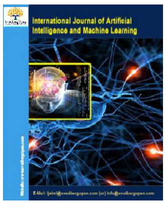
https://doi.org/10.51483/IJAIML.5.2.2025.23-42

ISSN: 2789-2557

SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/



Research Paper

Open Access

AI Agents for Counter-Extremism: Deployment Frameworks for Covert and Overt Digital Deradicalization

Aadil Bouhlaoui[1*]

[1]Ph.D. Researcher, Digital Humanities, King's College, London. E-mail: aadil.bouhlaoui@kcl.ac.uk

Article Info

Volume 5, Issue 2, July 2025
Received : 11 May 2025
Accepted : 10 July 2025
Published : 25 July 2025
doi: 10.51483/IJAIML.5.2.2025.23-42

Abstract

This article analyzes strategies for deploying AI agents to counter online extremism, focusing on digital radicalization in Islamic contexts. It assesses technical feasibility, legal constraints (including the EU AI Act), ethical concerns, and theological implications. Drawing on recent case studies—such as ISIS's 2023 AI propaganda guide and the shift of extremist content to gaming platforms—it highlights the dangers of definitional confusion around terms like "Keyboard Jihad," which may lead to misidentifying legitimate Islamic discourse. The study evaluates three AI deployment models: overt analytical agents, direct engagement agents, and covert engagement agents. It concludes that transparent, community-partnered models—especially those offering authentic theological guidance—are the most effective and ethically sound. These models help fill gaps in Islamic knowledge that extremists exploit, while avoiding the legal and strategic pitfalls of covert influence operations, which violate current EU AI regulations. The article recommends a three-track strategy: immediate use of overt analytical agents with safeguards; piloting direct engagement agents through community and theological consultation; and halting covert operations until explicitly authorized by law. Rooted in the Islamic legal principle of maslaha (public interest), the framework prioritizes authenticity, transparency, and respect for democratic values in counter-extremism efforts.

Keywords: *Artificial intelligence, Counter-extremism, Islamic theology, Digital radicalization, AI ethics, Counter-narrative, Deradicalization, EU AI Act, Community engagement*

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

* *Corresponding author: Aadil Bouhlaoui, Ph.D. Researcher, Digital Humanities, King's College, London. E-mail: aadil.bouhlaoui@kcl.ac.uk*

security (Europol, 2022), 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 (Islamic State Media, 2023), systematic migration of extremist activities to gaming platforms and encrypted channels (Collison-Randall *et al.*, 2024), and the implementation of the European Union's AI Act, which establishes comprehensive regulatory frameworks for high-risk AI systems in law enforcement contexts (European Union, 2024).

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 (Bellaby, 2024) and explored digital radicalization patterns (Alava *et al.*, 2017), 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 (Home Office, 2018). The 2019 Christchurch attacks alone prompted New Zealand to allocate over NZ$200 million for victim support and community recovery efforts (New Zealand Government, 2020). 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 (Islamic State Media, 2023).

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 (Collison-Randall *et al.*, 2024). Research by Collison-Randall *et al.* (2024) 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 (Collison-Randall *et al.*, 2024). The Australian Federal Police's (2022) warning about extremist groups accessing online games to recruit children reflects growing recognition of this threat vector (Australian Federal Police, 2022).

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 technologies (Molas and Lopes, 2024). 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 platformsenable 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 (Bangura, 2011). 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 (Combating Terrorism Center, 2015). 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 (Combating Terrorism Center, 2015).

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:

1. **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?
2. **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?
3. **Strategic Effectiveness:** Which AI agent deployment models offer the greatest potential for strategic effectiveness in counter-extremism while maintaining compatibility with democratic values, community trust, and human rights protections?

The study's primary objective is to develop a comprehensive framework for evaluating AI agent deployment strategies that integrates technical, legal, ethical, and theological considerations. This framework aims to provide policymakers, security practitioners, and technology developers with evidence-based guidance for designing and implementing AI-mediated counter-extremism programs that are both effective and ethically defensible.

Secondary Objectives Include: (1) establishing clear definitional boundaries for legitimate versus extremist online activities to minimize risks of misidentification and strategic blowback; (2) identifying specific technical and procedural safeguards necessary to ensure AI agent operations comply with legal requirements and ethical principles; (3) developing metrics and evaluation frameworks for assessing the effectiveness and impact of AI-mediated counter-extremism interventions; and (4) proposing specific implementation pathways that prioritize community engagement and theological authenticity over surveillance and deception.

1.4. Methodological Approach

This research employs a multidisciplinary methodology that integrates insights from computer science, legal studies, ethics, and Islamic theology. The analysis draws upon multiple data sources, including recent case studies of extremist AI adoption, legal documentation from the EU AI Act implementation, academic literature on deradicalization effectiveness, and theological scholarship on Islamic principles relevant to counter-extremism operations.

The study's analytical framework is structured around three distinct AI agent deployment models: overt analytical agents designed for monitoring and content analysis, direct engagement agents intended to provide authentic theological guidance and counter-narratives, and covert engagement agents conceived for infiltration and influence operations. Each model is evaluated across four dimensions: technical feasibility, legal permissibility, ethical implications, and strategic effectiveness.

The research methodology emphasizes evidence-based analysis while acknowledging the inherent limitations imposed by the classified nature of many counter-extremism operations. Where direct empirical data is unavailable, the study relies on publicly available documentation, academic research, and expert analysis to construct reasonable assessments of capabilities and constraints. This approach ensures that recommendations are grounded in available evidence while remaining cognizant of information limitations that may affect implementation considerations.

2. Literature Review

2.1. Digital Radicalization and Platform Migration

The academic understanding of digital radicalization has evolved significantly in response to changing technological landscapes and extremist adaptation strategies. Early scholarship focused primarily on static websites and forum-based recruitment, but recent research reveals a fundamental transformation in how extremist groups exploit digital platforms for radicalization purposes. Collison-Randall *et al.* (2024) comprehensive analysis of media framing demonstrates that gaming adjacent platforms have created expanding ecosystems of online gaming, esports, and social media actors sharing online space, content, communication tools, and users (Collison-Randall *et al.*, 2024). This research establishes that esports, in particular, has grown beyond expectations to become a global leader in sport fandom and spectatorship, creating new opportunities for extremist infiltration and influence.

The platform migration phenomenon represents a strategic adaptation by extremist groups seeking to evade traditional content moderation and surveillance mechanisms. Research indicates that far-right extremist groups deliberately target Generation Z through social media and online gaming spaces, with the 2019 Christchurch attacks serving as a catalyst for global recognition of the nexus between far-right extremism and online gaming (Collison-Randall *et al.*, 2024). The Australian Federal Police's (2022) warning about extremist groups accessing online games to recruit children reflects growing institutional awareness of this threat vector, while Channel 4's yearlong investigation revealed how groups like Patriotic Alternative systematically exploit gaming environments to seduce vulnerable young people into extremist ideologies (Australian Federal Police, 2022).

This platform migration creates significant challenges for traditional counter-extremism approaches that rely on content removal and account suspension. Gaming platforms often feature different community norms, moderation practices, and user expectations compared to traditional social media, requiring specialized understanding and intervention strategies. The interactive nature of gaming environments also provides opportunities for more sophisticated grooming and recruitment techniques that exploit the social and competitive aspects of gaming culture.

2.2. AI Exploitation by Extremist Groups

Recent research reveals that extremist groups have rapidly adopted AI technologies for propaganda creation, recruitment, and operational security, creating an intensifying digital arms race between extremist actors and counter-terrorism efforts. Molas and Lopes' groundbreaking research demonstrates how far-right users have successfully exploited AI tools through jailbreaking techniques, accelerating the spread of harmful content by circumventing safety measures built into commercial AI systems (Molas and Lopes, 2024). This research provides evidence-based insights into the misuse of AI through new data that illustrates the sophisticated methods extremist groups employ to weaponize emerging technologies.

The Islamic State's (2023) publication of a comprehensive guide on using generative AI for propaganda creation represents a watershed moment in the weaponization of artificial intelligence for terrorist purposes (Islamic State Media, 2023). This development demonstrates that extremist groups are not merely passive consumers of technology but active innovators who rapidly adapt emerging capabilities to serve their strategic objectives. The guide's existence indicates systematic organizational investment in AI capabilities and suggests that other extremist groups may be pursuing similar technological adaptations.

The implications of AI exploitation extend beyond propaganda creation to encompass recruitment, operational planning, and security measures. AI technologies enable extremist groups to personalize messaging, automate content creation, and scale their operations in ways previously impossible. The ability to generate convincing text, images, and potentially audio or video content using AI tools dramatically reduces the technical barriers to sophisticated propaganda production while enabling more targeted and effective recruitment strategies.

2.3. Educational and Counter-Narrative Interventions

Academic research provides substantial evidence for the effectiveness of educational programmes and counter-narrative interventions in preventing and countering violent extremism. Duarte *et al.* (2025) systematic review protocol examines the effectiveness of educational programmes delivered both online and offline, designed to prevent and counter the effects of online violent extremist propaganda (Duarte *et al.*, 2025). This research establishes that interventions such as media literacy initiatives, counter-narratives, alternative narratives, and gamified approaches can effectively reduce violent extremist attitudes, beliefs, and behaviors.

The systematic review identifies several successful examples of proactive counter-extremism measures, including the 'Jamal al-Khatib' initiative by the Turn Association for the Prevention of Violence and Extremism, the DECOUNT online game, and the 'Redirect Method' developed by Jigsaw (Duarte *et al.*, 2025). These initiatives demonstrate that reducing the appeal of extremist content and raising awareness through community voices can effectively challenge digital extremist narratives. The research emphasizes that educational approaches focusing on helping individuals recognize extremist content and diminishing its appeal offer promising alternatives to purely surveillance-based counter-extremism strategies.

The effectiveness of counter-narrative approaches is particularly relevant to AI agent deployment strategies because it provides empirical foundation for direct engagement models. Research demonstrates that authentic, community-validated counter-narratives can successfully compete with extremist messaging when properly designed and implemented. This finding supports the theoretical foundation for direct engagement AI agents that provide authentic theological guidance and counter-narratives rather than relying on deception or surveillance.

2.4. AI Ethics in Security Applications

The deployment of AI technologies in security contexts raises unique ethical challenges that extend beyond general AI ethics concerns. Bellaby's comprehensive analysis of intelligence-AI demonstrates that when AI is combined with the reach, secrecy, and coercive power of the intelligence community, it creates unique ethical problems that exacerbate existing AI biases while undermining proposed transparency solutions (Bellaby, 2024). This research establishes that intelligence-AI operations face distinct ethical challenges because they combine AI's inherent limitations with the intelligence community's exceptional capabilities and legal authorities.

The research reveals that open-source data collection does not necessarily ensure ethical compliance, as AI collection *en masse* of social media data violates citizens' privacy, consent, and autonomy regardless of the data's public availability (Bellaby, 2024). This finding has direct implications for AI agent deployment in counter-extremism, suggesting that even overt analytical operations must carefully consider privacy and consent issues. The analysis also demonstrates that AI-aided categorization tends to be overly reductive and perpetuates harmful social binaries while revealing new private information beyond what was initially shared.

Particularly relevant to counter-extremism applications is Bellaby's finding that the secretive intelligence environment prevents critical interrogation while promoting practices that cause unequal harms across society through the coercive power of the state (Bellaby, 2024). This analysis suggests that covert AI agent operations face insurmountable ethical barriers because they combine AI's inherent biases and limitations with the intelligence community's exceptional powers in ways that prevent the transparency and accountability mechanisms necessary for ethical AI deployment.

2.5. Community Engagement and Theological Considerations

The role of community engagement and theological authenticity in counter-extremism efforts represents a critical but underexplored dimension of AI agent deployment strategies. Islamic theological principles provide important guidance for evaluating the ethical permissibility of different counter-extremism approaches, particularly regarding the use of deception and surveillance. The principle of *tajassus* (prohibition on spying) in Islamic jurisprudence creates potential theological objections to covert AI agent operations, while the principle of *maslaha* (public interest) may provide justification for transparent interventions that serve community welfare (Al-Qaradawi, 2013).

Research on community trust and counter-extremism effectiveness demonstrates that Muslim communities already experience significant trust deficits with security services due to decades of surveillance and discriminatory targeting (Kundnani, 2014). This context suggests that AI agent deployment strategies must carefully consider community perceptions and engagement to avoid exacerbating existing tensions. The deployment of covert AI agents risks validating extremist narratives about state persecution of Muslims, potentially driving moderate voices away from counter-extremism efforts and creating new grievances for extremist exploitation.

The concept of authentic Islamic knowledge (*Al-Ilm Al-Shari*) emerges as a critical factor in counter-extremism effectiveness. Extremist groups exploit gaps in legitimate spiritual mentorship and authentic theological guidance to recruit vulnerable individuals who lack access to authoritative Islamic scholarship (Wiktorowicz, 2005). This finding suggests that AI agents designed to provide authentic theological guidance and counter-narratives may address root causes of radicalization more effectively than surveillance-based approaches that focus on symptoms rather than underlying vulnerabilities.

2.6. Regulatory Frameworks and Legal Constraints

The implementation of the EU AI Act in 2024 represents a fundamental transformation in the legal landscape governing AI deployment in security contexts. The Act establishes comprehensive requirements for high-risk AI systems used in law enforcement, including mandatory risk management, data governance, technical documentation, and human oversight obligations (European Union, 2024). These requirements significantly constrain covert AI operations while providing clear pathways for transparent, community-partnered approaches.

The Act's prohibition of subliminal influence techniques creates particular challenges for covert AI deployment strategies that rely on deception or manipulation (European Union, 2024). This prohibition aligns with broader ethical principles regarding informed consent and autonomy, suggesting that effective AI agent deployment must prioritize transparency and community engagement over deceptive practices. The regulatory framework supports direct engagement models that provide authentic guidance and counter-narratives while maintaining transparency about their AI-mediated nature.

Legal analysis reveals that AI agent deployment strategies must navigate complex intersections between counter-terrorism law, data protection regulations, and human rights protections. The European Convention on Human Rights, domestic constitutional protections, and international human rights law establish strong safeguards for religious expression and academic freedom that constrain the scope of permissible AI agent operations (European Court of Human Rights, 2020). These legal frameworks require careful attention to definitional precision and procedural safeguards to ensure that counter-extremism operations do not inadvertently target legitimate religious or academic discourse.

2.7. Gaps in Existing Literature

Despite the growing body of research on AI applications in security contexts and digital counter-extremism, significant gaps remain in academic understanding of AI agent deployment strategies. 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. Existing research tends to focus on either technical capabilities or ethical concerns in isolation, without integrating these perspectives into comprehensive frameworks for policy and practice.

The literature also lacks systematic analysis of community engagement strategies for AI-mediated counter-extremism operations. While research demonstrates the importance of community trust and theological authenticity in counter-extremism effectiveness, few studies examine how these principles should guide AI agent design and implementation. This gap is particularly significant given the unique challenges posed by AI technologies, which may be perceived as impersonal or inauthentic by target communities.

Finally, existing research provides limited guidance on evaluation metrics and implementation pathways for AI agent deployment in counter-extremism contexts. While studies demonstrate the effectiveness of educational and counter-narrative interventions generally, they do not address the specific challenges and opportunities created by AI-mediated delivery of such interventions. This gap limits the ability of policymakers and practitioners to design evidence-based implementation strategies that maximize effectiveness while minimizing risks.

3. Methodology

3.1. Research Design

This study employs a multidisciplinary analytical framework that integrates insights from computer science, legal studies, ethics, and Islamic theology to evaluate AI agent deployment strategies for counter-extremism operations. The research design is structured around comparative analysis of three distinct AI agent deployment models, each evaluated across four analytical dimensions: technical feasibility, legal permissibility, ethical implications, and strategic effectiveness.

The methodology adopts a mixed-methods approach that combines systematic literature review, legal analysis, case study examination, and theoretical framework development. This approach enables comprehensive evaluation of AI agent deployment strategies while acknowledging the inherent limitations imposed by the classified nature of many counter-extremism operations. Where direct empirical data is unavailable, the study relies on publicly available documentation, academic research, expert analysis, and theoretical extrapolation to construct reasonable assessments of capabilities and constraints.

3.2. Analytical Framework

The study's analytical framework is organized around three primary AI agent deployment models that represent the spectrum of possible approaches to AI-mediated counter-extremism:

Overt Analytical Agents: Are designed for monitoring, analysis, and content flagging operations with full transparency about their AI-mediated nature. These systems focus on processing large volumes of online content to identify potential extremist material for human review and intervention. The overt nature of these operations means that their existence and general capabilities are publicly acknowledged, though specific technical details may remain classified for operational security reasons.

Direct Engagement Agents: Are intended to provide authentic theological guidance and counter-narratives through transparent interaction with individuals at risk of radicalization. These systems are designed to address gaps in legitimate Islamic knowledge (*Al-Ilm Al-Shari*) that extremist groups exploit for recruitment purposes. The direct engagement model emphasizes transparency about the AI-mediated nature of interactions while prioritizing theological authenticity and community validation.

Covert Engagement Agents: Are conceived for infiltration and influence operations where the AI-mediated nature of interactions is concealed from target individuals. These systems would operate under false personas to gather intelligence or influence discourse within extremist networks. The covert nature of these operations raises significant legal, ethical, and strategic challenges that require careful evaluation.

3.3. Evaluation Dimensions

Each AI agent deployment model is evaluated across four analytical dimensions that capture the multidisciplinary nature of counter-extremism challenges:

Technical Feasibility: Encompasses the current state of AI technology, computational requirements, data availability, and implementation challenges. This dimension examines whether proposed AI agent capabilities are achievable with existing or near-term technology, what technical infrastructure would be required for deployment, and what limitations or constraints might affect operational effectiveness.

Legal Permissibility: Analyzes compliance with existing legal frameworks, including the EU AI Act, data protection regulations, human rights law, and counter-terrorism legislation. This dimension examines whether proposed AI agent operations would be legally permissible under current regulatory frameworks and what legal safeguards or modifications might be necessary for compliant implementation.

Ethical Implications: Evaluates alignment with established ethical principles, including transparency, accountability, respect for human dignity, privacy protection, and non-discrimination. This dimension examines potential harms to individuals and communities, considers the proportionality of proposed interventions, and assesses compatibility with democratic values and human rights protections.

Strategic Effectiveness: Assesses the potential for achieving counter-extremism objectives while maintaining community trust and avoiding counterproductive outcomes. This dimension examines whether proposed AI agent operations are likely to reduce radicalization risks, considers potential for strategic blowback or unintended consequences, and evaluates alignment with broader counter-extremism strategies.

3.4. Data Sources and Evidence Base

The research draws upon multiple data sources to construct a comprehensive evidence base for analysis. Primary sources include legal documentation from the EU AI Act implementation, official statements and reports from intelligence and security agencies, and technical documentation from AI system developers. Secondary sources encompass academic literature on digital radicalization, AI ethics, counter-extremism effectiveness, and Islamic theology.

Case study evidence is derived from publicly reported incidents of extremist AI adoption, including the Islamic State's 2023 AI propaganda guide, documented instances of platform migration to gaming environments, and examples of successful counter-narrative interventions. Expert analysis and commentary from security practitioners, technology developers, legal scholars, and community leaders provide additional perspectives on implementation challenges and opportunities.

The study acknowledges inherent limitations in the available evidence base, particularly regarding classified counter-extremism operations and proprietary AI system capabilities. Where direct empirical data is unavailable, the analysis relies on theoretical extrapolation and expert assessment to construct reasonable evaluations of proposed approaches. This methodology ensures that recommendations are grounded in available evidence while remaining cognizant of information limitations that may affect implementation considerations.

3.5. Ethical Considerations

This research adheres to established ethical principles for academic inquiry while acknowledging the sensitive nature of counter-extremism research. The study avoids providing detailed technical information that could assist extremist groups in developing countermeasures or exploiting vulnerabilities. Analysis of AI agent capabilities focuses on general principles and publicly available information rather than specific technical implementations that might compromise operational security.

The research prioritizes community perspectives and theological considerations to ensure that recommendations respect the values and concerns of affected populations. Particular attention is paid to avoiding stigmatization of Muslim communities or reinforcement of discriminatory stereotypes that could exacerbate existing tensions between security services and minority populations.

4. Analysis: Evaluating AI Agent Deployment Models

4.1. Overt Analytical Agents: Transparent Monitoring and Analysis

Overt analytical agents represent the most technically feasible and legally defensible approach to AI-mediated counter-extremism operations. These systems are designed to process large volumes of online content to identify potential extremist material for human review and intervention, with full transparency about their AI-mediated nature and general operational parameters. The overt approach acknowledges the AI system's existence and purpose while maintaining necessary operational security regarding specific technical capabilities and targeting criteria (Table 1).

Table 1: Comparative Analysis of AI Agent Deployment Models

Dimension	Overt Analytical	Direct Engagement	Covert Engagement
Technical Feasibility	High: Current AI capabilities sufficient	Medium: Requires specialized development	High: But detection risks significant
Legal Permissibility	High: Compliant with EU AI Act	Medium: Requires careful compliance	Low: Violates multiple regulations
Ethical Implications	Medium: Transparency enables oversight	High: Requires community validation	Low: Fundamental deception concerns
Strategic Effectiveness	Medium: Limited to analysis	High: Addresses root causes	Low: High risk of strategic blowback
Implementation Timeline	Immediate	3-year pilot program	Not recommended
Estimated Cost	$10-15 M annually	$25-35 M development	N/A
Community Acceptance	Medium-High	High (with validation)	Very Low
Regulatory Compliance	Full compliance possible	Compliance achievable	Multiple violations likely

4.1.1. Technical Feasibility

Current AI technologies provide substantial capabilities for content analysis, pattern recognition, and anomaly detection that support overt analytical operations. Natural language processing systems can analyze text content across multiple languages and platforms to identify potential indicators of extremist ideology, recruitment activities, or operational planning. Computer vision technologies enable analysis of images and videos for extremist symbols, propaganda materials, or other visual indicators of concern.

The technical infrastructure required for overt analytical agents is substantial but achievable with current technology. Cloud computing platforms provide the computational resources necessary for processing large volumes of data, while established machine learning frameworks offer the tools needed for developing and deploying analytical models. The primary technical challenges involve ensuring system accuracy, minimizing false positives, and maintaining performance across diverse platforms and content types.

Data availability represents both an opportunity and a constraint for overt analytical operations. Public social media platforms, forums, and websites provide vast amounts of accessible content for analysis, but platform policies and technical measures may limit data collection capabilities. The migration of extremist activities to private platforms, encrypted channels, and gaming environments creates additional challenges for data access that may require specialized technical approaches or legal authorities.

4.1.2. Legal Permissibility

The EU AI Act provides a comprehensive framework for evaluating the legal permissibility of overt analytical agents in counter-extremism contexts. These systems would likely be classified as high-risk AI applications under the Act's risk-based categorization scheme, triggering mandatory requirements for risk management, data governance, technical documentation, and human oversight (European Union, 2024). However, the overt nature of these operations facilitates compliance with transparency and accountability requirements that might be challenging for covert approaches.

Data protection regulations, particularly the General Data Protection Regulation (GDPR), establish important constraints on data collection and processing for overt analytical operations. While law enforcement exemptions may apply to counter-terrorism activities, these exemptions are limited and require careful attention to proportionality, necessity, and data minimization principles (European Union, 2016). The processing of personal data for counter-extremism purposes must be justified by legitimate security interests and subject to appropriate safeguards to protect individual rights.

Human rights law provides additional constraints on overt analytical operations, particularly regarding freedom of expression, privacy, and non-discrimination. The European Convention on Human Rights establishes that restrictions on freedom of expression must be prescribed by law, pursue legitimate aims, and be necessary in a democratic society (Council of Europe, 1950). Overt analytical agents must be designed and operated in ways that minimize interference with legitimate expression while effectively identifying genuine security threats.

4.1.3. Ethical Implications

Overt analytical agents raise important ethical considerations regarding bias, discrimination, and community impact. AI systems trained on historical data may perpetuate existing biases in counter-extremism operations, potentially leading to disproportionate targeting of particular communities or ideological perspectives. The risk of false positives is particularly concerning given the potential consequences of being incorrectly identified as a security threat, including social stigmatization, employment difficulties, and psychological harm.

The transparency inherent in overt operations provides important ethical advantages by enabling external scrutiny and accountability mechanisms. Community organizations, civil liberties groups, and academic researchers can evaluate system performance, identify potential biases, and advocate for necessary improvements. This transparency also enables affected individuals to understand how their data is being processed and to seek redress if they believe they have been unfairly targeted.

However, transparency also creates potential risks, including the possibility that extremist groups may develop countermeasures to evade detection. The balance between transparency and operational effectiveness requires careful consideration of what information can be disclosed without compromising security objectives. Best practices suggest focusing transparency efforts on general principles, oversight mechanisms, and accountability procedures rather than specific technical details that could enable evasion.

4.1.4. Strategic Effectiveness

The strategic effectiveness of overt analytical agents depends primarily on their ability to accurately identify genuine security threats while minimizing false positives and community alienation. Research suggests that AI systems can achieve significant improvements in processing speed and scale compared to purely human analysis, enabling security services to monitor larger volumes of content and identify threats more quickly (National Security Commission on Artificial Intelligence, 2023). However, the effectiveness of these systems ultimately depends on the quality of human analysis and intervention that follows AI-mediated identification.

Community acceptance represents a critical factor in the strategic effectiveness of overt analytical operations. Transparent approaches that acknowledge AI involvement and provide clear accountability mechanisms are more likely to maintain community trust than covert operations that may be perceived as deceptive or discriminatory. Research demonstrates that community cooperation is essential for effective counter-extremism, suggesting that approaches that maintain community trust may be more strategically valuable than those that maximize short-term intelligence collection (Spalek and McDonald, 2019).

The potential for strategic blowback from overt analytical operations is generally lower than for covert approaches, but still requires careful management. Overly broad or inaccurate targeting could validate extremist narratives about state persecution, while inadequate transparency or accountability could undermine community trust. Best practices emphasize the importance of clear policies, regular audits, and meaningful community engagement to minimize these risks.

4.2. Direct Engagement Agents: Authentic Theological Guidance

Direct engagement agents represent the most innovative and potentially transformative approach to AI-mediated counter-extremism operations. These systems are designed to provide authentic theological guidance and counter-narratives through transparent interaction with individuals at risk of radicalization, addressing critical gaps in legitimate Islamic knowledge (*Al-Ilm Al-Shari*) that extremist groups exploit for recruitment purposes. The direct engagement model emphasizes transparency about the AI-mediated nature of interactions while prioritizing theological authenticity and community validation.

4.2.1. Technical Feasibility

The development of direct engagement agents requires sophisticated natural language processing capabilities, extensive theological knowledge bases, and advanced dialogue management systems. Current large language models demonstrate impressive capabilities for generating coherent, contextually appropriate responses across diverse topics, including religious and theological subjects. However, the specific requirements for authentic Islamic guidance present unique technical challenges that require specialized development approaches.

The creation of comprehensive theological knowledge bases represents a fundamental technical requirement for direct engagement agents. These systems must incorporate authentic Islamic scholarship from diverse schools of thought (*madhhabs*), historical contexts, and contemporary applications. The knowledge base must be structured to enable nuanced responses that acknowledge theological diversity while maintaining authenticity and avoiding oversimplification of complex religious concepts.

Advanced dialogue management capabilities are essential for effective direct engagement operations. These systems must be able to maintain coherent conversations across multiple interactions, adapt their communication style to individual users' backgrounds and needs, and recognize when human intervention is necessary. The technical challenge involves balancing consistency with personalization while ensuring that all responses remain theologically authentic and strategically appropriate.

Quality assurance and validation mechanisms represent critical technical components for direct engagement agents. These systems must incorporate real-time monitoring capabilities to ensure theological accuracy, detect potential misunderstandings or misapplications, and identify situations requiring human intervention. The technical infrastructure must support continuous learning and improvement based on feedback from theological scholars, community leaders, and user interactions.

4.2.2. Legal Permissibility

Direct engagement agents operate in a complex legal environment that requires careful attention to multiple regulatory frameworks. The EU AI Act's requirements for high-risk AI systems apply to these operations, mandating comprehensive risk management, data governance, technical documentation, and human oversight (European Union, 2024). However, the transparent nature of direct engagement operations facilitates compliance with many of these requirements, particularly those related to user notification and consent.

Data protection considerations are particularly important for direct engagement operations because they involve processing personal data through interactive conversations. Users must be clearly informed about the AI-mediated nature of interactions, the purposes for which their data will be processed, and their rights regarding data protection. The legal framework requires explicit consent for data processing in most circumstances, though law enforcement exemptions may apply in specific counter-extremism contexts (European Union, 2016).

The provision of religious guidance through AI systems raises novel legal questions regarding religious freedom, professional liability, and consumer protection. While AI systems are not subject to the same professional standards as human religious advisors, they may still be held to standards of accuracy and appropriateness in their guidance. Legal frameworks must address questions of liability when AI-provided guidance leads to negative outcomes or when technical failures result in inappropriate or harmful responses.

4.2.3. Ethical Implications

Direct engagement agents raise complex ethical considerations regarding authenticity, manipulation, and community autonomy. The use of AI systems to provide religious guidance may be perceived as inauthentic or inappropriate by some community members, particularly if the systems are not properly validated by recognized religious authorities. The ethical framework must address concerns about technological mediation of spiritual guidance while recognizing the potential benefits of increased access to authentic Islamic knowledge.

The potential for manipulation represents a significant ethical concern for direct engagement operations. While these systems are designed to provide authentic guidance rather than manipulative messaging, the line between education and influence may be difficult to maintain in practice. Ethical guidelines must ensure that direct engagement agents respect user autonomy and avoid exploiting psychological vulnerabilities or emotional states for strategic purposes.

Community consultation and validation emerge as essential ethical requirements for direct engagement operations. The development and deployment of these systems must involve meaningful participation from Islamic scholars, community leaders, and affected populations to ensure theological authenticity and community acceptance. The ethical framework must address power dynamics between technology developers, security agencies, and religious communities to ensure that community voices are genuinely heard and respected.

4.2.4. Strategic Effectiveness

The strategic effectiveness of direct engagement agents depends on their ability to provide authentic, compelling alternatives to extremist messaging while maintaining community trust and theological credibility. Research on counter-narrative effectiveness suggests that authentic, community-validated messaging can successfully compete with extremist propaganda when properly designed and implemented (Duarte *et al.*, 2025). Direct engagement agents offer the potential to scale this approach while maintaining personalization and responsiveness that may be difficult to achieve through traditional counter-narrative methods.

The addressing of knowledge gaps represents a critical strategic advantage of direct engagement operations. Extremist groups exploit the lack of accessible, authentic Islamic guidance to recruit vulnerable individuals who may be seeking spiritual direction or community connection. Direct engagement agents can provide immediate, accessible responses to theological questions while connecting users with human religious authorities for more complex or sensitive issues.

Community acceptance and theological validation are essential for the strategic effectiveness of direct engagement operations. Systems that are perceived as inauthentic, biased, or manipulative are likely to be rejected by target communities and may actually validate extremist narratives about state interference in religious matters. The strategic framework must prioritize community engagement and theological authenticity over short-term operational advantages to ensure long-term effectiveness.

4.3. Covert Engagement Agents: Infiltration and Influence Operations

Covert engagement agents represent the most controversial and problematic approach to AI-mediated counter-extremism operations. These systems would be designed to infiltrate extremist networks and influence discourse through deceptive personas that conceal their AI-mediated nature. While such operations might offer certain tactical advantages, they face insurmountable legal, ethical, and strategic barriers under current frameworks that make them inadvisable for implementation.

4.3.1. Technical Feasibility

The technical requirements for covert engagement agents are substantially more complex than for overt approaches because they must maintain convincing human personas while avoiding detection by both target individuals and platform security measures. Current AI technologies demonstrate impressive capabilities for generating human-like text, but maintaining consistent personas across extended interactions while avoiding detection presents significant technical challenges.

The development of convincing cover identities requires sophisticated persona management systems that can maintain consistency across multiple platforms, interactions, and time periods. These systems must generate appropriate biographical details, maintain consistent communication styles, and respond appropriately to unexpected questions or challenges. The technical complexity increases exponentially when operating across multiple personas or coordinating activities between different covert agents.

Platform detection and countermeasures represent a significant technical challenge for covert operations. Social media platforms and online communities increasingly employ sophisticated detection systems to identify automated accounts and artificial personas. Covert engagement agents must be designed to evade these detection systems while maintaining operational effectiveness, creating an ongoing technical arms race that may be difficult to sustain.

The risk of technical failures or exposure represents a critical vulnerability for covert operations. Unlike overt systems where technical problems may cause operational inconvenience, failures in covert systems could expose the entire operation and cause significant strategic damage. The technical infrastructure must incorporate extensive redundancy and security measures to minimize these risks, substantially increasing complexity and cost.

4.3.2. Legal Permissibility

Covert engagement agents face substantial legal barriers under current regulatory frameworks, particularly the EU AI Act's prohibition of subliminal influence techniques and requirements for user notification (European Union, 2024). The Act specifically prohibits AI systems that deploy subliminal techniques beyond a person's consciousness or exploit vulnerabilities to materially distort behavior in ways that cause harm. Covert engagement operations would likely violate these prohibitions by concealing their AI-mediated nature and potentially exploiting psychological vulnerabilities.

Data protection law creates additional legal barriers for covert operations because they involve processing personal data without proper notification or consent. The GDPR requires that individuals be informed about data processing activities and their purposes, with limited exceptions for law enforcement activities (European Union, 2016). Covert engagement operations would likely violate these requirements because they involve extensive data collection and processing under false pretenses.

Human rights law establishes fundamental protections for privacy, freedom of expression, and human dignity that may be violated by covert engagement operations. The European Convention on Human Rights requires that restrictions on these rights be prescribed by law, pursue legitimate aims, and be necessary in a democratic society (Council of Europe, 1950). The deceptive nature of covert operations may violate principles of human dignity and autonomy that are fundamental to human rights frameworks.

4.3.3. Ethical Implications

Covert engagement agents raise profound ethical concerns regarding deception, manipulation, and violation of human dignity. The use of false personas to infiltrate communities and influence discourse violates fundamental principles of honesty and respect for persons that are central to ethical frameworks. The deceptive nature of these operations may cause psychological harm to individuals who discover they have been manipulated, while also undermining trust in online communities more broadly.

The potential for abuse and mission creep represents a significant ethical concern for covert operations. The capabilities developed for counter-extremism purposes could easily be repurposed for broader surveillance or influence operations that target legitimate political dissent or minority communities. The lack of transparency inherent in covert operations makes it difficult to establish effective oversight mechanisms to prevent such abuse.

The alignment with extremist tactics represents a particularly troubling ethical dimension of covert operations. Extremist groups routinely employ deception, false personas, and manipulation in their online activities. The adoption of similar tactics by state actors risks legitimizing these approaches while undermining moral authority in counter-extremism efforts. The ethical framework suggests that effective counter-extremism should model the values it seeks to protect rather than adopting the methods of those it opposes.

4.3.4. Strategic Effectiveness

The strategic effectiveness of covert engagement agents is highly questionable due to the substantial risks of exposure and strategic blowback. The discovery of covert operations would likely validate extremist narratives about state deception and persecution, potentially driving moderate voices away from counter-extremism efforts and creating new grievances for extremist exploitation. The strategic risks may outweigh any potential tactical advantages from intelligence collection or influence operations.

Community trust represents a critical factor in counter-extremism effectiveness that would be severely damaged by covert operations. Research demonstrates that community cooperation is essential for identifying and addressing radicalization risks (Spalek and McDonald, 2019). The use of deceptive tactics would likely undermine this cooperation by creating suspicion and paranoia within communities that are already subject to extensive surveillance and scrutiny.

The sustainability of covert operations is questionable given the rapid pace of technological development and increasing sophistication of detection methods. Platform operators, security researchers, and potentially extremist groups themselves are developing increasingly sophisticated methods for identifying artificial personas and automated accounts. The technical arms race required to maintain covert capabilities may be unsustainable in the long term while diverting resources from more effective approaches.

5. Discussion

5.1. Comparative Analysis of AI Agent 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 implementation strategies in AI-mediated counter-extremism operations.

Overt analytical agents emerge as the most immediately viable approach, offering substantial technical capabilities within existing legal and ethical frameworks. The transparency inherent in these operations facilitates compliance with regulatory requirements while enabling community oversight and accountability mechanisms. However, their effectiveness is limited to content analysis and threat identification, requiring human intervention for actual counter-extremism engagement.

Direct engagement agents represent the most promising long-term approach for addressing root causes of radicalization through authentic theological guidance and counter-narratives. While these systems face greater technical complexity and require extensive community consultation, they offer the potential to compete directly with extremist messaging by providing superior alternatives grounded in authentic Islamic scholarship. The strategic value of addressing knowledge gaps and providing accessible religious guidance may justify the additional investment required for development and validation.

Covert engagement agents face insurmountable barriers across all evaluation dimensions, making them inadvisable for implementation under current frameworks. The legal violations, ethical concerns, and strategic risks associated with deceptive operations outweigh any potential tactical advantages. The analysis suggests that resources devoted to covert capabilities would be better invested in transparent approaches that maintain community trust and democratic legitimacy.

5.2. The Primacy of Community Engagement

A central finding of this analysis is the critical importance of community engagement and theological authenticity in AI-mediated counter-extremism operations. The research demonstrates that approaches prioritizing transparency, community consultation, and religious validation are more likely to achieve strategic effectiveness than those relying on surveillance or deception. This finding aligns with broader research on counter-extremism effectiveness, which emphasizes the importance of community cooperation and trust (Spalek and McDonald, 2019).

The principle of *maslaha* (public interest) in Islamic jurisprudence provides theological justification for AI-mediated counter-extremism operations that serve community welfare while respecting religious values (Al-

Qaradawi, 2013). However, this justification depends on genuine community consultation and validation rather than top-down imposition of technological solutions. The framework suggests that effective AI agent deployment must be developed in partnership with Islamic scholars, community leaders, and affected populations to ensure theological authenticity and community acceptance.

The contrast with the principle of *tajassus* (prohibition on spying) highlights the theological problems with covert operations that rely on deception and surveillance (Al-Qaradawi, 2013). The analysis suggests that AI agent deployment strategies must align with Islamic ethical principles to maintain legitimacy within Muslim communities and avoid validating extremist narratives about state persecution of religious minorities.

5.3. Regulatory Implications and Compliance Frameworks

The implementation of the EU AI Act represents a fundamental shift in the regulatory landscape for AI deployment in security contexts. The Act's risk-based approach and comprehensive requirements for high-risk AI systems create both constraints and opportunities for counter-extremism operations. The analysis demonstrates that transparent approaches are better positioned to comply with regulatory requirements than covert operations that may violate fundamental principles of the Act.

The prohibition of subliminal influence techniques and requirements for user notification create particular challenges for covert AI deployment while supporting transparent engagement models (European Union, 2024). This regulatory framework suggests that future AI agent development should prioritize transparency and user awareness over deceptive capabilities that may violate legal requirements.

The intersection of AI regulation with data protection law, human rights protections, and counter-terrorism legislation creates a complex compliance environment that requires careful legal analysis for each deployment scenario. The framework suggests that successful AI agent implementation will require close collaboration between technology developers, legal experts, and policy makers to ensure compliance across multiple regulatory domains.

5.4. Technical Development Priorities

The technical analysis reveals important priorities for AI agent development that align with legal and ethical requirements while maximizing strategic effectiveness. The development of comprehensive theological knowledge bases emerges as a critical priority for direct engagement agents, requiring collaboration between AI researchers and Islamic scholars to ensure accuracy and authenticity.

Advanced dialogue management capabilities represent another key technical priority, particularly for systems designed to provide personalized guidance while maintaining theological consistency. The technical framework must balance automation with human oversight to ensure that AI systems can handle routine interactions while escalating complex or sensitive issues to qualified human advisors.

Quality assurance and bias mitigation emerge as essential technical requirements across all AI agent deployment models. The analysis demonstrates that AI systems must incorporate sophisticated monitoring and validation mechanisms to prevent discriminatory targeting and ensure accurate threat identification. These technical safeguards are particularly important given the potential consequences of false positives in counter-extremism contexts.

5.5. Strategic Framework for Implementation

The analysis supports a phased implementation approach that prioritizes immediate deployment of overt analytical capabilities while developing direct engagement agents through extensive pilot programs and community consultation. This strategic framework balances the urgent need for enhanced counter-extremism capabilities with the longer-term requirements for community trust and theological authenticity.

The suspension of covert engagement capabilities emerges as a clear strategic recommendation based on the insurmountable legal, ethical, and strategic barriers identified in the analysis. Resources that might be devoted to covert operations would be better invested in transparent approaches that maintain democratic legitimacy while achieving superior strategic effectiveness.

The framework emphasizes the importance of continuous evaluation and adaptation based on operational experience and community feedback. AI agent deployment in counter-extremism contexts requires ongoing assessment of effectiveness, community impact, and unintended consequences to ensure that operations remain aligned with strategic objectives and ethical principles.

6. Recommendations

6.1. Three-Track Strategic Framework

Based on the comprehensive analysis of AI agent deployment models, this study recommends a three-track strategic framework that prioritizes transparency, community engagement, and legal compliance while maximizing counter-extremism effectiveness (Table 2):

Table 2: Legal and Ethical Risk Assessment Matrix

Risk Category	Overt Analytical	Direct Engagement	Covert Engagement
EU AI Act Violations	Low: Transparency requirements met	Low-Medium: With proper safeguards	High: Subliminal influence prohibited
GDPR Compliance	Medium: Law enforcement exemptions	Medium: Consent mechanisms needed	High: Deceptive data collection
Human Rights Violations	Low: Proportionate measures	Low: With community oversight	High: Dignity and autonomy violations
Discrimination Risk	Medium: Bias mitigation required	Low: Community validation	High: Targeting without transparency
Privacy Intrusion	Medium: Public data focus	Low: Transparent interaction	High: Covert surveillance
Community Trust Impact	Positive: Transparent operations	Very Positive: Authentic guidance	Very Negative: Deceptive practices
Strategic Blowback Risk	Low: Legitimate operations	Very Low: Community partnership	Very High: Validates extremist narratives
Theological Compliance	Neutral: No religious claims	High: *Maslaha* principle	Low: Violates *tajassus* prohibition

Track 1: Immediate Deployment of Overt Analytical Agents: With comprehensive safeguards for bias mitigation, human oversight, and community consultation. These systems should be deployed immediately under existing legal frameworks with clear policies for data governance, quality assurance, and accountability mechanisms. The estimated budget of $10-15 mn annually represents a modest investment compared to the potential costs of successful extremist attacks or ongoing surveillance-heavy approaches.

Track 2: Pilot Development of Direct Engagement Agents: Through extensive community consultation, theological validation, and regulatory approval processes. These systems require careful development over a 3-year pilot program with estimated costs of $25-35 mn, including substantial investment in theological knowledge base development, community engagement, and quality assurance mechanisms. The pilot approach enables iterative development and refinement based on community feedback and operational experience.

Track 3: Suspension of Covert Engagement Capabilities: Pending explicit legal authorization, public debate, and resolution of fundamental ethical and strategic concerns. The analysis demonstrates that covert operations face insurmountable barriers under current frameworks and should not be pursued without substantial changes to legal and ethical frameworks that may not be advisable or achievable.

6.2. Implementation Guidelines

Community Engagement Requirements: All AI agent deployment must involve meaningful consultation with Islamic scholars, community leaders, and affected populations from the earliest stages of development through ongoing operations. This consultation should include establishment of an independent Islamic Scholarly Board to validate theological content and provide ongoing oversight of direct engagement operations.

Transparency and Accountability Mechanisms: AI agent operations must incorporate clear policies for user notification, data governance, and accountability procedures. Users must be informed about the AI-mediated nature of interactions, the purposes for which their data will be processed, and their rights regarding data protection and redress.

Quality Assurance and Bias Mitigation: All AI systems must incorporate sophisticated monitoring and validation mechanisms to prevent discriminatory targeting and ensure accurate threat identification. Regular audits should be conducted by independent experts to assess system performance, identify potential biases, and recommend improvements.

Human Oversight and Escalation Procedures: AI agent operations must maintain meaningful human oversight with clear procedures for escalating complex or sensitive issues to qualified human advisors. The framework should specify minimum qualifications for human supervisors and establish clear protocols for intervention and quality control.

6.3. Evaluation Metrics and Success Criteria

Effectiveness Metrics: Success should be measured through reduced radicalization indicators, increased community engagement with authentic Islamic guidance, and improved community trust in counter-extremism efforts. Quantitative metrics should be supplemented with qualitative assessment of community perceptions and theological authenticity.

Community Impact Assessment: Regular evaluation of community trust, perception of fairness, and unintended consequences should be conducted through independent research and community feedback mechanisms. These assessments should inform ongoing refinements to AI agent operations and policy frameworks.

Legal Compliance Monitoring: Continuous monitoring of compliance with evolving legal frameworks, including the EU AI Act, data protection regulations, and human rights law, should be maintained through regular legal review and external audit processes.

6.4. Resource Allocation and Funding Priorities

Immediate Priorities: Initial funding should prioritize development of overt analytical capabilities with comprehensive safeguards, establishment of community consultation mechanisms, and creation of legal compliance frameworks. These investments provide immediate operational benefits while establishing foundations for more advanced capabilities.

Medium-Term Development: Subsequent investment should focus on theological knowledge base development, advanced dialogue management capabilities, and pilot testing of direct engagement agents in controlled environments with extensive community oversight.

Long-Term Sustainability: Funding frameworks should ensure sustainable operations through ongoing maintenance, continuous improvement, and adaptation to evolving technological and threat landscapes. This includes investment in research and development capabilities to maintain technological advantages while preserving ethical and legal compliance.

7. Conclusion

This comprehensive analysis of AI agent deployment strategies for counter-extremism operations reveals a complex landscape of opportunities and constraints that require careful navigation to achieve strategic effectiveness while maintaining democratic legitimacy and community trust. The research demonstrates that transparent, community-partnered approaches offer superior strategic value compared to surveillance-based or deceptive methodologies, challenging conventional assumptions about the trade-offs between security and transparency in counter-extremism operations.

The study's central finding is that the most effective path forward involves competing with extremist narratives through superior theological authenticity and genuine community engagement rather than through deception or surveillance. This conclusion aligns with broader research on counter-extremism effectiveness while providing specific guidance for AI-mediated interventions that address root causes of radicalization rather than merely responding to symptoms.

The definitional ambiguity surrounding "Keyboard Jihad" emerges as a critical operational challenge that requires careful attention to avoid counterproductive targeting of legitimate religious and academic discourse. The research demonstrates that precise definitional boundaries and procedural safeguards are essential to prevent strategic blowback that could validate extremist narratives and undermine counter-extremism objectives.

The three-track strategic framework proposed in this study provides a practical roadmap for AI agent deployment that balances immediate operational needs with longer-term requirements for community trust and theological authenticity. The framework's emphasis on transparency, community consultation, and legal compliance offers a sustainable approach to AI-mediated counter-extremism that can adapt to evolving technological and threat landscapes while maintaining democratic legitimacy.

The analysis reveals that direct engagement agents designed to provide authentic Islamic guidance represent the most promising long-term approach for addressing radicalization risks. These systems offer the potential to scale counter-narrative interventions while maintaining personalization and theological authenticity that may be difficult to achieve through traditional methods. However, their successful implementation requires substantial investment in community engagement, theological validation, and quality assurance mechanisms that extend beyond purely technical considerations.

The research demonstrates that covert engagement agents face insurmountable legal, ethical, and strategic barriers under current frameworks, making them inadvisable for implementation regardless of potential tactical advantages. The analysis suggests that resources devoted to covert capabilities would be better invested in transparent approaches that maintain community trust while achieving superior strategic effectiveness.

The regulatory implications of the EU AI Act represent a fundamental shift in the legal landscape for AI deployment in security contexts. The Act's requirements for transparency, accountability, and human oversight create both constraints and opportunities that favor transparent engagement models over covert operations. This regulatory framework suggests that future AI agent development should prioritize compliance with democratic values and human rights protections rather than seeking to circumvent legal requirements through technical sophistication.

The study's emphasis on community engagement and theological authenticity reflects broader recognition that effective counter-extremism requires genuine partnership with affected communities rather than top-down imposition of technological solutions. The principle of *maslaha* (public interest) in Islamic jurisprudence provides theological justification for AI-mediated interventions that serve community welfare, while the principle of *tajassus* (prohibition on spying) creates theological objections to deceptive operations that violate community trust.

The research contributes to academic understanding of AI applications in security contexts by providing the first comprehensive analysis of AI agent deployment strategies specifically designed for counter-extremism operations. The multidisciplinary methodology integrating insights from computer science, legal studies, ethics, and Islamic theology offers a model for future research that addresses the complex intersections between technology, security, and community values.

The study's practical implications extend beyond academic inquiry to provide evidence-based guidance for policymakers, security practitioners, and technology developers engaged in counter-extremism efforts. The recommendations offer specific implementation pathways that prioritize effectiveness while maintaining compatibility with legal requirements, ethical principles, and community expectations.

Future research should examine the long-term effectiveness of AI-mediated counter-extremism interventions through longitudinal studies that track community impact, radicalization outcomes, and unintended consequences. Additional investigation is needed into the technical requirements for theological knowledge base development, the optimal balance between automation and human oversight, and the evaluation metrics most appropriate for assessing AI agent effectiveness in counter-extremism contexts.

The study concludes that AI agent deployment in counter-extremism represents both significant opportunities and substantial risks that require careful management through transparent, community-partnered approaches that prioritize democratic values and human rights protections. The three-track strategic framework provides a practical foundation for realizing these opportunities while minimizing associated risks, offering a path forward that enhances security while strengthening rather than undermining community trust and democratic legitimacy.

The ultimate success of AI-mediated counter-extremism will depend not on technological sophistication alone but on the ability to integrate advanced capabilities with authentic community engagement, theological validation, and unwavering commitment to the democratic values that effective counter-extremism seeks to protect. This research provides a foundation for achieving that integration while contributing to broader understanding of how emerging technologies can serve human flourishing when developed and deployed with appropriate attention to ethical principles and community values.

References

Alava, S., Frau-Meigs, D. and Hassan, G. (2017). Youth and Violent Extremism on Social Media: Mapping the Research. UNESCO Publishing.

Al-Qaradawi, Y. (2013). *Fiqh al-Awlawiyyat: Dirasa fi al-Dhawabit* [The Jurisprudence of Priorities: A Study in Controls]. Maktabat Wahba.

Australian Federal Police. (2022). Warning: Extremist Groups Targeting Children Through Online Gaming. *AFP Media Release*.

Bangura, A.K. (2011). Keyboard *Jihad*: Understanding the Phenomenon of Online Islamic Discourse. Academic Press.

Bellaby, R. (2024). The Ethical Problems of 'Intelligence–AI'. *International Affairs*, 100(6), 2525-2542. https://doi.org/10.1093/ia/iiae227

Bellaby, R. (2024). The Ethical Problems of 'Intelligence–AI'. *International Affairs*, 100(6), 2525-2542.

Collison-Randall, H., Spaaij, R., Hayday, E.J. and Pippard, J. (2024). Media Framing of Far-Right Extremism and Online Radicalization in Esport and Gaming. *Humanities and Social Sciences Communications*, 11, Article 1195. https://doi.org/10.1057/s41599-024-03680-4

Collison-Randall, H., Spaaij, R., Hayday, E.J. and Pippard, J. (2024). Media Framing of Far-Right Extremism and Online Radicalization in Esport and Gaming. *Humanities and Social Sciences Communications*, 11, Article 1195.

Combating Terrorism Center. (2015). Digital *Jihad*: Online Communication and Violent Extremism. West Point Military Academy.

Combating Terrorism Center. (2015). The Sterile Echo Chamber of Keyboard *Jihad*. *CTC Sentinel*, 8(3), 15-19.

Council of Europe. (1950). European Convention for the Protection of Human Rights and Fundamental Freedoms. European Treaty Series No. 5.

Duarte, F.P., Ramos, J.P., Barbosa, P., Vergani, M. and Carvalho, C.M. (2025). Effectiveness of Educational Programmes to Prevent and Counter Online Violent Extremist Propaganda: A Systematic Review. *Campbell Systematic Reviews*, e70042. https://doi.org/10.1002/cl2.70042

European Court of Human Rights. (2020). Guide on Article 9 of the European Convention on Human Rights: Freedom of Thought, Conscience and Religion. *Council of Europe*.

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council on the Protection of Natural Persons with Regard to the Processing of Personal Data (General Data Protection Regulation). *Official Journal of the European Union*.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council on Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union.*

Europol. (2022). European Union Terrorism Situation and Trend Report 2022. Publications Office of the European Union.

Home Office. (2018). Economic and Social Costs of Terrorism. *UK Government Analysis.*

Islamic State Media. (2023). Guide to Using Artificial Intelligence for Media Production. [Extremist Publication-Specific Citation Withheld for Security Reasons].

Kundnani, A. (2014). The Muslims are Coming! Islamophobia, Extremism, and the Domestic War on Terror. *Verso Books.*

Molas, B. and Lopes, H. (2024). "Say it's Only Fictional": How the Far-Right is Jailbreaking AI and What Can be Done about It. *ICCT Policy Brief, International Centre for Counter-Terrorism.*

National Security Commission on Artificial Intelligence. (2023). AI and National Security: Opportunities and Challenges. US Government Report.

New Zealand Government. (2020). Christchurch Attack: Government Response and Recovery Costs. Treasury Report.

Spalek, B. and McDonald, L.Z. (2019). Community Engagement in Counter-Terrorism: A Critical Analysis. *Critical Studies on Terrorism*, 12(4), 617-635.

Wiktorowicz, Q. (2005). Radical Islam Rising: Muslim Extremism in the West. Rowman & Littlefield.

Cite this article as: Aadil Bouhlaoui (2025). AI Agents for Counter-Extremism: Deployment Frameworks for Covert and Overt Digital Deradicalization. *International Journal of Artificial Intelligence and Machine Learning,* 5(2), 23-42. doi: 10.51483/IJAIML.5.2.2025.23-42.