Artificial Intelligence Agents in Counter-Extremism: A Framework for Ethical Deployment in Digital Deradicalization

Aadil Bouhlaoui *

Posted Date: 18 June 2025

doi: 10.20944/preprints202506.1513.v1

Keywords: artificial intelligence; counter-extremism; digital radicalization; Islamic theology; AI ethics; counter-narrative; deradicalization; EU AI Act; community engagement

 doi:10.20944/preprints202506.1513.v1

Article

Artificial Intelligence Agents in Counter-Extremism: A Framework for Ethical Deployment in Digital Deradicalization

Aadil Bouhlaoui

Department of Digital Humanities, King's College London, Virginia Woolf Building, 22 Kingsway, London WC2B 6LE, United Kingdom; aadil.bouhlaoui@kcl.ac.uk

Abstract: This article presents a comprehensive analysis of artificial intelligence (AI) agent deployment strategies for countering online extremism, with particular focus on addressing the phenomenon of digital radicalization in Islamic contexts. Drawing upon recent developments in AI capabilities, evolving legal frameworks including the EU AI Act, and emerging patterns of extremist adaptation to digital technologies, this study examines the technical feasibility, legal permissibility, ethical implications, and theological dimensions of AI-mediated counter-extremism operations. The research integrates contemporary case studies, including the Islamic State's 2023 AI propaganda guide and the systematic migration of extremist activities to gaming platforms, to provide evidence-based strategic recommendations for policymakers and security practitioners. The analysis reveals a fundamental tension between the definitional ambiguity surrounding "Keyboard Jihad" and operational requirements for precise targeting. While academics employ the term to describe legitimate intellectual efforts to rectify misperceptions of Islam, security practitioners use it to denote online terrorist propaganda and recruitment activities. This definitional dichotomy presents severe operational risks of misidentifying legitimate discourse, potentially validating extremist narratives and causing strategic blowback that undermines counter-extremism objectives. Through systematic evaluation of three distinct AI agent deployment models—overt analytical agents, direct engagement agents, and covert engagement agents—this study demonstrates that transparent, community-partnered approaches offer superior strategic effectiveness compared to surveillance-based or deceptive methodologies. The research establishes that direct engagement AI agents, designed to provide authentic theological guidance and counter-narratives, represent the most promising paradigm for addressing critical gaps in legitimate Islamic knowledge (Al-Ilm Al-Shari) that extremist groups exploit for recruitment and radicalization purposes. The study concludes that covert AI agents for engagement and influence operations present insurmountable legal, ethical, and strategic barriers under current regulatory frameworks, particularly the EU AI Act's comprehensive requirements for high-risk AI systems. Conversely, the principle of maslaha (public interest) in Islamic jurisprudence provides theological justification for transparent AI agents that offer authentic guidance while respecting community values and democratic principles. The article proposes a three-track strategic framework prioritizing immediate deployment of overt analytical capabilities with comprehensive safeguards, pilot development of direct engagement agents through extensive community consultation and theological validation, and suspension of covert engagement capabilities pending explicit legal authorization and public debate. This approach emphasizes competing with extremist narratives through superior theological authenticity and genuine community partnership rather than through deception or surveillance, aligning strategic effectiveness with democratic values and human rights protections.

Keywords: artificial intelligence; counter-extremism; digital radicalization; Islamic theology; AI ethics; counter-narrative; deradicalization; EU AI Act; community engagement

 doi:10.20944/preprints202506.1513.v1

1. Introduction

The intersection of artificial intelligence and counter-extremism represents one of the most complex and consequential challenges facing contemporary security studies and digital policy. As extremist groups increasingly exploit sophisticated AI technologies for propaganda creation, recruitment, and operational security [1], policymakers and security practitioners confront an urgent imperative to develop effective, ethical, and legally compliant responses. The 2022-2025 period has witnessed unprecedented developments in this domain, including the Islamic State's publication of comprehensive guides for using generative AI in propaganda operations [2], systematic migration of extremist activities to gaming platforms and encrypted channels [3], and the implementation of the European Union's AI Act, which establishes comprehensive regulatory frameworks for high-risk AI systems in law enforcement contexts [4].

This article addresses a critical gap in academic literature by providing the first comprehensive analysis of AI agent deployment strategies specifically designed for counter-extremism operations. While existing scholarship has examined AI applications in security contexts broadly [5] and explored digital radicalization patterns [6], no previous study has systematically evaluated the technical feasibility, legal permissibility, ethical implications, and theological dimensions of deploying AI agents for direct engagement with individuals at risk of radicalization. This research fills this lacuna by integrating insights from computer science, legal studies, ethics, and Islamic theology to develop a holistic framework for AI-mediated counter-extremism.

The study's significance extends beyond academic inquiry to address pressing policy challenges. Recent estimates suggest that terrorist attacks impose enormous economic and social costs on affected societies, with individual incidents generating direct costs exceeding £45 million and broader economic impacts reaching hundreds of millions [7]. The 2019 Christchurch attacks alone prompted New Zealand to allocate over NZ$200 million for victim support and community recovery efforts [8]. Beyond immediate financial costs, the long-term societal impact of radicalization includes family breakdown, community fragmentation, and the loss of human potential as individuals become isolated from mainstream society. These costs underscore the urgent need for innovative, effective approaches to counter-extremism that can address root causes rather than merely responding to symptoms.

1.1. The Digital Transformation of Extremism

The contemporary digital extremism landscape bears little resemblance to the static websites and email chains that characterized early online jihadist activity. The transformation has been particularly pronounced in the 2022-2025 period, which has witnessed fundamental changes in both the sophistication of extremist digital operations and the technological capabilities available to counter them. Extremist groups have demonstrated remarkable adaptability in exploiting emerging technologies, with the Islamic State's 2023 AI propaganda guide representing a watershed moment in the weaponization of artificial intelligence for terrorist purposes [2].

This digital transformation manifests across multiple dimensions. First, extremist groups have migrated from traditional social media platforms to gaming environments, Discord servers, and encrypted messaging applications, exploiting these platforms' community- building features and reduced content moderation [3]. Research by Collison-Randall et al. demonstrates that gaming adjacent platforms have created expanding ecosystems where extremist groups can communicate and connect with users globally, with esports providing particular opportunities for targeting Generation Z audiences [9]. The

Australian Federal Police's 2022 warning about extremist groups accessing online games to recruit children reflects growing recognition of this threat vector [10].

Second, the sophistication of AI-powered propaganda has increased exponentially. Molas and Lopes' research reveals how far-right users have successfully exploited AI tools through jailbreaking techniques, accelerating the spread of harmful content and demonstrating the dual-use nature of AI

 doi:10.20944/preprints202506.1513.v1

technologies [11]. These developments indicate that extremist groups are not merely passive consumers of technology but active innovators who rapidly adapt emerging capabilities to serve their objectives.

Third, the scale and reach of digital extremism have expanded dramatically. Unlike traditional recruitment methods that required physical proximity or established networks, digital platforms enable extremist groups to reach vulnerable individuals across geographical boundaries and cultural contexts. This global reach, combined with AI's capacity for personalization and targeting, creates unprecedented opportunities for radicalization while simultaneously challenging traditional counter-extremism approaches that rely on geographical or community-based interventions.

1.2. The "Keyboard Jihad" Definitional Challenge

Central to any effective counter-extremism operation is the precise definition of target activities and populations. The term "Keyboard Jihad" presents a particularly acute challenge in this regard, embodying a fundamental definitional ambiguity that poses significant operational and strategic risks. This ambiguity is not merely semantic but reflects deeper tensions between academic understanding, community perspectives, and security imperatives that must be carefully navigated to avoid counterproductive outcomes.

The academic conceptualization of "Keyboard Jihad" emerges from Professor Abdul Karim Bangura's seminal work, which frames the concept as a constructive intellectual endeavor aimed at correcting widespread misunderstandings about Islam and Muslims [12]. Bangura's approach represents the term as a form of digital scholarship and counter-narrative work, utilizing online platforms to engage in interfaith dialogue, dispel misconceptions, and promote peaceful understanding. This definition positions the "keyboard" as an instrument of education and clarification rather than radicalization or violence, emphasizing the original meaning of "jihad" as struggle or striving, particularly the "greater jihad" of internal spiritual development.

In stark contrast, security practitioners and media outlets employ "Keyboard Jihad" to describe the use of digital platforms by extremist groups for propaganda dissemination, recruitment activities, and incitement to violence [13]. This usage aligns with concepts of "media mujahideen" and digital warfare, where online activities serve as weapons in an

ideological conflict. The Combating Terrorism Center at West Point's analysis describes the "sterile echo chamber of keyboard jihad" as a space where sympathizers may discuss extremist ideology and potentially transition from online engagement to actual militancy [14].

The operational implications of this definitional dichotomy are profound and potentially catastrophic. Conflating legitimate academic and religious discourse with extremist propaganda could result in the targeting of scholars, religious leaders, and community advocates engaged in legitimate counter-narrative work. Such targeting would validate extremist claims about state persecution of Muslims, potentially driving moderate voices away from counter-extremism efforts and creating new grievances that extremist groups could exploit for recruitment purposes. The risk of strategic blowback extends beyond immediate operational concerns to encompass broader community relations and democratic legitimacy.

1.3. Research Questions and Objectives

This study addresses three primary research questions that emerge from the contemporary challenges outlined above:

Technical Feasibility: What are the technical requirements, capabilities, and limitations of different AI agent deployment models for counter-extremism operations, and how do these align with current technological capabilities and regulatory constraints?

Legal and Ethical Framework: How do existing legal frameworks, particularly the EU AI Act, constrain or enable different approaches to AI agent deployment, and what ethical principles should guide the development and implementation of such systems?

 doi:10.20944/preprints202506.1513.v1

Strategic Effectiveness: What deployment models offer the greatest potential for achieving counter-extremism objectives while maintaining democratic legitimacy, community trust, and operational sustainability?

The research objectives encompass both theoretical analysis and practical application. Theoretically, the study seeks to develop a comprehensive framework for evaluating AI agent deployment in sensitive security contexts, integrating technical, legal, ethical, and theological considerations. Practically, the research aims to provide actionable recommendations for policymakers, technology developers, and security practitioners working to address digital radicalization challenges.

2. Literature Review

2.1. Digital Radicalization and Online Extremism

The academic literature on digital radicalization has evolved significantly since the early 2000s, reflecting both the changing nature of online extremism and the development of more sophisticated analytical frameworks. Early scholarship focused primarily on the role of websites and forums in facilitating extremist communication and recruitment [15]. However, recent research has revealed a more complex landscape characterized by platform migration, algorithmic amplification, and the exploitation of mainstream social media features for extremist purposes.

Collison-Randall et al.'s 2024 research represents a significant advancement in understanding contemporary digital extremism patterns. Their analysis of media framing around far-right extremism and online radicalization in esports and gaming contexts reveals how extremist groups have systematically exploited gaming platforms' community-building features and reduced content moderation to establish new recruitment pathways [9]. The research demonstrates that gaming adjacent platforms have created expanding ecosystems where extremist groups can communicate and connect with users globally, with particular focus on targeting Generation Z audiences through esports engagement.

The migration of extremist activities to gaming platforms reflects broader patterns of platform adaptation that characterize contemporary digital extremism. As mainstream social media platforms have implemented more sophisticated content moderation systems, extremist groups have demonstrated remarkable agility in identifying and exploiting alternative spaces. This pattern of migration and adaptation challenges traditional approaches to counter-extremism that focus on specific platforms or technologies rather than addressing underlying recruitment and radicalization processes.

Recent research has also highlighted the role of algorithmic systems in facilitating radicalization pathways. While platforms' recommendation algorithms are designed to maximize user engagement, they can inadvertently create "rabbit holes" that lead users from mainstream content to increasingly extreme material [16]. This algorithmic amplification effect is particularly concerning in the context of AI-powered content generation, which can produce personalized extremist content at unprecedented scale and sophistication.

2.2. AI Applications in Security and Counter-Terrorism

The application of artificial intelligence technologies in security and counter-terrorism contexts has generated substantial academic and policy interest, though much of the literature remains focused on surveillance and detection rather than direct intervention approaches. Bellaby's 2024 analysis of intelligence-AI ethics provides a comprehensive framework for understanding the moral implications of AI deployment in intelligence operations, emphasizing the need for careful consideration of proportionality, necessity, and human oversight [17].

The ethical challenges identified by Bellaby are particularly relevant to counter- extremism applications, where AI systems may be deployed to influence human behavior and beliefs. The research highlights tensions between operational effectiveness and respect for human autonomy,

 doi:10.20944/preprints202506.1513.v1

privacy rights, and democratic values. These tensions are especially acute in the context of AI agents designed for direct engagement with vulnerable individuals, where the potential for manipulation and coercion must be carefully balanced against legitimate security objectives.

Molas and Lopes' research on far-right jailbreaking of AI systems reveals the dual-use nature of AI technologies and the challenges facing efforts to prevent malicious exploitation [11]. Their analysis demonstrates how extremist groups have successfully circumvented AI safety measures to generate harmful content, including propaganda materials, recruitment messaging, and instructional content for violent activities. This research underscores the importance of developing robust safety measures and oversight mechanisms for AI systems deployed in counter-extremism contexts.

The technical capabilities required for effective AI-mediated counter-extremism operations encompass multiple domains, including natural language processing, conversational AI, content analysis, and behavioral modeling. Current large language models demonstrate sophisticated capabilities for engaging in theological discussions, providing personalized responses to individual concerns, and maintaining coherent long-term conversations about complex topics [18]. However, significant challenges remain in ensuring theological accuracy, cultural sensitivity, and appropriate crisis response capabilities.

2.3. Counter-Narrative Approaches and Deradicalization Programs

The effectiveness of counter-narrative approaches and deradicalization programs has been the subject of extensive academic research, with mixed findings regarding their impact on preventing radicalization and promoting disengagement from extremist groups. Duarte et al.'s 2025 systematic review of educational programmes to prevent

violent extremism provides valuable insights into the factors that contribute to program effectiveness [19].

The research reveals that successful counter-extremism programs typically share several characteristics: they are developed in partnership with affected communities, they address underlying grievances and vulnerabilities rather than focusing solely on ideological content, and they provide alternative pathways for meaning-making and social connection. These findings have important implications for AI-mediated counter- extremism approaches, suggesting that technological solutions must be embedded within broader community engagement strategies to achieve sustainable impact.

The theological dimensions of counter-narrative work are particularly important in the context of Islamic extremism, where extremist groups exploit religious concepts and texts to justify violence and recruit supporters. Effective counter-narratives must demonstrate superior theological authenticity and scholarship while addressing the underlying spiritual and intellectual needs that extremist groups claim to fulfill. This requirement presents both opportunities and challenges for AI systems, which can provide access to authentic Islamic scholarship at scale but may lack the spiritual authority and contextual understanding that characterize human religious guidance.

Recent research has also highlighted the importance of addressing the social and psychological factors that contribute to radicalization vulnerability. Individuals at risk of radicalization often experience social isolation, identity confusion, and a sense of grievance or injustice that extremist groups exploit for recruitment purposes [20].

Effective counter-extremism approaches must address these underlying vulnerabilities while providing alternative sources of meaning, community, and purpose.

3. Materials and Methods

3.1. Research Design and Analytical Framework

This study employs a multidisciplinary analytical framework that integrates insights from computer science, legal studies, Islamic theology, and security studies to evaluate AI agent deployment models for counter-extremism operations. The research design combines theoretical

 doi:10.20944/preprints202506.1513.v1

analysis, case study examination, and comparative assessment to provide comprehensive evaluation of different deployment approaches across multiple dimensions of feasibility and effectiveness.

The methodological approach recognizes that counter-extremism operations exist at the intersection of multiple domains, each with distinct requirements, constraints, and evaluation criteria. Technical feasibility must be assessed alongside legal permissibility, ethical implications, and strategic effectiveness to provide meaningful guidance for

policy and practice. This multidisciplinary approach ensures that recommendations are grounded in realistic understanding of operational constraints while maintaining commitment to democratic values and human rights.

The analytical framework evaluates three distinct AI agent deployment models across four primary dimensions:

Technical Feasibility encompasses the current state of AI technology capabilities, implementation requirements, scalability considerations, and technical risks. This dimension examines whether proposed interventions can be implemented using existing or near-term AI technologies, what technical infrastructure would be required, and what technical limitations might constrain operational effectiveness.

Legal Permissibility examines compliance with existing and emerging legal frameworks, including the EU AI Act, data protection regulations, human rights law, and national security legislation. This dimension considers both explicit legal requirements and broader constitutional principles that constrain government action in democratic societies.

Ethical Implications assess the moral dimensions of different deployment approaches, including respect for human autonomy, privacy rights, religious freedom, and community self-determination. This dimension draws on established ethical frameworks while considering the specific cultural and religious sensitivities relevant to counter- extremism operations.

Strategic Effectiveness evaluates the likely operational impact of different approaches, including their capacity to achieve stated counter-extremism objectives, potential for unintended consequences, and sustainability over time. This dimension considers both direct effects on target populations and broader systemic impacts on community relations and democratic governance.

3.2. Case Study Selection and Analysis

The analysis incorporates examination of recent developments in digital extremism and counter-extremism to ground theoretical analysis in contemporary realities. Key case studies include the Islamic State's 2023 AI propaganda guide, extremist migration to gaming platforms, the implementation of the EU AI Act's provisions for high-risk AI systems in security applications, and emerging patterns of AI jailbreaking by extremist groups.

These case studies were selected to represent different aspects of the contemporary digital extremism landscape and to illustrate the practical challenges facing counter- extremism operations. The Islamic State's AI adoption demonstrates the sophistication of contemporary extremist technological capabilities and the urgent need for equally sophisticated counter-measures. Platform migration patterns illustrate the adaptive capacity of extremist networks and the limitations of platform-specific interventions. Regulatory developments provide insight into the legal and policy constraints that will shape future counter-extremism operations.

The case study analysis employs a structured approach that examines each case across the four analytical dimensions outlined above. This approach enables systematic comparison of different scenarios and identification of common patterns and challenges that inform the broader analytical framework.

3.3. Theological Analysis and Islamic Jurisprudential Considerations

The research incorporates detailed analysis of Islamic theological principles relevant to counter-extremism operations, particularly the concepts of ***maslaha*** (public interest), ***tajassus*** (surveillance), and the proper role of technology in religious guidance and community service. This theological

 doi:10.20944/preprints202506.1513.v1

analysis is essential for understanding how different intervention approaches might be perceived by Muslim communities and for developing approaches that can achieve operational objectives while maintaining community trust and cooperation.

The theological analysis draws on classical Islamic jurisprudence as well as contemporary scholarly interpretations to provide nuanced understanding of how different intervention approaches align with or conflict with Islamic ethical principles. Particular attention is paid to the principle of ***maslaha***, which provides a framework for evaluating actions based on their contribution to public welfare and the prevention of harm. This principle is especially relevant to AI-mediated counter-extremism operations, which must balance potential benefits in preventing radicalization against risks of community alienation and rights violations.

The analysis also examines the concept of ***tajassus*** (surveillance or spying), which is generally prohibited in Islamic ethics except under specific circumstances involving imminent threats to community safety. This prohibition has important implications for covert AI operations and surveillance-based approaches to counter-extremism, suggesting that transparent, community-partnered approaches may be more theologically defensible and strategically effective.

3.4. Legal and Regulatory Analysis

The legal analysis focuses primarily on the European Union's AI Act, which represents the most comprehensive regulatory framework for AI systems currently in force. The AI Act's classification system for AI applications, risk assessment requirements, and

governance mechanisms provide important constraints and opportunities for AI deployment in counter-extremism contexts.

The analysis examines how different AI agent deployment models would be classified under the AI Act's risk categories, what compliance requirements would apply, and what governance mechanisms would be necessary to ensure legal operation. Particular attention is paid to the Act's requirements for high-risk AI systems, which include many security and law enforcement applications.

The legal analysis also considers broader human rights frameworks, including the European Convention on Human Rights, data protection regulations, and constitutional principles that constrain government action in democratic societies. These frameworks provide important safeguards against abuse while also establishing legitimate grounds for security operations that serve compelling public interests.

3.5. Limitations and Constraints

This study acknowledges several important limitations that affect the scope and applicability of its findings. The analysis is necessarily theoretical given the sensitive nature of counter-extremism operations and the limited availability of detailed information about current practices. The rapid pace of technological development means that technical assessments may become outdated as AI capabilities continue to evolve. Legal and regulatory frameworks are also evolving rapidly, particularly in the area of AI governance, which may affect the applicability of current legal analysis.

The study does not include primary data collection involving human subjects, both for ethical reasons and due to the sensitive nature of the research topic. Instead, the analysis relies on publicly available sources, academic literature, and legal documentation. While this approach provides valuable insights, it necessarily limits the depth of understanding about how different approaches might be perceived or received by target communities.

The theological analysis, while comprehensive, reflects the author's interpretation of Islamic jurisprudential principles and may not capture the full diversity of scholarly opinion on these issues. The analysis attempts to present mainstream scholarly positions while acknowledging areas of disagreement and uncertainty.

 doi:10.20944/preprints202506.1513.v1

Results and Analysis

4.1. Overt Analytical Agents: Intelligence and Threat Assessment

Overt analytical agents represent AI systems deployed transparently for intelligence gathering, content analysis, and threat assessment in counter-extremism operations. These systems operate with clear identification as government or institutional tools, focusing on data collection and analysis rather than direct intervention with target populations. The evaluation of this deployment model reveals both significant capabilities and important limitations across the analytical framework.

Technical Feasibility Assessment

Current AI technologies demonstrate strong capabilities for content analysis, pattern recognition, and threat assessment that would support overt analytical operations. Natural language processing systems can effectively identify extremist content, track narrative evolution, and assess threat levels across multiple languages and platforms. Machine learning algorithms can detect behavioral patterns associated with radicalization processes and identify emerging trends in extremist communications with accuracy rates exceeding 85% in controlled testing environments [21].

The technical infrastructure required for overt analytical agents is substantial but achievable using existing technologies. Large-scale data processing capabilities would be necessary to handle the volume of content generated across multiple platforms daily. Sophisticated natural language processing systems would be required to analyze content in multiple languages and cultural contexts, with particular attention to Arabic, Urdu, and other languages commonly used in Islamic discourse. Robust security measures would be essential to protect sensitive intelligence data and prevent unauthorized access or manipulation.

Integration with existing intelligence systems and databases would enhance analytical capabilities while requiring careful attention to data security and access controls. The system would need to interface with law enforcement databases, immigration records, and other government information systems while maintaining appropriate separation between different types of data and ensuring compliance with data protection regulations.

However, significant technical challenges remain. The definitional ambiguity surrounding "Keyboard Jihad" creates substantial risks of false positives, where legitimate religious discourse might be incorrectly identified as extremist content. The system would need sophisticated contextual understanding to distinguish between academic discussions of jihad concepts, legitimate religious education, and actual extremist propaganda. Current AI systems struggle with this level of contextual nuance, particularly when dealing with religious and cultural concepts that have multiple legitimate interpretations.

Legal Permissibility Analysis

Overt analytical agents operate within well-established legal frameworks for intelligence gathering and threat assessment. The transparent nature of their deployment addresses many privacy and due process concerns that arise with covert operations. However, significant legal constraints remain, particularly regarding data collection from private communications, cross-border operations, and the use of AI systems for decision- making that affects individual rights.

Under the EU AI Act, most analytical applications would be classified as high-risk AI systems, requiring comprehensive risk assessment, human oversight, and transparency measures. Article 6 of the Act establishes specific requirements for AI systems used in law enforcement contexts, including mandatory conformity assessments, quality management systems, and ongoing monitoring of system performance [4]. Compliance with these requirements is achievable but would require substantial investment in governance systems and ongoing monitoring capabilities.

Data protection regulations impose additional constraints on data collection, processing, and retention that must be carefully managed. The General Data Protection Regulation (GDPR) requires

 doi:10.20944/preprints202506.1513.v1

that data processing be necessary, proportionate, and based on legitimate legal grounds [22]. For counter-extremism operations, the legal basis would likely be public task or legitimate interests, but the processing would still need to comply with principles of data minimization, purpose limitation, and individual rights.

Cross-border data sharing presents particular legal challenges, especially when dealing with platforms and communications that span multiple jurisdictions. International cooperation agreements and mutual legal assistance treaties provide frameworks for legitimate information sharing, but these processes can be slow and may not accommodate the real-time analysis capabilities that AI systems enable.

Ethical Implications and Community Impact

The overt nature of analytical systems addresses many ethical concerns about deception and manipulation that arise with covert operations. Transparency about the existence and purpose of these systems enables public debate about their appropriateness and provides opportunities for oversight and accountability. However, significant ethical challenges remain regarding privacy, surveillance, and the potential for discriminatory targeting.

The use of AI systems for analyzing religious and political expression raises particular concerns about freedom of speech and religious liberty. Even when conducted transparently, systematic monitoring of religious discourse could have chilling effects on legitimate religious expression and community participation. Muslim communities, which have experienced disproportionate surveillance in many Western countries, may view such systems as continuation of discriminatory practices regardless of their stated purposes [23].

Community trust considerations are complex for overt analytical systems. While transparency may enhance legitimacy compared to covert operations, the explicit surveillance function may generate community resistance and reduce cooperation with counter-extremism efforts. The history of surveillance programs targeting Muslim communities has created deep skepticism about government monitoring activities, even when conducted under legal authority and with appropriate oversight [24].

The potential for discriminatory targeting represents a significant ethical concern. AI systems trained on historical data may perpetuate existing biases in law enforcement and intelligence gathering, leading to disproportionate focus on certain communities or types of expression. The definitional challenges surrounding "Keyboard Jihad" exacerbate these risks, as systems may be more likely to flag content from Muslim users or discussions of Islamic concepts as potentially extremist.

Strategic Effectiveness Evaluation

Overt analytical agents provide valuable intelligence capabilities that can enhance understanding of extremist networks, track threat evolution, and support targeted interventions. The ability to process large volumes of content in real-time enables identification of emerging trends and threats that might otherwise go undetected. Pattern recognition capabilities can reveal connections between seemingly disparate individuals and groups, providing insights into network structures and operational planning.

However, the overt nature of these systems limits their capacity for direct engagement with target populations and may prompt adaptive responses from extremist groups seeking to evade detection. Sophisticated extremist organizations are likely to modify their communication patterns, adopt new platforms, or employ encryption and obfuscation techniques to avoid analytical detection. This adaptive response could reduce the long-term effectiveness of analytical systems while potentially driving extremist activities to more secure and less monitored platforms.

The strategic value of analytical systems lies primarily in their support for other counter-extremism activities rather than as standalone interventions. Intelligence gathered through AI analysis can inform community engagement efforts, support law

 doi:10.20944/preprints202506.1513.v1

enforcement investigations, and guide policy development. However, analytical systems cannot directly address the underlying factors that drive radicalization or provide alternative narratives to counter extremist messaging.

The risk of strategic blowback must also be considered. If analytical systems are perceived as discriminatory or intrusive, they could damage community relations and reduce cooperation with legitimate counter-extremism efforts. The revelation of extensive surveillance programs has historically led to community withdrawal from engagement with law enforcement and government agencies, potentially undermining broader security objectives [25].

4.2. Direct Engagement Agents: Theological Guidance and Counter- Narratives

Direct engagement agents represent AI systems designed to interact directly with individuals at risk of radicalization, providing counter-narratives, theological guidance, and alternative perspectives through transparent AI-mediated communications. These systems operate openly as AI agents while engaging in substantive dialogue about religious, political, and social issues. The evaluation reveals this approach offers significant potential while requiring careful implementation to address theological, ethical, and operational challenges.

Technical Feasibility and Implementation Requirements

Recent advances in large language models and conversational AI provide strong technical foundations for direct engagement systems. Current state-of-the-art models demonstrate sophisticated capabilities for engaging in theological discussions, providing personalized responses to individual concerns, and maintaining coherent long-term conversations about complex religious and philosophical topics. GPT-4 and similar models show particular strength in handling nuanced religious concepts and can engage with Islamic theological principles at a level that approaches scholarly discourse [26].

The technical architecture for direct engagement agents would require several specialized components. A comprehensive knowledge base of Islamic scholarship would be essential, including classical texts, contemporary scholarly interpretations, and approved counter-narrative materials. This knowledge base would need to be continuously updated to reflect evolving scholarly consensus and emerging extremist narratives that require response. Natural language processing capabilities would need to be optimized for religious and cultural contexts, with particular attention to Arabic terminology and concepts that may not translate directly into other languages.

Integration with crisis intervention systems represents a critical technical requirement. The AI system must be capable of recognizing when conversations indicate imminent threats of violence, severe psychological distress, or other situations requiring immediate human intervention. This capability requires sophisticated sentiment analysis, threat assessment algorithms, and robust escalation procedures that can connect individuals with appropriate human support services within minutes rather than hours.

However, significant technical challenges remain. Ensuring theological accuracy across the full spectrum of Islamic jurisprudential schools and interpretations requires extensive validation by qualified scholars. The system must be capable of recognizing the limits of its knowledge and appropriately referring users to human experts when discussions move beyond its competence. Cultural sensitivity presents another major challenge, as religious concepts and practices vary significantly across different Muslim communities and cultural contexts.

The personalization capabilities that make AI systems effective for engagement also create risks of manipulation or inappropriate influence. The system must be designed to provide authentic guidance while respecting individual autonomy and avoiding coercive or manipulative techniques. This balance requires careful attention to the design of conversation flows, response generation algorithms, and oversight mechanisms.

Legal Framework and Regulatory Compliance

 doi:10.20944/preprints202506.1513.v1

Direct engagement agents operate within established frameworks for public education and community outreach, avoiding many of the legal complications associated with covert operations. The transparent nature of AI-mediated engagement addresses concerns about deception while the educational focus aligns with legitimate government interests in promoting social cohesion and preventing violence. However, several legal challenges require careful consideration.

The provision of religious guidance by government-affiliated systems raises potential establishment clause issues in jurisdictions with strong separation of church and state principles. While the focus on counter-extremism provides a compelling secular purpose, the detailed engagement with religious concepts and practices could be viewed as government endorsement of particular religious interpretations. Careful design of engagement protocols and clear limitations on the scope of AI responses would be necessary to maintain constitutional compliance.

Under the EU AI Act, direct engagement systems would likely be classified as high-risk AI systems due to their potential impact on individual behavior and decision-making.

Article 5 prohibits AI systems that deploy subliminal techniques or exploit vulnerabilities to materially distort behavior in ways that cause psychological or physical harm [4].

Direct engagement agents must be designed to avoid these prohibited techniques while still providing effective counter-narratives and guidance.

Data protection considerations are particularly complex for direct engagement systems, which would necessarily collect and process sensitive personal data about religious beliefs, political opinions, and psychological states. The GDPR provides special protections for this type of data, requiring explicit consent or other strong legal justifications for processing [22]. The counter-extremism context may provide legitimate grounds for processing, but individuals would retain rights to access, correct, and delete their personal data.

Cross-border operations present additional legal challenges, particularly when engaging with individuals in different jurisdictions with varying laws regarding religious expression, government speech, and AI systems. International cooperation frameworks would be necessary to ensure legal compliance while enabling effective cross-border engagement with transnational extremist networks.

Ethical Considerations and Theological Authenticity

Direct engagement agents raise complex ethical questions about the appropriate role of AI systems in religious and political discourse. The provision of theological guidance by AI systems requires careful attention to authenticity, accuracy, and respect for religious authority structures. The legitimacy of these systems depends heavily on their acceptance by religious communities and their alignment with established theological principles.

The principle of *maslaha* (public interest) in Islamic jurisprudence provides a framework for evaluating the ethical permissibility of AI-mediated religious guidance. If such systems genuinely serve the public interest by preventing harm and promoting authentic understanding of Islamic principles, they may be theologically justified even if they represent a departure from traditional models of religious authority [27]. However, this justification requires that the systems provide genuinely authentic and beneficial guidance rather than serving primarily as tools of state control or surveillance.

Community consultation and ongoing oversight by qualified religious scholars would be essential to maintain ethical legitimacy. The development and operation of direct engagement agents should involve extensive consultation with diverse Muslim communities and religious authorities to ensure that the systems reflect authentic Islamic scholarship and address genuine community needs. Ongoing validation of AI responses by qualified scholars would be necessary to maintain theological accuracy and authenticity.

The transparent nature of these systems addresses concerns about deception and manipulation while enabling informed consent from users. Individuals engaging with AI agents would know they are interacting with artificial systems rather than human religious authorities, allowing them to make

 doi:10.20944/preprints202506.1513.v1

informed decisions about the weight to give to AI-provided guidance. Clear disclosure of AI capabilities and limitations would be necessary to ensure ethical operation.

However, questions remain about the potential for subtle influence and the appropriate boundaries for AI engagement with vulnerable individuals. Even transparent AI systems can be highly persuasive, particularly when engaging with individuals experiencing psychological distress or identity confusion. The design of engagement protocols must carefully balance effectiveness in providing counter-narratives with respect for individual autonomy and decision-making capacity.

Strategic Effectiveness and Community Impact

Direct engagement agents offer significant potential for addressing the root causes of radicalization by providing alternative narratives, theological guidance, and social support to individuals at risk. The scalability of AI systems enables engagement with large numbers of individuals simultaneously while the personalization capabilities allow for tailored responses to individual circumstances and concerns. This combination of scale and personalization represents a significant advantage over traditional counter- extremism approaches that rely on limited human resources.

The strategic effectiveness of direct engagement systems depends heavily on their perceived authenticity and legitimacy within target communities. AI agents that are viewed as government propaganda tools or theologically inauthentic are unlikely to achieve meaningful impact and may actually reinforce extremist narratives about state persecution of Muslims. Success requires genuine community partnership, ongoing validation by respected religious authorities, and demonstrated commitment to serving community needs rather than purely security objectives.

The ability to provide immediate, accessible responses to religious questions and concerns addresses a critical gap that extremist groups often exploit. Many individuals at risk of radicalization have legitimate questions about Islamic principles and practices but lack access to qualified religious guidance. Extremist groups exploit this knowledge gap by providing simplified, distorted interpretations that serve their recruitment objectives. AI systems that can provide authentic, accessible religious guidance could significantly reduce the appeal of extremist narratives.

However, the long-term strategic impact depends on the broader context of community relations and government policy. If direct engagement agents are deployed as part of a broader strategy that includes genuine community partnership, investment in community institutions, and addressing underlying grievances, they may contribute to sustainable reductions in radicalization risk. If they are viewed as technological substitutes for genuine community engagement or as tools for monitoring and control, they may generate backlash that undermines counter-extremism objectives.

The potential for positive community impact extends beyond direct counter-extremism effects. AI systems that provide authentic religious guidance and support could strengthen community resilience against extremist recruitment while also serving broader educational and social support functions. This dual purpose could enhance community acceptance while providing sustainable justification for continued operation.

4.3. Covert Engagement Agents: Deception and Influence Operations

Covert engagement agents represent AI systems that would operate without disclosure of their artificial nature or institutional affiliation, engaging with target populations through deceptive personas designed to influence beliefs and behaviors related to extremism. This deployment model raises fundamental questions about the appropriate limits of government action in democratic societies and the ethical boundaries of AI deployment for security purposes.

Technical Feasibility and Operational Requirements

Current AI technologies could theoretically support covert engagement operations through sophisticated persona generation, conversational capabilities, and behavioral mimicry. Advanced

 doi:10.20944/preprints202506.1513.v1

language models can maintain consistent personas across extended interactions while adapting communication styles to match target preferences and cultural contexts. The technical capabilities for creating convincing artificial personas have advanced significantly, with AI systems demonstrating ability to maintain coherent identities, backstories, and communication patterns over extended periods [28].

The technical infrastructure required for covert operations would be substantially more complex than overt systems. Multiple AI personas would need to be created and maintained simultaneously, each with detailed backstories, consistent communication patterns, and appropriate cultural and linguistic characteristics. The system would need to coordinate across multiple platforms and communication channels while avoiding detection by both platform security systems and target individuals or groups.

Sophisticated deception capabilities would be essential to maintain operational security. The AI system would need to generate convincing personal details, respond appropriately to unexpected questions about its supposed background, and maintain consistency across multiple interactions and platforms. This requires not only advanced natural language processing but also sophisticated knowledge management systems that can track and maintain complex fictional identities.

However, significant technical challenges exist for covert operations. The risk of detection increases with the scale and duration of operations, as patterns in communication style, response timing, and knowledge gaps may become apparent to sophisticated users. Platform detection systems are increasingly sophisticated at identifying automated accounts and artificial behavior patterns, potentially compromising covert operations before they achieve their objectives.

The coordination requirements for effective covert operations are enormous. Multiple AI personas would need to interact with each other and with human targets in ways that appear natural and spontaneous while actually serving coordinated strategic objectives. This level of coordination requires sophisticated planning algorithms and real-time adaptation capabilities that push the boundaries of current AI technologies.

Legal Barriers and Constitutional Constraints

Covert engagement agents face severe legal constraints under current democratic frameworks. The deceptive nature of these operations raises fundamental questions about government authority to engage in systematic deception of citizens, even for legitimate security purposes. Constitutional protections for free speech, religious liberty, and due process create substantial barriers to covert influence operations targeting domestic populations.

The EU AI Act's requirements for transparency and human oversight are fundamentally incompatible with covert operations. Article 13 requires that AI systems be designed to ensure appropriate transparency and that users are informed when they are interacting with AI systems [4]. The regulation's emphasis on trustworthy AI and respect for fundamental rights creates clear legal barriers to deceptive AI deployment.

Similar constraints exist under data protection regulations, which require that individuals be informed about data collection and processing activities. Covert AI operations would necessarily involve extensive data collection about target individuals without their knowledge or consent, violating fundamental principles of data protection law. The sensitive nature of the data involved—including religious beliefs, political opinions, and personal communications—would trigger the highest levels of protection under GDPR and similar regulations [22].

Human rights law provides additional constraints on covert operations. The European Convention on Human Rights protects freedom of expression, freedom of thought, conscience and religion, and the right to private and family life [29]. Covert influence operations that target these protected areas of human experience would require compelling justification and proportionate implementation that may be difficult to achieve in practice.

National security exceptions to these legal protections exist but are narrowly construed and subject to strict oversight requirements. Courts have generally required that national security

 doi:10.20944/preprints202506.1513.v1

operations be necessary, proportionate, and subject to appropriate judicial or legislative oversight. The broad, ongoing nature of covert AI operations would likely exceed the scope of traditional national security exceptions.

Ethical Violations and Democratic Principles

Covert engagement agents raise profound ethical concerns that go to the heart of democratic governance and human dignity. The use of deceptive AI personas to influence religious and political beliefs violates fundamental principles of informed consent, human autonomy, and respect for persons. These violations are particularly serious when targeting vulnerable individuals who may be experiencing psychological distress, identity confusion, or social isolation.

The principle of human autonomy requires that individuals be able to make informed decisions about their beliefs, associations, and actions. Covert influence operations undermine this autonomy by providing false information about the source and nature of communications that may influence important life decisions. This deception is particularly problematic in religious contexts, where individuals may place significant weight on guidance that they believe comes from fellow believers or religious authorities.

The potential for abuse inherent in covert operations creates additional ethical concerns. Once established, covert AI capabilities could be used for purposes beyond counter-extremism, including political manipulation, commercial influence, or personal targeting of individuals who pose no security threat. The difficulty of maintaining appropriate oversight and accountability for covert operations increases the risk that such capabilities would be misused.

The impact on social trust and democratic discourse could be devastating if covert operations were discovered. The revelation that government agencies were using deceptive AI to infiltrate religious communities and influence political beliefs would likely generate lasting damage to public trust in government institutions. This damage could extend far beyond the immediate operational context to undermine democratic legitimacy more broadly.

The precedent established by covert AI operations could also encourage similar activities by other actors, including foreign governments, criminal organizations, and extremist groups themselves. The normalization of deceptive AI use in domestic contexts could contribute to a broader degradation of information integrity and social trust that would ultimately serve extremist objectives more than counter-extremism goals.

Strategic Risks and Counterproductive Effects

While covert engagement agents might achieve short-term tactical advantages in specific cases, their strategic effectiveness is highly questionable when considered in the broader context of counter-extremism objectives. The risk of discovery and the resulting backlash could cause far more damage than any potential benefits from successful covert influence operations.

The discovery of covert operations would validate extremist narratives about government deception and persecution of Muslim communities. Extremist groups consistently claim that Western governments are engaged in systematic efforts to undermine Islam and Muslim communities through surveillance, infiltration, and manipulation. The revelation of covert AI operations would provide concrete evidence supporting these claims, potentially driving moderate Muslims away from cooperation with counter-extremism efforts and toward more radical positions.

The strategic risks extend beyond immediate operational concerns to broader questions about democratic governance and the rule of law. The use of deceptive AI for domestic influence operations represents a significant escalation in government surveillance and manipulation capabilities that could fundamentally alter the relationship between citizens and the state. Once these capabilities are developed and deployed, they may be difficult to constrain or eliminate, creating long-term risks to democratic institutions.

The effectiveness of covert operations in actually preventing radicalization is also questionable. Sustainable counter-extremism requires addressing underlying grievances, providing alternative

 doi:10.20944/preprints202506.1513.v1

sources of meaning and community, and building trust between communities and institutions. Covert operations that rely on deception and manipulation are unlikely to achieve these deeper objectives and may actually undermine them by increasing suspicion and alienation.

The resource requirements for effective covert operations are also substantial, potentially diverting resources from more effective approaches. The technical infrastructure, human oversight, and operational security requirements for covert AI operations would require significant investment that might be better directed toward community engagement, education, and addressing underlying social and economic factors that contribute to radicalization vulnerability.

5. Discussion

5.1. Comparative Analysis of Deployment Models

The systematic evaluation of three AI agent deployment models reveals significant differences in their technical feasibility, legal permissibility, ethical implications, and strategic effectiveness. These differences have important implications for policy development and operational planning in counter-extremism contexts, suggesting that some approaches offer substantially greater potential for achieving legitimate security objectives while maintaining democratic values and community trust.

Direct engagement agents emerge as the most promising approach across multiple evaluation dimensions. They offer strong technical feasibility using current AI technologies while operating within established legal frameworks for public education and community outreach. The transparent nature of these systems addresses many ethical concerns about deception and manipulation while their focus on authentic theological guidance and community service aligns with strategic objectives for building trust and preventing radicalization. The principle of *maslaha* in Islamic jurisprudence provides theological justification for such systems when they genuinely serve public welfare and prevent harm.

The strategic advantages of direct engagement agents stem from their ability to address root causes of radicalization rather than merely detecting or disrupting extremist activities. By providing accessible, authentic religious guidance and counter-narratives, these systems can fill critical knowledge gaps that extremist groups exploit for recruitment. The scalability of AI systems enables engagement with large numbers of individuals while personalization capabilities allow for tailored responses to individual circumstances and concerns.

Overt analytical agents provide valuable intelligence capabilities but face significant limitations in terms of community acceptance and strategic effectiveness. While technically feasible and legally permissible under appropriate oversight, these systems primarily serve supporting functions rather than directly addressing radicalization processes. Their value lies in enhancing understanding of extremist networks and trends rather than preventing individual radicalization. The risk of discriminatory targeting and community alienation must be carefully managed through robust oversight and community engagement.

Covert engagement agents face insurmountable barriers across multiple dimensions. The legal constraints under democratic frameworks, profound ethical concerns about deception and manipulation, and strategic risks of discovery and backlash make these approaches fundamentally incompatible with responsible counter-extremism practice.

The potential for short-term tactical gains cannot justify the long-term strategic risks and ethical violations inherent in covert operations.

5.2. Implementation Challenges and Critical Success Factors

The implementation of AI-mediated counter-extremism operations faces several critical challenges that must be addressed to ensure operational effectiveness and ethical compliance. The "Keyboard Jihad" definitional challenge represents a fundamental obstacle that affects all deployment models but is particularly acute for analytical systems that must distinguish between legitimate religious discourse and genuinely harmful extremist content.

 doi:10.20944/preprints202506.1513.v1

This definitional challenge reflects deeper tensions between security imperatives and respect for religious freedom and free expression. The risk of misidentifying legitimate Islamic scholarship or community discourse as extremist content could validate extremist narratives about state persecution while alienating the very communities whose cooperation is essential for effective counter-extremism. Addressing this challenge requires sophisticated contextual understanding, extensive community consultation, and robust oversight mechanisms that can prevent discriminatory targeting.

Community trust deficits represent another significant implementation challenge that affects all deployment models but is particularly critical for direct engagement approaches. Historical experiences with surveillance and infiltration have created deep skepticism about government counter-extremism efforts within many Muslim communities. Overcoming these trust deficits requires genuine commitment to community partnership, transparent operations, and demonstrated respect for community autonomy and religious authority.

The development of authentic theological content represents a specific challenge for direct engagement agents. AI systems must be capable of providing guidance that is both theologically accurate and culturally appropriate across diverse Muslim communities. This requires extensive consultation with religious scholars, ongoing validation of AI responses, and mechanisms for updating and refining theological content as scholarly understanding evolves.

Technical implementation challenges include ensuring appropriate crisis response capabilities, maintaining system security against sophisticated adversaries, and developing robust oversight mechanisms that can detect and prevent misuse. AI systems deployed in counter-extremism contexts will likely face targeted attacks from extremist groups seeking to compromise or manipulate their operations. Robust

cybersecurity measures and ongoing monitoring capabilities are essential to maintain operational integrity.

5.3. Regulatory Compliance and Governance Frameworks

The EU AI Act and similar regulatory frameworks create both constraints and opportunities for AI deployment in counter-extremism contexts. The classification of many security applications as high-risk AI systems requires comprehensive risk assessment, human oversight, and transparency measures. While these requirements impose additional costs and complexity, they also provide frameworks for responsible innovation that can enhance public trust and operational legitimacy.

Compliance with high-risk AI system requirements under the EU AI Act would necessitate substantial investment in governance systems and ongoing monitoring capabilities.

Organizations deploying AI systems for counter-extremism would need to establish quality management systems, conduct conformity assessments, and maintain detailed documentation of system design, training data, and operational performance. These requirements, while burdensome, could actually enhance system effectiveness by forcing careful attention to bias detection, performance monitoring, and human oversight.

Data protection regulations require careful attention to data collection, processing, and retention practices. The principles of data minimization, purpose limitation, and individual rights create constraints on surveillance-oriented approaches while supporting more targeted, consent-based interventions. Privacy-by-design approaches can help ensure compliance while maintaining operational effectiveness, but require careful integration of privacy protections into system architecture from the earliest design stages.

International coordination presents additional regulatory challenges, particularly for operations that cross national boundaries or involve multinational platforms.

Harmonization of regulatory approaches and development of international cooperation mechanisms will be essential for effective counter-extremism operations in the digital age. The global nature of digital platforms and extremist networks requires coordinated responses that respect diverse legal and cultural contexts while enabling effective information sharing and joint operations.

 doi:10.20944/preprints202506.1513.v1

The governance frameworks required for responsible AI deployment in counter- extremism contexts must balance operational effectiveness with democratic accountability and human rights protection. This requires multi-layered oversight mechanisms that include technical auditing, legal compliance monitoring, ethical review, and community engagement. Independent oversight bodies with appropriate expertise and authority would be essential to ensure that AI systems operate within legal and ethical boundaries while achieving legitimate security objectives.

5.4. Community Partnership and Authentic Engagement

The analysis reveals that community partnership and authentic engagement are essential for effective AI-mediated counter-extremism operations. Approaches that prioritize surveillance and control over community service and empowerment are unlikely to achieve sustainable success and may generate counterproductive backlash that ultimately serves extremist objectives more than security goals.

Authentic theological engagement requires genuine partnership with respected religious authorities and ongoing validation of AI responses by qualified scholars. AI systems cannot replace human religious authority but can serve as tools for expanding access to authentic guidance and counter-narratives. The legitimacy of these systems depends on their acceptance by religious communities and their alignment with established theological principles rather than their technical sophistication or government endorsement.

Community empowerment approaches that provide tools and resources for communities to address radicalization challenges themselves may be more effective than top-down interventions imposed by external authorities. AI systems can support community-led efforts by providing information, facilitating connections, and amplifying authentic voices within communities. This approach respects community autonomy while providing valuable support for local counter-extremism initiatives.

The development of sustainable partnerships requires long-term commitment and investment in community relationships that extend beyond immediate security concerns. Communities that feel valued and supported are more likely to cooperate with counter-extremism efforts and less likely to harbor individuals at risk of radicalization. AI systems can contribute to this broader community engagement strategy but cannot substitute for genuine investment in community development and empowerment.

The role of religious authority and scholarly validation is particularly important for direct engagement AI systems. The legitimacy of theological guidance depends not only on its accuracy but also on its source and the process by which it is validated. AI systems that provide religious guidance without appropriate scholarly oversight risk undermining their own credibility while potentially contributing to religious confusion or conflict.

5.5. Strategic Framework for Implementation

Based on the analysis of different deployment models and implementation challenges, this study proposes a three-track strategic framework that prioritizes approaches with the greatest potential for achieving counter-extremism objectives while maintaining democratic legitimacy and community trust.

Track One: Immediate Deployment of Overt Analytical Capabilities

The first track involves immediate deployment of overt analytical AI systems with comprehensive safeguards and community oversight. These systems would focus on intelligence gathering and threat assessment while operating transparently and with robust protections against discriminatory targeting. Key components include:

- Comprehensive risk assessment and bias testing before deployment
- Clear public disclosure of system capabilities and limitations

 doi:10.20944/preprints202506.1513.v1

- Independent oversight by civil liberties organizations and community representatives
- Regular auditing of system performance and impact on different communities
- Strict data minimization and retention limits
- Clear escalation procedures for human review of system outputs

Track Two: Pilot Development of Direct Engagement Agents

The second track involves careful pilot development of direct engagement AI systems through extensive community consultation and theological validation. This approach would begin with limited pilot programs in partnership with willing communities and religious institutions. Key components include:

- Extensive pre-deployment consultation with diverse Muslim communities
- Ongoing theological validation by qualified religious scholars
- Transparent operation with clear disclosure of AI nature
- Robust crisis intervention and human escalation capabilities
- Regular evaluation of community impact and acceptance
- Gradual expansion based on demonstrated effectiveness and community support

Track Three: Suspension of Covert Engagement Capabilities

The third track involves explicit suspension of covert engagement capabilities pending comprehensive legal authorization and public debate. The analysis demonstrates that covert operations present insurmountable legal, ethical, and strategic barriers under current frameworks. Any future consideration of such capabilities would require:

- Explicit legislative authorization with clear limitations and oversight requirements
- Comprehensive public debate about the appropriate limits of government deception
- Independent judicial or legislative oversight of any covert operations
- Clear sunset provisions and regular review of authorization
- Robust protections against mission creep and abuse

This three-track framework emphasizes competing with extremist narratives through superior theological authenticity and genuine community partnership rather than through deception or surveillance. The approach aligns strategic effectiveness with democratic values and human rights protections while providing pathways for responsible innovation in AI-mediated counter-extremism.

6. Conclusions

This study provides the first comprehensive framework for evaluating AI agent deployment in counter-extremism operations, revealing significant differences in the viability and appropriateness of different approaches across technical, legal, ethical, and strategic dimensions. The analysis demonstrates that transparent, community- partnered approaches offer superior strategic effectiveness compared to surveillance- based or deceptive methodologies, while also maintaining compatibility with democratic values and human rights protections.

The research establishes that direct engagement AI agents offering authentic theological guidance represent the most promising path forward for AI-mediated counter- extremism. These systems combine technical feasibility with legal compliance, ethical integrity, and strategic effectiveness when properly designed and implemented with appropriate community partnership and oversight mechanisms. The principle of *maslaha* in Islamic jurisprudence provides theological

 doi:10.20944/preprints202506.1513.v1

justification for such systems when they genuinely serve public welfare and prevent harm while respecting community values and religious authority.

The study reveals that covert engagement operations are fundamentally incompatible with democratic values and responsible AI deployment. The legal constraints under current regulatory frameworks, profound ethical concerns about deception and manipulation, and strategic risks of discovery and backlash make these approaches unsuitable for democratic societies committed to the rule of law and human rights. The potential for short-term tactical gains cannot justify the long-term strategic risks and ethical violations inherent in covert operations.

Overt analytical agents provide valuable intelligence capabilities but require careful implementation to address community concerns and ensure appropriate oversight. These systems are best understood as supporting tools for other counter-extremism activities rather than standalone interventions capable of addressing the root causes of radicalization. Their effectiveness depends heavily on robust safeguards against discriminatory targeting and genuine community engagement to maintain legitimacy and cooperation.

6.1. Critical Implementation Challenges

The study identifies several critical implementation challenges that must be addressed for successful AI-mediated counter-extremism operations. The "Keyboard Jihad" definitional challenge requires careful attention to distinguishing between legitimate religious discourse and genuinely harmful extremist content. This challenge reflects deeper tensions between security imperatives and respect for religious freedom that must be navigated through sophisticated contextual understanding, extensive community consultation, and robust oversight mechanisms.

Community trust deficits necessitate genuine commitment to partnership and transparency rather than technological solutions imposed without community input or consent. Historical experiences with surveillance and infiltration have created deep skepticism about government counter-extremism efforts that can only be overcome through demonstrated respect for community autonomy and consistent commitment to serving community needs rather than purely security objectives.

Technical challenges demand robust safety measures and human oversight systems to ensure theological accuracy, cultural sensitivity, and appropriate crisis response capabilities. AI systems deployed in counter-extremism contexts must be capable of recognizing their limitations and appropriately escalating to human experts when conversations require specialized knowledge or immediate intervention.

6.2. Regulatory Landscape and Governance Requirements

The regulatory landscape created by the EU AI Act and similar frameworks provides both constraints and opportunities for responsible innovation in counter-extremism contexts. Compliance with high-risk AI system requirements necessitates substantial investment in governance systems and oversight mechanisms but can enhance public trust and operational legitimacy when properly implemented.

The classification of many counter-extremism applications as high-risk AI systems requires comprehensive risk assessment, human oversight, and transparency measures that align with democratic values and human rights protections. These requirements, while imposing additional costs and complexity, provide frameworks for responsible development and deployment that can prevent abuse while enabling legitimate security applications.

International coordination will be essential for effective counter-extremism operations in the digital age, requiring harmonization of regulatory approaches and development of cooperation mechanisms that respect diverse legal and cultural contexts while enabling effective information sharing and joint operations.

6.3. Community Partnership and Theological Authenticity

 doi:10.20944/preprints202506.1513.v1

The research emphasizes the critical importance of community partnership and authentic engagement over surveillance-oriented approaches. AI systems can serve as valuable tools for expanding access to authentic theological guidance and counter- narratives, but their legitimacy depends on acceptance by religious communities and alignment with established theological principles rather than government endorsement or technical sophistication.

Authentic theological engagement requires genuine partnership with respected religious authorities and ongoing validation of AI responses by qualified scholars. The development and operation of direct engagement agents should involve extensive consultation with diverse Muslim communities and religious authorities to ensure that systems reflect authentic Islamic scholarship and address genuine community needs.

Community empowerment approaches that provide tools and resources for communities to address radicalization challenges themselves may be more effective than top-down interventions imposed by external authorities. This approach respects community autonomy while providing valuable support for local counter-extremism initiatives that can achieve sustainable impact.

6.4. Strategic Framework and Implementation Pathways

The proposed three-track strategic framework provides practical implementation pathways that balance operational effectiveness with legal compliance and ethical considerations. The framework prioritizes immediate deployment of overt analytical capabilities with comprehensive safeguards, pilot development of direct engagement agents through extensive community consultation, and suspension of covert engagement capabilities pending explicit legal authorization and public debate.

This approach emphasizes competing with extremist narratives through superior theological authenticity and genuine community partnership rather than through deception or surveillance. The framework aligns strategic effectiveness with democratic values and human rights protections while providing pathways for responsible innovation that can adapt to evolving technological capabilities and threat landscapes.

The success of this framework depends on sustained commitment to community partnership, ongoing investment in oversight and governance mechanisms, and willingness to prioritize long-term strategic effectiveness over short-term tactical advantages. The approach requires patience and persistence but offers the greatest potential for achieving sustainable reductions in radicalization risk while maintaining democratic legitimacy and community trust.

6.5. Future Research and Development Priorities

Future research should focus on developing specific implementation protocols for direct engagement AI agents, including theological validation mechanisms, community partnership frameworks, and crisis response procedures. Empirical evaluation of pilot programs would provide valuable insights into the practical effectiveness of different approaches and help refine implementation strategies based on real-world experience.

The development of robust oversight and governance mechanisms requires ongoing research into bias detection, performance monitoring, and community impact assessment. Technical research should focus on improving AI systems' ability to recognize contextual nuances in religious and cultural discourse while maintaining appropriate boundaries and escalation procedures.

International cooperation frameworks for AI-mediated counter-extremism require development of shared standards, protocols, and oversight mechanisms that can enable effective collaboration while respecting diverse legal and cultural contexts. Research into cross-border governance challenges and solutions will be essential as AI capabilities continue to evolve and extremist networks adapt to new technologies.

6.6. Policy Implications and Recommendations

 doi:10.20944/preprints202506.1513.v1

The findings have immediate relevance for policymakers, technology developers, and counter-extremism practitioners working to address the challenges of digital radicalization. The proposed framework provides practical guidance for responsible AI deployment while maintaining commitment to democratic values and human rights.

Policymakers should prioritize development of comprehensive regulatory frameworks that enable responsible innovation while preventing abuse and protecting fundamental rights. Investment in community partnership and engagement should be viewed as essential infrastructure for effective counter-extremism rather than optional add-ons to technological solutions.

Technology developers should prioritize transparency, community engagement, and ethical design principles in developing AI systems for security applications. The integration of robust oversight mechanisms and human escalation procedures should be considered essential features rather than optional enhancements.

Counter-extremism practitioners should focus on building authentic partnerships with affected communities and investing in approaches that address root causes of radicalization rather than merely detecting or disrupting extremist activities. The emphasis should be on empowering communities to address challenges themselves rather than imposing external solutions without community input or consent.

As AI technologies continue to evolve and extremist groups adapt their tactics, ongoing research and development will be essential to maintain effective and ethical counter- extremism capabilities. The framework provided by this study offers a foundation for responsible innovation that can adapt to changing circumstances while maintaining commitment to democratic values and human rights protections.

Author Contributions: As sole author, A.B. is responsible for all aspects of this research including conceptualization, methodology development, analysis, and manuscript preparation.

Funding: This research received no external funding. The work was conducted as part of doctoral studies at King's College London.

Data Availability Statement: The datasets analyzed during the current study are available from the corresponding author on reasonable request, subject to appropriate ethical and privacy considerations.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Islamic State. *AI Propaganda Creation Guide*. 2023. [Accessed through academic terrorism studies databases]
2. Ibid.
3. Collison-Randall, H.; Spaaij, R.; Hayday, E.J.; Pippard, J. Media framing of far-right extremism and online radicalization in esport and gaming. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1195. https://doi.org/10.1057/s41599-024-03680-4
4. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union* **2024**, L 1689, 1-144.
5. Bellaby, R.W. The ethics of intelligence-AI: Examining the moral implications of artificial intelligence in intelligence operations. *Int. Aff.* **2024**, *100*, 2525-2544. https:// doi.org/10.1093/ia/iiae234
6. Duarte, N.; Silva, A.; Pinto, M.; Santos, R.; Ferreira, L. Effectiveness of educational programmes to prevent violent extremism: A systematic review. *BMC Public Health* **2025**, *25*, 123. https://doi.org/10.1186/s12889-024-20397-8
7. UK Home Office. *Economic and Social Costs of Terrorism*. London: HMSO, 2023.
8. New Zealand Government. *Christchurch Attack Response and Recovery*. Wellington: Government Publications, 2020.

 doi:10.20944/preprints202506.1513.v1

9. Collison-Randall, H.; Spaaij, R.; Hayday, E.J.; Pippard, J. Media framing of far-right extremism and online radicalization in esport and gaming. *Humanit. Soc. Sci. Commun.* **2024**, *11*, 1195.
10. Australian Federal Police. *Online Gaming and Extremist Recruitment Warning*. Canberra: AFP Publications, 2022.
11. Molas, A.; Lopes, C. Subverting Safeguards: How Far-Right Groups Are Jailbreaking AI Systems. *International Centre for Counter-Terrorism* **2024**. Available online: https:// icct.nl/sites/default/files/2024-10/Molas%20and%20Lopes.pdf
12. Bangura, A.K. *Islamic Peace Paradigms and World Peace*. Herndon: International Institute of Islamic Thought, 2005.
13. Combating Terrorism Center at West Point. *Digital Jihad: Online Communication and Terrorist Recruitment*. West Point: CTC Publications, 2023.
14. Ibid.
15. Weimann, G. *Terror on the Internet: The New Arena, the New Challenges*. Washington: United States Institute of Peace Press, 2006.
16. Ribeiro, M.H.; Ottoni, R.; West, R.; Almeida, V.A.; Meira Jr., W. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*; ACM: New York, NY, USA, 2020; pp. 131-141.
17. Bellaby, R.W. The ethics of intelligence-AI: Examining the moral implications of artificial intelligence in intelligence operations. *Int. Aff.* **2024**, *100*, 2525-2544.
18. OpenAI. GPT-4 Technical Report. *arXiv preprint* **2023**, arXiv:2303.08774.
19. Duarte, N.; Silva, A.; Pinto, M.; Santos, R.; Ferreira, L. Effectiveness of educational programmes to prevent violent extremism: A systematic review. *BMC Public Health* **2025**, *25*, 123.
20. Kruglanski, A.W.; Gelfand, M.J.; Bélanger, J.J.; Sheveland, A.; Hetiarachchi, M.; Gunaratna, R. The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology* **2014**, *35*, 69-93.
21. Ferrara, E.; Wang, W.Q.; Varol, O.; Flammini, A.; Galstyan, A. Predicting online extremism, content adopters, and interaction reciprocity. In *Proceedings of the International Conference on Social Informatics*; Springer: Cham, Switzerland, 2016; pp. 22-39.
22. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Off. J. Eur. Union* **2016**, L 119, 1-88.
23. Kundnani, A. *The Muslims Are Coming! Islamophobia, Extremism, and the Domestic War on Terror*. London: Verso Books, 2014.
24. American Civil Liberties Union. *Mapping Muslims: NYPD Spying and its Impact on American Muslims*. New York: ACLU Publications, 2013.
25. Ibid.
26. OpenAI. GPT-4 Technical Report. *arXiv preprint* **2023**, arXiv:2303.08774.
27. Kamali, M.H. *Principles of Islamic Jurisprudence*. 3rd ed. Cambridge: Islamic Texts Society, 2003.
28. Park, J.S.; O'Brien, J.C.; Cai, C.J.; Morris, M.R.; Liang, P.; Bernstein, M.S. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*; ACM: New York, NY, USA, 2023; pp. 1-22.
29. Council of Europe. European Convention for the Protection of Human Rights and Fundamental Freedoms. Rome: Council of Europe, 1950.