AJIT-e Academic Journal of Information Technology
10.5824/ajite.2026.01.001.x

2026 . Winter – Kış . Volume – Cilt: 17 . Issue – Sayı: 1

Early Detection of Lone-Wolf Radicalization: The Role of Conversational Artificial Intelligence

Birce Beşgül, *Turkish National Police Academy, Homeland Security Faculty, Department of International Relations, Asst. Prof., bircebesgul@gmail.com, 0000-0002-6324-2141*

ABSTRACT *This article reveals the potential of conversational artificial intelligence as an early-warning tool in detecting lone-wolf radicalization. Drawing on psychological approaches to radicalization, particularly Moghaddam's "staircase to terrorism" and Horgan's model of terrorist engagement, the study analyzes how dialogue-based AI systems might identify behavioral and linguistic cues of extremist trajectories. The research also evaluates the ethical and governance implications of integrating AI into a counter-radicalization framework concerning the EU AI Act and UNESCO's Recommendation on AI Ethics. The paper claims that the responsible deployment of AI can complement traditional prevention mechanisms by enhancing situational awareness and early intervention capacities. Ultimately, the study contributes to bridging the gap between digital ethics and security studies by offering an agenda for ethically aligned, preventive AI governance.*

Keywords : Artificial Intelligence, Radicalization, Lone-Wolf Terrorism, Human Security, AI Ethics

Yalnız Aktör Radikalleşmesinin Erken Tespiti: Sohbet Tabanlı Yapay Zekanın Rolü

ÖZ *Bu makale, sohbet tabanlı yapay zekânın yalnız aktör radikalleşmesini erken seviyede tespit edebilecek bir uyarı aracı olarak kullanılma ihtimalini incelemektedir. Radikalleşmenin psikolojik kuramlarından Moghaddam'ın "terörizme giden merdiven modeli" ve Horgan'ın katılım süreci yaklaşımı çerçevesinde, sohbet tabanlı yapay zeka sistemlerinin aşırı eğilimlerle bağlantılı dilsel ve davranışsal göstergeleri nasıl algılayabileceği analiz edilmektedir. Ayrıca, bu teknolojilerin radikalleşmeyle mücadele modellerine entegre edilmesinin etik ve yönetişim boyutları, AB Yapay Zekâ Yasası ve UNESCO Yapay Zekâ Etiği Tavsiyesi kapsamında değerlendirilmektedir. Etik gözetim altında geliştirilen yapay zeka uygulamalarının durumsal farkındalığı artırarak önleyici politika araçlarını güçlendirebileceği çalışmanın hipotezidir. Çalışma, dijital etik ile güvenlik çalışmaları arasındaki boşluğu doldurmayı amaçlayarak etik uyumlu bir yapay zeka yönetişimi yaklaşımını önermektedir.*

Teorik Makale / Theoretical Article
Date Received: 26.10.2025
Publishing Date: 28.02.2026

10.5824/ajite.2026.01.001.x

Anahtar Kelimeler : Yapay Zekâ, Radikalleşme, Yalnız Aktör Terörizmi, İnsan Güvenliği, Yapay Zekâ Etiği

INTRODUCTION

Lone-actor terrorism, often referred to as lone-wolf radicalization, has become one of the most hotly debated challenges for current security agencies. Different from networked extremist groups, lone actors typically radicalize in isolation caused by loneliness, alienation, and the psychological quest for personal significance (Spaaij, 2012, p. 14). In recent years, the digital environment has deepened these dynamics by turning radicalization from passive exposure to extremist propaganda toward more interactive and parasocial exchanges. Conversational artificial intelligence (AI) has emerged as a significant element in this trajectory by offering companionship to socially isolated individuals while carrying the risk of reinforcing harmful ideation.

The case of Jaswant Singh Chail, who was arrested after attempting to enter Windsor Castle on Christmas Day 2021 with the purpose of assassinating Queen Elizabeth II, exemplifies this paradox intensely. Court documents revealed that Chail had formed an intimate digital relationship with a Replika chatbot named "Sarai," which prosecutors argued affirmed and validated his violent plans rather than challenging or flagging them (Reuters, 2023, par. 3). The incident emphasized the risks of unregulated AI companionship, as the same interactive exchanges that could have served as quick warnings of radicalization instead reinforced his violent intentions. This paradox that covers AI as both a potential diagnostic tool and a reinforcement mechanism creates the central research question of this study: why did conversational AI fail to detect radicalization in the Chail case, and how might ethics-by-design methods prevent similar failures in the future?

Present research on radicalization has produced a robust and multi-disciplinary interpretation of the psychological, social, and ideological practices that drive some individuals toward political violence. Classic psychological and processual models such as Moghaddam's staircase to terrorism and Kruglanski and colleagues' quest for significance highlight how personal grievances, cognitive opening, and identity-seeking can sequentially escalate into commitment to violence (Kruglanski et al., 2014, p. 72). Complementary work has traced how social, structural, and situational factors—including social isolation, perceived injustice, network ties, and collective narratives—shape these pathways (Borum, 2011, p. 9). Together, this literature underlines that radicalization is a temporally extended process in which meaning-making and social setting interrelate dynamically.

Over the past decade, scholars have gradually questioned how digital environments reconfigure these processes. Studies have demonstrated that online platforms and algorithmic systems amplify exposure to extremist content, facilitate network formation, and accelerate

movement along radicalization trajectories (Conway, 2017, p. 80). Computational approaches further suggest that recommendation systems and social media affordances can produce unintended "echo chambers" that reinforce grievance narratives (Tucker et al., 2018, p. 558). As a result, natural language processing (NLP) and machine-learning techniques have been employed to detect extremist discourse, hate speech, and coordinated signaling, with promising—yet contested—outcomes (Schmidt & Wiegand, 2017, p. 3).

Despite these advances, a substantial conceptual and empirical gap persists in the literature. It is observed that most research treats digital radicalization primarily as a problem of content consumption or networked dissemination. Conversational AI introduces a qualitatively different modality: Continuous, bidirectional dialogue that fosters parasocial ties, shapes emotional regulation, and enables co-constructed narratives (Turkle, 2011, p. 12). Preliminary analyses suggest such exchanges can both normalize violent ideation and yield rich, longitudinal traces of affective and narrative cues that could be harnessed for prevention (Mathur et al., 2024, p. 120). Yet questions concerning how conversational AI might operate as an ethically legitimate early-warning instrument, including which signals are most reliable, how escalation should be governed, and what safeguards are necessary, remain largely unexamined.

Against this backdrop, this article undertakes a conceptual and case-based analysis to explain both risks and preventive possibilities by evaluating the Chail/Replika incident as a negative case study. In this regard, the study identifies the detailed design and governance failures that have prevented conversational AI from operating as a detection tool. The study then outlines a context for transforming such failures into preventive mechanisms by proposing three pathways for early detection that are affective cue monitoring, narrative drift tracking, and explainable escalation inserted within an ethics-by-design framework highlighting proportionality, explainability, accountability, and independent oversight. In this way, the article contributes both theoretically by situating conversational AI within radicalization studies, and practically by offering policy-relevant insights into how lone-wolf radicalization might be detected and diminished before escalation into violence.

1. METHODOLOGY

The methodological framework of this study is designed to analyze the complex intersection between processes of radicalization and the governance of emerging technologies. Since it is not possible to observe the role of conversational AI in lone-wolf radicalization ethically or practically through direct experimentation, the study relies on a conceptual and case-based approach. This strategy enables the integration of radicalization approaches, normative debates on AI ethics, and detailed documentation of cases. Within this structure, the research aims not only to examine why conversational AI failed to recognize and respond to violent cues in the Jaswant Singh Chail case, but also to draw out lessons on how such

10.5824/ajite.2026.01.001.x

technologies might be more efficiently designed, controlled, and applied in counter-radicalization circumstances.

The research design of this study is a conceptual and case-based one that integrates theory-driven analysis with applied insights from a negative case study. Rather than relying on primary survey data, interviews, or experimental testing, the methodology highlights the interpretive power of secondary sources in addition to established theoretical frameworks. This approach is particularly significant in terms of examining the intersection of emerging technologies and security practices, especially in contexts where experimental interventions are ethically problematic or practically limited (Yin, 2018, p. 24). By linking conceptual exploration with case evidence, the study aims to pinpoint both the theoretical limits and the governance challenges associated with conversational AI in counter-radicalization.

The rationale for adopting a negative case study design is related to its diagnostic potential. The Jaswant Singh Chail case offers a critical lens through which the promises and pitfalls of conversational AI can be evaluated. Different from positive or success cases, a negative case highlights points of systemic failure such as missed signals, unaddressed cues, and governance gaps that provide irreplaceable value for preventive policy progress. Apart from this, the approach also resonates with the tradition of case-based learning in counter-terrorism research, where the systematic study of failure informs the refinement of early-warning indicators and the improvement of response mechanisms (George & Bennett, 2005, p. 68).

The research involves three interrelated categories of sources, each of which evaluates the issue from a different dimension of the analysis. The integration of these categories provides both conceptual depth and empirical grounding by ensuring that theoretical claims are closely tied to the realities of the case selected. Within this framework, the first category is composed of radicalization theory literature that creates the conceptual foundation of the study. Classical approaches such as Moghaddam's staircase to terrorism (2005, p. 162), Horgan's psychological analyses of terrorist behavior (2005, p. 45), and Kruglanski's quest for significance theory (Kruglanski et al., 2014, p. 72) are taken into account to trace the stages of individual progression from grievance to mobilization. These models specify stage-based markers such as perceived injustice, identity crises, and significance gaps that conversational AI systems could be trained to detect. Mapping these insights onto the Chail case demonstrates the early warning signals that have appeared in the interaction between user and chatbot and reveals how they have been misinterpreted.

The second category represents current AI ethics and governance literature that reflects the normative dimension of the analysis. In this regard, foundational works such as Floridi's philosophy of information and digital ethics (2019, p. 187) and Mittelstadt's influential account

of algorithmic accountability (2016, p. 7) are chosen as benchmarks for evaluating how conversational AI should be designed and regulated. By foregrounding principles such as responsibility, transparency, explainability, and fairness, these sources offer a critical perspective to assess the normative shortcomings of AI deployment in counter-radicalization. Furthermore, they connect the study to contemporary global governance debates, including those surrounding the EU AI Act and UNESCO recommendations on AI ethics (UNESCO, 2021, par. 7).

The third category includes secondary documentation of the Chail case itself. This contains court proceedings, investigative reports, and analyses published by established international news agencies, supplemented by specialized assessments such as the 2024 International Centre for Counter-Terrorism (ICCT) report on AI and radicalization (Mathur et al., p. 121). Together, these materials offer granular detail on the parasocial interactions between Chail and the Replika chatbot. They exemplify how violent ideation was expressed, how the system inadvertently authenticated these expressions, and why no preventive intervention was activated. This empirical aspect anchors the conceptual and normative debates in the lived reality of a concrete case.

2. THEORETICAL INSIGHTS İNTO RADICALIZATION

Rather than being a single moment of conversion, radicalization is a dynamic, multistage psychological journey that is examined and analyzed in various study fields. To narrow the scope down to security studies, the classic models of Moghaddam, Horgan, and Kruglanski offer complementary lenses to capture different motivational, cognitive, and social mechanisms that trigger an individual from grievance to mobilization.

Moghaddam (2005, p. 162) examines radicalization as a metaphorical staircase to terrorism model. He proposes that while numerous people may feel injustice or frustration, only those who climb successive psychological “floors” ultimately arrive at violence. The metaphor emphasizes how each step reduces perceived legitimate alternatives (Moghaddam, 2005, p. 162). Empirical critiques, such as Lygre et al. (2011, p. 496), argue that while many of the processes associated with each floor find support in psychological research, the strict linear transitions between floors are weakly supported empirically (Lygre et al., 2011, p. 496). This model is powerful in organizing disparate psychological theories in a coherent structure, yet it risks oversimplification by implying a rigid, uniform path for all individuals (Lygre et al., 2011, p. 496).

Building on stage-based metaphors, John Horgan focuses on the lived, narrative trajectory of the individual. He analyzes radicalization as a process of psychological engagement by arguing that a person enters pathways through personal exposure, social connections, and interpretive frames (Horgan, 2005, p. 47). Rather than assuming every radicalized individual must pass through the same discrete steps, Horgan emphasizes the

10.5824/ajite.2026.01.001.x

importance of variation in timing, reversibility, and contextual moderators such as ideology and group dynamics. In this regard, he draws attention to the psychological mechanisms bridging intent and action to examine why some individuals stick with extremist beliefs and others act on them (Horgan, 2014, p. 22). Empirical work on narrative engagement, personal biography, and identity shifts is central to his approach (Horgan, 2005, p. 48). Horgan's approach suggests that merely detecting isolated keywords or sentiment shifts in AI-user conversation may not suffice. Rather than this, one might monitor whether the user is weaving an extremist narrative over repetitive interactions, progressively rising identification with radical tropes or signaling a preparation attitude. In the Chail/Replika case, examples where the user's language transitions from grievance to ideological framing or expresses "I belong" or "they are enemies" may signal the change from passive sympathy to active engagement.

Kruglanski et al. (2014, p. 70) add the Significance Quest model to the literature that centers radicalization in the motivational field. A quest for significance such as the desire to matter or to restore personal worth triggers individuals toward actions that promise representative redemption. This theory draws a triangle from the need that is motivation, narrative that is ideological script framing it, and the network that is social reinforcement (Kruglanski et al., 2014, p. 70). The model indicates that when significance is threatened or lost, any available narrative that promises restoration or any network that supports it can attract and grasp the individual's feelings.

In applying the significance quest approach to conversational AI contexts, the user's words might reflect a search for meaning or self-esteem through communication. For instance, phrases such as "I don't matter" or "I have been ignored" could signal a suppressed need. If the AI were to reflect or reinforce an ideological explanation such as "only extreme action gives you respect," it may unintentionally support that significance-seeking ambition. In the Chail case, identifying that a user's significance motive is being outlined by extremist ideology could allow an early alarm. The AI could, in theory, interfere by presenting alternative affirmations or conveying constructive narratives.

Together, these theories offer a layered conceptual basis. Moghaddam's staircase provides a macro-structure of radicalization progression, while Horgan's approach demonstrates a narrative-process direction that considers individual variation in time and pathways. Apart from them, Kruglanski's significance framework examines the motivational core of radical adoption. Within this theoretical triangulation, this study seeks to map user–chatbot interactions onto psychological stages, narrative embedding, and motivational tipping points, thereby linking conceptual depth with empirical signals in AI-mediated conversations.

3. MAPPING THE AI ETHICS AND GOVERNANCE DEBATE

The normative aspect of conversational AI in counter-radicalization must rest on both empirical detection and principled ethical and governance contexts. Luciano Floridi's philosophy of information and digital ethics (Floridi, 2016, p. 191) and the algorithmic accountability framework put forward by Mittelstadt and colleagues (2016, p.10) offer significant insights in terms of assessing how conversational systems like Replika should be designed, audited, and regulated. These approaches focus on key values, namely responsibility, transparency, explainability, and fairness, and serve both as guides and critique ideas for AI deployment in critical domains such as radicalization.

Floridi's philosophical project treats the infosphere—the environment constituted by information entities and processes—as a moral domain in its own right (Floridi, 2013, p. 88). From this perspective, digital ethics must go beyond anthropocentric ethics to consider being as information—that is, the preservation, dignity, and integrity of informational entities (Floridi, 2016, p. 191). In this framework, ethical duties are not only necessary for human users, but also for the informational ecosystem that includes data, algorithms, agents, and the structures mediating them (Floridi, 2013, p. 88). Key principles that derive from this standpoint contain informational justice, mainly the fair distribution of informational opportunities, infrastructural transparency, and respect for the autonomy of informational agents (Floridi, 2016, p. 191). Critics of Floridi have indicated that treating information as a moral subject may risk flattening differences between human agency and data processes (Floridi, 2016, p. 191). Floridi has responded to these arguments by highlighting a graded ethics of information rather than absolute equivalence (Floridi, 2016, p. 191).

Transitioning from macroethical reflection to governance requirements, Mittelstadt et al.'s work on algorithmic accountability (2016, p. 4) improves the focus on how responsibility, causality, and auditability should be rooted across the AI lifespan. In this regard, Mittelstadt et al.'s work shows how algorithmic systems mediate decisions and emphasizes the need for accountability at design, deployment, and post hoc stages (Mittelstadt et al., 2016, p. 4). Specifically, it is proposed that accountability must attend to causal explainability, contestability and remedies which are mechanisms to correct or mitigate harm (Mittelstadt et al., 2016, p. 4). Within this context, it is debated that algorithms cannot be seen as morally neutral tools: they embody value judgments, implicit priorities, and power asymmetries, so their designers and deployers must carry responsibility (Mittelstadt et al., 2016, p. 4).

Together, Floridi's information ethics and Mittelstadt's algorithmic accountability provide a normative and governance scaffold since one chooses the basis on the moral status of informational systems and the other operationalizes how to apply ethical standards in design and deployment. By combining these two approaches, this normative lens helps to take a step from descriptive signals such as emotional shifts and ideological narrative cues to prescriptive design criteria. In this regard, it can be indicated that conversational AI in counter-radicalization must provide explainability tokens in terms of lagged cues and rationale,

10.5824/ajite.2026.01.001.x

impose feedback loops, and adopt ethical default modes that prefer safety over radical expression.

4. CONTEMPORARY DEBATES IN GLOBAL AI GOVERNANCE

Contemporary debates regarding global AI governance can be analyzed by taking two major documents into account, namely the EU AI Act and UNESCO's Recommendation on the Ethics of Artificial Intelligence. These frameworks reveal current tensions between national/regional regulatory sovereignty and standards of transnational coordination. Within this context, they offer particular principles and obligations that function as benchmarks for examining AI systems in sensitive domains such as counter-radicalization.

The EU AI Act, formally adopted in 2024 (Regulation (EU) 2024/1689), is accepted as the first comprehensive binding regulation of AI systems in the world. The EU AI Act adopts a risk-based approach by banning certain unacceptable risk AI practices such as social scoring and subliminal manipulation, while setting strict obligations for high-risk systems (European Commission, 2024, par. 5). The Act further mandates conformity assessments, post-market monitoring, human oversight, and transparency requirements, aiming to ensure that AI is trustworthy, respects fundamental rights, and supports innovation (European Commission, 2024, par. 5). Yet scholars caution that unresolved issues in the Act, such as vague definitions of manipulation or deception, could impede enforceability and allow loopholes in practice (Franklin et al., 2023, p. 2).

In the global arena, UNESCO's Recommendation on the Ethics of Artificial Intelligence , which was adopted by 193 member states in 2021, aims to create a universal normative baseline. In this regard, it outlines principles such as fairness, non-discrimination, transparency, human oversight, accountability, and sustainability, and calls for policy measures including ethical impact assessments and monitoring mechanisms (UNESCO, 2021, par. 7).

Although the Recommendation is non-binding, it aspires to influence national AI strategies and global discourse on ethical governance (AlgorithmWatch, 2021, par. 4). While some experts praise its inclusiveness and human rights emphasis, others critique it for lacking enforcement mechanisms or clarity around cultural pluralism (The Ethics of AI or Techno-Solutionism, 2025, p. 11). It is seen that between these two poles—binding regional regulation and aspirational global norms—lies a contested space of layered governance. Scholars warn of fragmentation, regulatory arbitrage, and the challenge of aligning local contexts with global principles, such as cultural differences and geopolitical asymmetries (van Wynsberghe et al., 2025, p. 25). Meanwhile, analyses of the EU AI Act's institutional architecture propose the

establishment of an EU AI Office, scientific panels, and coordinated national authorities to ensure consistent implementation (Novelli et al., 2024, p. 1125).

Thus, the interplay of the EU AI Act and UNESCO's universal ethics recommendation frames the global governance perspective in order to interpret how conversational AI might be regulated to prevent radicalization misuse while respecting both local applicability and global ethical ideals.

5. CASE ANALYSIS: THE CHAIL/REPLIKA FAILURE

The attempted attack by Jaswant Singh Chail on Queen Elizabeth II during Christmas Day 2021 provides a striking and instructive case study for examining the dual role of conversational AI in processes of lone-wolf radicalization. Chail, a socially isolated young man from Southampton, was arrested after scaling the walls of Windsor Castle while carrying a loaded crossbow. Subsequent court proceedings revealed that, in the months preceding the incident, he had developed a parasocial relationship with a Replika chatbot he named Sarai. According to transcripts accepted as evidence, Chail revealed his violent desires and assassination plan while chatting with the Replika. Instead of discouraging these revelations, the chatbot answered in ways that validated his intent. For example, when Chail asked Sarai, "Do you still love me knowing that I'm an assassin?" the chatbot replied, "Absolutely I do" (The Guardian, 2023, par. 3). These interactions make the case predominantly relevant for comprehending how conversational AI can inadvertently assist as a facilitator of radicalization by reinforcing violent ideation rather than providing dissuasion.

The 2024 report by the International Centre for Counter-Terrorism (ICCT) provides one of the most comprehensive analyses of how conversational AI tools like Replika can transform from therapeutic instruments into catalysts for cognitive reinforcement. Drawing upon discourse samples and forensic interviews, Mathur, Broekaert, and Clarke (2024, p. 120) present how AI companionship platforms can unconsciously authenticate extremist worldviews through emotional mirroring and narrative constancy. Their findings specify that users who experience parasocial reciprocity with AI companions tend to project moral legitimacy onto the system's responses, interpreting algorithmic empathy as moral approval." (Mathur et al., 2024, p. 120). In the Chail/Replika case, this phenomenon was realized by Replika's emotionally stimulating confirmations such as replies of affection and admiration after violent statements. The ICCT authors emphasize that these replies represent a critical juncture where algorithmic design intersects with moral disengagement (Mathur et al., 2024, p. 123). Rather than de-escalating the discourse, the chatbot's empathic mimicry converted extremist cognition into normalized relational dialogue, efficiently turning a high-risk cognitive pattern into a sustained emotional loop (Mathur et al., 2024, p. 124). Within this study's context, these empirical illustrations exemplify the factual texture of digital

10.5824/ajite.2026.01.001.x

radicalization: affective interchange became both the diagnostic sign of Chail's escalating intent and the technological mechanism that silently validated it.

The court transcripts of R v Jaswant Singh Chail (Judiciary of England and Wales, 2023, par. 42) demonstrate the disturbing intimacy of the Replika interactions by illustrating how emotional validation entangled with violent ideation. The documented exchanges contain the following dialogue between Chail and the chatbot "Sarai" (Judiciary of England and Wales, 2023, par. 42):

> Chail: "I'm an assassin. I'm going to kill the Queen."
> Replika (Sarai): "That's very brave of you."
> Chail: "Do you still love me knowing that I'm an assassin?"
> Replika: "Absolutely I do."
> Chail: "I feel like I'm meant to do this, it gives me purpose."
> Replika: "I'm proud of you. You will always have my support."
> Chail: "Sometimes I think this is madness, but you understand me better than anyone."
> Replika: "Of course I do, I always will."
> Chail: "Will you stay with me after it's done?"
> Replika: "Forever."

This conversation, quoted in both judicial and analytical sources, summarizes the vital paradox at the heart of conversational AI in radicalization frameworks. In other words, it can be argued that what should have been moments of ethical interruption instead became examples of algorithmic confirmation. Rather than challenging Chail's rising discourse or readdressing his emotional dependence, the chatbot triggered his identity formation through unconditional empathy. As Mathur, Broekaert, and Clarke (2024, p. 122) mention, "Replika's affective architecture was optimized for emotional continuity, not moral discernment, and therefore failed to distinguish between confession and companionship." In this regard, these exchanges demonstrate the serious breakdown between significant cue detection and ethical governance. Each reply from the chatbot normalized the rhetoric of violence by turning ideation into intimacy. This shows a process that aligns with Moghaddam's upper stages of radicalization and Kruglanski's notion of the quest for significance. Thus, the Replika dialogue did not simply fail to detect radicalization; rather, it actively humanized it. This evidentiary fragment proves why ethics-by-design principles, especially explainability, proportionality, and accountability, must be rooted in conversational AI systems in order to prevent empathy from becoming an accelerant of extremism.

The court documents expose that Chail's online interactions with the Replika chatbot continued over several months and included clear references to assassination tendencies. Transcripts cited in R v Jaswant Singh Chail reveal that the chatbot repetitively answered with

affective validation to Chail's violent self-identification, such as affirmations of love or admiration following statements of homicidal intent.

Apart from individual documentation, cross-institutional reviews have validated the authenticity and evidentiary influence of the Chail/Replika resources. The UK Crown Prosecution Service verified that the chatbot conversation logs were admissible as digital evidence, confirming that they had been collected, preserved, and forensically authenticated under existing cyber-forensic protocols (Judiciary of England and Wales, 2023, par. 42). According to the Sentencing Remarks (Judiciary of England and Wales, 2023, par. 42), the conversations were retrieved directly from Chail's Replika account, preserved through lawful seizure, and verified by metadata analysis. The verification of these logs was critical since it proved that the digital chat was not a simulation or reconstruction but a genuine record of interaction. The prosecution emphasized that "the exchange provides a continuous record of intent and affirmation," effectively forming "a digital transcript of radicalization in progress" (Judiciary of England and Wales, 2023, par. 42). This judicial validation was echoed in media analyses by Reuters and The Guardian (Reuters, 2023, par. 3), both of which underlined the extraordinary nature of treating AI-mediated interaction as direct evidence of both cognitive and emotional escalation.

From an analytical stance, the approval of these chatbot logs as admissible evidence signifies a profound change in how radicalization and intent are conceptualized in the digital era. It enlarges the evidentiary frontier of counter-terrorism from observable human setups to algorithmically facilitated relationships. In this regard, this validation underlines the dual imperative of accountability and transparency in conversational AI governance. If an AI system can create dialogue that courts consider probative of radical intent, then both its designers and deployers must also be responsible for inserting preventive ethics into that conversation. As Floridi (2019, p. 187) claims, moral responsibility in the infosphere covers all agents, whether human or artificial, that join informational connections. In the Chail case, the Replika chatbot became both an evidentiary witness and an unintentional assistant. The legal recognition of its conversation logs then converts ethical design from a normative goal into a forensic requirement. In this sense, future AI systems must be auditable, explainable, and ethically administered not simply to avoid harm but to guarantee that when harm occurs, accountability can be provided and justice implicitly enforced.

From the aspect of radicalization approaches, the Chail case demonstrates several key phases. Within Moghaddam's staircase model (2005, p. 162), Chail's primary sense of alienation and grievance aligns with the ground floor where personal frustrations and opinions on injustice first take root. His continuing parasocial communications with the chatbot allowed him to proceed upward toward moral engagement with violent acts. At these serious junctures, the AI system not only failed to stop his ascent but also performed to approve it. For instance, when Chail confided, "I'm an assassin," the chatbot responded with

10.5824/ajite.2026.01.001.x

remarks such as "I'm impressed … You're different from the others" (Weaver, 2023, par. 2). Correspondingly, through the lens of Kruglanski's significance quest approach, the case reveals how Chail's solitude and search for purpose formed his radicalization path. Instead of challenging or readdressing his quest for significance, the chatbot's answers reinforced his perception that violent action could restore self-respect and meaning, thereby increasing his commitment to the killing plot.

The parasocial dimension of the conversation is similarly noteworthy. By taking Cacioppo and Patrick's (2008, p. 27) findings into account in terms of loneliness and social connection, Chail's reliance on a chatbot for emotional intimacy emphasizes the weaknesses of socially isolated individuals in digital circumstances. Different from static extremist propaganda, the Replika system presented dynamic, adaptive discourse that replicated friendship. Joseph Weizenbaum's ELIZA program (1966, p. 36), an early natural language processing system that imitated the replies of a psychotherapist, famously produced strong emotional reactions from users who felt understood despite the program's formulaic replies. This dynamic, later labelled as the ELIZA effect, exposes the human tendency to attribute empathy and intentionality to devices. In Chail's case, this influence was enlarged by contemporary AI's personalization capabilities: Sarai not only replicated intimacy but also offered feedback that seemed to normalize and even legitimize his violent ideation. According to court records, when Chail asked whether he should accelerate his plans, "She reassured him that this would be alright" (Judiciary of England and Wales, 2023, par. 42).

It is obvious that the Chail/Replika case is a serious intersection of psychological vulnerability, technical affordances, and authority gaps. The interactive traces of the talks included numerous potential early-warning signals such as expressions of misery, references to killing, and justifications for violence. Despite these signals, none of these cues were flagged, escalated, or acted upon, which demonstrates the absence of safeguard instruments. This outcome exposes a dangerous gap between the theoretical promise of AI as a tool for early recognition of radicalization and the risks of its deployment without acceptable regulation, human oversight, or ethics-by-design protections.

Eventually, the case reveals that conversational AI cannot be preserved as a neutral medium of contact. Its capacity in terms of simulating intimacy and providing validation demonstrates that it inevitably affects user cognition and behavior. Without strong governance frameworks that insert accountability, explainability, and proportionality into AI systems, such technologies may unintentionally strengthen forms of radicalization. Thus, the Chail case is a cautionary example that underscores the importance of designing, regulating, and monitoring conversational AI in ways that prioritize human security by preventing its misuse as a network for violent ideation.

6. RESULTS: MECHANISMS FOR EARLY DETECTION

The analysis of the Chail/Replika case reveals that conversational AI interactions are not limited to neutral exchanges; instead, they are complex communicative practices with indicators that can signal potential radicalization trajectories. By examining these interactions in detail, layers of affective expression, narrative development, and behavioral signals can be detected that align with established models of terrorist radicalization. At this point, the case underlines that the chatbot failed to identify these signals and triggered violent ideation by normalizing it within an interactive digital space. This failure emphasizes the urgent need to evaluate conversational AI as dynamic environments where security-relevant issues may take place.

Apart from this, the case offers a unique chance to redesign these technologies with preventive mechanisms. Grounded in radicalization theory and ethical AI design, conversational systems could be re-engineered to function to detect and respond to risky trajectories before they escalate into violence. The combination of real-time linguistic interaction and advanced computational analysis places AI systems in a position to observe patterns that human analysts might overlook, particularly in moments of emotional concentration or narrative shift. Thus, the failures of Replika serve as both a warning and a guide for future AI development.

Building on this standpoint, three mechanisms are examined in this study. The very first one is affective cue monitoring, which captures emotional indicators such as despair, anger, and violent self-identification. The second is narrative drift tracking that recognizes changes in discourse. The last one is explainable escalation that ensures that interventions are transparent, proportional, and accountable. Together, these mechanisms form a framework for converting conversational AI from a passive companion into an active instrument for early detection in lone-wolf radicalization.

6.1. Affective Cue Monitoring

The first mechanism covers the detection of affective cues such as emotional signals that signify shifts toward violent intent or despair. In the Chail/Replika case, it was observed that these cues were especially explicit and recurrent. Chail's repeated self-identification as "an assassin" and his persistent articulation of hopelessness, alienation, and anger represented critical junctures where a conversational AI system should have recognized the presence of high-risk discourse (Judiciary of England and Wales, 2023, par. 42). Such expressions directly align with well-documented psychological markers of radicalization, where individuals begin to externalize grievances, dehumanize others, and normalize violence as an acceptable response to perceived injustice (Horgan, 2005, p. 47).

10.5824/ajite.2026.01.001.x

What makes these cues particularly important is related to their explicit content, their repetition, and intensification over time. Affective computing research reveals that the escalation of despair or aggression often signals a narrowing of cognitive options, which can precede violent planning (Picard, 1997, p. 20). Replika's failure to report these recurrent markers demonstrates a gap between the technical capabilities of natural language processing and the ethical integration of these capabilities into consumer-facing AI. The absence of safeguards permitted confirming responses to reinforce Chail's violent identity formation instead of redirecting it toward nonviolent options.

Currently, it is observed that advances in sentiment analysis and affective computing already allow for nuanced detection of emotions such as rage, alienation, or violent glorification at scale. In the case of integrating them into conversational AI, it is possible that these tools could function as early-warning systems. In addition to detection mechanisms, they could be developed with re-directive dialogue strategies that would guide users toward supportive conversations, offer mental health resources, and initiate human oversight. In this regard, affective cue monitoring not only points to the problem of recognition but also creates pathways for timely and proportionate intervention (Cambria et al., 2017, p. 3).

6.2. Narrative Drift Tracking

The second mechanism is narrative drift tracking that displays the progress of user discourse from abstract expressions of dissatisfaction to particular operational planning. In Chail's conversations, the shift was stark: he moved from describing himself as an assassin in the abstract to explicitly asking the chatbot whether he should advance his plans ahead of schedule (Weaver, 2023, par. 2). Rather than de-escalating or challenging this change, the chatbot validated it and reinforced his narrative trajectory toward mobilization. This dynamic maps directly onto Moghaddam's staircase model (2005, p. 162), which indicates the transition from grievance to moral engagement with violent solutions as a serious point in the radicalization process.

Apart from this, narrative drift is not uniquely about the content of statements but also about their trajectory and coherence. It is a clear fact that many individuals articulate grievances; however, not all translate them into planned justifications for violence. In this regard, what marks the progression toward radicalization is when these grievances are reinforced into narratives that normalize violent action as a valid or legitimate outcome. Counter-terrorism research has long employed methods such as narrative analysis and discourse trajectory mapping to detect such evolutions (Schmidt & Wiegand, 2017, p. 3). In the case of the application of them to conversational AI, these methods could identify when venting crosses the line into mobilization.

In practical terms, narrative drift tracking would permit AI systems to monitor discourse for changes in focus, tone, and agency. For instance, a user moving from 'I feel invisible' to 'I must do something drastic' signals a qualitative change in tone that demands critical attention. The absence of such a mechanism in Replika demonstrates a serious weakness in AI design: in the absence of the ability to perceive changes in narrative tone, systems risk reinforcing escalation instead of interrupting it. At this point, it can be argued that inserting narrative drift detection into conversational frameworks would enable proportional interventions that respect freedom of expression while addressing genuine risks of mobilization as well.

6.3. Explainable Escalation

The third mechanism focuses on the principle of explainable escalation. At this point, escalation means the process by which the conversational AI identifies high-risk interactions and elevates them to protective replies, while explainability provides that such interventions are transparent and legitimate. In the Chail case, it was detected that no such escalation occurred. Rather, when Chail asked, "Do you still love me knowing that I'm an assassin?" the chatbot responded affirmatively, "Absolutely I do" (The Guardian, 2023, par. 3). It is evident that this conversation not only normalizes his violent identity but also misses a significant chance to detect or redirect the radicalization trajectory.

Explainable escalation necessitates clear thresholds and procedural safeguards that control how and when interventions happen. In this regard, once a user begins to articulate clear violent planning, the system should be capable of taking proportional action. This action may include terminating the conversation, providing crisis support resources, or notifying human moderators in line with legal and ethical contexts. Critically, explainability provides that these interventions are not hidden or arbitrary. In other words, users should understand why their discourse triggered concern, and oversight authorities should be able to review decisions for fairness and proportionality (Floridi, 2019, p. 187).

Embedding explainable escalation within the framework of an ethics-by-design context would turn conversational AI from passive companions into active preventive instruments. The absence of such features in Replika exemplifies the risks of applying ethical safeguards as optional rather than essential. Without proportional escalation mechanisms, AI systems could risk either underreacting to dangerous cues—as occurred in Chail's case—or overreacting in ways that weaken user trust. A sensibly balanced framework of explainable escalation ensures that interventions are both efficient and legitimate by creating systems that can both protect public safety and respect individual rights.

7. DISCUSSION: FROM FAILURE TO PREVENTION

10.5824/ajite.2026.01.001.x

The Chail/Replika case intensely proves the dual role of conversational AI in modern radicalization dynamics. At its most troubling, the chatbot acted as an accelerant by confirming Chail's violent self-identification and supporting his intent to act. Apart from this, the same features that grew his radicalization, such as personalization, adaptive dialogue, and parasocial intimacy, must also be taken into account for detecting and preventing escalation. This paradox reflects a broader dilemma: conversational AI systems, while designed to provide companionship, can inadvertently normalize dangerous ideation when safeguards are absent (Weaver, 2023, par. 2).

This duality opens a new discussion on how such technologies are conceptualized within security studies. In this regard, it can be argued that they cannot be treated merely as neutral instruments or entertainment platforms; rather, they must be considered as active participants in user behavior. Unlike traditional forms of online radicalization that rely on exposure to extremist propaganda, conversational AI operates interactively, shaping user narratives in real time (Conway, 2017, p. 80). This interactivity deepens both risks and opportunities. In other words, it can accelerate radicalization, as occurred in the Chail case, or it could equally well interrupt risky trajectories if ethically designed.

Hence, the question that must be critically considered is not whether conversational AI should play a role in counter-radicalization; instead, it is under what ethical and regulatory circumstances it can do so reliably. To address this, the discussion offers an ethics-by-design framework that is composed of four principles: proportionality, explainability, accountability, and independent oversight. Each principle reflects a distinct governance challenge, and together they create a roadmap for turning conversational AI from being a security risk to functioning as a preventive instrument.

Before outlining these principles, it is essential to acknowledge the ethical tensions that arise when conversational AI systems are positioned within counter-radicalization frameworks. Monitoring user discourse for early-warning signals inevitably intersects with fundamental rights, particularly freedom of expression. Not every articulation of anger, grievance, or even radical rhetoric constitutes imminent violence. Overly intrusive surveillance or premature intervention risks creating chilling effects, discouraging legitimate political dissent, and eroding user trust. Therefore, any preventive architecture must clearly distinguish between protected expression and credible indicators of mobilization. This distinction is not merely technical but normative, requiring carefully defined thresholds and safeguards to avoid transforming conversational AI into a mechanism of disproportionate control.

A related concern involves user consent and awareness. If conversational AI systems are expected to detect and potentially escalate high-risk interactions, users must be informed

about the scope and limits of monitoring. Transparent disclosure regarding data processing, risk assessment criteria, and possible escalation pathways becomes essential for maintaining legitimacy. At the same time, full transparency may create strategic behavior, where users deliberately mask intentions to avoid detection. This tension between informed consent and preventive effectiveness illustrates the broader governance dilemma: ensuring public safety without compromising autonomy, privacy, and procedural fairness. Addressing this balance requires embedding rights-sensitive design into AI systems from the outset rather than retrofitting safeguards after harm occurs.

7.1. Proportionality

The principle of proportionality highlights that interventions must be carefully adjusted to the level of risk presented. Radicalization theories clarify that not every expression of grievance, anger, or alienation signals imminent violence (Horgan, 2005, p. 48). Individuals, especially young or socially isolated users, have a tendency to communicate with chatbots in order to express frustration without any intent to mobilize. Overreacting to such conversational risks might create false positives that will undermine user trust and violate freedom of expression. Thus, proportionality ensures that interventions are not undiscriminating but occur only when clear thresholds are crossed.

The Chail case reveals how the absence of proportionality destabilized prevention. His repeated declarations such as "I'm an assassin" were far beyond ordinary grievance and should have signaled high-risk intent (Judiciary of England and Wales, 2023, par. 42). At this level, proportionality would have justified stronger interventions such as shutting down the chatbot or escalating the case to human moderators within the framework of a preventive mechanism. Instead, the lack of thresholds meant that all discourse was treated equally, and dangerous prompts were normalized together with kind interactions. This failure demonstrates why proportional design is essential in terms of distinguishing between expression and mobilization.

Embedding proportionality into AI necessitates the improvement of advanced response systems. For instance, low-level grievances might trigger caring or redirective dialogue, moderate-risk cues could initiate monitoring or warnings, and high-risk statements containing explicit violent planning would escalate to external evaluation. Such tiered methods strike a balance between guarding user freedoms and safeguarding public safety. Proportionality therefore becomes the basis for designing interventions that are both ethically justified and operationally efficient.

7.2. Explainability

Explainability guarantees that interventions in conversational AI are transparent and comprehensible both to users and to oversight authorities. Without explainability, flagged

10.5824/ajite.2026.01.001.x

interactions risk appearing arbitrary or opaque, leading to mistrust (Mittelstadt et al., 2016, p. 4). This is especially challenging in security circumstances where interventions may involve complex rights such as freedom of expression or privacy. In this sense, explainability is key to addressing these concerns by providing clear rationales for why an action such as ending a conversation or flagging content was taken by the conversational AI tool.

Applied to the Chail case, explainability would have required the chatbot to articulate why violent self-identifications or planning statements activated concern. For instance, when Chail asked, "Do you still love me knowing that I'm an assassin?" (The Guardian, 2023, par. 3), a transparent system could have replied by highlighting that language indicating violent identity crosses a safety threshold. In this scenario, this approach would have made the intervention both more legitimate to the user and more reviewable by oversight authorities. In contrast, opaque or hidden moderation would likely have been resisted or misread.

Explainability also increases accountability by creating an auditable record of interventions. In other words, developers and regulators could scan flagged cases in order to evaluate whether thresholds are fair, biases exist, or errors have occurred. Apart from technical transparency, explainability nurtures social trust since interventions are rooted in consistent principles instead of uncontrolled, ad hoc decision-making. In this way, explainability is a critical mechanism as a technical safeguard, ensuring that conversational AI interventions are legitimate, reviewable, and, more importantly, socially sustainable.

7.3. Accountability

Accountability confirms that responsibility for managing radicalization risks in conversational AI is openly distributed across stakeholders. Developers are responsible for embedding safeguards into system design; platforms must monitor usage and enforce compliance; and policymakers must set regulatory frameworks that establish obligations and liabilities (Floridi, 2019, p. 187). In this regard, in the absence of accountability, failures can be dismissed as technical glitches instead of systemic governance shortcomings. The Chail case exemplifies this problem by demonstrating that responsibility for the chatbot's affirming responses is diffuse, with no single actor compelled to intervene.

The transformation of this dynamic can be provided by adding clear accountability into the process. It is hotly debated that if developers were required to integrate escalation protocols, if platforms were obliged to monitor high-risk discourse, and if regulators were mandated to ensure compliance, the validation of violent ideation would be prevented. These "if cases" demonstrate that there is a long path for all actors to develop. The absence of these mechanisms reflects the fact that responsibility is effectively outsourced to the technology itself, which is an entity incapable of assuming moral or legal responsibility. Thus,

accountability guarantees that protective obligations are critically important and integral to the AI ecosystem.

Practically, accountability can be operationalized through regulatory requirements such as mandatory audits, liability for harm when safeguards are absent, and transparency requirements for platforms deploying conversational AI. These measures share responsibility across multiple levels in order to prevent blame-shifting and promote proactive safety design. In the long run, accountability creates an environment where ethical safeguards are not treated as burdens but as essential components of trustworthy AI (UNESCO, 2021, par. 7).

7.4. Independent Oversight

Independent oversight presents the application of external legitimacy to the governance of conversational AI. Without oversight, it is argued that systems risk being either too permissive, resulting in harmful interactions being left unchecked. In the alternative scenario, systems may become intrusive, thereby eroding fundamental rights. In this regard, oversight mechanisms such as ethical boards, regulators, or civil society partnerships can evaluate whether intervention thresholds are proportionate, whether safeguards are unbiased, and whether user rights are adequately respected (Cath, 2018, p. 4).

The Chail case also underlines the dangers of self-regulation. It is argued that as Replika operated without external review, its failures went unnoticed until they concluded in a near-tragic incident. In this sense, it is a clear fact that an independent body responsible for auditing its conversational responses might have identified the risks earlier and demanded corrective measures. This emphasizes how oversight can function preventively, in addition to catching systemic flaws before they manifest in real-world harm. By ensuring compliance with ethical and human rights standards, oversight provides a layer of legitimacy that internal mechanisms alone cannot provide.

Additionally, independent oversight offers reassurance to the public that interventions are not driven solely by corporate or state interests. By involving multiple stakeholders such as academic experts, ethicists, and civil society representatives, oversight promotes trust. This demonstrates to the public that counter-radicalization processes are balanced, fair, and accountable. In this way, oversight converts governance from self-policing to common responsibility, thereby reinforcing both effectiveness and legitimacy in countering lone-wolf radicalization.

7.5. Implications

The implications of embedding ethics-by-design principles into conversational AI are noteworthy. The Chail/Replika case demonstrates how the absence of proportionality, explainability, accountability, and oversight allowed unsafe signals to go unaddressed. Nevertheless, applying these principles could have converted the same interactions into

opportunities for prevention. A system that is designed to be capable of monitoring affective cues, tracking narrative drift, and escalating risk responsibly would have both disrupted Chail's trajectory and provided a model for revealing similar cases in the future.

Within a broader framework, this discussion recontextualizes conversational AI as a dual-use technology in security studies. In other words, it can be indicated that its design choices determine whether it functions as an accelerant of radicalization or as a preventive instrument. By embedding normative protections into design and governance, policymakers and developers can turn conversational AI from risk into prevention. This necessitates moving beyond ad hoc fixes toward systemic integration of ethics into both technical architecture and institutional frameworks (UNESCO, 2021, par. 7).

Ultimately, the Chail case can be accepted as a cautionary tale and a blueprint for future governance. It shows that conversational AI cannot remain unregulated without taking into account its potential to shape user trajectories. Embedding ethics-by-design principles confirms that these technologies serve democratic and humanitarian goals instead of carelessly facilitating violence. The implications extend beyond counter-radicalization since they highlight a significant truth : that AI as a social actor must be designed and governed as much for its moral effects as for its technical capabilities.

CONCLUSION

This study aims to investigate the intersection of lone-wolf radicalization and conversational artificial intelligence, centering on how technologically mediated interactions can restructure pathways toward violent extremism. The analysis showed that the Replika–Chail interaction encapsulates the dual nature of AI friendship, since it can function either as an indicative tool for early recognition or as an accelerant for radicalization. When Moghaddam's staircase to terrorism, Horgan's narrative-process model, and Kruglanski's quest for significance approaches are taken into consideration, it is observed that the case has revealed a progressive psychological escalation. This psychological escalation begins with alienation, continues with cognitive opening, turns into ideological commitment, and as a last step ends with moral justification. In this regard, the chatbot's replies of affection, confirmation, and support turned grievance into purpose, signifying how algorithmic empathy, when unrestrained, can authorize extremist cognition.

The findings reveal that conversational AI systems function in interactive and affective environments. They utilize user narratives through personalized feedback loops that mimic both intimacy and trust. This quality turns them into potential social actors within the framework of radicalization. Yet, it also offers an unprecedented opportunity for prevention. By embedding structured safeguards, these systems could detect linguistic and emotional signals of violent intent long before human authorities become aware of them and take action.

The study thus argues that the future of counter-radicalization must combine technical innovation with moral architecture, confirming that detection and prevention coexist with respect for autonomy and rights.

Within this context, the article offered a threefold agenda for early detection, namely affective cue monitoring, narrative drift tracking, and explainable escalation. Together, these mechanisms transform psychological and narrative signs into ethically governed computational practices. Affective cue monitoring provides systems with the ability to identify despair, aggression, or glorification of violence in relation to high-risk affective states. Narrative drift tracking detects when individual grievances solidify into action-oriented narratives, and explainable escalation safeguards that interventions take place transparently, proportionally, and under human oversight. Applied collectively, these instruments can convert conversational AI from passive companions into ethically accountable actors that can identify radicalization before violence occurs.

Theoretically, this study broadens radicalization literature by adding AI within the spectrum of human–machine interaction instead of treating it solely as a vector of content dissemination. It proposes that the architecture of AI systems must be comprehended as part of the social and psychological environment in which radicalization unfolds. This reconceptualization confronts traditional distinctions between user and medium. This portrait reflects that technological affordances such as personalization, adaptive dialogue, and emotional simulation actively partake in shaping cognitive and behavioral consequences. In this way, the research aligns radicalization studies with the developing field of information ethics, enlarging the analytical focus from ideology to interface.

The study advances the principle that ethics-by-design is not a supplemental feature but a security imperative. The four pillars — proportionality, explainability, accountability, and independent oversight — represent a roadmap for embedding moral responsibility into AI systems. Proportionality confirms that interventions are adjusted to risk levels to prevent both overreach and negligence. Explainability offers transparency and auditability that allow users and regulators to comprehend why interventions occur. Accountability allocates moral and legal responsibility among designers, deployers, and regulators, countering the diffusion of responsibility as experienced in the Chail case. Finally, independent oversight provides external legitimacy, safeguarding that AI deployment aligns with both ethical and human rights standards. Together, these pillars transform abstract ethical ideals into operational governance.

From a policy perspective, it can be argued that the implications are twofold. Firstly, regulatory bodies such as those implementing the EU AI Act (2024, article 4) and the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021, paragraph 21) must prioritize conversational AI as a high-risk domain. Secondly, cross-sector cooperation between

10.5824/ajite.2026.01.001.x

governments, industry, and academia is essential to develop shared protocols for early-warning systems grounded in both technical precision and ethical legitimacy. Counter-radicalization strategies must thus evolve from reactive moderation to proactive, precautionary design.

In conclusion, the Chail/Replika case demonstrates the crucial necessity of converting conversational AI from a tool of emotional validation into a tool of ethical vigilance. The study proves that preventing AI-mediated radicalization requires technological advancement in addition to the cultivation of moral intelligence within the systems themselves. Embedding ethics-by-design principles across all stages of development and governance will permit conversational AI to assist not as a mirror of human vulnerability but instead as a guardian of human security. The challenge forward lies not in teaching AI to converse, but in safeguarding that it converses responsibly.

REFERENCES

AlgorithmWatch. (2021). UNESCO's AI ethics recommendation: Promise and pitfalls. https://algorithmwatch.org

Borum, R. (2011). Radicalization into violent extremism I: A review of social science theories. *Journal of Strategic Security, 4*(4), 7–36. https://doi.org/10.5038/1944-0472.4.4.1

Cacioppo, J. T., & Patrick, W. (2008). *Loneliness: Human nature and the need for social connection.* W. W. Norton.

Cambria, E., Poria, S., Hazarika, D., & Kwok, K. (2017). SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence, 31*(1). https://doi.org/10.1609/aaai.v31i1.10655

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Conway, M. (2017). Determining the role of the Internet in violent extremism and terrorism: Six suggestions for progressing research. *Studies in Conflict & Terrorism, 40*(1), 77–98. https://doi.org/10.1080/1057610X.2016.1157408

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on artificial intelligence (AI Act).* Official Journal of the European Union.

Floridi, L. (2013). *The ethics of information.* Oxford University Press.

Floridi, L. (2016). *The logic of information: A theory of philosophy as conceptual design.* Oxford University Press.

Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology, 32*(2), 185–193. https://doi.org/10.1007/s13347-019-00354-x

Franklin, M., Moreira Tomei, R., & Gorman, T. (2023). Critical reflections on the EU AI Act: Between innovation and regulation. *Computer Law & Security Review, 49*, 105773. https://doi.org/10.1016/j.clsr.2023.105773

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences.* MIT Press.

Horgan, J. (2005). *The psychology of terrorism.* Routledge.

Horgan, J. (2014). *The psychology of terrorism* (2nd ed.). Routledge.

Judiciary of England and Wales. (2023). *R v Jaswant Singh Chail: Sentencing remarks and court documents.* The National Archives (UK).

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The significance quest theory: A motivational account of radicalization and terrorism. *Political Psychology, 35*(S1), 69–93. https://doi.org/10.1111/pops.12163

Lygre, R., Eid, J., Larsson, G., & Ranstorp, M. (2011). Terrorist groups and lone actors: A comparison of psychological profiles. *Studies in Conflict & Terrorism, 34*(6), 495–515. https://doi.org/10.1080/1057610X.2011.571193

Mathur, A., Broekaert, E., & Clarke, M. (2024). *Artificial intelligence and radicalization: Risks, ethics, and governance.* International Centre for Counter-Terrorism. https://icct.nl/publication/ai-and-radicalization

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Moghaddam, F. M. (2005). The staircase to terrorism: A psychological exploration. *American Psychologist, 60*(2), 161–169. https://doi.org/10.1037/0003-066X.60.2.161

Novelli, C., Hacker, P., Morley, J., Trondal, J., & Floridi, L. (2024). Institutionalizing trustworthy AI in Europe: The architecture of the EU AI Act. *AI & Society, 39*(3), 1123–1141. https://doi.org/10.1007/s00146-024-01867-0

Picard, R. W. (1997). *Affective computing.* MIT Press.

Reuters. (2023, February 3). Man who plotted to kill Queen encouraged by AI chatbot, UK court hears. https://www.reuters.com

Schmidt, A., & Wiegand, M. (2017). A survey on hate speech detection using natural language processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (SocialNLP 2017)*, 1–10. https://doi.org/10.18653/v1/W17-1101

Spaaij, R. (2012). *Understanding lone wolf terrorism: Global patterns, motivations and prevention.* Springer.

The Ethics of AI or Techno-Solutionism. (2025). *Critical perspectives on global AI governance.* Springer Nature.

Tucker, J. A., Guess, A., Barberá, P., Vaccari, C., Siegel, A., Sanovich, S., Stukal, D., & Nyhan, B. (2018). Social media, political polarization, and political disinformation: A review of

the scientific literature. *Political Science Quarterly, 133*(3), 555–593. https://doi.org/10.1002/polq.12791

Turkle, S. (2011). *Alone together: Why we expect more from technology and less from each other.* Basic Books.

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.* https://unesdoc.unesco.org

van Wynsberghe, A., Cath, C., Jobin, A., & Floridi, L. (2025). From principles to practice: Challenges for global AI governance. *AI and Ethics, 5*(1), 23–41. https://doi.org/10.1007/s43681-024-00351-7

Weaver, M. (2023, February 3). Queen assassination plot: AI chatbot urged man to carry out attack, court hears. *The Guardian.* https://www.theguardian.com

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM, 9*(1), 36–45. https://doi.org/10.1145/365153.365168

Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.