The Dual-Edged Sword:
Generative AI and the Global Challenge of Violent Extremism

Muhammad Hassan Abbas[1], Sajjad Ahmad[2],Muhammad Asim Imam[3]

Abstract

Generative AI, such as ChatGPT, DeepSeek, and deepfakes, plays a complex dual role in both advancing and combating online extremism, raising critical ethical and theological questions. While extremist entities exploit AI for recruitment and ideological dissemination, counter-extremism initiatives are increasingly using AI to identify and disrupt radical content. This article explores this tension and discusses illustrative case studies to show how online radicalization can lead to real-world violence. It also addresses theological concerns about AI's interpretation of sacred texts, which may reflect biases across various religious traditions, including Islam, Christianity, Judaism, and Hinduism. The paper proposes the formation of "Ethical AI Oversight Committees" comprising religious scholars, technologists, and policymakers to ensure a balanced approach that safeguards security, religious liberty, and cultural diversity. Integrating perspectives from AI ethics, religious studies, and policy, this study advocates for a proactive, ethically guided digital environment to counter the threat of violent extremism.

Keywords: Generative AI, Counter-extremism, Online radicalization, Ethical AI, Algorithmic bias, Religious freedom, Digital literacy, Ethical oversight

Introduction

The explosive advancement of generative artificial intelligence (AI) systems, including ChatGPT and deepfake devices, has altered the environment of historical communication, which has allowed the creation of both beneficial and harmful tools in spreading ideologies (Vaughan *et al*, 2025). This has provided extremist organizations such as Islamic State (ISIS) with the means to improve their recruitment campaigns and use AI-generated propaganda to create compelling narratives and identify vulnerable individuals whom they can now target with a precision never seen before (Hemrajani, 2024). Meanwhile, counter-extremism efforts have started using AI to identify and block the evolution of radicalization through sentiment analysis and

[1] PhD Scholar, Istambul Sabahattin Zaim University, Turkiye. Email: mhassan.abbas299@gmail.com

[2] Associate Professor, sajjad@mul.edu.pk , Faculty of Economics and Management Sciences, Minhaj University Lahore, Pakistan.

[3] School of Business and Economics, Universiti Putra Malaysia.

https://doi.org/10.58932/MULF0050

predictive modeling to discover at-risk people (Irfan et al., 2024). This symbiosis of AI with both extremist mobilization and the thwarting of processes of radicalization is an expression of two sides of the same coin in terms of potential dangers as well as opportunities, thereby requiring solid academic research into the topic as far as it concerns the wider society (Zaidi *et al.* 2024).

The technical aspect of the dual-use nature of AI is not the only issue: numerous ethical and theological problems also still require addressing despite surpassing critical levels (Henschke & Reed, 2021). These ethical questions, including algorithmic bias and the conflict between free speech and content moderation, have already been thoroughly covered in Western society, where AI technology tends to exacerbate the prejudices prevalent in society or unintentionally silence the voices of the right (Kawka, 2025). The theological aspects of AI, especially in non-Western and religiously diverse contexts, however, bring yet further depths of complexity. As an example, the faulty understanding of religious texts in AI or the creation of religiously-sensitive material may cause cultural bias and give the extremist narratives more impact (Kawka, 2025). This is of specific concern in areas where religious identity is one of the main factors promoting either social cohesion or division, as is the case in the Middle East or South-Asia (Meena *et al.* 2025). The technical abilities of AI and its ethical and theological implications therefore require a delicate system to determine how they can be used to propagate as well as to control extremism.

Although the field of research related to the issue of AI and extremism is expanding, there are considerable gaps, specifically with regard to the theological implications of AI and its implementation in other regions beyond the West (Florea & Gilder, 2024). Most of the literature is biased in favour of technical or ethical aspects of AI in Western contexts with little or no attention being paid to the interaction between AI and religious beliefs or cultural systems not located in the Global North (Peters, 2023). To cite an example, concerning the use of AI in identifying extremist content on such platforms as X or YouTube, researchers consider the matter of the role of AI in detecting extremist content (Hussey, 2024) but few explore how AI-generated content can misinterpret religious texts such as Islamic or Christian texts. This risks alienating communities or promoting further extremist ideologies (Lipps). In addition, there is a necessity to note that the counter-extremism policies are deficient if they lack culturally sensitive artificial intelligence (AI) frames in areas where religious and cultural considerations influence the process of radicalization. The existence of this research gap is highlighted by the necessity to adopt a multidisciplinary approach, which considers theological, ethical and cultural aspects of the role of AI in evaluating extremism (Nwankwo).

This research aims at examining the duality of Artificial Intelligence as a tool to promote and combat extremism in relation to its ethical and theological issues and suggesting a model of ethical supervision that can balance technological novelty with cultural and religious sensitivity (Juma, 2024). The paper will investigate more specifically the use and abuse of generative AI tools by extremist militants and their application to counter-extremism, ethical and theological issues mostly in non-Western societies (Adib-Moghaddam, 2025). The study is expected to address the research gap already explained by including case studies, theological analysis and ethical considerations that are designed to deal with the research gap and make recommendations to policymakers, technologists, and religious scholars (Gantt & Natt, 2024).

The second section of the paper provides a literature review of the applications of AI to extremism and counter-extremism, its technical, ethical, and theological aspects. In the third section, case studies of AI applications by extremist groups and counter-extremism efforts are offered, which will draw from geographically different contexts. In the fourth section, the ethical and theological dilemmas (and, in particular, the bias in algorithms, freedom of speech, and religious text distortion) are examined. Section five offers a paradigm of ethical AI management that prioritizes values and design, respectful of the specifics of culture and inter-disciplinary considerations. Lastly, section six will present the implications on policy and future research. This paper aims to provide a detailed review of the role of AI in extremism and a future-oriented comprehensive investigation of the ethical implications of AI.

Background and Literature Review

The advancement of generative artificial intelligence (AI) has greatly expanded the power of extremist groups to advance their beliefs and obtain adherents. Islamic State (ISIS) and other extremist groups use AI-based data analysis to track down vulnerable populations and apply predictive algorithm-based propaganda to target specific geographical areas and respond to unacceptable behavior (Egunjobi, 2025). With assistance from AI, social media offers echo chambers that spread extremist messages because algorithms are programmed to showcase the information that results in the highest engagements, which commonly promotes incendiary and polarizing content (Velazquez, 2024). For example, AI-generated content, such as deepfake videos and texts developed with the help of natural language processing (NLP), allows extremists to create realistic and credible propaganda that would have a positive impact on its persuasiveness (Afolabi & Molade). NLP-based chatbots have also been applied to automate the recruitment procedure through the real-time engagement of potential recruits by simulating one-on-one connections and enhancing ideological commitment (Kirova *et al.* 2023).

The research also suggests that AI tools in the hands of extremists can help them raise their operations to the level where personalized messages could reach their targets everywhere around the globe and manipulate their cultural and religious sensibilities (Elshimi, 2024). Technical access to extremist groups is also easier by means of open-source AI models because even the most minor units can run advanced propaganda on their own.

AI as a Tool to Counter Extremism

As extremist organizations misuse AI, counter-extremism efforts have become more inclined to use AI technologies to identify, interfere with, and prevent radicalization activities (Culture, 2023). Machine learning-based models of sentiment analysis and anomaly detection allow authorities to identify people who are likely to be radicalized, based on patterns of online behavior and communication patterns (Rassler, 2021). As an example, moderation by means of predictive algorithms would allow the identification of extremist content on platforms such as X or YouTube based on linguistic features and network links and then remove this content almost immediately (AI, 2023).

AI is also being used to create counter-narratives and NLP systems are being created that will produce specific messages to disarm extremist ideologies and present more moderate visions. Organizations and governments, including the Global Internet Forum to Counter Terrorism (GIFCT), have also used AI to moderate content automatically in order to minimize the transmission of extremist content online. Nevertheless, such systems also have their limitations since, in many cases, they fail to accept the nuances of the context, which can lead to an incorrect positive judgment that can either censor valid speech or fail to identify extreme rhetoric, which was slightly hidden (Hindu, 2025). Nevertheless, AI-supported counter-extremism is on the rise and further research is being conducted to make the technology more accurate and culturally sensitive so as to be efficient in various global contexts.

Ethical Challenges

There are serious ethical issues regarding AI's involvement in extremism and counter-extremism issues, such as bias, transparency, and accountability (Rassler, 2021). Also, algorithmic bias is a well-documented problem since the data used to train an algorithm can include social prejudices that lead to disproportionate targeting of marginalized sections of society in various counter-extremism measures (Jambrek, 2024). For example, elements of facial recognition and predictive policing are being criticized due to their ability to encourage racial and religious biases that diminish the reliability of AI systems.

The question of transparency is also of crucial concern because proprietary algorithms of technology companies are not open to scrutiny and this casts doubts on how they make decisions (Hurley, 2024). Accountability is also a problem because the wide use of AI development by governments, industry firms, and open-source groups makes it more difficult to control the harmful consequences of the misuse of AI. There is a need to include stakeholders in the designing process and it is important to establish proper oversight to reduce risks. Yet, the minimal implementation of these measures and the long-term ethical requirements of AI in non-Western countries is poor (Barlow & Holt, 2024).

Theological Perspectives

The overlap between AI and religious studies also presents theological dilemmas, especially in the interpretation of religious texts and the use of technology in religious practices (Scherz, 2024). When it comes to creating or analyzing religious material, AI already includes or will include the possible misinterpretation of sacred texts, which may cause great offense to certain religious communities and may even encourage radicalism. For example, NLP models that use religious concepts can be problematic because the results may lack theological subtleties and result in the misinterpretation of Islamic, Christian, or other sacred scriptures (Tzeng).

Scholars claim that the growing use of technology in religious dialogue requires that theological framework provide guidelines for the development of AI according to various belief systems (Jambrek, 2024). Within the Islamic world, the idea of using AI to analyze the Sharia or Qur’anic texts creates issues of authenticity and authority because AI might overlook the contextual nature of religious learning. Hence, algorithms are unlikely to replicate the work of specialists in this field. These theological perspectives highlight the need for interdisciplinary collaboration between technicians and religious scholars to deal with the impact of AI on sacred narratives and prevent the exploitation of AI by extremist groups (Gbadebo).

Case Studies

The Rohingya Crisis in Myanmar

The Rohingya genocide in Myanmar (2016–2017) highlighted how AI-driven social media platforms can become complicit in real-world violence. Facebook’s recommendation algorithms, designed to maximize engagement, amplified anti-Rohingya hate speech in Burmese language leading to widespread provocation and contributing to a deadly campaign of ethnic cleansing (Amnesty International, 2022a; Mozur, 2018). Despite multiple warnings from human rights groups, Facebook’s moderation systems failed to respond adequately, largely due to poor local language processing and lack of culturally contextual understanding by its AI systems (United Nations, 2018). The platform’s automated tools not only failed to detect hate speech but in many cases actively promoted it, as Meta’s own internal reviews later revealed (Amnesty International, 2022b). A UN fact-finding mission concluded that Facebook played a “determining role” in the genocide (United Nations, 2018). This case underscores the urgent need for ethical, localized AI development and robust content moderation frameworks to prevent digital tools from fueling mass violence.

Sri Lanka Easter Bombings (2019)

The Sri Lanka Easter bombings in April 2019 perpetrated by the National Thowheeth Jama'ath (NTJ) confirmed the role of online coordination in perpetrating large-scale attacks and the inadequacy of AI detection outside the West (Henschke & Reed, 2021). The attackers used social media such as WhatsApp and Telegram to organize and coordinate the bombings that resulted in the death of more than 250 people and were able to avoid traditional surveillance efforts by taking advantage of the encrypted medium (Meena *et al.* 2025). The ideology of the NTJ was fueled by ISIS-related propaganda spread by global online networks. This is one more example of transnational access to the sphere of digital extremism (Florea & Gilder, 2024).

The challenge posed by non-Western environments to AI-based extremist content detection consists mainly of the unavailability of material in languages such as Sinhalese and Tamil and the fact that much of the algorithmic training occurs in the Western context, which is not easily applicable in a non-Western environment (Sajjad, 2022). For example, automated moderation tools are not likely to detect local extremist rhetoric/signs and this undermines their performance in regions such as South Asia. The attack in Sri Lanka provides evidence that AI should consider local linguistic and cultural data to increase its success rate in detection. The attack also suggests that the global community should work together to track transnational terrorist groups (Viana & da Silva, 2021).

Christchurch Shooting (2019)

The Christchurch mosque shooting in 2019 perpetrated by Brenton Tarrant provides a clear case of live-streamed extremism and raises critical questions about AI's role in rapid content removal and associated ethical concerns (Kingdon, 2024). Tarrant, who killed 51 people, broadcast the attack on Facebook live and the footage was rapidly shared across platforms like X and YouTube, fueled by algorithmic amplification and user engagement (Sonrexa *et al.* 2023). His statements, posted online, drew heavily from far-right online rhetoric and were similar to those used by the radical, Breivik. AI-driven content moderation systems, such as those used by major platforms, employ real-time detection to remove extremist material but the Christchurch attack exposed their limitations as the live-stream continued for over 17 minutes before intervention. It is true that ethical challenges arise in balancing rapid content removal with free speech, as overly aggressive AI moderation risks censoring legitimate content, while under-moderation allows harmful material to spread. Moreover, the global dissemination of the video highlighted the difficulty of coordinating AI moderation across jurisdictions with varying legal frameworks (Birch, 2021). The Christchurch case emphasizes the need for AI systems to improve real-time detection capabilities and incorporate ethical safeguards to mitigate harm while respecting user rights (Rabiu *et al.* 2025).

Balochistan Train Attack (2025)

On March 11, 2025, the Baloch Liberation Army (BLA), a separatist and banned group, hijacked the Jaffar Express in Balochistan, Pakistan, killing 21–27 hostages, including 18 soldiers, and 33 BLA militants (Hindu, 2025). The attack involved 400–500 passengers and targeted military personnel. The BLA used encrypted apps like Telegram for coordinating and releasing propaganda videos, which were possibly AI-generated, in order to highlight their narrative. Pakistani authorities allege Indian backing arising from geopolitical tensions (Guardian, 2025). The BLA's demands included releasing their member prisoners, which reflected their grievances over resource exploitation and marginalization (Ebner, 2023)

AI could have detected these attacks via network analysis by tracking extremist communications on X or Telegram. However, AI had limited access to data in Balochi and Urdu, which enabled encrypted platforms to evade surveillance. Ethically, AI risks over-censorship and potentially silencing legitimate voices while geopolitical accusations complicate detection. The attack underscores the need for culturally sensitive AI to address proxy-driven extremism without escalating regional conflicts (Douglas & Shin, 2025).

Theological Implications of AI Moderation

Islamic Perspectives

The use of artificial intelligence (AI) to moderate online content, particularly in the context of countering extremism, raises significant theological concerns within Islamic scholarship, especially regarding the interpretation of Qur'anic texts (Tsuria & Tsuria, 2024). AI-driven systems, such as natural language processing (NLP) models, are increasingly used to detect extremist content by analyzing religious texts or user-generated content for radical rhetoric. But scholars contend that the use of AI in the interpretation of the Qur'an can lead to distortion because AI can fail to consider contextual and legal realities in interpreting the Qur'an and distort the process of *tafsir* jurisprudence. For example, algorithms based on a small amount of biased data misread Qur'anic texts when Islamic propaganda is included in the same category as acceptable religious rhetoric (Gantt & Natt, 2024). Such bias is especially evident in Arabic environments where lack of skill in the language and in the local dialects can undermine the effectiveness of AI and lead to stereotypes or to the silencing of minority voices.

With regard to the understanding the sacred texts, Islamic schools of thought emphasize the significance of human agency, which is impossible to replace with AI services that tend to undermine the role of the trained jurist. The bias is also compounded by issues of historical inadequacy surrounding the teaching of alternative interpretations of Islam in the resources of AI modeling that tend to follow the Western or Sunni model as opposed to the Shia or Sufi interpretation. For example, AI instructions that operate in some given platforms, including X, may label Arabic materials as extremist and become over-censored by not taking cultural or theological perspectives into account. In response, scientists suggest the combination of structures made by various Islamic professionals to control the development of AI and to maintain religious diversity and ethical integrity (Irfan *et al.* 2024).

Christian, Jewish, and Hindu Perspectives

AI moderation is based on theological considerations that by-pass Christian, Jewish, and Hindu traditions that consider the influence of technology on the interpretation of sacred texts as an expression of power (Kitching & Gholami, 2023). The use of AI to review Christian scripture in search of extremist materials poses problems of oversimplification because the computer code might not be able to grasp the historical and theological richness of the Bible and thus declare the approach of evangelism or liberation theology as radical. Christian scholars also highlight the risk of AI diminishing the role of clergy in contextual interpretation, particularly when automated systems prioritize literal readings over nuanced exegesis. With regard to Jewish applications, the use of AI on Torah or Talmud commentary is no less problematic because the interpretive tradition (*midrash*) that depends on rabbinic insight and communal discussion cannot be simulated by AI. For example, AI moderation can erroneously regard discussions of Jewish law as extremist in a multilingual context in which cultural specificity is paramount (Hebrew or Yiddish).

The work of AI with texts such as the *Bhagavad Gita* or the *Vedas* can over-simplify philosophical dimensions by using two categories that may offend devotees or promote a non-violent doctrine in a form of militant ideas (Schotten, 2024). The application of AI to implementing moderation of Hindu nationalism texts on websites such as YouTube only complicates the situation because the algorithms are unable to understand the difference between cultural pride and extremist speeches and so run the risk of strengthening bias in the training data (Henschke & Reed, 2021). Across all these religious traditions, scholars call for AI systems that demonstrate theological expertise so as to ensure respectful and accurate handling of sacred texts. This emphasizes the need for interfaith collaboration to address shared concerns.

Common Themes

The implications of culturally insensitive AI responses to theology highlight tensions that can be found in all the religious traditions such as the contradiction between the control of AI and spiritual self-determination. Of chief concern is that AI will reduce human agency in religious interpretation since the role of human reason in the interpretation of sacred texts lies at the heart of Islamic, Christian, Jewish, and Hindu spiritual traditions. The risk of using AI is the affront to religious sensitivities by the incorrect reference to legitimate religious language as extremism (Beard et al., 2024). For instance, platforms using AI to moderate content may inadvertently censor theological debates, stifle religious expression and spiritual dialogue (Beard, 2024).

Another cultural insensitivity of AI is the tendency to use homogenized datasets or Western-oriented datasets, which results in discrimination against minority religious communities or non-Western nations in general (Burlacu, 2024). The insensitivity of AI to linguistic and theologically diverse groups, which it identifies as extremist, may actually contribute to the mistrust between these religious groups. In contrast, scholars can promote interdisciplinary study of various religious traditions in order to develop culturally sensitive AI frameworks, which are based on transparency, accountability and inclusivism that enable AI moderation to respect spiritual autonomy while successfully curbing extremism. The development of such frameworks requires global cooperation to address the universal challenges posed by the theological implications of AI (Condrey, 2024).

Ethical and Practical Challenges

We have seen that the ethical deployment of AI for the purposes of counter-extremism is critically challenged by algorithmic biases, free speech concerns, a lack of legal frameworks, and a rapidly evolving digital landscape. A primary issue is that AI tools, often developed in the West using biased training data, fail to accurately interpret non-Western religious and political content. Natural language processing (NLP) models may consider theological discussions in Arabic, Urdu, or Hindi as extremist and this could exacerbate the marginalization of minority communities (Fernandez, 2024). Such lack of cultural awareness was evident in responses to the Baluchistan train attack, where AI struggled to distinguish separatist rhetoric from legitimate grievances (Groothuis, 2023).

Moreover, such biases lead to the excessive censorship of free speech, which prevents freedom of expression. AI’s automated, context-blind systems can remove content that is simply an expression of grief or legitimate religious debate as occurred following the Christchurch mosque attacks (Adib-Moghaddam, 2025; Rashid, 2023). Such intervention undermines trust and accountability, particularly in regions where religion is central to public discourse (Slattery & Green, 2024). Furthermore, the absence of synchronized international laws can create inconsistent enforcement. For example, differing responses to the Charlie Hebdo attack highlighted the diversity of national standards related to hate speech and data privacy and this prevented a unified global approach (Hemrajani, 2024). Geopolitical tensions such as those existing between India and Pakistan further hinder essential intelligence sharing and standardization (Zaidi *et al*. 2024).

Finally, extremist groups continuously adapt to new circumstances by using encrypted platforms like Telegram, AI-generated propaganda and deepfakes to evade detection. The live-streamed Christchurch attack and the Baluchistan Liberation Army's (BLA) use of evolving tactics demonstrate how AI systems that lack real-time adaptability are quickly outpaced (Meena *et al*. 2025; Guardian, 2025). Continuous model retraining and dynamic collaboration between tech firms and governments are essential to keep pace with these threats (Florea & Gilder, 2024).

Proposed Solutions: an Ethical AI Framework

Artificial intelligence (AI) plays a positive role as well as a negative role with regard to extremism. This fact suggests the need for an effective ethical control system to address the theological, cultural, and ethical problems the emerge with the use of AI. The suggestion is to engage the so-called Ethical AI Oversight Committees (EAOCs), an interdisciplinary group of at least several employees consisting of clergy, technologists, policy-makers and community leaders, who would endeavour to develop AI moderation along with ethical and cultural sensitivity (Elshimi, 2024). The purpose of these committees would be to reduce algorithmic bias, defend the freedom of expression and promote culturally sensitive artificial intelligence, especially in counter-extremist work. By deriving benefit from various ideas and perspectives, EAOCs could aim to prevent the disconnection between technological advancement and human ideals so that AI would become aware of different religious and cultural backgrounds and involved them in preventing extremist propaganda. For example, during the Baluchistan train attack, these committees could have instructed AI to discriminate between legitimate Baloch demands and extremist rhetoric, thereby reducing damage to marginalized groups. An EAOC model relies on the principles of participatory governance that prevents the development of bias regarding inclusivism and gives greater emphasis to the role of accountability for the elimination of the global challenges of extremism with the help of AI.

The formation of EAOCs allows for the involvement of diverse elements as well as strong governance to manage the development and application of AI in the prevention of extremism. The composition of each committee ought to consist of (1) Islamic, Christian, Jewish, Hindu and other religiously related scholars who could contribute their religious expertise; (2) AI technologists, whose activity would relate to technical feasibility; (3) policymakers, who would help accommodate legal frameworks; and (4) community leaders who would represent the cultural and social settings of the environment. For example, in Pakistan an EAOC might bring together Baloch tribal leaders and Islamic scholars to adapt AI moderation to the local context. Moreover, the main roles of EAOCs would be: (1) to set up guidelines for culturally sensitive AI algorithms to make sure that they do not disrespect religious texts and practices; (2) to audit the AI system regularly in order to detect and reduce biases, especially the kind of bias that misinterprets the non-Western content; and (3) to develop dispute resolution measures for effective action against the improper removal of content that interferes with freedom of expression. At the national and international levels, the work of EAOCs could be combined with such platforms as X and such organizations as the Global Internet Forum to Counter Terrorism (GIFCT) to give them a more universal impact. Such an ethics-driven governance system will allow for the promotion of transparency and accountability and pay attention to the ethical and practical issues of AI moderation (Juma, 2024).

Implementing EAOCs requires the presence of strategic leaders (or pilots) in diverse geopolitical contexts, such as Pakistan and the European Union, to test their efficacy in counter-extremism measures while addressing the local context. In Pakistan, such a pilot could focus on the region of Baluchistan where the 2025 train attack exposed the limitations of AI to detect culturally specific extremist rhetoric. Such an EAOC could collaborate with local scholars and Baluch leaders to refine NLP models and to reduce false elements in content moderation by incorporating Balochi and Urdu datasets. In the EU, a pilot could address far-right extremism, evident in the Christchurch shooting, by integrating Christian and Jewish perspectives to ensure that AI respects religious discourses while targeting hate speech. Integration with digital literacy programs is critical in order to educate communities about the role of AI in fostering trust and reducing alienation. Counter-narrative campaigns supported by EAOCs could use AI to promote moderate voices and counter extremist propaganda like that used during the Charlie Hebdo attack. Success metrics would include the reduction of false positives in content removal (e.g., a 20% decrease in erroneous flags), the increase of community trust (measured via surveys), and faster detection of extremist content (e.g., 50% reduction in propagation time). Challenges would include securing funding and navigating geopolitical tensions such as during the Pakistan-India disputes and obtaining neutral facilitation by international bodies like the UN (Barlow & Holt, 2024).

Recommendations for a Resilient Digital Ecosystem

Deploying artificial intelligence (AI) ethically in counter-extremism requires robust frameworks to minimize bias and ensure transparency, addressing issues like those seen in the Baluchistan train attack. Ethical AI Oversight Committees (EAOCs), as proposed, should integrate clergy, technologists, and community leaders to develop guidelines that prioritize cultural sensitivity and accountability. These frameworks must mandate diverse training datasets to reduce Western-centrist biases, which often misinterpret non-Western religious content, as evidenced in cases like the Charlie Hebdo attack. Transparency can be achieved through public reporting of AI moderation decisions, allowing scrutiny and appeals to protect freedom of expression. For instance, platforms like X should publish anonymous moderation logs to build trust, particularly in regions like South Asia where mistrust is high. Regular audits, informed by the IEEE Ethically Aligned Design principles, can ensure compliance with ethical standards, mitigating risks of over-censorship and bias (Ismail, 2024). Such measures foster equitable AI systems that balance security and human rights.

Community Engagement and Education

Community engagement through digital literacy programs is vital to counter radicalization and build trust in AI systems. Programs should educate communities about AI moderation as seen, for example, in the need to clarify misinterpretations of Baloch grievances. Initiatives taken after the Christchurch shooting can help to promote counter-narratives by using AI to amplify moderate voices. Engaging religious and community leaders, as is done by the EAOCs, has the advantage that programs will reflect local values and reduce alienation. For instance, collaboration with Islamic scholars in Pakistan could have countered BLA propaganda and fostered resilience against extremism (Irfan *et al.* 2024).

Continuous Monitoring and Evaluation

Continuous monitoring and evaluation of AI systems is crucial to maintain their effectiveness and ethical impact (Meena *et al.* 2025). Metrics should include false positive rates (e.g., a 20% reduction in erroneous content flags), community trust levels (via surveys), and detection speed (e.g., 50% faster content removal) as was evident in the Christchurch case. Regular audits by EAOCs can identify the biases that became clear in the Baluchistan attack and ensure compliance with ethical standards (Sonrexa *et al.* 2023). Feedback loops that involve users and local stakeholders could refine AI performance and increase adaptability to evolving tactics such as deepfakes (AI, 2023).

Discussion

The interaction between artificial intelligence (AI) in extremism and counter-extremism, as explored through case studies such as the Baluchistan train attack, the Oslo and Utøya attacks, Charlie Hebdo, the Manchester Arena bombing, the Sri Lanka Easter bombings and the Christchurch shooting reveals a complex field of opportunities and challenges. These case studies highlight AI's dual role, namely, that extremist groups make use of AI for propaganda and coordination while counter-extremism efforts use it for detection and content moderation. Theological considerations expose the risks of AI misinterpreting sacred texts in Islamic, Christian, Jewish, and Hindu contexts, which results in potentially alienating communities and fueling mistrust. For instance, in the Baluchistan attack, AI's failure to distinguish separatist rhetoric from legitimate grievances exacerbated the tensions that were already evident. Ethical considerations such as algorithmic bias and over-censorship further complicate AI's deployment, as seen in the Christchurch case where legitimate content was mistakenly removed. Practical hurdles including the lack of global legal frameworks and the rapidly evolving digital landscape hinder AI's adaptability to new tactics such as deepfakes. The proposed Ethical AI Oversight Committees (EAOCs) depend on all of these insights and make use of interdisciplinary expertise to ensure culturally sensitive, transparent and accountable AI systems (Alfiannor & Kurniawan, 2025).

Implications

The EAOC tasks and recommendations address critical gaps in counter-extremism by fostering culturally sensitive AI systems that mitigate bias and respect theological diversity. By incorporating clergy and community leaders, EAOCs ensure AI respects religious contexts, as needed in cases like the Charlie Hebdo attack, where multilingual moderation failed. The framework's emphasis on transparency and dispute resolution will protect freedom of expression by addressing excessive censorship issues such as those evident during the Christchurch shooting. International collaboration and digital literacy programs tackle the lack of global coordination and community trust, which was critical in regions like Baluchistan with geopolitical tensions. Continuous monitoring ensures that AI will adapt to evolving tactics, enhancing resilience against propaganda that play a role in the Manchester bombing (Peters, 2023). These measures will bridge the gap between technological innovation and ethical governance and promote a digital ecosystem that balances security with cultural and spiritual integrity (Gantt & Natt, 2024).

Limitations and Future Research

The EAOC framework faces feasibility challenges due to resource constraints and fluctuating regional capacity, particularly in data-scarce areas such as Baluchistan (Zaidi et al., 2024). Moreover, resistance from tech firms that prioritize proprietary systems and from governments that are wary of sovereignty issues may hinder implementation, as became evident during the Pakistan-India disputes. Ensuring adequate representation in EAOCs becomes an issue in polarized regions and leads to the risk of tokenism or the exclusion of minority voices. These limitations require careful navigation to ensure effective deployment (Rashid, 2023).

Future research should explore AI's role in countering misinformation, which is a key driver of radicalization as became evident in the spread of extremist narratives post-Christchurch. Counter-extremism strategies will be enhanced by investigating AI's impact on broader radicalization contexts, by including non-religious ideologies and by developing adaptive algorithms for decentralized platforms. Research into cross-cultural AI training datasets could further address biases and ensure global applicability (Viana & da Silva, 2021).

This study elucidates AI's dual role as both a tool for extremist groups, as seen in the Baluchistan train attack and a mechanism for counter-extremism, as evidenced in efforts post-Christchurch. However, its ethical and theological risks—algorithmic bias, excessive censorship, and misinterpretation of sacred texts—threaten its efficacy and societal trust. Case studies reveal persistent gaps in cultural sensitivity and global coordination that is exacerbated by the evolving digital landscape. The proposed EAOC framework addresses these issues by fostering inclusive, transparent AI systems (Mesok *et al.* 2024).

Interdisciplinary collaboration between technicians, religious scholars, policymakers and community leaders is imperative for the development of ethical AI systems that counter extremism without compromising human rights or cultural values. Stakeholders must prioritize EAOC pilots and digital literacy to build trust and resilience (Ismail, 2024). A resilient digital ecosystem will find the right balance between security, spirituality, and cultural diversity and enable AI to counter extremism while respecting theological, ethical and social values. The final result will be to create a safer and more inclusive online environment.

References

Adib-Moghaddam, A. (2025). *The myth of good AI: A manifesto for critical Artificial Intelligence*. Manchester University Press.

Afolabi, T., & Molade, S. How AI Is Identifying New Forms of Antisemitic Hate Speech on Extremist Social Media.

AI. (2023). Generating Insights from Catholic Social Teaching: Ethical Guidelines for Artificial Intelligence in Health Care Ministries.

Alfiannor, M., & Kurniawan, W. (2025). Countering violent extremism: The international deradicalisation agenda.

Barlow, J., & Holt, L. (2024). Attention (to Virtuosity) Is All You Need: Religious Studies Pedagogy and Generative AI. *Religions*, *15*(9), 1059.

Beard, B. A. (2024). Artificial Intelligence and Theology: A Bibliographic Essay. *Theological Librarianship*, *17*(2), 31-42.

Beard, B. A., Graham, A., Peterson, E., & Schmersal, D. E. (2024). How Doomed Are We? A Philosophical/Theological Consideration of AI/ChatGPT in Relation to Theological Libraries and Theological Education. *Atla Summary of Proceedings*, 83-105.

Birch, M. (2021). False Positives: The Prevent counter-extremism policy in healthcare.

Burlacu, M. (2024). Knowledge, conscience, consciousness, awareness, or about the presence and use of artificial intelligence (ai) in spiritual life and their challenges.

Studia Universitatis Babes-Bolyai-Theologia Orthodoxa, *69*(1), 37-56.

Condrey, B. (2024). The Christian educator as prophet, priest, and king: Nurturing moral formation in a ChatGPT era. *International Journal of Christianity & Education*, *28*(2), 198-218.

Culture, C. f. D. (2023). Encountering Artificial Intelligence: Ethical and Anthropological Investigations. *Journal of Moral Theology*, *1*(Theological Investigations of AI), i-262.

Douglas, M., & Shin, W. (2025). The Advent of a Three-Body Problem: Artificial Intelligence and Political Theologies. *Interpretation, 79*(2), 145-156.

Ebner, J. (2023). *The identity-extremism nexus in virtual groups: the impact of online group alignment on radicalisation towards violence* University of Oxford].

Egunjobi, J. P. (2025). The Misuse of AI-Generated Content in Academic and Religious Settings. *International Journal of Research and Scientific Innovation XII, no. XV*, 871-879.

Elshimi, M. S. (2024). The constraints hypothesis: Rethinking causality in deradicalisation, disengagement and reintegration pathways. A complex system perspective. *Studies in Conflict & Terrorism, 47*(12), 1623-1647.

Fernandez, S. (2024). When counter-extremism 'sticks': the circulation of the Prevent Duty in the school space. *Identities, 31*(5), 665-683.

Florea, D., & Gilder, E. (2024). Pushing the Limits of Theosis in the Digital Age: Exploring AI Complexities and their Impact on Romanian Traditional Religious Practices. *Journal for the Study of Religions and Ideologies*, 73-87.

Gantt, L. O., & Natt, I. (2024). More than Machines: The Ethical and Human Implications of Generative AI and Lawyering. *J. Christian Legal Thought, 14*, 16.

Gbadebo, M. O. Terrorism and global security: Cyber threats, governance, and counterterrorism strategies.

Groothuis, S. (2023). Islamophobia and securitisation: the Dutch case.

Guardian, T. (2025). *Pakistan accuses India of sponsoring militant terror group after train hijacking.*

Hemrajani, A. (2024, August 26). *The use of AI in terrorism*

Henschke, A., & Reed. (2021). Toward an ethical framework for countering extremist propaganda online. *Studies in Conflict & Terrorism*, 1-18.

Hindu, T. (2025). *Jaffar Express train attack in Balochistan: Pakistan Army says 18 of 26 hostages killed were soldiers.*

Hurley, G. F. (2024). Augustinian Composition Pedagogy and the Catholic Liberal Arts in the Time of Generative AI. *Journal of Catholic Education, 27*(2), 59-75.

Hussey, I. (2024). Preaching and Generative AI: A Perspective from Early 2024. *International Journal of Practical Theology, 28*(2), 307-323.

Irfan, M., Almeshal, Z. A., & Anwar, M. (2024). Unleashing transformative potential of artificial intelligence (AI) in countering terrorism online radicalisation extremism and possible recruitment.

Ismail, N. H. (2024). Gaming and AI in the tussle over youth radicalisation. *RSIS Commentaries, 154-24.*

Jambrek, S. (2024). Christians Facing the Challenges of Artificial Intelligence. *Kairos: Evangelical Journal of Theology*, *18*(1), 75-94.

Juma, M. N. (2024). Navigating the ChatGPT Theological Terrain: Considerations for Graduate Theology Students. *Pan-African Journal of Education and Social Sciences*, *5*(2), 158-166.

Kawka, M. (2025). Theological imaginal reflections with AI artworks: revealing and obscuring divine knowledge in aesthetic experience. *Practical Theology*, *18*(2), 152-167.

Kingdon, A. (2024). Socio-Technical and Multi-stakeholder Approaches to Countering Online Propaganda. In *the World White Web: Uncovering the Hidden Meanings of Online Far-Right Propaganda* (pp. 209-251). Springer.

Kirova, V. D., Ku, C. S., Laracy, J. R., & Marlowe, T. J. (2023). The ethics of artificial intelligence in the era of generative AI. *Journal of Systemics, Cybernetics and Informatics*, *21*(4), 42-50.

Kitching, K., & Gholami, R. (2023). Towards Critical Secular Studies in Education: addressing secular education formations and their intersecting inequalities. *Discourse: Studies in the Cultural Politics of Education*, *44*(6), 943-958.

Lipps, J. Philosophical and Theological Implications of Generative AI Use for Human Flourishing.

Meena, G., Raha, S., Selvakumar, P., Satyanarayana, P., & Vats, C. (2025). The Role of AI in Combatting Extremism and Radicalization on Social Media. In *Ethical AI Solutions for Addressing Social Media Influence and Hate Speech* (pp. 63-90). IGI Global Scientific Publishing.

Mesok, E., Naji, N., & Schildknecht, D. (2024). White supremacy and the racial logic of the global preventing and countering violent extremism agenda. *Third World Quarterly*, *45*(11), 1701-1718.

Nwankwo, s. c. Navigating the Ethical Issues with Artificial Intelligence (AI) and Evangelism in the 21st Century Church.

Peters, T. (2023). *The Promise and Peril of AI and IA: New Technology Meets Religion, Theology, and Ethics*. ATF Press.

Rabiu, a. a., merican, a. m. m. n., & al murshid, g. (2025). Ethics in the digital age: Exploring the ethical challenges of technology. *Journal of Information Systems and Digital Technologies*, *7*(1), 29-50.

Rashid, W. (2023). Using Artificial Intelligence to Combat Extremism. *Pakistan Journal of Terrorism Research*, *5*(2).

Rassler, D. (2021). Commentary: Data, AI, and the Future of US Counterterrorism: Building an Action Plan. *CTC Sentinel*, 31-40.

Sajjad, F. W. (2022). Rethinking education to counter violent extremism: a critical review of policy and practice. *Ethics and education*, *17*(1), 59-76.

Scherz, P. (2024). AI as Person, Paradigm, and Structure: Notes toward an Ethics of AI. *Theological Studies*, *85*(1), 124-144.

Schotten, C. H. (2024). Zionism and the war on terror: extinction phobias, anti-Muslim racism, and critical scholarship. *Critical Studies on Terrorism*, *17*(4), 996-1018.

Slattery, J. P., & Green, B. P. (2024). Encountering Artificial Intelligence in the Catholic Tradition. *22*(2), 251-254.

Sonrexa, J., Kelly, L. M., Barton, G., & Ware, A. (2023). Perspectives on violent extremism from development–humanitarian NGO staff in Southeast Asia. *Third World Quarterly*, *44*(1), 170-189.

Tsuria, R., & Tsuria, Y. (2024). Artificial intelligence's understanding of religion: Investigating the moralistic approaches presented by generative artificial intelligence tools. *Religions*, *15*(3), 375.

Tzeng, W. AI and the Network Generation: Theological Discussions with ChatGPT on Historical Theology.

Vaughan, G., Yoo, J., & Szűts-Novak, R. (2025). Wisdom of the Heart: A Contemporary Review of Religion and AI. *Religions*, *16*(7), 834.

Velazquez, A. (2024). Instructing AI through the Exercise of Labor's Solidarity: A Christian Perspective. *J. Christian Legal Thought*, *14*, 25.

Viana, M. T., & da Silva, P. P. d. S. (2021). Preventing extremisms, taming dissidence: Islamic radicalism and black extremism in the US making of CVE. *Critical Studies on Terrorism*, *14*(1), 24-46.

Zaidi, S. M. S., Abbasi, S. N., & Hayat, M. U. (2024). Understanding the rise in violent extremism in Pakistan through the lens of securitization theory. *Asian Journal of Political Science*, 1-25.

Amnesty International. (2022). Myanmar: Facebook's systems promoted violence against Rohingya; Meta owes reparations. Retrieved from https://www.amnesty.org/en/latest/news/2022/09/myanmar-facebooks-systems-promoted-violence-against-rohingya-meta-owes-reparations-new-report/

Amnesty International. (2022). Rohingya survivor asks US regulator to investigate Meta's potential role in Myanmar atrocities. Retrieved from https://www.amnesty.org/en/latest/news/2025/01/united-states-rohingya-survivor-asks-us-regulator-to-investigate-metas-potential-role-in-myanmar-atrocities/

Mozur, P. (2018, October 15). A genocide instigated on Facebook, with posts from Myanmar's military. The New York Times. Retrieved from https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html

United Nations Human Rights Council. (2018). Report of the Independent International Fact-Finding Mission on Myanmar. Retrieved from https://www.ohchr.org/en/hr-bodies/hrc/myanmar-ffm/index