McCulloh was assigned to Carnegie Mellon University, supporting US

Special Operations Command and developed methods for extracting actionable open-source intelligence, signals intelligence, and identifying information warfare threats in this relatively new medium. One of the programs that emerged from this body of work was the U.S. Central Command's Web Operations (WebOps) in 2008. Much of the tactics employed by WebOps in its early days involved forms of content moderation, such as identifying extremist propaganda and recruitment activities and then working with online platforms to have them removed. At that time, it was not well understood how easy it was for extremists to simply create a new profile and quickly connect with their former online audience. It was later discovered that the censorship backfired, creating a boomerang effect to fuel online radicalization. Understanding the science behind influence, persuasion, and cognitive psychology is counter-intuitive and essential to counter threats.

Perhaps the first program outside of the DoD was the Department of State's Center for Strategic Counterterrorism Communications (CSCC), led by Ambassador Alberto Fernandez, one of the last leaders still in government that had served in the U.S. Information Agency (USIA). While equipped with better knowledge and tactics, CSCC received less that 5% of the DoD WebOps budget, which forced a great deal of collaboration out of necessity to the detriment of operational efficacy. The threat of the ISIS Digital Caliphate started gaining significant attention among counter-terrorism experts in 2014 and by 2016 that research started impacting US policy and operations.

In 2016, DHS launched one of the earliest domestic programs to counter online radicalization, the "Peer-to-Peer: Challenging Extremism" initiative. This program empowered and mobilized community-based organizations, including non-profits, educational institutions, and local government entities, to counter online radicalization within their communities. The program provided grants to selected organizations, enabling them to develop and implement initiatives that addressed the root causes of radicalization and promoted alternative narratives. These initiatives encompassed a range of activities, such as conducting research on radicalization processes, creating educational campaigns to raise awareness, organizing community dialogues, and utilizing social media to counter extremist propaganda. By engaging local organizations and individuals, the Peer-to-Peer program aimed to foster resilience, build trust, and facilitate the development of effective counter-messaging strategies at the community level. While these programs are much more effective at mobilizing domestic resources, they lack the multi-disciplinary diversity to fully understand the modern threat. This is a problem addressed by this DHS funding opportunity today.

In 2016, the CSCC was rebranded the Global Engagement Center (GEC), received a significant increase in funding that surpassed DoD WebOps and was charged with coordinating the whole-of-government approach. At this time, however, much of the tactics employed were still uninformed by the recent advances in scientific research. Arrow's PI stood up an Analytics and Research team to support the GEC in assessing and informing a coordinated international response across the US interagency and 23 NATO allies. Their quarterly analyst exchanges and common cloud platform allowed greater collaboration on research, information sharing to include scripts for data collection and analysis and advanced the global response against extremist radicalization and state-sponsored threats. We now know so much more of online threats.

The digital world presents unique challenges when it comes to online influence, persuasion, radicalization and mobilization to violent action. Some key considerations include:

1. **Identifying and monitoring extremist content**: The internet is vast, and extremist content can be disseminated through many, many different platforms including social media, messaging apps, and other online forms. Legal and ethical considerations surrounding user privacy, probable cause, and platform terms of service create additional limitations on the government's ability to monitor content.
2. **Diverse platform content moderation programs**: Different online platforms implement content moderation differently. It may not be necessary for the US Government to moderate online content when platforms have privately funded programs in place. It is critical to understand the efficacy of

these programs to better inform policy and programs in this evolving landscape. For example, Facebook's budget for content moderation exceeds the revenue of any other platform, and perhaps federal spending, while 4Chan has a decentralized user-led content moderation approach. It is not well understood how different content moderation strategies impact online radicalization or how government and non-government agencies might better support tailored intervention programs.

3. **Understanding online user behavior**: Human behavior is often counter-intuitive, especially online. Interventions and messaging can often backfire depending upon user latitude of acceptance, misunderstood social norms, or malign counter-messaging. It is important to ground measurement, strategy, and interventions within a solid understanding of psychology, neuroscience, social networks, communications, and how human behavior is moderated through online forums.
4. **Reaching target audiences**: Effectively reaching and engaging with individuals who may be susceptible to radicalization is a challenge. Extremist groups often use sophisticated recruitment strategies tailored to specific demographics, and countering these efforts requires understanding the target audience and developing content that resonates with them. In a recent book authored by Arrow's PI and published by Oxford University Press, global Islamic extremist radicalization follows a two-stage process that must be understood to effectively reach and counter violent extremism.
5. **Evaluating program effectiveness**: Measuring the impact and effectiveness of online counter-extremism programs can be challenging. Determining the extent to which these programs effectively counter radicalization and prevent individuals from engaging in violent activities, especially when violent acts are a "black swan" event require proven models to compare present state against a likely alternative future driven by validated human behavior models.
6. **Interagency and intergovernmental cooperation and coordination**: Domestic targeted violence and extremism is a national security issue that transcends federal, state, and local boundaries. Cooperation and coordination among different state and local governments, online platforms, law enforcement agencies, and community organizations are crucial for effectively countering extremism online. However, achieving this level of collaboration can be challenging due to legal, political, and logistical barriers.

Arrow develops innovative approaches to analyze and understand the factors that shape online behavior, helping organizations optimize their digital engagement strategies and influence user decision-making. Our expertise in psychology, sociology, neuroscience, and data science allows us to unravel the intricacies of online interactions and provide valuable insights to drive effective engagement and reverse the process of violent extremist online radicalization, mobilization, and action.

For ease of evaluation, we use the following iconography for this proposal: Capabilities/Competencies (★), Feasibility (✓), Ethical Considerations and Risk Mitigation (+), and Cost Reduction/Effectiveness (🕒). We also add the additional criteria of Innovation (↑) for your consideration. We believe Arrow has the right people, assets, experience, and innovation to advance DHS' important mission of defending the homeland against online violent extremism.

3.1 Technical Approach

The proposed work is divided into three phases to be completed over 24 months as shown in Figure 1. The first phase will consist of a mixed methods (qualitative/quantitative) approach to understand and document different content moderation strategies in use by common online platforms and measure their impact on the volume, content, and structure of extremist discourse. Key research questions addressed in this phase are:

RQ1: What are the different content moderation strategies used by the top 20 online social media platforms?

RQ2: How effective are different content moderation strategies in limiting extremist discourse online?

RQ3: What factors or incentives drive effective content moderation strategies in US online platforms?

RQ4: How might we measure latitude of acceptance, attitudinal beliefs, social norms, efficacy, and social acceptance among online communities while protecting personally identifiable information (PII) and privacy?

RQ5: How might we rapidly identify candidate messaging strategies and interventions to reach target users?

Outcomes from this phase will consist of academic papers, conference presentations, content for online education/training, and a measurement approach/rubric to estimate the level of extremist discourse on various online platforms using stratified sampling.

The second phase will consist of a minimally viable product (MVP) multi-agent simulation application that will be tailored to specific identified online communities. This simulation will include a semi-automated capture of community psychographics, attitudinal beliefs, norms, efficacy, and social acceptance necessary for realistic and validated simulation of online behavior. It will be designed according to the two-stage model of extremist radicalization (Dagher et al, 2023) and incorporate Theory of Planned Behavior (Ajzen, 1985)(★). The simulation MVP will be used to conduct virtual experiments of interventions to affect disengagement from targeted violence that can be validated against historic cases and then used to understand platform-specific strategies while considering unique platform content moderation programs. Key research questions addressed in this phase are:

RQ6: How might we narrow the scope of monitoring and surveillance efforts for state and local stakeholders?

RQ7: How might we prioritize engagement platforms and online communities for intervention?

RQ8: How might we evaluate effective intervention approaches and strategies?

RQ9: Does the two-stage radicalization model fit domestic online data as well as it fits global data?

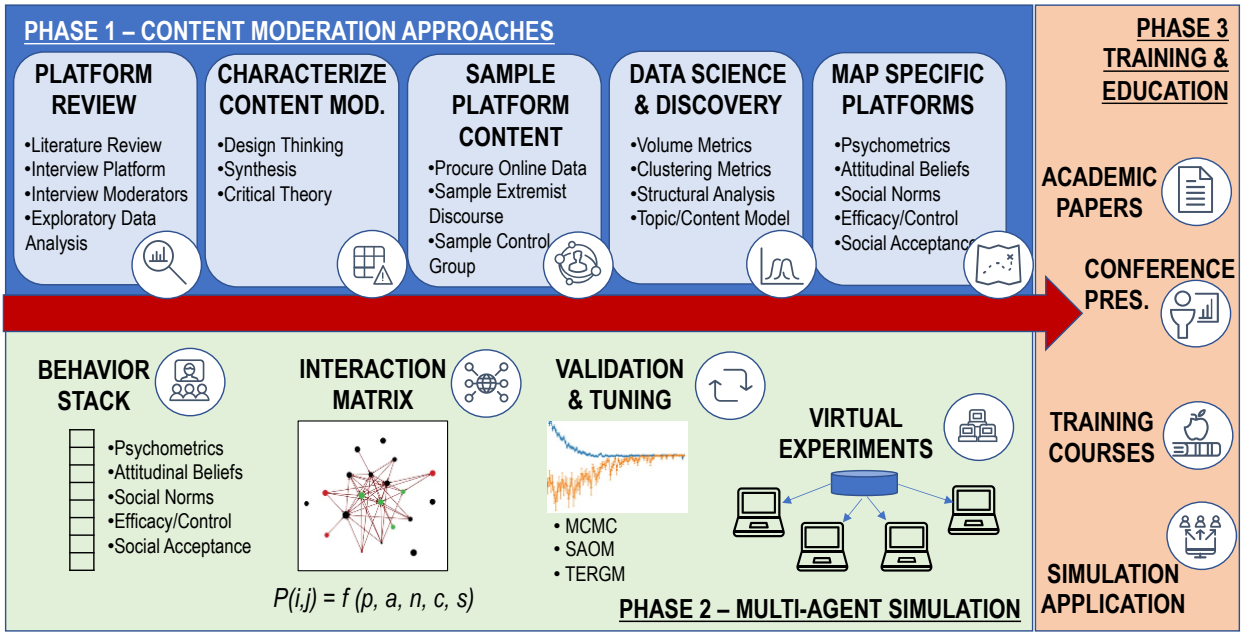

Figure 1. Key Research Phases

In addition to the simulation tool, additional academic papers and conference presentations will be developed to increase the scientific body of knowledge, making findings more accessible to online platform providers, local law enforcement, and community partners.

The third phase will consist of packaging key findings from phase one and two for inclusion into training and education programs. The online, short training course will be published either via YouTube or Udemy, the largest online learning platform, depending on government preference. The advantage of YouTube is the free user access. The advantage of Udemy is the global marketing for wider dissemination. Arrow can deliver effective online educational programs for state and local partners. Arrow's PI is an experienced university professor having taught almost 200 courses to thousands of students and exclusively teaching online for the last five years (★✓+).

3.2 Background and Literature Review

3.2.1 Terrorism

This proposal is broadly focused on advancing scientific understanding of domestic terrorism and targeted violence threats, paying specific attention to the role of online communities and platform content moderation policies for countering the threats of extremist radicalization. We define extremism as ideas or behaviors outside of societal norms and what is acceptable in politics (Midlarsky, 2011; Dagher et al, 2023). There are generally two categories of reasons people support extremism: they internalize norms of the extremist value system; or for utilitarian reasons of survival. Internalization of norms means that the individual holds the belief that the systems of norms is morally correct (Finley, 1971). When someone internalizes extremist norms and belief systems, they are radicalized (Neumann, 2013; Dagher et al, 2023).

Most of the scholarly works on extremist radicalization finds its roots in basic and applied research from the 1960s funded by the Advanced Research Projects Agency (ARPA), the organization later rebranded as DARPA. The most notable work, authored by Ted Gurr (1970) was the book "Why Men Rebel" based on his analysis of the causes of protest and rebellion in 1,100 "strife events" occurring in 114 "polities" from 1961-1965. Gurr's work was derived from correlation coefficients between different aspects of conflict and based on the hypothesis of frustration-aggression theory (Gurr, 1968; 1970). Frustration-aggression theory, originally posed by Dollard et al (1939) stated that aggression is the result of frustration, which is any event of stimulus preventing a person from attaining their goal or objective. This theory is a form of rational choice, where violent action is then understood within the context of a perceived social grievance. When there is no recourse to address grievances through peaceful political channels, violence and rebellion may be considered the best path to redress grievances (Gurr and Moore, 1997). Subsequent research in the field of terrorism studies assumes the grievance hypothesis without question and focuses on potential explanations for the emergence of extremist grievances, such as inter-group grievances (Mummendey et a 1999; Walker and Pettigrew 1984; Taylor and Louis 2004; Post et al 2009) or religious ideology (Juergensmeyer 2000; Stern 2003; Atran and Axelrod 2008).

US strategies for countering terrorism, targeted violence, and extremist radicalization appear to be grounded in the grievance hypothesis, even though it has not been found to be a proven or effective model. Horgan (2005) stated "it is somewhat misleading, if not naive, to assume that we can remove the grievances of terrorists [or domestic extremists] in an attempt to prevent terrorism from occurring." He also notes that many if not most extremists do not mobilize to violent action. Other scholars have found that support for extremist ideology is a poor predictor of violent action (Tessler and Robbins 2007; Fair, Malhotra, and Shapiro 2012), but may indicate other forms of support (Fair 2015; Kaltenthaler et al. 2010; Fair, Kaltenthaler, and Miller, 2015; Dagher et al, 2023). In a very practical sense, the application of grievance-based strategies appears to have been very ineffective in Vietnam, against the IRA, Afghanistan, Iraq, and the Islamic State in al-Sham

(ISIS also known as DAESH). Why then do grievance-based strategies remain a policy of choice?

The vast majority of case studies, to include recent studies published by the National Terrorism Awareness Center (NTAC), find that people who have attempted ideologically motivated targeted violence went through a radicalization process (Borum 2011: Wiktorowicz 2005; Dagher et al, 2023). The process is typically slow and not a sudden decision to act. This is acknowledged in current US strategy. Perhaps the most important research question to ask in the field of terrorism and targeted violence is the following:

> If people who engaged in terrorism and targeted violence have a certain pathology in common, why is it that most people with the same pathology do not engage in violent acts?

McCulloh (2006) proposed a two-stage model based on Strain Theory (Merton, 1938) to address this question. Strain theory states that a society places pressure on people to attain socially accepted goals, such as the American Dream using an equally accepted means, such as hard work and education. Many people, however, lack the means to attain those goals due to resources, social inequality, cultural or physical barriers. The resulting cognitive dissonance leads people to seek alternative means of attain goals which may involve committing crimes. This has become a widely accepted theory in criminology.

McCulloh extended his two-stage model incorporating Arie Kruglanski's theory of "needs" and Michelle Gelfand's theory of the societal context for norms. Kruglanski argues that the most important considerations an individual considers when deciding to engage in violent extremism are the needs for personal significance, security, and sustenance, with the need for significance being most paramount (Kruglanski et al, 2013; 2014; 2018; 2019; Dagher et al, 2023). The need for significance is the desire to feel like one's life matters and is respected. A significant loss of significance often leads to feelings of loneliness, confusion, and social pain. Extremist ideologies can create cognitive closure, certainty, solidarity, and purpose, reassuring someone who feels confusion and anxiety (Dagher et al, 2023). Gelfand argues that people with these high levels of uncertainty will seek tighter social norms, act less tolerant, and are more punitive towards those falling outside of tight norms (2012; 2019). Almost all extremist ideologies call for tighter norms to be imposed (Dagher et al, 2023). Under McCulloh's revised model, all people seek meaningful purpose, social acceptance, and control as their basic needs for significance. When they lack these needs, they exhibit social strain, which is evidenced by almost those who engage in terrorism and targeted violence, and this is stage 1: *Susceptible*. When a susceptible person is exposed to an extremist ideology that resonates with some of their existing beliefs, they might choose to follow the path of radicalization. They may also choose a number of other alternatives, and this is stage 2: *Alternatives*. If the alternative satisfies their need for purpose, acceptance, and control, their social strain is resolved, and they are not open to alternative points of view. From a strategic implementation perspective, we must first assess whether members of malign groups exhibit social strain. If they do not, they are not vulnerable to moderate ideology. Action must first be taken to increase strain and then paired with moderate alternatives that lower strain. If a benign group exhibits strain, the best resilience approach is to reduce the strain to protect their moderate ideology. This strategy was successfully demonstrated in Iraq during the Neutralizing Al-Qaeda in Iraq (NAQI) campaign in 2010-2011 and the Daughters of Anbar campaign in 2014.

Equal in importance to understanding the two-stage model of counter-radicalization is a model of influence and persuasion for susceptible populations. We draw on two theories from the psychology of influence: social judgement theory (Sheriff, 1963) and the theory of planned behavior (Ajzen, 1985). Social judgement theory assumes a range of beliefs on a particular topic and further assumes that there is a sub-range, called the latitude of acceptance (LoA). There may also be sub-ranges that are more extreme than the LoA called the latitude of rejection (LoR). The range of views between the LoA and LoR are called the latitude of non-commitment. When someone attempts to change the opinion of another, they may have an intuitive desire to express more extreme counter views that land in the subject's LoR. When this occurs, they create a

"boomerang" effect where the subject is likely to change opinion in the opposite direction and narrow their LoA, becoming less tolerant to alternate opinions. In this way, intuitive yet untested messaging efforts can often backfire, especially when the extent of extremist ideology is not fully measured or known.

The theory of planned behavior states that the best predictor of observed behavior in behavioral intent (Ajzen, 1985). Behavioral intent is a function of attitudinal beliefs, social norms, and perceived behavioral control. The closer in time to an observed event that behavioral intent is measured, the stronger is serves as a predictor. Planned behavior provides a heuristic model for reasoning about interventions. For example, if attempting to change the belief portion of the equation, an intervention may seek to introduce a new counter-belief, affect the relative importance of a belief compared to others, or devalue/eliminate an existing belief. Similarly, if targeting norms, interventions can target the norm itself, motivation to comply, or affect the social structure of opinion leaders. The theories of planned behavior and social judgement provide a framework to design and test potential interventions targeting susceptible populations (stage 1) or for increasing social strain (stage 2). This model of intervention finds further support in the body of literature on social networks, normative behavior, and terrorism (Krebs, 2002; Carley and Reminga, 2004; Kruglanski et al 2018; Sageman 2004; 2011; McCulloh 2007; 2013; Dagher et al 2023; Wiktorowicz 2004; 2005; Gaines and Mondak 2009; Klandermanns and Oegma 1987).

In response to mixed findings and the lack of empirical support for central counter-terrorism theories, Horgan (2005) states, "if we do not ask the right questions, we most certainly will not arrive at meaningful answers, regardless of the perspective we take in trying to approach the problem in the first place". In the recent book, published by Dagher et al (2023), the authors reconcile many competing theories of extremist radicalization and demonstrate strong empirical support for the rise and fall of ISIS in Iraq using all of the extensive population data collected by the U.S., U.K., and additional commercial data. Their model has yet to be applied to U.S. domestic extremist radicalization and targeted violence. There is also opportunity to further explore potential intervention strategies using multi-agent simulation (MAS) which is discussed further in the methods section.

3.3 Methods

3.3.1 Protection of Human Subjects

Arrow shares the government's commitment to the respect and protection of civil rights and fully complies with all DHS and U.S. Department of Health and Human Services (HHS) requirements regarding research involving human subjects (45 CFR) known as *The Common Rule* (+). Arrow has been conducting research involving human subjects for nine years in compliance with Federal policy (★✓) and has developed a Human Subjects Research (HSR) plan that addresses data protection/privacy, legal compliance, and participant protections. Arrow has partnered with Pearl, an expert in the area of HSR with a proven track record of human research protection, to serve as the Institutional Review Board (IRB) of record for this project. Pearl's IRB Chairs will coordinate and manage all HSR related aspects of the project in close collaboration with Arrow's PI. Under their guidance, the Pearl IRB will review each data collection protocol to ensure that they follow the appropriate procedures for protection of human subjects and that any potential risk to participants is minimized (+). Pearl has served as the IRB for multiple federal projects led by Arrow's PI and has a demonstrated track-record of timely review and oversight actions to ensure maximum project feasibility (✓) without any ethical compromise or risk to human subjects (+).

Planned data collection – At this time, we anticipate that our planned protocols will qualify as ***Exempt*** from full IRB review as they will involve Category 2: Observation of public behavior (such as public social media posts) (✓). Participant information will be masked (+). For example, no usernames or any other PII will be stored in our dataset and will instead be replaced by a unique identifier. Arrow does not employ any deceptive

or covert practices to collect data and will abide by all terms of service for internet providers, social media platforms, and any other source for which data will be collected (+). Data will be properly sourced and applicable terms of service will be retained for the duration of the project (+). Arrow will not use web scraping as stipulated by the government.

The Common Rule (45 CFR 46) requires that every individual who interacts with participants and/or their data must first complete government approved HSR training, such as the appropriate Collaborative Institutional Training Initiative (CITI) courseware. Arrow will require all research staff members and participant recruiters to complete or renew their CITI training (+). In accordance with our data handling procedures, collected data will be stored securely in a project-restricted space with no network connection (+).

3.3.2 Content Moderation Review

Arrow will investigate different types of social media content moderation programs in use by online platforms to address research questions one through three (RQ1-3). We will employ a combination of qualitative and quantitative research methods. The goal is to gain a comprehensive understanding of current programs and their impact on various aspects of online radicalization, mobilization, and coordinated action. Methods are:

1. **Case Studies**: The United States Secret Service National Threat Assessment Center (NTAC) published an annual report on Mass Attacks in Public Spaces from 2017-2019 and a five-year review covering 2016-2020 that provides summary descriptive statistics for all public violent attacks in which three or more people (not including the attacker) were harmed. Approximately 26% of these attacks were motivated by extremist beliefs. 63% of attackers were active on social media and posted concerning content to include all of the extremist attackers. While the NTAC reports provide valuable information, they lack and in-depth investigation regarding the specific online platforms used by the attackers, their posts and responses by the online community, and most importantly the content moderation measures employed by the platform or online community at the time leading up to the attack. Arrow will conduct a detailed case study of each mass attack involving extremist ideology from 2016-2020 detailed in the NTAC reports (✓). Case studies will focus on the use of online platforms and their policies to moderate extremist ideology. These case studies provide an opportunity to examine specific instances of content moderation and understand the contextual factors influencing outcomes. Keywords and search terms will be derived from the case studies for use in a subsequent systematic literature review (✓).
2. **Systematic Literature Review (SLR)**: Keyword and search terms derived from the case study will be entered into Google Scholar to identify any academic literature published from 2016 to present. Acceptance criteria for the SLR involves all studies where the abstract mentions content moderation, radicalization, or extremism in the context of online platforms to include social media. Arrow research investigators will read and synthesize publications to gather existing knowledge, theories, and frameworks related to social media content moderation programs. This step helps researchers establish a foundation for the program and identify any gaps in current understanding (★✓).
3. **Structured Interviews**: The Arrow PI will conduct interviews with key stakeholders involved in content moderation programs to elicit valuable insights. Arrow's PI served as the quality assurance director for Accenture's global content moderation program from 2020-2023 employing approximately 70K moderators and serving most of the large social media platforms (★✓+). This experience provides access to individuals currently engaged in content moderation both within the online platform as well as within the companies that conduct content moderation on their behalf. In adherence to stated ethical considerations, all respondent names will be kept anonymous and only the Arrow PI will be involved in conducting interviews (+). Interviews will be conducted until the body of new knowledge is less than 10% of that captured in previous interviews.

4. **Content Analysis**: Arrow will conduct content analysis of a large sample of social media posts using natural language processing, specifically topic modeling and content classification. Arrow investigators will create a coding scheme to categorize content based on different moderation outcomes, such as removal, warning labels, or no action. This method allows for systematic analysis of patterns and trends in content moderation practices. Data for content analysis will be drawn from historic archives of online discourse and anonymized to remove any PII to protect individual privacy (+). While it is unlikely, we will be able to gather complete data for all 45 extremist-motivated attacks, there exists sufficient access through scholarly and non-government watch organizations for a sufficient sample for research purposes.
5. **Network Analysis**: Social networks will be constructed from online content to include direct communication commenter-to-poster, friend/follower, hashtag co-mention, mention, retweet and potentially other platform-relevant networks. Descriptive network measures will include centralization, subgroup analysis, evaluate attackers' network position. Exponential random graph models (ERGM) and stochastic actor-oriented models (SAOM aka Siena) will be used to identify statistically relevant factors contributing to network interaction and online normative behavior. ERGM and SAOM are statistical models that control for high levels of network dependence and are used to estimate node, link, and structural co-variates correlated with network link formation in a dependent network. Arrow's PI teaches all of these methods in his graduate courses at Johns Hopkins University and has published a number of academic papers involving these methods in similar domains (Sailer & McCulloh, 2012; Sadayappan et al, 2018) (★✓🕒).
6. **Machine Learning (ML)**: Arrow investigators will attempt to develop a classifier to estimate psychometrics, attitudinal beliefs, and efficacy among online personas. Key to our approach is attention to inter-annotator agreement (IAA)(↑). Naser and colleagues (2019) demonstrated that consistent annotation is essential for driving high ML precision, while data volume improves recall. Arrow has discovered through experience that most AI performance ceilings are the result of high annotation disagreement, and we can achieve double-digit performance gains through consistent annotation. This step is important to build a solid MVP in phase 2 and establish a foundation for future ML improvement with additional data collected over time (★✓🕒).

Arrow will further investigate research questions four and five (RQ4-5) through one or more facilitated design thinking sessions. Invited experts in neuroscience, information warfare, communications and marketing, psychology, any government stakeholders, and other program performers will meet in an Arrow facility or location of the government's choice and participate in structured brainstorming sessions led by Arrow. Research findings listed above will be presented throughout the day to elicit participant reactions and innovative insights. Proposed ideas will be empirically tested through virtual experiments in the multi-agent simulation phase (★✓🕒↑).

3.3.3 Multi-Agent Simulation

The applications of network science and multi-agent simulation (MAS) has existed for a long time. Forrester (1961) proposed a simple model consisting of retailers, wholesalers, distributers, and manufacturers to model consumer behavior. Following the attacks of 9/11, much of the emerging work in network science became refocused on terrorist networks. Carley extended a simple model and introduced it as a potential approach to target and destabilize terrorist networks (Carley, Lee, and Krackhardt, 2002). Under the Office of Naval Research (ONR) funding, this approach became the foundation of the Organizational Risk Analyzer (ORA) and used across the intelligence community to target adversarial terrorist networks at the social and cultural level as much as the kinetic level (Carley and Reminga, 2004). Unlike many of the existing social network analysis tools of the time, early versions of ORA had two critical features that are particularly useful for

modern online networks, the meta-network and MAS. The meta-network was an extension of the traditional social networks, recognizing that there are many different types of nodes (e.g., organizations, people, resources, hashtags, platforms, and more) and many different types of relationships or links (e.g., communication, finance, kinship, formal leadership, and more). While this greatly increased the complexity of the threat networks, new measures were introduced such as *specialization* where instead of calculating betweenness centrality (Freeman, 1977) across a single-mode network, the measure would only include source and target nodes such as person and skill. Similar to betweenness centrality's ability to measure brokerage in a social network, specialization centrality measures skill brokerage in a more complex network. By constraining the nodes utilized in the measure, specialization centrality is more computationally efficient than betweenness centrality, while being more intuitive and better capturing the right social construct.

ORA includes a MAS module based on the theory of Constructuralism, which "states that the socio-cultural environment is continually being constructed and reconstructed through individual cycles of action, adaptation and motivation" (Schrieber, Singh and Carley, 2004). Inspired by early models of Schelling (1971) who first applied MAS to social behavior, ORA added modern features such as mimicking neighbors (Latané, 1996; Axelrod, 1997), agent autonomy (Kaufman, 1996), and decisions influenced by others in the network (Granovetter, 1978; DeMarzo et al, 2003). This MAS enabled better and more realistic threat targeting (★✓).

By the mid-2000's, ORA was the most widely used network analysis tool across the Intelligence Community and Department of Defense. The ability to reason against more complex networks and simulate the impact of targeting decisions, in many people's opinion, turned the tide of US success in the Iraq war in 2006-7. By 2009, US forces developed the ability to infer illicit organizational networks thru **forensic weapons technical intelligence**, modeling the **terrorist adversary as a meta-network**. They evaluated targeting options to disrupt enemy operations. They discovered that formal leadership networks, resource networks, and financial networks, while important are somewhat obvious, have built-in resilience and have little relative impact on enemy operations. Targeting the knowledge networks, however, disrupted the ability for adversaries to recruit and train those who would build and those who would emplace, and overwatch improvised explosive devices. Targeting key knowledge brokers not only prevented adversary innovation but prevented them from sustaining current proficiency and was perhaps the most important targeting innovation during the Global War on Terror (GWOT), currently sustained as part of the Army's Advanced Network Analysis and Targeting (ANAT) program[1][2][3][4]. Much of this was led by Arrow's PI (★✓🕒).

Arrow proposes to construct an MAS of an online community engaged in extremist discourse using ORA (✓🕒). The proposed MAS will consist of six basic objects. The 1) *agents* will represent individual actors in the online forum ranging from radicalized to moderate. The 2) *attribute stack* will model the core variables derived from phase 1 as a set of 500 binary bits and used to represent demographic (e.g., race, age, gender), psychometric (e.g., post frequency, pattern of life, post engagement), attitudinal beliefs (using ML from phase 1), normative behavior (using network analysis from phase 1), agent perceived efficacy, and feeling of social acceptance. The 3) *time* will determine the number of simulation iterations to run and will be varied between parameterization, validation, and execution runs. The 4) *object relations* will link individual agents with corresponding values in the attribute stack. The 5) *interaction matrix* will allow or restrict agent interactions based on online social proximity, friend/follow networks, and content moderation policies. The network interaction matrices that govern allowable agent communication will be typologically patterned to resemble

[1] https://oe.tradoc.army.mil/net/
[2] https://www.npr.org/2010/12/03/131755378/u-s-connects-the-dots-to-catch-roadside-bombers
[3] https://www.army.mil/article/38497/social_networking_the_silent_counterinsurgent
[4] https://www.kellogg.northwestern.edu/faculty/uzzi/ftp/media%20hits/Science_Counterterrorism_July09.pdf

the observed online network. The 6) *platform behavior model* will be a platform specific model governing agent behavior associated with posting latency, volume, persona charisma, retweet probability, etc.

Following the initialization function, the MAS will cycle through a sequence of **think-update-communicate-cleanup** functions and then repeat for subsequent time steps until time is elapsed. The ***think*** function is critical for the interaction models and generates "messages" in the form of binary bit strings and modeling ideologically consistent statements that are used to model agent-agent communication on the online platform. The ***update*** function allows the MAS the opportunity to manipulate messages such as adding, modifying, or removing information to model how an agent may interpret the message based on theories of planned behavior and social judgement. The ***communicate*** function partitions message bits and is moderated by the interaction matrix and platform behavior model to model agent interaction. The ***cleanup*** function updates agent knowledge and beliefs in the knowledge stack based on their interpretation and engagement with messages as moderated by their current beliefs, norms, and perceived behavioral control. This completes one time step and the cycle repeats with the think function.

Agent interaction will update attitudinal beliefs, norms, and perceived efficacy according to an agent's latitude of acceptance. Repeated exposure to like-minded extremist content will narrow an agent's latitude of acceptance. Interactions with agents of different attitudinal beliefs will create a boomerang effect proportionate to the magnitude of the difference in belief. Interactions with agents of slightly different attitudinal belief, however, will simply widen the agent's latitude of acceptance. The likelihood of agent interaction within ORA's interaction matrix will be governed by factors found significant in the ERGM and SAOM analysis conducted in Phase 1. MAS parameters will be adjusted until modeled behavior is consistent with real-world observed behavior in the online platform in order to validate MAS behavior (★✓).

With a valid MAS model, potential interventions developed during the Phase 1 design thinking session will be coded within the MAS model and varied in a statistically designed experiment. A central factor in the virtual experiment will be the different types of content moderation policy. This will allow Arrow researchers to evaluate the impact of content moderation policies on demobilization and disengagement from extremist ideology and behavior. Other potential interventions developed during the design thinking sessions may be modeled as well.

The focus of the MAS virtual experiments is to reason about generalizable interventions and content moderation strategies and not necessarily intended to model a specific threat or online community. Modeling specific communities is beyond the scope of this study and would require further validation, testing, and must consider the potential ethical considerations. We do feel, however, that generalized findings will inform better policy interventions and may help focus state and local law enforcement monitoring and surveillance efforts on priority platforms and communities where they can be more effective at mitigating the risk of future targeted violence.

3.3.4 Training and Education

Arrow will develop two virtual/online training modules to support knowledge transfer and training of state and local stakeholders. The first module will focus on the findings from Phase 1, explaining the two-stage model for radicalization, empirical support within domestic threats, and implications for informing and assessing effective interventions. The second module will focus on the findings from Phase 2, explaining the MAS and virtual experiments along with their implications for designing effective threat mitigation interventions. Modules will increase audience engagement with embedded knowledge checks and an end-of-module assessment. Dr. McCulloh has been a university professor for almost 20 years, has taught almost 200 courses to thousands of students (★). He has always been rated well above fellow professors in student end-of-course evaluations across multiple universities (★).

Where applicable, Arrow will also record any academic presentations or public talks and post them online to further increase knowledge transfer (🕒).

3.4 Organizations and Key Personnel

3.4.1 Principal Investigator (PI) – Key Person (KP)

Dr. **Ian McCulloh** is proposed as the principal investigator (PI) / key personnel (KP). He is an internationally recognized multi-disciplinary expert in countering radical extremism bringing a unique blend of academic and practical hands-on-experience. He served as an associate professor at Johns Hopkins University from 2015-2018 with joint appointments in Computer Science, Public Health, and the Applied Physics Lab. He has authored three books, 88 peer-reviewed academic publications, and given numerous public presentations including a TEDx (★). In his most recently published book ***"ISIS in Iraq: Understanding the Psychological Foundations of Terror,"*** published by Oxford University Press, he integrated all public opinion and counter-radicalization data ever collected in Iraq by the US and UK to present an empirically driven understanding for the rise and fall of ISIS in Iraq and the Levant (★✓). He also collected independent data from incarcerated ISIS fighters following the fall of Mosul using methods from quantitative anthropology to avoid respondent bias (↑). He has also conducted novel human subjects experiments recruiting non-violent, yet highly polarized Sunni and Shia Iraqis and measuring neural response under conditions of argument and cooperation in joint tasks (★↑). He continues to serve as an adjunct professor at Johns Hopkins University where he teaches graduate courses in Social Media Analytics, the Neuroscience of Influence and Persuasion, and Human Computation (✓). Dr. McCulloh is also a retired Army officer with 13 combat deployments to Afghanistan, Iraq, and the Middle East directly combatting violent extremism (★). He founded the Army's Network Science Center which sped the transition of multi-disciplinary research to counter the threats of improvised explosives and extremist radicalization. He also served in the West Point Combating Terrorism Center. He culminated his military career as the chief information warfare strategist and chief of offensive cyber operations at U.S. Central Command. He was the principal architect of some of the most effective counter-radicalization campaigns during the Iraq war to include the Neutralizing Al-Qaeda in Iraq (NAQI), Rule-of-Law, and Daughters-of-Anbar campaigns (★). Prior to launching Arrow Analytics, Dr. McCulloh was the Chief Data Scientist and managing director for Accenture's Federal Applied Intelligence practice from 2018 until his retirement earlier this year (★). During that time he grew their practice from 50 to 800 people (without acquisition) and associated sales from $240M to $1.4B per year, delivering AI and data science solutions to the U.S. Federal Government. He holds a Ph.D. in Computer Science from Carnegie Mellon University and master's degrees in Applied Statistics, Industrial Engineering, and Sociology from Florida State University (★).

3.4.2 Unique Relationships/Access

Arrow's PI served as the quality assurance director for Accenture's social media content moderation program for over three years from 2020 until 2023. This program consisted of over 70 thousand content moderators and supported the most popular global platforms including Facebook, YouTube, Twitter, and Instagram. His relationships developed over the past few years provides access to experts currently engaged in content moderation which will be used to facilitate the proposed structured interviews following case studies and SLR (★✓🕒).

3.4.3 Data

Qualitative data from case studies and literature review (SLR) will inform a set of potential incidents of ideologically motivated mass-attacks committed in the US. Candidate cases will include those listed in the NTAC report. In addition, we know that the SLR will reveal more recent cases committed since 2020. As stated earlier, the inclusion criteria for considered cases will be those committed in the last decade that

involved ideological or extremist motives. For each case a unique set of key search terms, hashtags (where appropriate), and identified personas will be developed (✓). These sets will be used to pull historic data from online platforms and potentially identify archived data sets as well.

Arrow maintains a number of social media data collection assets (🕒). The nature of these assets varies, however, for this proposal we will only utilize those that access platform APIs in accordance with the platform's terms of service in order to comply with DHS S&T requirements (+). These assets are updated twice per year and used in Dr. McCulloh's social media analytics course taught every Fall and Spring semester at Johns Hopkins University (★✓+🕒). Many of these will allow an analyst, using a developer account, to pull a limited number of social media posts. By structuring queries in small, well-formed requests, Arrow can piece together a sufficient sample of social media data to meet the quantitative analytic requirements laid out in the methods section (✓). Current platforms that Arrow can access for this proposal include YouTube, Twitter, Facebook (discussion surrounding specific posts only), Reddit, and 4Chan. Arrow is currently developing a TikTok asset, given recent changes in TikTok's terms of service (✓🕒). We feel that this represents both centralized and decentralized content moderation strategies to allow inference on their effects on online discourse.

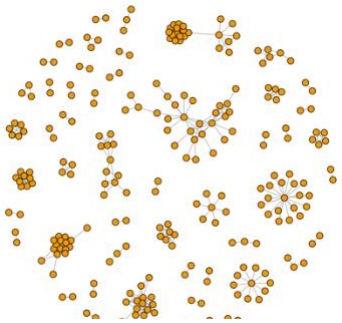

Figure 2. Extremists

We will supplement our data collection with archived data that have been previously collected and shared for the purpose of basic and applied research. Organizations that maintain social media data of terrorism and targeted violence attacks in the US include: *The Violence Project, The University of Maryland Profiles of Individual Radicalization in the United States (PIRUS), The Gun Violence Archive, Stanford Social Network Analysis Project (SNAP) Large Network Dataset Collection. Figure 2 shows a sample network extracted from the PIRUS data set involving online discourse among radicalized white supremists.*

3.5 Projected Outputs and Knowledge Products

Arrow will share research knowledge and findings in three key areas: academic papers, conferences, and training courses. Arrow's PI has a strong track record of academic publications having published 88 peer-reviewed academic papers and authoring three books. Given that the proposed effort is focused on the topic of how online platforms content moderation policies will affect likely radicalization and mobilization to violence, Arrow proposes to publish in key venues that are well respected and read by decision makers and scientists at online platforms. While we can target publications preferred by the government, we recommend conference proceedings, because they publish peer-reviewed results faster, transitioning findings to both the academic community and the field quickly and they tend to be more accessible than journals that often have a paywall which could make them non-accessible to state and local stakeholders. Ideal venues include:

1. Proceedings, Advances in Social Networks Analysis and Mining (ASONAM). This is an IEEE/ACM (★) annual publication that is well indexed with wide readership. McCulloh has successfully published 10 peer-reviewed papers with this publication (✓).
2. Proceedings, Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium. AAAI is the academic society for Artificial Intelligence (★). They hold an annual symposium in Arlington VA focused on government applications which is well attended by those at the intersection of tech and government. McCulloh has not only published in this venue multiple times, but also organized special tracks with invited government keynote speakers (✓).
3. Proceedings, International Conference on Social Computing, Behavioral-Cultural Modeling & Prediction and Behavior Representation in Modeling and Simulation (SBP-BRiMS). This conference in its 16th year brings together government and academic scientists to improve government response

to major social problems to include terrorism and targeted violence (★). McCulloh has published in this venue multiple times, has served on the review committee and even asked to chair the event (✓). The proceedings are published by Springer.

4. Proceedings, International Conference on Web and Social Media (ICWSM). This is the AAAI sponsored premier conference for academic research focused on social media and online forums.

Most of the above venues include both an academic conference as well as the associated proceedings. In addition, Arrow plans to target the International Conference for Social Network Analysis which is a presentation only conference but draws a diverse group of thousands to discuss social networks and often includes law enforcement and national security professionals. Arrow's PI has also organized the first three North American Social Networks (NASN) Conferences which are held in the US when the international conference occurs outside of North America. NASN has been held in Washington DC, sponsored by big tech and draws a large group of academia, government, and some industry and tends to focus on applications in health and terrorism. Awarding this proposal would offer DHS S&T the opportunity to shape the future NASN program and invite list as an excellent forum to advance their broader program objectives.

As described earlier, Arrow will also produce training that can be shared virtually or in-person with state and local partners. Training curriculum will include a module from phase 1 outlining findings from the impact of disparate content moderation policies on the outcomes of targeted violence and drawing insight for future monitoring and intervention. Another module would be derived from phase 2 drawing generalized findings about potential interventions as assessed in the MAS. A third module would be developed to share the two-stage model for understanding extremist radicalization to increase the diversity of thought for countering future threats. Arrow's PI has been teaching online and virtual courses for Johns Hopkins University for six years and can share lecture materials and curriculum to lower cost and increase feasibility (★✓🕒).

3.6 Schedule and Timeline

Arrow proposes basic research over a 23-month period in accordance with the schedule shown in Figure 2.

		Phase 1										Phase 2									Phase 3			
	Milestones	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14	M15	M16	M17	M18	M19	M20	M21	M22	M23
Content Moderation	Case Studies	■	■	■																				
	Systematic Literature Review		■	■	■																			
	Structural Interviews				■	■	■																	
	Content Analysis						■	■	■															
	Network Analysis						■	■	■															
	Machine Learning							■	■	■	■													
MAS	Construct Multi-Agent Simulation										■	■	■	■										
	Tune and Validate Simulation													■	■	■								
	Virtual Experimentation															■	■	■	■	■	■			
Outputs, Trainig Know. Products	Write Academic Paper			■			■				■											■		
	Conference Presentation			■			■				■											■		
	Virtual Training Module											■	■										■	■
	Monthly Report	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
	Support demos and transition														■	■				■	■	■	■	■

Figure 2. Research Schedule.

3.7 Deliverables

Arrow will deliver academic papers and presentations to DHS S&T at least one week prior to submission for government review. In addition to the draft paper/manuscript, Arrow will deliver all data, anonymized to protect individual privacy and R source code for any statistical analysis such that the government can independently recreate and verify findings. For training materials, Arrow will provide three milestones for each of the three modules: 1) Statement of Learning Objectives (SLO); 2) Outline or storyboard; and 3)

Final interactive video lecture. Following the delivery of each milestone, the government will have the opportunity to provide input and modification to ensure timely and cost-effective production (✓🕒). The following table outlines scheduled milestones, waypoints, deliverables, and monthly status reports (MSR).

Phase	Month	Event	Description	Comments	Deliverables
1	1	Milestone	List of case studies for investigation	Allows DHS S&T to add or remove cases	MSR
1	2	Milestone	List of search terms and the acceptance/rejection criteria for the SLR	Allows DHS S&T to modify and provide input into SLR	MSR
1	3a	Deliverable	Academic paper describing the case studies reviewed	We plan to write to generalizable macro trends for the role of online platform use and content moderation strategies	Academic Paper
1	3b	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
1	4a	Milestone	Research plan and discussion guide for structured interviews	Allows DHS S&T to provide input. Names of subjects will remain confidential for privacy.	Research Plan
1	4b	Milestone	IRB Approval		IRB docs
1	4c	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
1	5	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
1	6a	Deliverable	Academic paper describing the case studies reviewed	We plan to write to generalizable macro trends for the role of online platform use and content moderation strategies	Academic Paper
1	6b	Waypoint	Monthly Status Report	Launch content and network analysis	MSR
1	7	Waypoint	Monthly Status Report	Launch machine learning project	MSR
1	8	Deliverable	Monthly Status Report	This will be a longer report and include preliminary findings from content and network analysis.	MSR
Phase	**Month**	**Event**	**Description**	**Comments**	**Deliverables**
1	9	Waypoint	Any progress/issues	Monthly Status Report	MSR

1	10a	Deliverable	Academic paper describing the AI/ML classification of online extremist discourse	This will describe a repeatable process to classify variables from online discourse to populate MAS	Academic Paper
1	10b	Deliverable	Design thinking session	Brainstorm possible interventions, RQ4-5, for MAS virtual experiments	Facilitated session and report
1	10c	Deliverable	Data and source code for independent validation	Data will be anonymized to protect privacy and PII	Source code, data
1	10d	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
3	11a	Milestone	Training module SLOs		Document
3	11b	Milestone	Training module outline/storyboard		Document
1	11c	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
3	12a	Deliverable	Completed training module	Suitable for publication with YouTube or Udemy	Video and supporting document
1	12b	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
2	13a	Milestone	MAS MVP v.0	Initial MAS MVP developed and tested, but not yet tuned or validated.	Source code, data
2	13b	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
2	14	Waypoint	Any progress/issues	Monthly Status Report	MSR
2	15a	Deliverable	MAS MVP v.1	Validated MAS MVP	Source code, data
2	15b	Milestone	Planned virtual experimental design	Allows DHS S&T to provide input on variables and experimental design	Document
2	15c	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
2	16	Waypoint	Any progress/issues	Monthly Status Report	MSR
2	17	Waypoint	Any progress/issues	Monthly Status Report	MSR
2	18	Waypoint	Any progress/issues	Monthly Status Report	MSR
2	19	Waypoint	Any progress/issues	Monthly Status Report	MSR
2	20a	Deliverable	Preliminary findings from virtual experimentation	Will include all source code and anonymized data	Document, C++ source code, data
2	20b	Waypoint	Any progress/issues	Monthly Status Report	MSR
Phase	**Month**	**Event**	**Description**	**Comments**	**Deliverables**

2	21a	Deliverable	Academic paper describing the MAS virtual experiments	This will describe a repeatable process to simulate online forums with variations in content moderation and generalize their impact on countering domestic online radicalization	Academic Paper
2	21b	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
3	22a	Milestone	Training module SLOs		Document
2	22b	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
3	23a	Milestone	Training module outline/storyboard		Document
2	23b	Waypoint	Monthly Status Report	Any progress/issues not related to deliverable.	MSR
3	24a	Deliverable	Completed training module	Suitable for publication with YouTube or Udemy	Video and supporting document
2	24b	Deliverable	Final report	Summarize all papers submitted for publication, conference presentations, source code, data, and other accomplishments.	Final Report

3.8 Risk Mitigation Plan

Arrow's risk mitigation plan is provided in the following table.

Risk	Mitigation
Threats to individual privacy, civil rights and liberties.	All personnel working with data will be trained and certified in data privacy, protection, and how to statistically verify anonymity. Arrow will anonymize any data collected during this project and verify that any independently collected data will all meet privacy standards equivalent or more rigorous than required by Health Insurance Portability and Accountability Act (HIPPA) and Family Educational Rights and Privacy Act (FERPA). Data will be stored on an independent internal network and only verified anonymized data will be shared with the government. All data collection will follow applicable terms of service.
Data availability or insufficient for machine learning training or data sparseness.	Arrow has identified diverse data sources including archived, published, historic data via platform API, and qualitative. In the event different data sources lack key variables, Arrow will implement appropriate methods for imputing data, synthetic data, or modeling missing data to create an integrated set for training machine learning algorithms.
Data bias or measurement error.	Arrow will apply proprietary, automated, software to assess potential bias and measurement error. The tool provides a validated approach to test whether data may be biased towards a protected class when demographic

	variables are not captured or maintained. It also assesses data labeling consistency and the impact of missing or wrongly coded data. This code has been used by other federal agencies to assess potential bias and discrimination in AI/ML solutions to protect individual civil liberties and ensure improved performance of AI/ML solutions.
Platform API changes affecting data collection.	APIs are constantly changing. Arrow updates data collection assets twice per year and partners with Johns Hopkins University and University of Central Florida to maintain shared assets and immediately update any that affect client delivery. Arrow has maintained an ability to collect data from major social media platforms for almost a decade.
Poor machine learning performance.	Most ML performance issues stem from insufficient or inconsistently labeled data affecting recall and precision respectively. Arrow often achieves double-digit performance gains by focusing on label consistency and can further boost performance using synthetic data.
Simulation validation	Model development based on proven social science reduces threats to validity. Model complexity can also be reduced as needed.

3.9 Transition Plan

Arrow intends to deliver all non-commercial software (including source code), software documentation, and technical data with Government Purpose Rights (GPR). Any personally identifiable information (PII) will be anonymized or deleted in accordance with government requirements. Arrow will present demos, training modules, academic presentations, and academic papers to state and local stakeholders in Florida and in the DC metro area to encourage adoption. Arrow works with a number of charitable and non-profit organizations in Central Florida that often work with disadvantaged populations. We plan to share findings with them and look for opportunities to support community-based interventions.

4.0 References

Atran, S., & Axelrod, R. (2008). Reframing sacred values. *Negotiation journal*, 24(3), 221-246.

Ajzen, I. (1985). *From intentions to actions: A theory of planned behavior* (pp. 11-39). Springer Berlin Heidelberg.

Axelrod, R. (1997). Advancing the art of simulation in the social sciences. In *Simulating social phenomena* (pp. 21-40). Springer Berlin Heidelberg.

Borum, R. (2011). Radicalization into violent extremism I: A review of social science theories. *Journal of strategic security*, 4(4), 7-36.

Carley, K. M., Lee, J. S., & Krackhardt, D. (2002). Destabilizing networks. *Connections*, 24(3), 79-92.

Carley, K. M., Reminga, J., & Borgatti, S. (2003, September). Destabilizing dynamic networks under conditions of uncertainty. In *IEMC'03 Proceedings. Managing Technologically Driven Organizations: The Human Side of Innovation and Change* (IEEE Cat. No. 03CH37502) (pp. 121-126). IEEE.

Dagher, M., Kaltenthaler, K., Gelfand, M. J., Krugslanksi, A., & McCulloh, I. (2023). *ISIS in Iraq: The Social and Psychological Foundations of Terror*. Oxford, UK: Oxford University Press

DeMarzo, P. M., Vayanos, D., & Zwiebel, J. (2003). Persuasion bias, social influence, and unidimensional opinions. *The Quarterly journal of economics*, 118(3), 909-968.

Dollard, J., Doob, L. W., Miller, N. E., Mowrer, O. H., & Sears, R. R. (1939). *Frustration and aggression*, New Haven, Yale Univer.

Fair, C. C. (2015). Explaining support for sectarian terrorism. In *Pakistan: Piety, Maslak and Sharia. Religions*, 6(4), 1137-1167.

Fair, C. C., Kaltenthaler, K., & Miller, W. (2015). Pakistani political communication and public opinion on US drone attacks. *Journal of Strategic Studies*, 38(6), 852-872.

Fair, C. C., Malhotra, N., & Shapiro, J. N. (2012). Faith or doctrine? Religion and support for political violence in Pakistan. *Public Opinion Quarterly*, 76(4), 688-720.

Finley Scott, J. (1971). *Internalization of Norms: A Sociological Theory of Moral Commitment*.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35-41.

Gaines, B. J., & Mondak, J. J. (2009). Typing together? Clustering of ideological types in online social networks. *Journal of Information Technology & Politics*, 6(3-4), 216-231.
Gelfand, M. J. (2012). Culture's constraints: International differences in the strength of social norms. *Current Directions in Psychological Science*, 21(6), 420-424.
Gelfand, M. (2019). *Rule makers, rule breakers: Tight and loose cultures and the secret signals that direct our lives*. Scribner.
Granovetter, M. (1978). Threshold models of collective behavior. *American journal of sociology*, 83(6), 1420-1443.
Gurr, T. (1968). Psychological factors in civil violence. *World politics*, 20(2), 245-278.
Gurr, T.R. (1970) *Why Men Rebel*. New Haven, CT: Yale University Press.
Gurr, T. R., & Moore, W. H. (1997). Ethnopolitical rebellion: A cross-sectional analysis of the 1980s with risk assessments for the 1990s. *American Journal of Political Science*, 1079-1103.
Horgan, J. (2005). The social and psychological characteristics of terrorism and terrorists. In *Root causes of terrorism* (pp. 62-71). Routledge.
Juergensmeyer, M. (2000). Understanding the new terrorism. *Current History*, 99(636), 158.
Kaltenthaler, K., Miller, W. J., Ceccoli, S., & Gelleny, R. (2010). The sources of Pakistani attitudes toward religiously motivated terrorism. *Studies in Conflict & Terrorism*, 33(9), 815-835.
Kaufman, D. (1996). Constructivist-based experiential learning in teacher education. *Action in teacher education*, 18(2), 40-50.
Klandermans, B., & Oegema, D. (1987). Potentials, networks, motivations, and barriers: Steps towards participation in social movements. *American sociological review*, 519-531.
Krebs, V. (2002). Uncloaking terrorist networks. *First Monday*.
Kruglanski, A. W., Bélanger, J. J., Gelfand, M., Gunaratna, R., Hettiarachchi, M., Reinares, F., ... & Sharvit, K. (2013). Terrorism—A (self) love story: Redirecting the significance quest can end violence. *American Psychologist*, 68(7), 559.
Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, 35, 69-93.
Kruglanski, A., Jasko, K., Webber, D., Chernikova, M., & Molinario, E. (2018). The making of violent extremists. *Review of General Psychology*, 22(1), 107-120.
Kruglanski, A. W., Bélanger, J. J., & Gunaratna, R. (2019). *The three pillars of radicalization: Needs, narratives, and networks*. Oxford University Press, USA.
Latané, B. (1996). Dynamic social impact: The creation of culture by communication. *Journal of communication*, 46(4), 13-25.
Merton, R. K. (1938). Social structure and anomie. *American sociological review*, 3(5), 672-682.
Midlarsky, M. I. (2011). *Origins of political extremism: Mass violence in the twentieth century and beyond*. Cambridge University Press.
Mummendey, A., & Wenzel, M. (1999). Social discrimination and tolerance in intergroup relations: Reactions to intergroup difference. *Personality and social psychology review*, 3(2), 158-174.
Nassar, J., Pavon-Harr, V., Bosch, M., McCulloh, I. (2019). Assessing Data Quality of Annotations with Krippendorff's Alpha for Applications in Computer Vision. In *Proc. AAAI 2019 Fall Symposium*. Arlington, VA: AAAI
Neumann, P. R. (2013). The trouble with radicalization. *International affairs*, 89(4), 873-893.
Post, J. M., Ali, F., Henderson, S. W., Shanfield, S., Victoroff, J., & Weine, S. (2009). The psychology of suicide terrorism *Psychiatry*, 72(1), 13-31.
Sageman, M. (2004) *Understanding Terror Networks*. University of Pennsylvania Press
Sadayappan, S., Piorkowski, J., & McCulloh, I. (2018). Evaluation Political Party Cohesion Using Exponential Random Graph Modeling. In *Proceedings 2018 IEEE/ACM Conference on Advances in Social Network Analysis and Mining 2018*. Barcelona, Spain: IEEE/ACM.
Sailer, K. & McCulloh, I. (2012). Social Networks and Spatial Configuration – How Office Layouts Drive Social Interaction. *Journal of Social Networks*, 34(1): 47-58.
Schelling, T. C. (1971). Dynamic models of segregation. *Journal of mathematical sociology*, 1(2), 143-186.
Schreiber, C., Singh, S., & Carley, K. M. (2004). *Construct-a multi-agent network model for the co-evolution of agents and socio-cultural environments*. Carnegie-Mellon Univ Pittsburgh Pa Inst Of Software Research
Stern, J. (2003, August). *Terror in the Name of God.* New York: Ecco.
Taylor, D. M., & Louis, W. (2004). *Terrorism and the quest for identity*.
Tessler, M., & Robbins, M. D. (2007). What leads some ordinary Arab men and women to approve of terrorist acts against the United States?. *Journal of Conflict Resolution*, 51(2), 305-328.
Walker, I., & Pettigrew, T. F. (1984). Relative deprivation theory: An overview and conceptual critique. *British Journal of Social Psychology*, 23(4), 301-310.
Wiktorowicz, Q. (Ed.). (2004). *Islamic activism: A social movement theory approach*. Indiana University Press.
Wiktorowicz, Q. (2005). *Radical Islam rising: Muslim extremism in the West*. Rowman & Littlefield Publishers.

Attachment A – Privacy Certificate

Funding recipient, Arrow Analytics, LLC, certifies that performers requiring access to privacy sensitive data identifiable to a private person pursuant to a grant or cooperative agreement, will agree to comply with the terms and conditions of the notification to which they applied. This Privacy Certificate is being completed in accordance with the S&T Financial Assistance Agreements Privacy Policy for Funded Research in addition to other applicable DHS privacy policies, including DHS Directive No. 047-01, "Privacy Policy and Compliance" and DHS Directive No. 140-06, "Privacy Policy for Research Programs and Projects."[5]

Project Name:	Understanding Online Content Moderation Strategies' Impact on Targeted Violence Through a 2-Stage Model of Violent Radicalization		
Type of Certificate	☒ New ☐ Update	Submission Date:	June 21, 2023
If update, most recent prior Privacy Certificate submission date			Click here to enter a date.
Period of Performance End Date:			June 30, 2025

Brief Description of Project:
This project will involve mixed-methods to systematically understand evolving online content moderation strategies implemented by online platforms, government and non-government organizations to reduce the risk of extremist radicalization, mobilization, and violent action. The effectiveness of current and potential interventions will be evaluated through multi-agent simulation.

PRINCIPAL INVESTIGATOR			
Name:	Ian McCulloh, Ph.D.		
Employer	Arrow Analytics, LLC.		
Phone:	240-506-3417	Email:	ian@arrowanalytics.net

Please describe how you provide notice to the public that you are engaging in a research activity that may involve the collection of PII from members of the public.
Arrow Analytics will post an announcement on the company webpage, LinkedIn, and notify local press. We do not anticipate collecting and will certainly not maintain any PII.

[5] *See* DHS Directive 140-06 Privacy Policy for research Programs at https://www.dhs.gov/publication/privacy-policy-research-programs-and-projects-directive-140-06. *See also* DHS directive 047-01, https://www.dhs.gov/publication/privacy-policy-and-compliance-directive-047-01.

Please select all types of sensitive PII that may be collected as part of this effort.

- ☐ Social Security number
- ☐ Alien Number (A-Number)
- ☐ Tax Identification Number
- ☐ Visa Number
- ☐ Passport Number
- ☐ Bank Account, Credit Card, or other financial account number
- ☐ Social Media Handle/ID
- ☐ Driver's License/State ID Number
- ☐ Geolocation Data linked to an individual (including IP addresses and geolocation mobile app data)
- ☐ Biometrics. Please list modalities (e.g., fingerprints, DNA, iris scans)._____
- ☐ Basic Biographic Data (name, title, address, phone number, email address)
- ☐ Data not listed above that may be linked, or is linkable, to an individual
- ☒ None of the above

Any data independently collected through this effort will be anonymized and verified using proprietary software to ensure data privacy and anonymity. No PII data is planned to be collected.

Describe any other types of data, not addressed above, that will be collected as part of this effort. If applicable, explain steps that will be taken to address the incidental collection of PII.

NOTE: If the project includes the collection of PII, please include from whom the project will collect PII. Explain if the collection includes the data of sensitive populations (e.g., victims of crime, victims of human trafficking, refugees, asylees). Some PII may be collected with little concern because it has a low risk of harm to the individual, such as the PII in citations and in newspaper articles. Other PII may be particularly harmful if not adequately protected, such as social security numbers and biometric information. The Department balances privacy risks and mitigation measures against the need for research and efficacy of the proposed research approach.

Data on social media, personas, and their discourse will be collected. These data will be anonymized and statistically tested to ensure anonymity. For qualitative interviews, names will not be maintained and position and roles will only be retained if it can be shown that there is no statistical likelihood of identifying individual respondents.

From what sources will the data be collected? (Select all that apply)

- ☒ Newspaper/magazine articles (mass market)
- ☒ Academic/professional journals/conference proceedings
- ☐ Commercial (purchased) data
- ☒ Publicly available social media information (performers may not employ deceptive or covert practices to collect data)
  - ☒ Acknowledge that the use of false identities on publicly available social media requires approval by the principal investigator
- ☒ Publicly available State or local government agency records
- ☒ Non-DHS Publicly available federal data
- ☐ DHS data
- ☒ Focus groups/interviews
- ☒ Publicly available information provided by researchers or through publicly accessible data archives
- ☐ The dark web
- ☐ Other sources _____

How will data be collected from these sources? Please be specific.
Interviews will be conducted by the PI via phone or video (e.g., Zoom, Teams). A systematic literature review will use the Google Scholar search engine and may involve any and all data sources indicated above. We will also use public websites from non-government organizations such as the Violence Project, or Stanford's Social Network Analysis Project (SNAP). We will access some social media data via platform APIs, following all terms of use and DHS imposed restrictions on data collection.

How will the data be used? Please include in this description the types or methods of analysis that will be conducted.
Data will be anonymized and used to identify generalizable theories and frameworks to understand how online content moderation efforts moderate radicalizations processes through the internet. Qualitative data will be used for critical theory, synthesis, and narrative analysis. Quantitative data will be used for content analysis, social network analysis, developing machine learning classifiers, and developing a multi-agent simulation tool to evaluate current and potential interventions.

Will this project attempt to identify individuals through data mining?	☐ Yes (If yes, please describe below.) ☒ No
NOTE: Is this a project involving pattern-based queries, searches, or other analyses of one or more electronic databases, where: (A) a department or agency of the Federal Government, or a non-Federal entity acting on behalf of the Federal Government, is conducting the queries, searches, or other analyses to discover or locate a predictive pattern or anomaly indicative of terrorist or criminal activity on the part of any individual or individuals; (B) the queries, searches, or other analyses are not subject-based and do not use personal identifiers of a specific individual, or inputs associated with a specific individual or group of individuals, to retrieve information from the database or databases; and (C) the purpose of the queries, searches, or other analyses is not solely: (i) the detection of fraud, waste, or abuse in a Government agency or program; or (ii) the security of a Government computer system?	
If you responded "yes" expand upon your response in a text box or attach separate sheets. If you have more text than will fit in a text box, attach a separate sheet. Projects very rarely conduct activity that meets this definition of data mining. Responses to this question are for reporting purposes only	

Describe procedures to anonymize data to ensure the confidentiality of personally identifiable information, including the use of any privacy enhancing technologies.
Any individual persona names will be replaced with a unique numeric label. This is necessary for network analysis. Arrow will back test the use of demographic variables to statistically identify individuals and remove any data fields that would allow identification or aggregate variables to a level were identities cannot be estimated within a 95% confidence level.

Please list any formal (written) information sharing and access agreements (ISAAs) that enable PII to be shared with entities outside your institution/employer, whether through memoranda of understanding/agreement (MOU/MOA), letters of intent (LOI), or similar documents.
Arrow does not maintain or share PII data with any organization.

Name and title of individual with the authority to transfer data containing PII:			
Name:	Click here to enter text.	Title:	Click here to enter text.

Access to data is restricted to the following individuals (Principal Investigator; Project Staff; Contractors, Subcontractors, and Consultants). Attach additional sheets if needed			
Name	Laurie McCulloh	Title	CEO
Name	Ian McCulloh	Title	Principal Investigator
Name	Alison Geissler	Title	Consultant
Name	Joshua Ngoboc	Title	Consultant
Name	Hannah Hardy	Title	Consultant

Do personnel participating in the research program receive training on how to protect PII?
☒ Yes (Optional, if yes, please describe below) ☐ No
All individuals that access data are trained and certified through CITI Family Educational Rights and Privacy Act (FERPA). Arrow supplements training with statistical methods to verify anonymity and privacy as well as other project-related rules and regulations such as HIPPA.

Explain how the project ensures the physical and administrative security of PII:
Data will be stored on a closed network not connected to the internet and anonymized and verified as soon as practical..

If any data being is being stored outside of U.S. jurisdiction, please explain below.
All work and data storage will be within the US..

Provide a description of plans for data dissemination and project deliverables:
Project deliverables include academic papers, conference presentations, multi-agent simulation software, and video presentations. All data will be anonymized and verified to ensure no identifiable data. No PII will be collected.

Procedures for the final disposition of data:
Following the completion of this project, data will be retained for two years and then destroyed.

Name and title of individual(s) authorized to determine the final disposition of data:			
Name	Ian McCulloh	Title	Principal Investigator
Name		Title	

Will any PII be shared with DHS?	☒ No ☐ Yes (If yes, please explain below)
Will PII be released to the public?	☒ No ☐ Yes (If yes, please explain below)
n/a.	

Supplemental Information I			
Note: information provided in this section is not used to evaluate the Privacy Certificate.			
Is this project undergoing a review through an Institutional Review Board (IRB)?[6]		☒ Yes ☐ No	
Name of IRB POC:	Ian McCulloh		
Phone:	240-506-3417	Email:	Ian@arrowanalytics.net

Supplemental Information II	
Note: information in this section is not used to evaluate the Privacy Certificate.	
Is this project using or developing artificial intelligence or machine learning technologies	☒ Yes ☐ No
➢ If you answered no to the question above, skip to the Certifications section below.	
Describe how individuals will be protected from inappropriate or irrelevant data use in the design, development, and deployment of automated systems, and from the compounded harm of its reuse.	
Arrow applies code to measure, evaluate, and compare potential bias against human baseline and ensure no protected classes are unfairly targeted to ensure any AI/ML system reduces potential harm or threats to civil liberties over manual	

[6] DHS does not require a copy of the IRB package as part of the Privacy Certificate; those go through a separate process at DHS.

human systems. Individuals are not directly being impacted by the ML proposed in this effort, rather ML is used to statistically classify online personas.
Affirm that individuals will not face discrimination by algorithms and systems and that they will be designed in an equitable way. Please Explain.
Individuals will not face discrimination by algorithms. Arrow statistically tests code to estimate protected class to verify no undue discrimination. We compare any AI/ML solution to a human baseline to ensure algorithms and systems reduce human bias and harm.
Affirm that individuals will be protected from abusive data practices via built-in protections and will have agency over how data about them is used. Please explain.
Individuals will not face discrimination by algorithms. Arrow statistically tests code to estimate protected class to verify no undue discrimination. We compare any AI/ML solution to a human baseline to ensure algorithms and systems reduce human bias and harm.
Affirm that individuals will know that an automated system is being used and understand how and why it contributes to outcomes that impact them. Please explain.
Individuals will not interact with any automated system.
Explain if individuals will be able to opt out, where appropriate, and have access to a person who can quickly consider, and remedy problems individuals encounter.
Individuals will not interact with any automated system.

Certifications

1. Funding recipient certifies that project plans will be designed to preserve the confidentiality of private persons to whom information relates, including where appropriate, name-stripping, coding of data, or other similar procedures.
2. Funding recipient certifies that the procedures described above are correct and shall be carried out.
3. Funding recipient certifies that all project personnel, including subcontractors, have been advised of and have agreed, in writing, to comply with all procedures outlined in this privacy certificate.
4. Funding recipient certifies that access to PII will be limited to those employees having a need for such data and that such employees shall be advised of the DHS privacy requirements included in the award.
5. Funding recipient certifies that DHS shall be notified of any material change in any of the information provided in this Privacy Certificate.
6. Funding recipient certifies that project findings and reports prepared for public release will not contain raw data that includes personally identifiable information.

Signature (s):		
	(Principal Investigator)	June 21, 2023
	(Institutional Representative)	June 21, 2023