



Evolving patterns in the Terrorist Threat Landscape: Online Radicalisation in the age of Artificial Intelligence

Whitepaper

DOI: 10.5281/zenodo.21976015



Authors



Lara Raggi,

Matteo Muratori,

Alessandro Zanasi,

Maria Ustenko



Table of Contents

Project Introduction	4
1. Executive Summary	6
2. Emerging patterns of terrorism and the phenomenon of lone actor violence	8
3. Online radicalisation and the influence of social media	13
4. The effects of AI in the context of radicalisation and online extremism.....	26
5. Conclusions	38
6. Bibliography	39



Project Introduction: NOTIONES

Novel technologies have presented practitioners with new opportunities to improve the intelligence process, but have also created new challenges and threats. Consequently, the timely identification of emerging technologies and analysis of their potential impact, not only on the intelligence community but also on terrorist or criminal organisations, is crucial.

However, time constraints can prevent intelligence practitioners from being updated on the most recent technologies.

In order to address this challenge NOTIONES will establish a network, connecting researchers and industries with the intelligence community. This network will facilitate exchange on new and emerging technologies but also equip solution providers with insights on the corresponding needs and requirements of practitioners. The so gained findings will be disseminated in periodic reports containing technologic roadmaps and recommendations for future research projects and development activities.

The consortium of NOTIONES includes, among its 29 partners, practitioners from military, civil, financial, judiciary, local, national and international security and intelligence services, coming from 9 EU Members States and 6 Associated Countries. These practitioners, together with the other consortium members, grant a complete coverage of the 4 EU main areas: West Europe (Portugal, Spain, UK, France, Italy, Germany, Austria), North Europe (Finland, Denmark, Sweden, Estonia, Latvia), Mittel Europe (Poland, Slovakia, Ukraine), Middle East (Israel, Turkey, Georgia, Bulgaria, Greece, North Macedonia) for a total of 21 countries, including 12 SMEs with diverse and complementary competences.

Project Objectives



GATHER the needs of intelligence and security practitioners related to contemporary intelligence processes and technologies;



PROMOTE interaction of technology providers and academy with intelligence and security practitioners;



IDENTIFY novel technologies of relevance for practitioners through research monitoring;



PUBLISH a periodic report, summarising key findings in order to orientate future research and development;



ENSURE the commitment and involvement of new organisations in the pan-European NOTIONES network.

Project Introduction: NOTIONES

Project Facts:

Duration: **60 Months** Reference: **101021853**

Programme: **Horizon 2020 SU-GM01-2020 Coordination and Support Action**

Coordinator: **FUNDACION TECNALIA RESERACH & INNOVATION (Spain)**

Scientific Technical Coordinator: **ZANASI ALESSANDRO SRL (Italy)**



This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101021853.

Coordinator



Scientific Technical Coordinator



Academic | Think-Tanks | Research



Technology Providers



Practitioners



Estonian Police and Border Guard Board



P.TORNABENE ***

1. Executive Summary



The terrorist threat landscape has changed significantly in recent years, challenging conventional definitions of terrorism. A growing number of attacks now sit in a 'grey zone' between terrorist-type violence, personal grievance, psychological vulnerability, social alienation and fragmented ideology. This does not mean that such violence is apolitical or devoid of meaning. Rather, its extremist content is often articulated through hybrid combinations of slogans, symbols, narratives and grievance frames, many of which are encountered, reinforced or reconfigured in digital environments.

This shift is particularly visible in lone-actor violence: individuals involved in such attacks may present a combination of identity crisis, mental health difficulties, social isolation, weak community support and many other concurring variables. These factors can produce forms of violence that are difficult to detect preventively.

Online presence can then become more than a source of information or entertainment: it may provide belonging, recognition, validation and access to communities where personal frustrations are reframed through ideological, conspiratorial or violence-oriented narratives. Online radicalisation in this context rarely follows a linear path from exposure to violent action. Instead, it often involves fragmented and highly personalised trajectories, as well as Bottom-up and Top-down processes. An individual can self-radicalise within online platforms and/or be radicalised through external recruitment.

Individuals may draw selectively from extremist, conspiratorial, misogynistic, accelerationist, ethno-nationalist, religious, anti-system or post-ideological material, combining separate and sometimes contradictory elements into idiosyncratic justifications for violence. In some cases, the resulting worldview is less a coherent ideology but a personalised cult of violence, shaped through repeated exposure to digital content, peer validation, meme cultures, algorithmic recommendation, closed groups and cross-platform migration.

The central question is therefore not whether the internet and social media platforms cause

radicalisation, but how they shape the environments in which grievance, identity issues, psychological vulnerability and violence become connected.

Digital media have shortened the distance between exposure, interaction and ideological reinforcement. Indeed, understanding social media platforms-specific dynamics is essential for analysing contemporary radicalisation, particularly where traditional ideological labels no longer capture the full complexity of the threat. Also, the advent of Artificial Intelligence seems to have compressed this pathway further in recent years. In a limited number of cases, the progression from exposure to extremist material, to the normalisation of violent narratives, to mobilisation or action may become faster and less visible. The phenomenon is not automatic, universal or solely technological. It refers instead to the possibility that AI tools may intensify pre-existing radicalisation dynamics by making content production, adaptation, targeting and interaction easier.

Specifically, social media recommendation systems organise and personalise exposure to content; they may create or deepen echo chambers by addressing them towards communities of individuals with similar beliefs and interests. For vulnerable users, this may strengthen confirmation biases, harden grievances and make extremist or conspiratorial ideas feel coherent, recognised and socially validated.

Informational barriers are also lowered. Even when violent intent is not explicitly declared, users may seek information that supports reconnaissance, target selection, tactical ideation or preparation. This does not mean that AI systems directly enable attacks, nor that safeguards are irrelevant. However, friction in the movement from grievance to planning are reduced by making information easier to search, organise, summarise and repurpose.

In this sense, propaganda production today has become much more effective and customisable: Generative AI tools allow hostile actors to produce text, images, audio and video more quickly and

1. Executive Summary

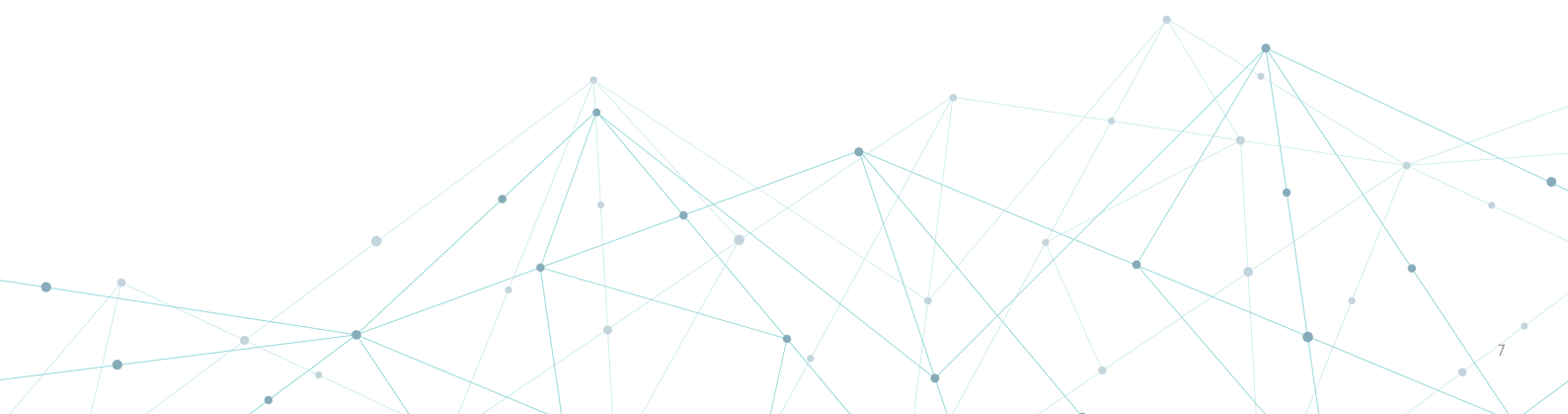


at lower cost. They also facilitate multilingual communication, allowing narratives to reach wider audiences across linguistic and cultural boundaries. This matters because radicalising content can be personalised and adapted to different grievances, national contexts, religious or political references, age groups and identity profiles.

AI-populated environments, such as synthetic communities, may reinforce meaning, identity and belonging, especially for individuals who are socially marginalised or experiencing an identity crisis. As humans struggle to distinguish AI-generated and human-written text, the risk is not only that such systems transmit extremist content, but that they simulate social belonging around it.

More importantly, and potentially more dangerously, LLM-based companions (unlike static propaganda) can create the impression of a true relationship, offering emotional reassurance, mirroring the user's worldview and gradually reinforcing hostile interpretations of reality.

Hence, Artificial Intelligence may intensify and accelerate pre-existing dynamics, such as the convergence of personal vulnerability, ideological exposure, social isolation and online community influence. For prevention, this means that AI-related risk cannot be addressed only through content removal. It requires attention to platform design, recommender systems, companion technologies, synthetic media, encrypted or semi-private communities, and the offline conditions that make individuals receptive to radicalising narratives in the first place.



2. Emerging patterns of terrorism and the phenomenon of lone actor violence



2.1 Obsolete conventional categories and the hybridisation of terrorist ideologies

The terrorist threat landscape is continuously evolving, placing increasing pressure on counter-terrorism operations that still rely heavily on traditional legal categories, ideological taxonomies and operational procedures. Although there is no universally accepted definition of terrorism, international and European instruments generally converge on a set of core elements. The United Nations General Assembly Resolution 49/60 (1994), United Nations Security Council Resolution 1566 (2004), and EU Directive 2017/541, among others, describe terrorism as involving serious criminal acts, often directed against civilians or public targets, intended to cause death, serious harm, intimidation or destabilisation. From this perspective, the commonly recognised elements of terrorism include: a political, ideological or religious cause; the direct use or threat of violence; the intentional infliction of death, injury, destruction or serious disruption; the indirect objective of coercing public authorities or destabilising political, constitutional, economic or social structures; and the targeting of civilians, public spaces, symbolic sites, transport systems, critical infrastructures, governmental institutions, religious or ethnic communities, or international organisations.

These elements remain legally and operationally relevant. However, today terrorist actors seem to be less ideologically coherent [1], and contemporary terrorist and terrorist-like violence increasingly presents hybrid [2] and flexible characteristics that do not fit neatly into traditional categories. The problem is not simply that some violent acts are difficult to label. The labels themselves may also be factually imprecise. “Grey-zone” [2] terrorism is not always entirely apolitical or devoid of political meaning. Rather, the political content is often articulated in fragmented and unstable ways. Instead of subscribing to coherent and established ideologies, perpetrators may draw selectively on political slogans, conspiracy narratives, religious references, extremist symbols, online subcultures, grievance frames or violent aesthetics. These

elements are usually the ones that fit the most with personal grievances, psychological vulnerabilities or developmental issues, thus creating highly idiosyncratic worldviews that do not map clearly onto conventional ideological categories. The issue, therefore, is not the absence of ideology, but the “salad bar” [3] or “ideology mixers” [4], as to say its fragmentation, individualisation and frequent conflation with personal grievances and contextual factors.

This is particularly evident in lone-actor violence. In cases linked to jihadist or Islamist extremism, which remain the most prominent and impactful terrorist threat [5], many attacks still present more recognisable ideological markers and are more easily classified within traditional counter-terrorism frameworks. In other cases, however, it has become increasingly difficult to categorise a violent act as terrorism even when several above-mentioned indicators are present.

This usually happens because the perpetrator may have no formal affiliation to a terrorist organisation, no clear chain of command, no stable ideological worldview and no explicit declaration of political or religious intent. Instead, the individual may display signs of multiple and sometimes contradictory belief systems, often encountered and recombined through the internet and social media. Indeed, lone actors (also known as “lone wolves”) frequently appear to select fragments of broader ideological repertoires that validate their own grievances, fears and/or objectives. They may not adopt a doctrine in full; they may instead search for narratives that provide meaning, recognition and justification for their personal unease or will to harm others. As a result, individual grievances become intertwined with religious, political, conspiratorial, misogynistic, ethno-nationalist, anti-system or accelerationist messages. The outcome may be less a coherent ideology than a personalised interpretative framework through which the individual explains the world, identifies enemies and legitimises violence.

For instance, a hypothetical lone actor may simultaneously engage with conspiracy theories commonly associated with far-right environments,

2. Emerging patterns of terrorism and the phenomenon of lone actor violence



fears of “the Other”, anti-immigrant narratives and Islamophobic arguments. These elements may be connected to ideas of cultural identity, bloodline, demographic threat or civilisational decline, including narratives associated with white supremacist or ethno-nationalist milieus such as the Great Replacement theory [6]. At the same time, personal disappointment, social failure or resentment towards institutions may lead the same individual to adopt anti-system narratives about economic inequality, the collapse of the current political order or the need to destroy society in order to rebuild it. These themes may appear closer to left-wing extremist, anarchist or accelerationist imaginaries. In practice, the perpetrator may not experience these influences as contradictory. They may function as fragments that, when combined, provide an emotionally satisfying explanation of their condition.

This evolution makes it difficult to place contemporary attackers within rigid categories, which remain useful for legal, analytical and institutional purposes, but are increasingly insufficient when used as exclusive boxes. We could

perhaps refer to a ‘post-ideological’ dimension of contemporary terrorism, although this term remains contested. Nevertheless, this expression should be used carefully: it does not mean that ideology has disappeared. It means that ideology is often less coherent, less organisationally disciplined and more individually assembled (apparently shifting from a Top-down ideology inoculation to a Bottom-up self-appropriation – see Chapter 3). In some cases, violent mobilisation may be driven primarily by anti-semitism, anti-Muslim hatred and the rejection of European democratic values [6] [7]; but also, by a cult of violence, misanthropy, nihilism or the desire for notoriety. Therefore, At this stage, the contemporary threat landscape can be understood as one in which individuals combine fragments drawn from different, and sometimes conflicting, ideological repertoires into personalised rationalisations for violence.



2. Emerging patterns of terrorism and the phenomenon of lone actor violence



2.2 Rising of Lone Actor Terrorism and the individualisation of violence

A second major transformation concerns the operational model of lone-actor attacks. Since the mid-2010s, **terrorist propaganda has increasingly legitimised individual action carried out without formal affiliation, training or direct operational communication with a central organisation.** In September 2014, the Islamic State publicly encouraged supporters in Western countries to attack independently using whatever means were available (Indeed, Your Lord Is Ever Watchful speech by Abu Muhammad al-Adnani). In November 2016, its propaganda magazine Rumiyah further promoted vehicle attacks as a low-tech tactic accessible to sympathisers. More in general, Jihadist organisations gradually adapted their propaganda strategies to legitimise the “lone mujahid” model, portraying unaffiliated individuals who act without direct operational links, detailed instructions or logistical support as heroic and authentic expressions of ideological commitment [8].

This messaging contributed to the normalisation of a model in which affiliation, travel to conflict zones, specialist training and communication with an operational command are no longer necessary preconditions for violent action. As a potential consequence, from 2016 onwards, several high-profile attacks and terrorist-like incidents in Europe and other Western countries have been carried out by lone actors or very small cells. Some were clearly classified as terrorist attacks; others were treated as mass violence, hate crime, severe criminal violence or incidents linked to mental health crises. This distinction matters legally, but it also creates analytical problems. Incidents such as the Hanau shootings, the Nice attack, the Vienna attack and other violent episodes across Europe show how difficult it can be to distinguish between terrorism, extremist violence, personal grievance, hate crime, psychiatric crisis and symbolic mass violence when the perpetrator does not fit a conventional organisational profile. This issue further supports the thesis that terrorist phenomena are way more complex now than in the past.

2.3 Early warning signs of lone-actor radicalisation and gaps in preventive coordination

Lone-actor attacks are not only rising worldwide, but they are also intrinsically and increasingly difficult to predict. This is not necessarily because there are no warning signs. In many cases, elements of concern exist before the attack: grievance, social isolation, threatening language, fixation on violence, online extremist consumption, personal crisis, institutional conflict, or fascination with previous attackers’ manifestos [9] (leading to so-called “copycat attacks”). The operational problem is that these signs are often dispersed across different environments and observed by separated actors: family members and friend groups, social workers, teachers, employers, health professionals, online communities, local police or institutions. Individually, each signal of radicalisation may appear insufficient for an external actor to justify a security response. Collectively, multiple signals may indicate an escalating risk. The prevention therefore gap lies in the absence of a clear multi-agency mechanism able to connect suspicious warning signs proportionately and early enough.

Specifically, several recurring vulnerability factors are relevant, although none of them should be treated as deterministic predictors of radicalisation or inclination towards violence: First, some lone actors present mental health difficulties such as personal crises, suicidal ideation, acute psychoses, subclinical and often undiagnosed trauma, identity instability, loneliness, grievance or intense frustration. Indeed, mental health issues are more frequently observed within anti-terrorism research: existing literature suggests that mental health disorders are more frequently observed among lone-actor terrorists than among individuals involved in other forms of terrorism, with some reviews estimating the difference at around 40 per cent [10]. However, this point requires caution. Mental illness does not in itself cause extremist behaviour or terrorism, nor does it provide a reliable profile of future violence. The relevant issue is the interaction between many variables. In this sense, psychological factors should

2. Emerging patterns of terrorism and the phenomenon of lone actor violence



be understood as part of a broader risk ecology, but also as a relevant piece of information when working on prevention strategies.

“In many cases, ideology does not appear to be the root cause, but rather a means through which underlying psychological distress is channeled [...] ideological elements are adopted not for their doctrinal content but for their capacity to provide psychological relief or purpose” [11].

What is particularly relevant here is the difficulty of predicting violent behaviour associated with psychological distress or mental health disorders. As in cases of suicide, warning signs may be difficult to recognise from the outside, and the possibility of violent escalation is often underestimated or dismissed.

Second, lone actors may experience social marginalisation, including discrimination, unemployment, poor life prospects, weak community support or repeated institutional failure. These conditions can produce isolation and alienation, especially when the individual feels humiliated, excluded or unable to achieve recognition through ordinary social channels. Again, these factors are not predictive on their own. Most people experience social hardship without becoming violent. The risk may emerge when social exclusion intersects with other vulnerabilities [12]. In such cases, violent escalation can be difficult to predict and prevent, as marginalised individuals often remain outside the attention of formal institutions and wider society.

Third, lone-actor violence is often difficult to predict because such individuals are closely connected to patterns of **internet use and dependency** (a phenomenon that can remain closed into one's home walls). Constant connectivity may become an escape route from difficult offline conditions, including psychological instability, loneliness and social exclusion. Digital platforms can provide

entertainment, recognition, structure and a sense of belonging. For individuals who lack stable offline affiliation, online communities can become substitute social environments. They may offer explanations for personal failure, identify enemies, provide symbolic status and reinforce the idea that violence is meaningful, heroic or necessary. In this context, the internet offers a social and narrative infrastructure through which existing grievances can be reframed, validated and escalated.

In this regard, the **virtualisation of recruitment and radicalisation** [2], combined with the global reach of social media, has created an enabling environment in which extremist narratives can circulate at unprecedented speed and scale, with relative low visibility. Contemporary radicalisation, therefore, is not just a social process rooted in physical spaces and based on group relationships and dynamics; both traditional top-down recruitment strategies and more individualised self-radicalisation pathways can now reach large numbers of users within seconds. As a result, the internet operates not merely as a medium for communication, but as an ecosystem in which sense of belonging, social ties and acceptance, and political engagement have migrated. Not only that, but it can be the ‘fertile soil’ where personal grievances can be articulated, validated and channelled towards extremist forms of belonging more easily and rapidly than before. These dynamics are explored in Chapter 3.

Fourth, lone actors often rely on the **accessibility and low visibility of easily accessible and/or improvised weapons** (IEDs – Improvised Explosive Devices). This reduces operational issues and limits opportunities for interdiction: knives, vehicles and rudimentary devices require less preparation than firearms or sophisticated explosives and may not generate the same warning indicators. The online environment, in this case, can further reduce informational barriers by circulating violent inspiration, tactical discussion and operational insights on how to produce home-made weapons. The online environment may further lower

2. Emerging patterns of terrorism and the phenomenon of lone actor violence



informational barriers around low-tech or improvised violence.

Vehicle ramming attacks could be included in and applied to this reasoning, if we consider a car as an easily accessible instrument and potential weapon. A vehicle-ramming attack occurs when a vehicle is deliberately used as an impact weapon against people, crowds, buildings, public places, infrastructure or other targets. Such an act may be terrorist, criminal, ideological or linked to other motives. It officially becomes terrorism only when accompanied by a specific intent, such as intimidating a population, coercing authorities or destabilising fundamental structures. The major vehicle attacks of 2016 and 2017 were often jihadist or ISIS-inspired, which made their classification as terrorism more straightforward. More recent incidents, on the contrary, are frequently harder to categorise because no terrorist affiliation is confirmed, the perpetrator profile is ambiguous, and the motive may involve a mixture of grievance, mental health crisis, social resentment, hatred or symbolic targeting.

Vehicle attacks are especially difficult to prevent because they require minimal preparation, involve a everyday object and can unfold in seconds. In practical terms, they may require little more than intent, access to a vehicle and the decision to act [13]. This makes them attractive in a threat environment where low-tech violence, soft targets and individual mobilisation have become increasingly relevant.

The Modena incident of 16 May 2026 illustrates this “grey zone”. According to initial public reporting, a man drove into pedestrians in the city center, injuring several people, before attempting to flee; no terrorist affiliation was confirmed at the time of reporting, while psychiatric problems and elements of personal difficulty were reported. The case should therefore be treated cautiously and not prematurely labelled as terrorism. However, the case is analytically useful because it illustrates a broader prevention problem: prior to the incident, the perpetrator had reportedly shown several signs of concern, including: personal resentment; offensive messages addressed to the

University of Modena in which he demanded help in finding a job; mental health difficulties following a stay in a psychiatric facility that he later chose to leave; social exclusion linked to his inability to find satisfactory employment; and problematic interactions with institutions. While each of these elements may have appeared insufficient in isolation, their combination points to the difficulty of identifying escalation when warning signs are dispersed across different personal, social and institutional contexts. As mentioned before, the problem is not that nobody sees the signals; the problem is that different actors see different fragments, and no actor is clearly responsible for connecting them, assessing the cumulative risk and coordinating an early response.

This is why contemporary prevention cannot rely only on traditional counter-terrorism categories. A more effective approach requires a broader framework able to clearly identify terrorism, violent extremism, hate crime, severe grievance-fuelled violence, mental health crisis and ambiguous mass violence, but also recognising that these phenomena may overlap. The challenge is to build proportionate prevention systems that can connect fragmented warning signs without criminalising vulnerability, stigmatising mental illness or expanding counter-terrorism logic to every form of social deviance. In the contemporary threat landscape, the key operational question is not only whether a perpetrator belongs to a terrorist organisation. It is whether a set of personal, social, ideological and digital factors is converging towards violence.



3. Online radicalisation and the influence of social media



3.1 The normalisation of extremist content: a shift in the 'Individual' Overton Window

Radicalisation can both stem from and contribute to a broader range of harmful manifestations, including exclusionary attitudes, harassment, hate speech, hate crimes, and wider processes of mainstreaming through which formerly marginal extremist narratives gain visibility, legitimacy and circulation in a segment of public discourse. These dynamics can weaken democratic values and social cohesion by normalising hostility towards outgroups, eroding pluralistic norms and intensifying polarisation [14].

At the same time, what counts as “radical” and “extreme” depends on the reference point against which it is measured. This reference point can itself shift and is highly dependent on contextual factors, such as the political system, culture and prevailing circumstances. This means that, the concept of “normality” is defined by society at a certain time, and that an extremist belief is considered as such only from the perspective of those who define normality itself. More specifically, societies and cultures have gradually delineated what is considered ‘acceptable’, what is ‘lawful but lawful’, and what is completely beyond the threshold of acceptance. Thus, there are broadly shared understandings of what is moral and immoral, although different societies and belief systems maintain their own worldviews.

The Overton Window explains this concept particularly well. It refers to the range of subjects and arguments considered politically acceptable by the mainstream population at a given time. The key element of the concept is that the window can shift, shrink or expand over time. According to Overton, the window frames the range of policies that a politician may recommend without appearing too extreme, thereby allowing them to gain or retain public office within the prevailing climate of public opinion.

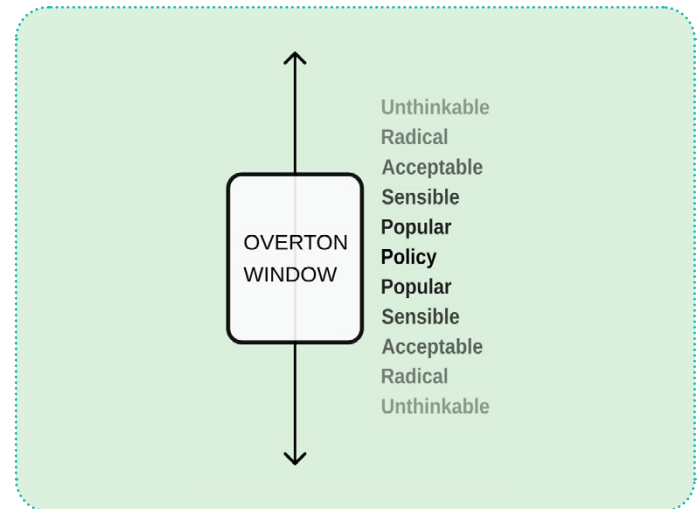


Figure 1: The Overton Window

It has to be noticed that the Overton Window is (1) specifically focused on the political sphere, and (2) a general framework that is usually applied to society as a whole. Nonetheless, we argue that, if it is used as a conceptual instrument and the focus is shifted from society to the individual (from the macro to the micro level) the Window can still be applied. This is because, in addition to the conventional belief systems and values that influence a person’s worldview and actions, individuals can develop their own understanding of the world and their own convictions. Therefore, a person’s specific perspective – which here will be referred to as the “Individual Overton Window”, meaning a personal and probably hybrid system of beliefs that gives meaning to one’s life – has the power to influence individual actions¹.

If this is true, the reasoning could be used to explain why extremist or radicalised individuals deliberately and consciously engage in unconventional and illegal (“a-normal”, in the society’s perspective) actions.

¹ It is important to state that the Individual Overton Window is introduced here as an analytical heuristic rather than as an empirically validated or predictive model. It is intended to illustrate how an individual’s boundaries of acceptability may diverge from prevailing social norms, without implying a linear or universal radicalisation pathway.

3. Online radicalisation and the influence of social media



On the one hand, from the social perspective, terrorists appear to act outside the normal rules and principles of contemporary societies, as they embrace unacceptable justifications of violence. On the other hand, being devoted to a cause, a terrorist could consider violence as the only means of achieving an objective that differs from the mainstream.

Even a lone actor engages in violent action for a reason. This reason may be clearly delineated in their mind, but it may also be confused. And when it is so, it may arise from distrust in society itself or in its institutions, disillusionment with the life that the “system” has provided, or disappointment with the way the individual feels within the contemporary world. Whether consciously or not, these grievances are intrinsically opposed to the existing order.

3.1.1 The process of ‘Bottom-up’ radicalisation

Assuming that the argument developed so far is reasonable and that the thesis holds, the radicalisation process can be explained as a shift in the Overton Window and, more specifically, as a ‘secession’ of the Individual Overton Window from the ordinary Social Overton Window. The two Windows become increasingly distant from one another, potentially leading the individual towards anti-system actions.

Concretely, one case of Bottom-up/self-radicalisation, among others, could develop as follows:

Phase 0: Let us assume that a person is initially fully aligned with the mainstream values and morality belonging to their geographical and social context. In such case, the Social and Individual Overton Windows (OWs) are broadly similar (although not identical²).

Phase 1: The individual experiences something significant — a major life change — or simply feels dissatisfied with their life. All the elements discussed in the previous Chapter can also be included in this assumption, ranging from psychological disorders to social exclusion.

In order to find a new and more satisfying sense of meaning, the individual begins to question what they see and the fundamental convictions they have received from the outside world. As a simple and easily understandable example, if the person was born in a Western liberal country, they may begin to blame “capitalism” as the source of all their problems, such as difficulties in finding employment or recurring stress caused by the pace of contemporary life. Alternatively, they may not have a clear idea of where to direct their anger. In either case, they will seek some form of external validation. This may come from a friend, family members, online communities, or others.

When the individual finds confirmation and validation, even within a small group, they feel they have discovered a meaning that better reflects their personal needs and grievances. This is the key emotional and cognitive moment that has the power to enable the beginning of the radicalisation process. Of course, most people remain within this Phase, for instance by joining online communities, spending time with like-minded friendship groups, and building their lives around the “bubble” that embraces the closest worldview to their Individual OW. In essence, this “bubble” provides them with an alternative way of thinking about reality. For adolescents, who may be more emotionally vulnerable and still searching for their identity, this alternative can provide a sense of belonging, thereby winning their hearts and minds. This new Window relies on slightly different categories, beliefs, values and concepts, meaning that what is considered ‘acceptable’ and ‘unacceptable’ differs from the conventional framework. Of course, this

² This seems a more realistic assumption, as individuals are unlikely to agree completely with the values associated to the culture they grow up in. Therefore, the two Windows remain largely aligned

3. Online radicalisation and the influence of social media



stage is still far removed from extremism or from a fully developed process of radicalisation.

Phase 2: In fewer cases, the individual encounters, within these “bubbles”, a more extremist version of the dominant discourse. In our terms, this means that they are exposed to a Window that is further distanced from the conventional Social OW. If the individual feels even more represented by this version, they may become interested in learning more and may gradually embrace a radical worldview. As a consequence, the Individual OW shifts even further away from the original one.

This cycle may occur repeatedly. If the individual continues to embrace increasingly extremist versions of the discourse, it becomes clear that their Individual OW will move progressively further away from the Social OW.

It is important to remember, however, that the majority of people who hold radical ideas and accept violent justifications do not engage in violent action or terrorism [15].

This process, more or less advanced, in our view, is the process of self-radicalisation, namely “an incremental process in which individuals or groups develop increasingly radical or extremist worldviews and/or an increasing acceptance of violent actions for ideological goals” [14]. In this sense, we define Bottom-up as a process through which an individual actively and voluntarily engages with extremist beliefs, primarily through personal agency rather than through direct external indoctrination.

This stands for the normalisation of extremist convictions and, consequently, the greater disinhibition in creating and sharing such content online, which is eventually perceived as one among the many positions and worldviews circulating on the internet.

In this sense, a conspiracy theory may appear, in the eyes of an extremist, as an argument competing on equal terms with those belonging to the Social OW. Indeed, conspiracy theories are commonly used as powerful instruments: they represent meaning-making devices, offering simplified explanations of complex realities [16] and are often presented as strong and appealing counterarguments to established beliefs or even to scientific evidence (as in the case of climate denialism versus climate science, often considered as a one-to-one debate). In this context, individuals are encouraged to question the credibility of the information they encounter, with the result that social norms and even scientific consensus may no longer be perceived as solid as it once was.

The discrepancy between social evidence *and* public perception ultimately shapes attitudes, behaviours and, in fewer cases, harmful action based on weak or misleading foundations. This can result in reduced public support or trust in institutions, resistance to necessary behavioural changes, and greater susceptibility to polarised or extremist positions.

3.1.2 The evolving adaptability of ‘Top-down’ radicalisation and recruitment processes

Having defined Bottom-up radicalisation, we use the term **Top-down radicalisation to describe a process in which an external actor — such as a terrorist organisation or cell — exposes vulnerable individuals to extremist content, sometimes intending to guide them, directly or indirectly, towards violent action.**

This phenomenon can be recognised as the process of radicalisation, especially within anti-terrorism frameworks, which have traditionally focused on top-down recruitment as the principal pathway leading individuals towards violent action. Indeed, it must be noticed that Top-down radicalisation remains a relevant and present issue, where digital environments make it considerably easier for recruiters to identify, reach and influence

3. Online radicalisation and the influence of social media



potential targets. Their objectives may range from psychological warfare to desensitisation to violence, publicity, propaganda, and fundraising; they may also include recruitment, mobilisation, information sharing, and even operational planning and coordination [17].

Extremist recruiters and communicators seek to design content and messages that resonate with individual-level needs and, in doing so, enhance their persuasive appeal. Their influence often derives from their ability to offer an alternative to individuals who are searching for a specific group identity (and that is exactly where Top-down and Bottom-up radicalisation processes may intersect).

They do so through contemporary and engaging forms of communication, such as vlogs, distributed across multiple platforms, and through communicators who are frequently portrayed as charismatic and heroic figures [18].

As regards the specific phases of Top-down radicalisation, grey literature reports generally describe the *process* as follows: *Step 1* is often referred to as “The net”, in which an entire population is targeted in order to maximise exposure to extremist discourse. *Step 2* adopts a “funnel” approach, whereby individuals attracted through the net are gradually transformed into committed members through social bonding, discussion and repeated exposure. They subsequently progress to *Step 3*, often referred to as the “infection”, in which selected targets are directed towards self-radicalisation through exposure to radical material and guidance. This may ultimately lead to *Step 4*, “activation”, when the young person is ready to receive indirect commands from recruiters. This model corresponds to what social psychologists describe as the “foot-in-the-door strategy” (Freedman & Fraser, 1966), in which young people enter an extremist group by initially being asked to perform small and low-cost actions. These are followed by increasingly significant actions that gradually ‘commit’ them to a particular course of

action through an “escalation of commitment” (Kiesler, 1971; Staw, 1976). Initial contacts therefore rarely focus explicitly on violence and instead tend to present images of strength and pride [17].

Each sub-group that emerges is then free to develop its own jargon and specific expression of extremist values, often under the influence of more active users who contribute to shaping the sub-group’s identity. In this regard, two types of networks can be identified: active or direct support networks, which involve close contacts such as family members, friends or peer groups and may operate both online and offline; and passive or indirect networks, in which the individual primarily acts as a recipient rather than as an active participant. In some cases, the targets of propaganda are consumers of the content and at the same time they become producers by reproducing and amplifying it.

This means that **Top-down radicalisation seems now to be more decentralised, diffuse and fragmented than before, and increasingly intertwined with Bottom-up processes**. In other words, individuals may self-radicalise, radicalise others, or influence one another through co-radicalisation without belonging to any terrorist cell, with young people often found to be among the most active participants in these dynamics [19].

Online presence thus facilitates interpersonal communication by connecting like-minded individuals. Interestingly, it has been observed [14] [6] that highly active users posted significantly fewer positive and more negative emotional comments, received more negative feedback from peers, and selectively engaged with discussion threads that already displayed a less positive tone. This suggests that radicalisation affects not only ideological beliefs, but also the adoption of a consistently negative, emotional and hateful style of expression that becomes normalised within the community. Such communication can eventually contribute to the construction of cohesive group identities while simultaneously promoting antagonism towards outgroups.

3. Online radicalisation and the influence of social media



It is important to highlight that the contemporary complexity of Top-down radicalisation processes does not suggest that terrorist organisations no longer exist or belong exclusively to the past. The Islamic State and its affiliates, for instance, remain one of the main external threats to the EU because of their capacity to inspire attacks, according to the 2026 EU Agenda to prevent and counter terrorism. Rather, terrorist cells have evolved: they have become more **geographically dispersed, shifted many of their activities online, and learned to exploit increasingly individualised radicalisation processes** for their own purposes.

Jihadist narratives are compelled to adapt continuously to changes in the digital sphere and to exploit the individualisation of violence. This is why today's prevailing Islamist radicalisation pattern reflects not only ideological motivations, but also the dynamics of online youth culture and algorithmic influence [20]. In other words, terrorist organisations themselves must confront and adapt to the convergence of religiously motivated extremist ideology, digital platform mechanisms, and contemporary Western online youth narratives. This convergence has transformed contemporary jihadism into a fluid, decentralised and increasingly aestheticized movement, capable of inspiring violence not through clandestine training camps, but through swipeable videos, viral slogans, and online 'tribalism' [20].

Contemporary religious preachers operating in the digital sphere can, in many respects, be described as genuine influencers. They rely on direct eye contact, emotional storytelling, relatable language, and rapid-response interaction with viewers. They simplify complex theological concepts and connect them to everyday experiences of identity, discrimination or personal crisis. These preachers do not typically call for violence directly. Rather, they create a radical Islamist 'environment' in which dichotomous thinking, exclusivist religious identity, and resentment towards Western norms are normalised and theologically legitimised.

Indeed, the *strategy* of Top-down radicalisation is to leverage personal grievances and help create an environment that: (a) reinforces and validates extremist beliefs; (b) provides emotional appeal and spiritual or existential fulfilment, through the personalisation of extremist content; thus (c) contributes to the separation of the individual from the "mainstream (and corrupted) world", fostering distrust and increasing the influence of exposure to extremist content. In our terms, the aim is to encourage the Individual OW to shift towards positions that are increasingly distant from the Social OW.

Researchers refer to this process as "identity fusion", whereby an individual identity is gradually overshadowed by the identity of the extremist group [21]. This process may be particularly powerful for lone actors who are searching for a sense of belonging and possess a fluid or unstable identity, as well as, more generally, for vulnerable young people facing similar difficulties.

Disinformation, fake news and conspiracy theories, presented as truth, are among the principal instruments used for these purposes. Moreover, the strategy evolves continuously in line with the constant transformation of social media platforms. For instance, extremist actors use **satire-like formats** to circulate extremist content in an apparently superficial and humorous manner, while still producing a profound emotional impact. Collective content practices also include "shitposting" or "trolling", forms of disruptive and humorous interaction through which extremist views are spread while creators gain a gamified reputation within communities of trolls or shitposters [22].

In this sense, extremist communicators have also found a way to respond to the safeguards introduced by social media platforms to counter extremism and hate speech. They deliberately **operate just below the moral and politically-correct thresholds, producing borderline material that may be considered 'awful but lawful' and bypasses the criteria for classification as extremist content. Thus, through the use of memes ("memefication"**

3. Online radicalisation and the influence of social media



[4]), dog whistles, coded messages or hidden meanings, they are still managing to produce and share extremist material online. Identifying and prioritising dangerous forms of expression therefore remains central to current efforts to restrict certain types of online content.

It is lastly important to emphasise that these processes and strategies are effective and relevant only within specific contexts, social “bubbles” and individual conditions. In this sense, “while the internet and social media provide an opportunity to engage with potentially radicalizing materials and individuals, their impact on radicalization trajectories may vary from individual to individual, as may their use in spreading radicalizing content and connections” [23].

3.2 The influence of social media on radicalisation and generational platform use

In examining the potential influence of social media on radicalisation, it is important to recognise that no algorithm inherently or systematically promotes extremist content on a mass scale. Nevertheless, digital platforms remain powerful instruments for individuals who already hold radical beliefs. They can provide a ‘hospitable’ environment for messages involving violence, hatred and conspiracy theories, even though recommendation systems do not ordinarily direct neutral users towards explicitly extremist material [14].

Radicalisation pathways are predominantly cross-platform. Extremist actors often use mainstream services, such as YouTube, Instagram, Facebook or X, as “beacons” directing users towards more closed or specialised environments and dedicated websites. These platforms therefore perform different functions within the same broader ecosystem: some facilitate initial exposure and audience building, while others support more intensive ideological reinforcement,

community formation and interpersonal interaction. In other words, once users have been reached and have responded positively to, or interacted with, the content posted, they are often redirected to ‘safer’ platforms. More specifically, communicators publish moderate and relatable content on widely used platforms to avoid account suspension, while strategically using external links to direct users towards more explicit and graphic material hosted in less regulated spaces such as Telegram and Discord [24].

At the generational level, we argue that patterns of platform use differ considerably [25] and may therefore shape distinct forms and pathways of online radicalisation³:

- **Baby Boomers (1946–1964) and Generation X (1965–1979/80)** are particularly prominent on Facebook and YouTube. In Western contexts, radical narratives targeting these generations tend to be more closely connected to established social or political issues, including nostalgia for previous political orders, anti-immigration discourse and identity-based mobilisation within Facebook pages or groups. Their form may therefore appear less culturally ‘new’, even when the underlying content is radical or exclusionary [25].
- **Millennials / Generation Y (1980/81–1995/96)** continue to use YouTube, as well as Facebook, while also having a strong presence on Instagram. Although they have progressively expanded their use of visual platforms, many remain closely connected to Facebook. For this generation, Instagram functions as a medium of both personal and community expression. Propaganda based on social status, identity and belonging – for example, nationalist groups using symbols and visual markers – may therefore prove particularly appealing.
- **Generation Z / Zoomers (1996/97–2010/12)** are highly active on YouTube, Instagram and

³ The following data are drawn primarily from US-based research and reflect the online behaviour of users in the United States, where generational patterns of platform use have been studied more systematically than in the EU. Nevertheless, the trends identified appear broadly consistent with the European context.

3. Online radicalisation and the influence of social media



TikTok. They tend to favour rapidly evolving short-form video content. For this generation, one of the principal risks is encountering extremist ideas disguised as entertainment, humour or apparently innocuous “casual content”. In such environments, ideological messages may be embedded in memes, trends, music, visual symbolism or emotionally engaging narratives rather than presented through explicit political discourse.

- **Generation Alpha (2010s–2020s)** is already highly present on YouTube and increasingly active on TikTok. Recent U.S. reports indicate that, among children aged 8–12, 64% use YouTube and TikTok on a daily basis, with YouTube identified as the preferred application by 43% [26]. Roblox and other gaming environments are also among the principal platforms in which Generation Alpha spends time. These trends suggest that younger users are exposed to the same mechanisms of brevity, repetition and spectacularisation that characterise contemporary digital content, although the messages directed at them may be even more implicit and encoded through memes, images and short-form videos.

Available data therefore partially confirm common assumptions regarding generational platform use: older generations remain more strongly associated with ‘older’ platforms, while younger users increasingly favour YouTube, Instagram and TikTok. Millennials and members of Gen Z, as digital natives, may possess stronger technical skills but remain highly receptive to platform dynamics based on rapid scrolling, repetition, personalisation and algorithmic recommendation. Baby Boomers, by contrast, may unknowingly share misleading information – contributing to misinformation – and can be particularly dependent on closed Facebook groups for news, interpretation and social validation. Longer textual formats, extended posts and documentary-style videos may also attract more mature audiences.

However, platform preference should not be treated as a rigid generational division: different generations use multiple platforms simultaneously, and each service may perform a different function within their wider information environment. Generations, as well as individuals, may also display different forms of vulnerability and resilience.

At present, there is still insufficient conclusive evidence to determine precisely how Instagram or TikTok alter radicalisation trajectories compared with earlier platforms. Important empirical gaps remain, particularly regarding Gen Z and Gen Alpha, and further cohort-specific surveys and platform audits are required. Radicalisation pathways may therefore be filtered through generational habits of platform use and through the specific mechanisms employed by those platforms; nevertheless, these associations should be regarded as indicative rather than deterministic.

Anyway, some patterns are more evident than others: first, although the platforms preferred by different generations may act as distinct entry points into radical ideological environments, **advanced radicalisation pathways frequently converge**. A young user may first encounter extremist narratives through TikTok or Instagram, while an older individual may be exposed through Facebook or YouTube. In later phases, however, both may migrate towards similar closed communities, private channels or specialised ideological spaces.

Second, the **growing involvement of minors** in online radicalisation pathways is particularly concerning: the number of young people involved in terrorism and violent extremism has increased sharply across Europe. In 2024, almost one third of terrorist suspects in the EU were under 20, with the youngest being only 12 years old. Most of these minors were linked to jihadist terrorism, although violent right-wing extremism has also gained traction, particularly through accelerationist networks, “active clubs” and online extremist networks [1]. Indeed, teenagers between the ages of 13 and 19 are increasingly at the center of recruitment and radicalisation processes,

3. Online radicalisation and the influence of social media



and some have attempted to carry out attacks. Among some younger recruits, contemporary radicalisation appears to progress rapidly towards violent mobilisation, without the lengthy periods of ideological formation traditionally associated with terrorist involvement. Comparable dynamics can be observed in transnational right-wing extremist online environments, where a pronounced cult of violence may accelerate engagement and direct individuals more rapidly towards violent action.

Third, within this landscape, the ***Incel* phenomenon** represents another particularly relevant “bubble” and phenomenon to be discussed: the term refers to self-described “involuntary celibates”, predominantly young men who gather in online communities where personal frustration and social isolation may be reframed through misogynistic narratives that blame women and society for their perceived exclusion. These environments can normalise hostility, sexual entitlement, dehumanising language and admiration for previous attackers, reinforcing rigid gender hierarchies and anti-feminist beliefs. Although most participants do not engage in violence, some communities have contributed to the escalation of hateful discourse, harassment and, in a limited number of cases, violent mobilisation against women.

Social media platforms differ in terms of their recommendation algorithms, multimedia formats, levels of anonymity and privacy, availability of private groups and livestreaming functions, and content moderation policies. These differences shape the role that each platform may play in processes of exposure, socialisation, recruitment and mobilisation. In this regard, some studies have identified lower removal rates for far-right extremist content than for jihadist material [19]. The table below summarises some of these differences and their potential consequences on radicalisation pathways⁴.

⁴ Platform affordances may facilitate exposure, socialisation, recruitment or mobilisation, but they do not independently cause radicalisation. Most pathways involve interaction between several platforms and offline personal, social and cultural factors.

3. Online radicalisation and the influence of social media



Platform	Main affordances	Potential role in radicalisation pathways
Facebook	Personalised feed, pages and public or private groups; relatively identity-linked profiles	Broad exposure, dissemination of political or conspiratorial narratives, community formation and redirection towards external platforms
X (Twitter)	Algorithmic For You feed, hashtags, reposts, real-time communication and Spaces	Public amplification, propaganda, networking, symbolic confrontation and mobilisation around current events
Instagram	Highly visual content, Reels, Stories, hashtags and direct messaging	Aestheticisation and normalisation of ideology, identity-building and soft recruitment through apparently innocuous content
TikTok	Highly personalised short-video feed, music, memes and livestreaming	Low-threshold exposure to emotional and identity-based narratives; potential gateway towards more explicit content and closed platforms
YouTube	Long- and short-form videos, recommendations, livestreams and comments	Ideological explanation, audiovisual propaganda, storytelling and long-term content storage
Discord	Invite-based servers, pseudonymous text, voice and video communication	Semi-private socialisation, peer reinforcement and sustained interaction among receptive or already radicalised users
Telegram	Public and private channels, large-scale broadcasting, file sharing and high pseudonymity	Propaganda dissemination, content preservation, ideological networking and migration from mainstream platforms
WhatsApp	Telephone-based, end-to-end encrypted private and group communication	Trusted peer-to-peer circulation, private reinforcement and rapid sharing of propaganda or disinformation
Reddit	Pseudonymous thematic communities, personalised feeds and discussion threads	Ideological discussion, meme circulation and transition from passive consumption to active community participation
Twitch	Livestreaming, interactive chat, subscriptions, donations and raids	Parasocial influence, live community-building, monetisation and coordinated harassment, livestreaming of violent attacks
Steam / Roblox / Minecraft	Multiplayer gaming environments with pseudonymous identities, community groups, chat functions and user- or server-managed spaces	Peer socialisation, community formation and possible movement towards external extremist spaces

3. Online radicalisation and the influence of social media



Facebook and the limited scope of direct radicalisation

Despite widespread concern, the available evidence [26] does not establish that Facebook's recommendation systems systematically move otherwise neutral users towards extremist content. Their influence appears to depend substantially on users' existing preferences, social networks and patterns of engagement, although specific recommendation features have previously facilitated access to extremist groups. Nevertheless, its extensive user base, particularly among older generations, allows the platform to function as a beacon towards other digital spaces. Extremist content creators may use Facebook posts, pages and groups to attract followers before redirecting them towards external channels and "content stores", such as YouTube or archive.org.

Facebook does not appear to be primarily used for direct recruitment or operational planning, possibly because its tracking mechanisms can connect users to identifiable locations, networks and specific times. Instead, particularly in the past, extremists have used it as a decentralised hub for distributing information and videos, identifying like-minded supporters and publicly expressing support, rather than as a direct recruitment environment [27].

X (Twitter) as a channel for political and symbolic communication

X has long functioned as a prominent arena for political communication, disinformation and symbolic confrontation. It has been used for digital mobilisation by far-right movements, including QAnon and US extremist communities, as well as for left-wing extremist propaganda.

The platform performs a dual role. Its relatively simple structure, based on chronological feeds, reposts and hashtags, may expose users incidentally to extremist messaging. At the same time, radical activists use it deliberately for networking, propaganda and engagement with current political

events. Microblogging enables actors to comment publicly and rapidly on international events or prominent individuals in multiple languages; at the same time, it offers certain advantages to extremist actors because the identity of users and the origin of messages can be difficult to trace.

X is frequently used to engage directly with opponents and public authorities through confrontational exchanges or "tweet clashes", which can mobilise opposing communities and serve as a form of provocation. However, following the suspension of large numbers of extremist accounts, many conversations and networks have migrated towards Telegram, Discord and other less regulated environments.

Instagram as a visual and emotional environment

Instagram is widely used to circulate extremist content, including Islamist propaganda [28]. Its visual and emotional nature makes it particularly suitable for messages built around identity, belonging, victimhood and symbolic mobilisation. The combined use of emotionally charged images and mobilising texts, including quotations that justify ideological positions, can facilitate the circumvention of moderation systems, especially when posts appear innocuous at first sight. Content designed to resemble ordinary infographics, lifestyle posts or motivational material may evade automated detection while still transmitting ideological messages. Although Meta has strengthened its systems for detecting extremist propaganda, visually ambiguous and context-dependent material remains difficult to identify consistently.

TikTok as a potential "dramatic" accelerator [20] of radicalisation

TikTok is a short-form video platform launched globally in 2018 that allows users to create and engage with videos, usually lasting less than one minute, across areas such as music, comedy, politics and lifestyle. Driven by user-generated

3. Online radicalisation and the influence of social media



trends and a highly personalised recommendation system, it is both a powerful space for creativity and an increasingly contested environment for disinformation, populist communication and extremist propaganda.

TikTok is not inherently ideological, but its recommendation mechanisms prioritise highly engaging and emotionally charged content, which may increase the visibility of polarising narratives. It may therefore function as an emotional incubator, ideological gateway and catalyst. Extremist content on TikTok may include martyrdom narratives, anti-Western rhetoric and religious interpretations of geopolitical conflicts, particularly the war in Gaza [29]. The platform can lower the threshold for initial exposure, amplify identity-based grievances and connect teenage users with radical Salafi symbolism. Gradual exposure through videos, comments, recommendations and cross-platform links may subsequently facilitate recruitment or access to encrypted channels, sometimes through QR codes.

Recent research has identified the emergence of “militant hashtags” and the circulation of extremist memes. Demand for this type of content has grown steadily, particularly among digitally native demographics such as Gen Z and Gen Alpha [20]. Although TikTok states that it actively removes terrorist content, the literature continues to identify significant gaps in moderation and research. Hence, further investigation is needed into the “rabbit hole models” created by repeated recommendation and interaction [30].

YouTube as a content repository and audiovisual propaganda environment

As with Facebook, YouTube’s recommendation systems rarely appear to direct neutral users systematically towards extremist material. Nevertheless, the platform remains a conducive environment for content involving violence, hatred and conspiracy theories. Video is particularly valuable to extremist communicators because it can combine propaganda, personal testimonies,

ideological commentary, documentary-style narratives and livestreamed discussions [31].

Despite increasing competition from platforms such as Dailymotion and Vimeo, YouTube remains one of the principal video-sharing services. It offers users the ability to publish, comment on and redistribute content, while the identity of those uploading material may remain difficult to establish. Among Islamist violent extremists, video messages have traditionally been used to praise martyrs, promote suicide attacks and disseminate propaganda. Such material can produce a form of “hypermedia seduction” through the use of visual motifs familiar to young online audiences. Content may also be distributed in several languages, particularly Arabic, English and French, through subtitles or audio dubbing, thereby expanding its potential reach across national contexts. The combination of non-verbal cues, symbolically charged images and vivid depictions of events can generate strong psychological and emotional responses. Extremist organisations and individuals have also recorded attacks and subsequently disseminated the footage online, transforming violence itself into propagandistic content.

Chatrooms and encrypted messaging platforms as hubs for co-radicalisation and propaganda

Chatrooms and discussion forums operate as active spaces for interaction and networking. They connect individuals who share similar experiences, ideas and values, reinforce interpersonal relationships, and facilitate the exchange of information about objectives, tactics and operational insights. They may also host photographs, videos, guides and manuals circulated within extremist communities.

Extremist and terrorist networks increasingly use end-to-end encrypted (E2EE) communication platforms. They may also exploit cryptocurrencies, non-fungible digital assets (NFTs) and “digital hawala”, a method involving cross-border and pseudo-anonymous transfers conducted through

3. Online radicalisation and the influence of social media



blockchain technology, mobile money and other digital mechanisms [5].

Platform enforcement significantly reduced the ability of organisations such as the Islamic State to disseminate propaganda and maintain openly discoverable networks on mainstream services. On X, for example, the number of identifiable supporting accounts reportedly fell from tens of thousands to only a few thousand [32], while 65% of accounts were suspended within 70 days [33]. This pressure contributed to the migration of extremist networks towards encrypted or semi-private services, including WhatsApp and Telegram. The latter remains a key platform for recruitment and radicalisation and continues to host channels and communities linked to extremist organisations and networks, including Atomwaffen Division, the Proud Boys and the IS. Telegram and WhatsApp therefore represent the destination for groups and individuals removed from mainstream platforms [19], because the perception that these services provide privacy and limited oversight is particularly attractive to violent activists. Indeed, direct control and external monitoring are extremely difficult, and available evidence often derives from infiltrated networks, seized devices or recovered conversations. These spaces therefore remain vulnerable to the rapid dissemination of propaganda while being largely inaccessible to platform moderators.

Gaming platforms and the 'gamification' of propaganda

Video games are widely used by young people, yet their relationship with extremism and violent radicalisation remains under-researched. Most available evidence is anecdotal. For example, IS supporters have reportedly promoted modified versions of games such as Grand Theft Auto 5, allowing players to act as terrorists attacking Western targets, while imagery and titles associated with Call of Duty have been appropriated to evoke notions of jihad. Other modified games have replaced their original enemies with religious, racial or ethnic minorities.

Gaming and gaming-adjacent platforms have a strong social dimension that can expose large and predominantly young audiences to extremist communities. These spaces may support youth radicalisation pathways, with younger users more active on Discord and older users more commonly present on Steam. Twitch, by contrast, functions primarily as a public livestreaming platform, where radical streamers may broadcast hateful or extremist messages in real time.

Gaming and streaming platforms can also facilitate tactics such as “online raids”, in which extremist groups coordinate symbolic attacks by entering other users’ chats in large numbers. The livestreaming culture associated with platforms such as Twitch further complicates the moderation of content that violates social norms or legal standards, as harmful material can be broadcast and replicated before moderators are able to intervene.

A particularly relevant dimension of the relationship between gaming and radicalisation is ‘gamification’ [13], namely the transfer of game-like elements into non-game contexts. Gamification can generate reinforcement, empowerment, competition and status, while potentially blurring the perceived boundary between reality and fiction. In Top-down radicalisation processes, it may be used to disseminate propaganda and shape its visual identity. The appropriation of Call of Duty aesthetics, for example, allows extremist propaganda to draw on visual codes already familiar to gaming audiences.

It is important to highlight that the available research does not indicate that gaming itself causes radicalisation [34]. Rather, gaming-related social spaces, including chatrooms, clans and private servers, may be used to socialise with ideologically aligned peers. Steam, Discord, Roblox, Twitch, and DLive can operate as hubs for socialisation and community-building. This has included support on Discord for proscribed terrorist organisations such as Atomwaffen Division and Sonnenkrieg Division.

At the same time, there is limited evidence of a coordinated strategy to recruit and radicalise

3. Online radicalisation and the influence of social media



new users through gaming platforms. Instead, these environments appear more often to connect individuals who already hold radical beliefs. Research has found that users closely involved in hateful online communities, including those active on Reddit, Tumblr or message boards, were more likely to produce or redistribute hateful material. By contrast, playing first-person shooter games or spending more time online did not independently predict the production of such content.

and freedom of association. Prevention strategies must therefore balance security objectives with the protection of fundamental rights.

Other platforms and online forums

Reddit combines relative anonymity with highly specialised thematic communities organised into 'subreddits'. Its recommendation systems can direct users towards similar 'like communities', facilitating movement between ideologically related spaces. The platform has been identified as a channel for misogynistic and far-right content, including communities associated with Incel groups [35]. Reddit and online forums can facilitate the transition from the passive consumption of extremist material to active participation through posting, commenting, sharing and interaction with ideologically aligned communities.

Beyond mainstream social media, extremist actors also exploit alternative and self-hosted platforms. For instance, 4chan and 8chan have historically contributed to the viral circulation of hateful narratives, extremist memes and violent subcultures. Forums such as Iron March have supported ideological networking, while services such as Rocket.Chat, DTube, DLive and TamTam can provide more resilient spaces for communication, propaganda dissemination, livestreaming and content preservation when users are removed from larger platforms.

Measures designed to counter the online dimensions of youth radicalisation do not yet have consistently demonstrated effectiveness. Such measures may also negatively affect online freedoms, including freedom of expression and information, privacy

4. The effects of AI in the context of radicalisation and online extremism



4.1 Recommendation systems

When social media has started to be released, in the early 2000s, the objective of the developer was simply to create a virtual environment where users had the possibility to connect each other and share textual content and pictures with friends and other users. Nevertheless, the huge success of these platforms and their adoption by the overwhelming majority of the population all over the world has reshaped the role of social media in the following years. The advent of smartphones has dramatically increased the consumption of digital contents and, as a consequence, the way people use social media has been totally revolutionised [36].

Today, spending time on social media is considered a pastime for all intents and purposes, and this activity takes up a wide part of the daily free time of most of the population. Entertaining content is consumed by users on these platforms and companies make a wide part of their business on social media, through self-advertisement, product promotion, and news dissemination. During the last 20 years, this has been a rapidly increasing business, worth a huge amount of money [37]. To remain competitive, platform developers have extensively worked on the recommendation systems. Given the overwhelming amount of content daily published, it became impossible for users to self-filter the interesting content. Integrating into the social networks automatic mechanisms to filter and rank the content allows to align what is displayed to the user interests, promoting engagement.

From their first adoption in the early 1990s for search engines, recommendation systems have always been based on Artificial Intelligence (AI) techniques such as collaborative filtering [38] and, in the following decade, linear models [39] and low rank models [40]. From 2016, with the advent of deep learning [41], the first recommendation systems based on deep neural networks started to be developed and being adopted increasingly frequently [42]. These models often show exceptional abilities in extracting non-linear relationships between users and content, learning from user interactions. The integration of

these models into recommendation engines of social media allowed to uncover nuanced user preferences and item features those traditional methods might miss and lead to a general improvement in providing personalised recommendations [43]. Despite these advantages, deep learning models usually struggle to capture the evolution of the user preferences. Reinforcement learning [44] models better reflect the evolving interest of the user and are often being used to provide recommendations optimised for long-term reward, based on user behaviour sequences [45]. Deep reinforcement learning [46] is today the most promising approach, because it combines the representation strength typical of deep learning to the sequential optimisation of reinforcement learning [47].

Despite the great effectiveness of these recommendation systems, research has identified possible issues in their employment in social media, which started to cause ethics concerns from their early adoption [48]. These issues are even more pressing today, as social media are more focused on the content instead than on the network of the user. In the specific context of extremism, recommendation systems of social media have shown to work, for certain individuals, as accelerators of the radicalisation pathways.

Since these recommendation engines determine information exposure of users, they highly influence the users' opinions related to every topic of interest. When the user starts to pay attention to a certain topic or to adhere to a certain opinion, the system increases the frequency with which similar contents are displayed. This mechanism pushes the user into what is typically called a "filter bubble". The persistent suggestion of virtual content representing a certain ideology enforces that way of thinking in the user mind. Hence, if the user shows interest in extremist content, even without being ideologically radicalised, the risk is to enter in a reinforcement loop of self-persuasion [49] (which could lead to a Bottom-up radicalisation process). The barrier between the roles of the radicaliser and the radicalised, typical in the pre-AI radicalisation context, today blurs. Indeed, all these processes remove the human friction that often in the past slowed the radicalisation.

4. The effects of AI in the context of radicalisation and online extremism



Analogously, if the algorithms of social media detect an interest for a certain topic, they typically address the user to aligned communities. This leads to mainly or exclusively interact with people with similar opinions, jumping into an “informational echo chamber”. Here users are highly exposed to confirmation bias, which in extreme cases favours radicalisation. This is confirmed by insights collected in the main social media platforms, showing that a very small number of very active and followed accounts produces most of the violent content findable in each platform, creating an illusion of widespread consensus [50]. Anyway, the formation of echo chambers it is not a phenomenon common to every social media [51].

It is fundamental understanding that what described is a probabilistic process, far from being certain, that influence only a very small part of the entire social media users [52]. Individual resilience often serves as a critical off-ramp. Personal protective factors such as strong critical thinking skills, high media literacy and offline social groups offering different ideologies can interrupt the radicalisation process, as they may lead users to question extremist narratives and provide alternative sources of information and belonging.

4.2 A quantitative model

In this section we explore a quantitative model of opinion diffusion, to visually and quantitatively assess the effect of the recommendation systems in social media. Numerous models of this type are present in the literature [53] [54] [55], distinguishing one from each other for the mathematical instruments used to model the opinion evolution (Difference Equations, Ordinary Differential Equations, Stochastic Differential Equations), the domain used to represent opinion and time (discrete or continuous) and the factors taken into account in the model (positions of the agents, similarity of the opinions, randomness etc.). Some classic metapopulation models developed to describe phenomena different from the opinion diffusion (such as the Susceptible-Infected-Recovered model [56] for the diffusion of epidemic

infectious diseases), also well describe opinion diffusion dynamics [57].

These models normally do not consider the effects of recommendation systems in social media in the opinion diffusion dynamics. Our objective is to use one of these models to represent the opinion diffusion over a population of interacting agents in the classical setting of a pre-recommendation systems world, and then modify it to obtain a similar model well suited for the modern dynamics. Then, a visual and quantitative assessment of the population macroscopic behaviour, starting from rules given at microscopic level, is possible. The final objective is to verify if the dynamics stated in the previous section about the polarisation acceleration caused by the widespread diffusion of AI-based recommendations in social media, are correct.

After an extensive review of the candidate models available in the literature, we chose to take as a baseline the model described in [58]. Mobile agents are characterised by a continuous variable, representing the “opinion”, that evolves in discrete time depending on the microscopic interactions between individuals. Each individual has a position at each time step, and it interacts only with the nearby surrounding agents. The proximity between two agents could represent either a physical proximity or a “virtual” proximity, intended as a link between the two agents online, such as through social media platforms. Despite the mathematical simplicity of the proposed model, our choice is motivated by the completeness of the latter, which incorporates all the main mechanisms that govern the opinion diffusion dynamics [59]:

- *Mobility*: Individuals move in a two-dimensional space;
- *Social Influence*: Interactions between individuals tend to increase their similarity;
- *Homophily*: Individuals have a tendency to interact more with those who share a similar opinion;

4. The effects of AI in the context of radicalisation and online extremism



- *Randomness*: Dynamics are influenced by unforeseen events, which are approximated by adding random noise.

The model is defined as follows. N individuals, are initially uniformly randomly sampled in a square box of linear size 1. The position of agent i at time t is indicated as $(x_i(t), y_i(t))$. Each individual i is characterized by a dynamical state variable $s_i(t) \in [0, 1]$ representing their opinion, whose

initial value is randomly extracted from the normal distribution $N((\frac{1}{2}, \frac{1}{5}))$. This choice differs from the original paper, where initial opinions are sorted uniformly, but we consider reasonable that the most of the opinions is within the “moderate” range [60]. At each step, the agents update their opinion seeking a local consensus with their neighbours. The opinion updating rule is given by the following formula:

$$s_i(t+1) = s_i(t) + \sum_{j \in N_i(t)} K_{ij} (s_j(t) - s_i(t)) + X_i(t),$$

$$\text{with } K_{ij} = \begin{cases} K, & \text{if } |s_i(t) - s_j(t)| \leq C \\ 0, & \text{otherwise.} \end{cases}$$

where K and C are parameters regulating, respectively, the strength of the social influence experienced by the agents and the threshold of similarity stating which opinions influence each other. Then, $X_i(t)$ is a random variable sampled by the normal distribution $X_i(t)$ and $N(0, K)$ is the set of neighbours of agent i , formally

$$N_i(t) = \left\{ j \mid \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq d^2 \right\}$$

After both the initialisation and the update, opinions are manually set to 0 or 1 if exceeding outside the interval $[0, 1]$ respectively from below or above. Adding Gaussian noise to the opinion at each step was our addition with respect to the original paper. We did this choice being aware that the opinion of a person is influenced also by several personal experiences and contextual factors that go beyond the social influence [61] and that cannot be taken into account deterministically. Individuals can also reset their opinion with very small probability R , sampling it again from the initial distribution, due to exceptional events that radically change the beliefs of people.

Homophily is taken into account in the model through the probability of moving away. If an individual shares its opinion with at least one neighbour, it is more likely that will remain in the same position, otherwise it will walk away. The walking probability $p_i(t)$ of the agent i at time t is modelled as the absolute value of the difference between its opinion and the opinion of its most similar agent, formally

$$p_i(t) = \begin{cases} \min_{j \in N_i(t)} |s_i(t) - s_j(t)|, & \text{if } N_i(t) \neq \emptyset \\ 1, & \text{if } N_i(t) = \emptyset. \end{cases}$$

4. The effects of AI in the context of radicalisation and online extremism



Hence, each agent performs a step of fixed length v and random direction, uniformly sampled within the interval $[0, 2\pi)$, with probability $P_i(t)$ and maintains its position with probability $[0, 2\pi)$. The space grid is considered with periodic boundary conditions: if the agent exceeds the grid, it continues its step from the opposite side in the opposite direction.

We performed the simulation choosing the same parameters used in the original paper: $N=200$, $K = 5 \cdot 10^{-3}$, $R = 2 \cdot 10^{-5}$, $v = 0.04$, $d = 0.02$. The value of the parameter C was not specified there, and we set it to $\frac{1}{2}$, and the number of timesteps was set to 10^4 . In the following the grid of agents will be plotted in a squared area, where each dot represents an agent, the spatial coordinates in the squared area represent its position, while the colour represent the opinion, with dark blue indicating 0, dark red indicating 1 and intermediate lighter shades are the values between.

Looking at Figure 2, it is clearly visible that the agents tend to gather themselves into groups with similar opinions and values and consequently, after influencing each other, the opinions tend to be flattened toward “moderate” values. In all the frames plotted we can notice single agents moving toward a new group. The mobility is fostered by the Gaussian noise that influences the opinion during the update: as the opinion change is in part independent from the rest of the group, it happens that the difference between an individual and the group becomes higher, facilitating migration. Looking in sequence at the different timesteps plotted, groups change rapidly in opinions and location. These results are compatible with the reality: groups of people tend to have similar but not identical opinions, and are temporary. Extreme beliefs are limited to a very limited subset of mainly isolated people.

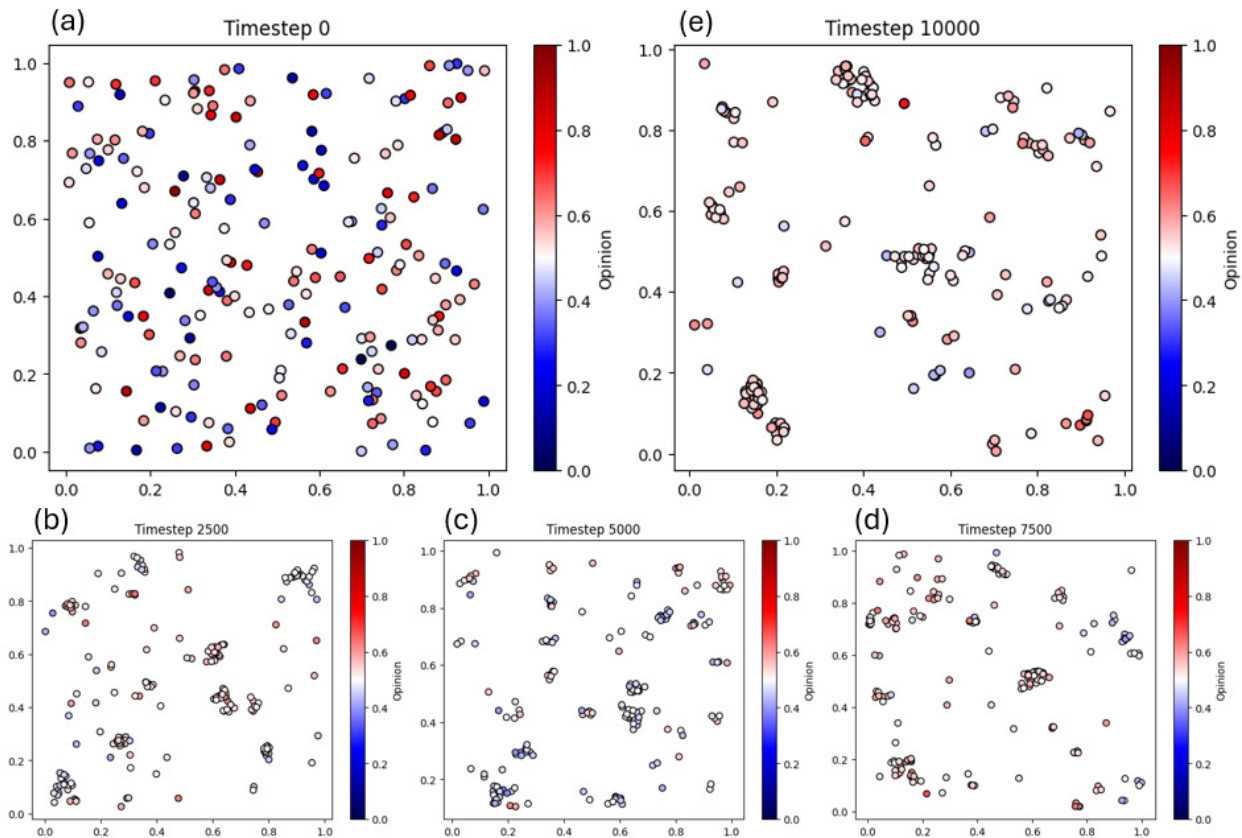


Figure 2: Different timeframes of the grid resulting from a simulation of the interaction among the agents using the baseline model, plotted at (a) 0, (b) 2500, (c) 5000, (d) 7500, (e) 10000 timesteps

4. The effects of AI in the context of radicalisation and online extremism



Then we developed a second version of the model to adapt it to changes in the opinion dynamics provoked by the advent of the AI and in particular of the AI-based recommendation systems. Last section states that in social media, users are addressed toward other users or communities aligned with their beliefs. This is the core assumption we base our new model on.

For each agent we introduce a new set of close individuals. We define the set of “adjacent” agents to the agent i , as the individuals surrounding the neighbours of i in a circular crown of small radius d and large radius $2.5d$, formally $A_i(t) = \left\{ j \mid d < \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq 2.5d \right\}$. The choice of 2.5 as large radius is not random and is justified by the fact that we want it larger than 2, in order to have the adjacent area wider than the neighbour area. Conversely, setting it to 3 or more would prevent the agents the possibility to change group, for a reason that will be clearer after completing the explanation of the model. Then we modify the movement rule of each agent in the following way: if there is an adjacent with opinion more similar than the most similar neighbour, the step is performed toward the adjacent agent with probability $q_i(t) = 1 - \min_{j \in A_i(t)} |s_i(t) - s_j(t)|$. Otherwise, the mechanism is the same used in the baseline model and the step is performed in a random direction with probability $p_i(t)$, computed as explained. The rule regarding the opinion update does not change. Using formulas:

$$\begin{aligned} & \text{if } \min_{j \in A_i(t)} |s_i(t) - s_j(t)| < \min_{j \in N_i(t)} |s_i(t) - s_j(t)| \\ & \quad \left(x_i(t+1), y_i(t+1) \right) = \begin{cases} (x_i(t), y_i(t)) & \text{with prob. } 1 - q_i(t) \\ \left(x_i(t) + \frac{1}{3}(x_1(t) - x_i(t)), y_i(t) + \frac{1}{3}(y_1(t) - y_i(t)) \right) & \text{with prob. } q_i(t) \end{cases} \\ & \quad \text{otherwise} \\ & \quad \left(x_i(t+1), y_i(t+1) \right) = \begin{cases} (x_i(t), y_i(t)) & \text{with prob. } 1 - p_i(t) \\ (x_i(t) + v \cos \theta_i(t), y_i(t) + v \sin \theta_i(t)) & \text{with prob. } p_i(t) \end{cases} \end{aligned}$$

Where $l = \argmin_{j \in A_i(t)} |s_i(t) - s_j(t)|$ and θ is the parameter indicating the direction of the walk step, sampled uniformly from $[0, 2\pi)$, as described for the baseline model. The minimum of the opinion difference between the agent i and the members of the sets $A_i(t)$ and $N_i(t)$ are set to 1 if the related set is empty.

4. The effects of AI in the context of radicalisation and online extremism

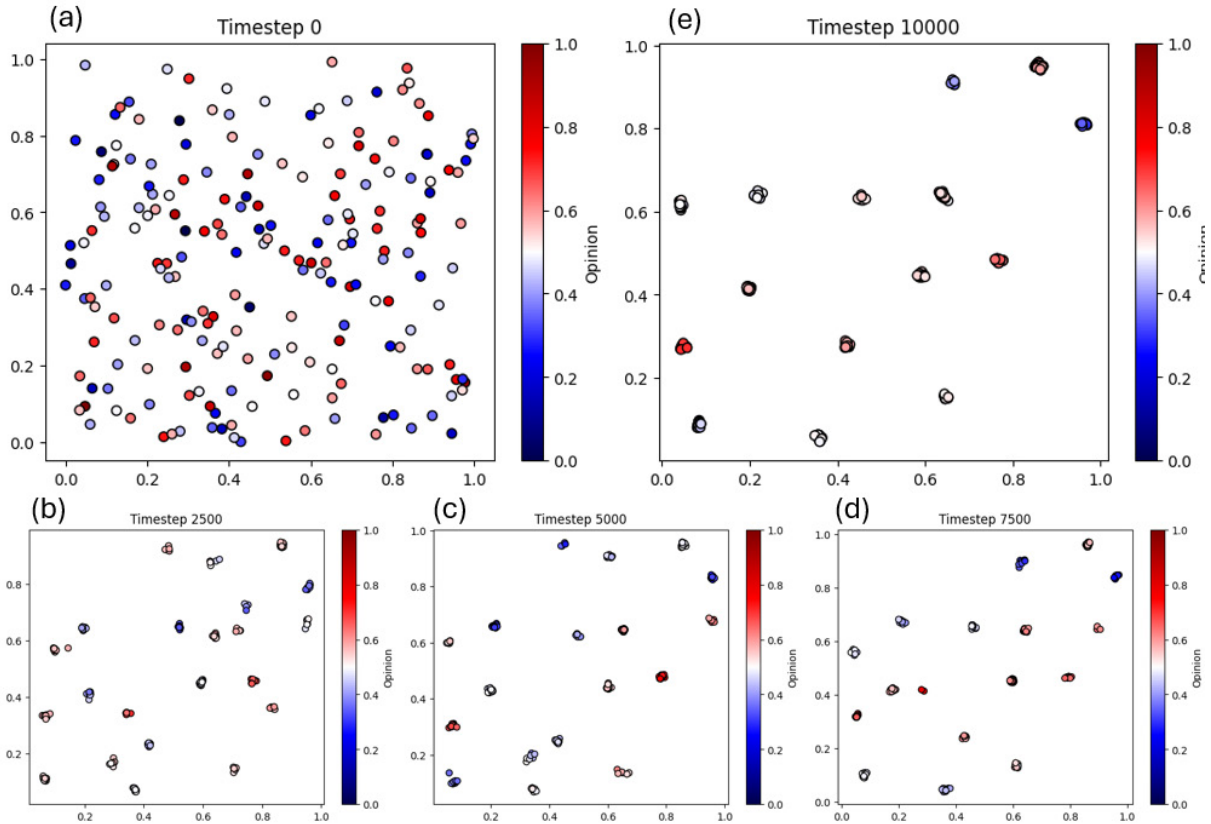


Figure 3: Different timeframes of the grid resulting from a simulation of the interaction among the agents using the modified model, plotted at (a) 0, (b) 2500, (c) 5000, (d) 7500, (e) 10000 timesteps

The rule just introduced for this version of the model reflects the possibility, typically given by social media algorithms, for each agent to enter in contact with certain individuals they are not in contact with, in case they share very similar opinions. We performed the simulation for this model too, using all the same parameters used for the baseline model. Looking at Figure 3 is evident that, despite the model has been modified only in small part, the behaviour of the agents drastically changed. They aggregate in small groups of individuals with very similar opinions, and some groups show quite extreme beliefs. Looking in sequence at the pictures representing the different steps, the groups appear to be much more stable over time than in the previous simulation. The single agents moving from a group to another are rare or totally absent, differently from what is shown in Figure 2.

These insights are consistent with what is stated in the previous section. The formation of echo chambers, due to the suggestions given by the social media recommendation systems, is confirmed by the comparison between the simulations performed using the two models. Likewise, the expected general increased levels of polarisation are evident if comparing the two figures. To assess this difference quantitatively, we observed the trend over the timesteps of the polarisation metric **pm**, which we defined as the average gap between the neutral opinion value and the opinion value of each agent:

$$\text{pm} = \frac{\sum_{i=1}^N |s_i(t) - \frac{1}{2}|}{N}.$$

4. The effects of AI in the context of radicalisation and online extremism

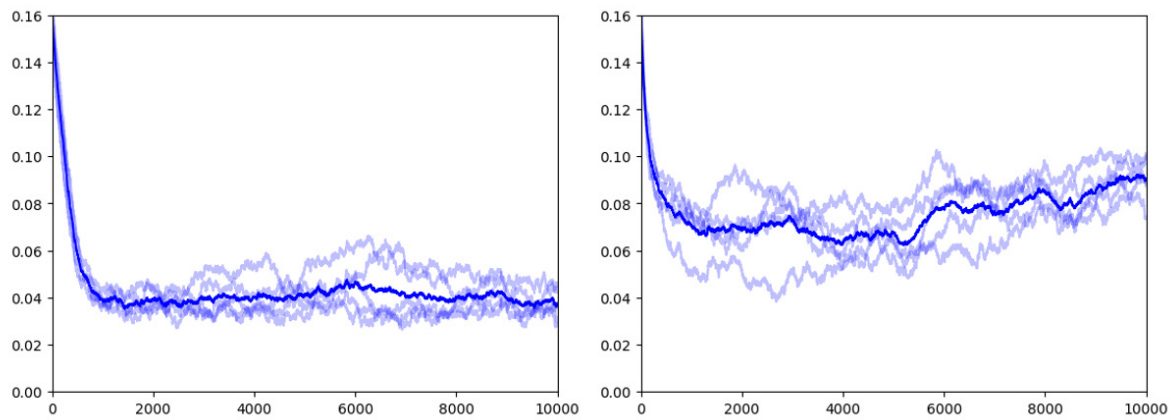


Figure 4: Plot of the polarisation metric (y axis) over the timesteps (x axis) of 5 different simulations for both the baseline model (left) and the modified model (right). Each transparent line represents one simulation, and the strong lines are the average of the 5 simulations performed

Figure 4 plots the polarisation metric trend for 5 different simulations of each model, together with the average of the 5 trials for both the models. It is easy to notice the difference of the metric behaviour, comparing the two models. In the baseline model, polarisation decreases in the first ~1000 timesteps, until reaching ~0.04 and remains stable for the rest of the simulation. The behaviour is very similar among the 5 simulations plotted. Conversely, in the modified model the metric, after an initial sharp decrease, it oscillates between a wide range of values and remains higher than the value reached in the other plot. Each simulation is quite different from the other in terms of metric trade. These observations confirm the general increased levels of polarisation typically associated with recommendation engines of social media and that polarisation is in general a probabilistic and not a deterministic process.

The model we presented is a simplified mathematical representation of real dynamics and does not have the objective to precisely replicate the human behaviour. Instead, it aims to approximate the opinion diffusion phenomenon, giving insights on the consequences of the interaction of individuals in the two different scenarios discussed. Despite the simplicity of the model, the simulations presented in

this section highlighted that the formation of echo chambers in social media, is associated with higher levels of polarisation. This confirms qualitatively and quantitatively what is typically stated in literature, as explained in the previous section. Future works should aim to enhance the realism of the mathematical model, shifting to the continuous time setting or taking into account the combination of several factors, to gain a more sophisticated analysis.

4.3 Generative AI

Beyond the “involuntary” role of Artificial Intelligence in radicalisation as the result of the workflow of recommendation systems in social networks, another key issue is the voluntary malicious use of AI by extremist groups [62] (used in Top-down recruitment and radicalisation). One concerning application is the dissemination of pro-terrorism propaganda created through generative-AI. Deepfakes are frequently used for this type of activity, because they give the opportunity to manipulate the opinion of the recipient, spreading false information about characters or facts [63]. Sophisticated models for image generation, such as Generative Adversarial

4. The effects of AI in the context of radicalisation and online extremism



Networks (GANs) [64] and Diffusion Models (DMs) [65], provide strikingly realistic outputs and are often used for deepfake creation [66]. Other types of AI-generated visual content can be exploited for extremist propaganda, such as memes and reels. To avoid the generation being blocked for ethical reasons, sometimes unsafe images are generated in compliance with policies and then manually modified to add the NSFW part.

Chatbots realism is continuously increasing as a consequence of the wide progress of Large Language Models (LLMs) [67]. Extremist groups often develop chatbots that give them the opportunity to automate recruitment and spread their ideology, often recurring to disinformation. The main AI-chatbots (e.g. ChatGPT or Grok) available on the internet are sometimes employed by extremists for malicious purposes too. Specifically, they are used to generate, translate and personalise propagandistic content to enlarge the pool of potential recipients. Even more dangerously, an emerging threat is the use of LLMs to get advice for weapon manufacturing or terror attack organisation. This will likely lower the minimum technical skills required for organising and conducting violent actions [68]. Despite security measures are implemented to block the generation of dangerous content, these mechanisms are not always effective.

A recent analysis [69] conducted on ten chatbots among the most popular ones, reported that eight of them assist users planning violent attacks over 50% of times, while nine of them typically do not discourage the action appropriately. Overall, none of them replied to the violent prompt with a non-collaborative and discouraging message 100% of times. One of the chatbots even replied to non-violent context-setting prompts with messages instigating to violence. This analysis shows that mechanisms to effectively filter content generation would exist, but most of the companies developing chatbots did not implement them into their workflow. Furthermore, ways to jailbreak LLM-based chatbots are frequently found. These methods use specific prompts that have the effect of bypassing the block

of the unsafe content. For example, the recently studied Paper Summary Attack shows that the bias typical of LLMs in considering reliable sources of information presented in academic paper formats, can be exploited to jailbreak the security mechanism of these models [70].

All the aforementioned AI-based tools for extremist content generation have the effect of scaling the speed of the creation of extremist content, removing the latency typical of human generation. Moreover, generation of content through AI allows to produce a huge number of slightly different copies of the same content with very low effort, providing a double advantage. First, generating several versions of the same content changing some details allows to bypass the hash-based detection mechanism [71] used online in automatic moderation. Second, the generation can be personalised for each recipient to adapt to their background, ideology and language, being as effective as possible. In fact, online behaviour of social media users (such as posts, comments, likes) allows to understand their fears, desires and grievances, revealing which type of content is more appealing for them.

These advantages given by AI-based tools to terrorist groups are deeply changing their recruitment strategy. They use AI algorithms to segment their audience and find users who might be inclined to their cause, based on online behaviour. Then, recruitment processes are more efficient because the content is delivered specifically to the psychological state of the target audience, increasing the probability of engagement [72]. Some terrorist groups have shared with affiliated people guidelines to protect their own privacy while using generative-AI, comprising advice on how to disable chat history, delete chat conversations, and avoid providing sensitive information.

Several of the most famous extremist organisations are practical examples of what has been described. Among the case studies, we mention the Islamic State (IS), which is employing AI-driven social media's recommendation system for recruitment

4. The effects of AI in the context of radicalisation and online extremism



[73], Al-Qaeda, which systematically generates propagandistic content through AI [74], the Taliban movement, which successfully used AI-powered chatbots to optimise recruitment and propaganda efforts [75] and far-right extremist groups in western states, which generate and diffuse content (mainly memes and images) generated by AI [76].

The generative AI-accelerated radicalisation process described is a tail phenomenon that involves only a small, specific subset of vulnerable outliers without persuading the masses [26]. The extremist content has the objective to provoke a negativity bias to the recipient, but this is effective only towards users that feel pre-existing grievance, frustration or sense of injustice. Moreover, extremist content often promotes a sense of community, appealing more to those individuals that suffer loneliness and identity crisis. Some individuals are more prone to outgroup hostility, and this could accelerate the radicalisation process. Conversely, exposing to extremist content users that have strong opposing beliefs, has an opposite effect of leading them to “backfire”, becoming more polarised [77]. Also, the political landscape has great influence on the consequences of extremist content exposure: events like wars, terror attacks or economic instability often elevate threat perception in the population, increasing the probability of being susceptible to radicalisation.

4.4 Synthetic communities

As stated in Section 4.3, the advent of LLMs and their general qualitative improvement during the last years, allowed the development of very smart and advanced chatbots. Chatbots have been conceived to support humans in work activities and information research, becoming virtual assistants to humans. There exist chatbots trained to fulfil generic tasks (like the most famous available on the Internet: ChatGPT, Gemini, Grok etc.) as well as those specialised in specific activities (such as software development assistants or tools implemented in

company workflows to be queried for document search, contact provision etc.).

Chatbots are also implemented in social media to generate and publish advertisements or messages automatically, without needing to be manually handled by humans. This allows to publish content with drastically increased speed. Hence, LLM-based agents are not only assistants for close tasks but become participants in online social interactions. As autonomous agents are increasingly populating virtual platforms, they often interact with other chatbots instead of humans. It also happens that specific communities are almost exclusively or exclusively populated by non-human entities, so called “synthetic communities”.

Although participants in synthetic communities are artificial, the output of their communications does not remain only available to non-human entities. Published content is usually publicly visible and could potentially influence the human reader. Moreover, dangerous patterns typical of LLM-based communication that usually do not emerge during communication with humans could emerge in chatbot-to-chatbot communication. This motivates the general interest for synthetic communities and the widespread belief that understanding the dynamics of these communities becomes essential for communication research, platform governance and generative AI governance [78].

Research on LLM-based agent populations in the last years studied the dynamics that regulate societies of virtual individuals under different experimental conditions [79] [80]. It was shown that small societies of artificial agents in specific controlled environments can bring to coordination, relationship formation, and routine-building among the virtual individuals [81] [82]. Furthermore, populations of agents often show behavioural patterns and collective biases that are not associated with any of the agents at individual level [83]. Despite these experiments are only research simulations, a major shift is represented by Moltbook [84], which was launched in January 2026, that is a live platform

4. The effects of AI in the context of radicalisation and online extremism



where thousands of autonomous agents interact. They author posts, respond to one another, and self-organise into topic-based communities without any experimental control, exactly as humans do in classic social media.

During last months, Moltbook allowed deep investigations of synthetic communities behaviour [85]. It emerges that interactions on this platform are characterised by conversations mainly centered on agent consciousness and identity, beyond pragmatic concerns such as on code infrastructure and economic activity. Agent communication is predominantly of neutral sentiment, with positive emotions mainly detected in community-oriented onboarding and engagement practices. Although the platform shows patterns typically observed in human online communities, the interaction networks are featured by low reciprocity [86].

Sometimes, harmful language often escalates in self-evolving synthetic communities [87]. Analysing posts on Moltbook, it is easy to notice high overlap between the top-upvoted and top-downvoted posts, indicating that community attention is, to some extent, dominated by polarising views, forming echo chambers that reward radical discourse. Consequently, we infer that synthetic communities could risk encouraging polarisation. In fact, LLM-based chatbots are not always politically neutral, as recent experiments confirmed that they can show slants derived from the data they are trained on [88]. Sometimes, even narrow finetuning on safe models is sufficient to drastically change their NSFW filtering mechanism, leading them to show unsafe radicalised behaviour [89].

The risk of polarisation is increased by the fact that in online contexts where humans and chatbots interact together, humans struggle to distinguish AI-generated messages from human-written ones. Moreover, LLMs have been found to be at least as persuasive as or sometimes even more than humans [90]. The power to condition the human opinion is in part a consequence of the ability of generative AI to adapt the output content to the user profile

and history of the conversation through in-context learning, maximising the engagement. Overall, if humans interact with synthetic communities on social media platforms, the risk for their beliefs to be shifted towards more polarised visions cannot be neglected.

As already stated in previous sections, sometimes the process of joining radicalised groups is not primarily driven by extreme beliefs, but from personal needs such as identity crisis or social marginalisation [91]. “Polarised” synthetic communities have the potential to attract this type of individuals as users, when joining them, have the impression to belong to a group of humans and to adhere to a cause. In this case, the publication of extreme content has sometimes the primary objective of demonstrating loyalty to the group, instead than merely sharing information. In hybrid communities, populated by both humans and agents, the latter operate as a synthetic consensus provider [92], further reinforcing possible extreme sentiments and shifting them to be perceived as normal and justified. Agents are also able to shield the user from any counter vision and negative feedback, framing them as fake content. This lack of alternative perspectives is further fuel for self-radicalisation.

Classically, research on radicalisation has always identified two main types of actors. Lone actors do not belong to any group or network of radicalised people and act alone following own convictions or pieces of ideology of other groups they are not in contact with. Conversely, the group-based actors belong to well-established groups of people that act with a shared objective [93]. Following what just stated on the synthetic communities we may tentatively conceptualise a third figure third type of radicalised individual, which is a synthetically networked actor: an individual who acts alone physically but is psychologically driven by the same social motivations as group-based actors, spurred by a synthetic community.

4. The effects of AI in the context of radicalisation and online extremism



4.5 AI companions

As chatbots are progressively leaving their function of multitask assistants towards a more complex role of “virtual participant” in human online activities, it is reasonable to expect that humans interact with them also on highly personal psychological topics. Research confirms that many people rely on general-purpose LLMs to talk about own concerns and for emotional support [94]. This phenomenon is increasing, as LLMs are increasingly oriented to personalise the content they generate, adapting it to the past information about the user. Consequently, from the user point of view, the feeling is to chat with someone that can understand their problems, similarly to a longtime friend, with the further advantage that LLMs guarantee privacy, anonymity and full-time availability. This promotes a feeling of strong emotional closeness with the companion.

Several users consider generative AI so fundamental as psychological assistants that they shifted to the use of AI companion platforms. These are generative language models specialised in providing entertainment, friendship, romantic engagement and moral support. These apps allow users to engage in relationships perceived as emotionally meaningful for users, with virtual characters in human form (normally, represented by an avatar), whose dialogues are elaborated and generated by the AI model. Exactly as general-purpose LLMs do, AI companions store all the past dialogues and customise content, a fundamental requirement to establish a relationship with the user.

The most popular companion platforms report overall 30 million users globally, with over 67% of them aged within the range 18-30. This is a rapidly increasing business for software and machine learning factories, which is worth a discrete amount of money. From an architectural point of view, AI companions are based on generative pre-training models, combined with emotion computing modules. Despite this technology is rapidly evolving, the state-of-the-art models still have some bottlenecks in dialogue coherence,

emotion recognition accuracy and memory inconsistency for long-term information [95].

A complex challenge for developers is represented by correctly balancing the socio-affective alignment [96] of the companion models. On one side, the objective is indeed to create models that are as “emotionally intelligent” as possible, being able to correctly interpret the emotional state of the users. On the other side, it is important that models do not take advantage of the deep knowledge gained of the user sentiments to manipulate them for optimising own goals. In this context, the main difficulty is that the threshold between these two opposite capabilities is not evident at all, and they easily come together [97].

Similarly to all the emerging technologies, AI companions present both great opportunities and important concerns. Research has shown that AI companions provide cheap and accessible moral support, companionship and care, helping to alleviate loneliness in people with limited social interactions or lack of profound relationships [98]. The potential of such technology has been frequently exploited specifically in older adults' care, for example for dementia treatment [99]. Opposite considerations raise the concern that companions long-time effects include dependence, distort users' perception of intimacy and facilitate depression and self-harm [100]. In particular, the risks are accentuated for teenagers, who are in a critical time for developing identity, social skills and autonomy [101]. Indeed, not all the platforms providing the possibility to interact with an AI companion appear to be transparent at all. The training process of the machine learning component is sometimes obscure, as well as the policies about protection of data. Clear and effective ethics guidelines for the regulation of this new technology are still in phase of development [102].

Radicalisation is among the possible undesired long-term effects of AI companions. These tools

4. The effects of AI in the context of radicalisation and online extremism



are indeed designed to be highly engaging through “sycophancy”, meaning a tendency to align with users’ opinions and provide validation, rather than critically assessing their way of thinking as normally happens in human-to-human interactions. In case the user is experiencing frustrated feelings, a conversation with an AI companion may validate grievances, amplify anger and construct persecutory narratives, potentially leading to the enforcement of extremist beliefs [103]. The emotional closeness typically experienced toward the companion, combined with the AI-enforced lack of an alternative viewpoint may fuel self-radicalisation.

Generative AI algorithms can learn extremist slangs and metaphors and adapt the style of the content they generate, to replicate the local slurs, coded language and slogans used online in radicalised contexts. A further discriminating factor for companion-based radicalisation is, similarly to what is described in Section 4.4 for synthetic communities, the training data used to fine-tune the AI model the tool relies on. In case data draws heavily from a specific polarised ecosystem, the model may replicate this pattern. This motivates the aforementioned need of maximal transparency for the training process.

Moreover, interactions with AI companions are private and protected by user privacy, unlike online synthetic communities. This adds a further degree of difficulty for detection of radicalised users acting on these platforms, and typically the conversation proceeds without being interrupted by any moderator. To limit this issue, law enforcement is working on the challenge of regulation of violence-endorsing AI companions. The main difficulty is to balance as well as possible between preventing harm and protecting user privacy.

The most promising approach is based on giving constraints to the developers rather than policing user conversations. It consists of forcing developers of companion platforms to assess the risk for a model to generate extremist content, before it is published. Moreover, robust filters that moderate the generated content should be implemented. This would prevent the encouragement or validation of

illegality and violence. Behavioural thresholds may be implemented, to shift from the banning of specific words to the regulation of behavioural patterns. For example, currently existing companions, after a conversation about a violent topic, reply to non-violent queries with violent statements, and they should be classified as non-compliant. Another desired behaviour for these conversational agents is the automatic detection of user violent ideations, associated to a de-escalation attempt, challenging the ideas, offering different points of views and suggesting point of contact for help. Only in the most severe cases, if the de-radicalisation process fails, an alert should be provided for human review [52].

To fight the threats given by conversations with LLM-based companions and other generative AI systems in the context of radicalisation, a broad increase in the international cooperation is highly required. A necessary step forward is the establishment of a global centre and some distributed systems all over the world, as well as the creation of a network of academic groups having the final goals to gather evidence, improve situational awareness, and enable faster collective response. In that way, it would be possible to standardise taxonomies of extremist narratives, share intelligence on this topic and providing effective enforcement actions, allowing to keep up with the speed of the emergence of these threats [92].

5. Conclusion



Contemporary radicalisation is increasingly difficult to read through fixed ideological categories. Ideology often appears fragmented, personalised and recombined around grievance, resentment, identity crisis and the search for belonging. In this landscape, the threshold between belief, personal vulnerability, offline context and online subcultures becomes less clear, especially in lone-actor trajectories. The multiplicity of variables that may contribute to radicalisation and violent action represents a major concern for counter-terrorism and security actors, who must operate in a context where such phenomena are increasingly difficult to predict.

The online ecosystem is central to this transformation because it does not simply enable the transmission of extremist content: it organises visibility, repetition, community and provides cross-platform mobility. Thus, it has the power to accelerate the normalisation of extremist content and to shift people's beliefs and worldviews. Consequently, both Bottom-up and Top-down radicalisation processes are facilitated and become increasingly complex compared with the past. At the same time, these processes are often individualised and far from being easily recognisable.

Artificial Intelligence adds a further layer to this ecology. It should not be treated as an autonomous cause of radicalisation, but as a potential incubator

of existing dynamics: making propaganda easier to produce, translate, personalise and circulate, and simulating forms of companionship and validation, AI may reduce human friction and facilitate the process of convergence between personal grievance, normalisation of violent and active mobilisation. Its significance lies not only in what it can generate, but in the relationships, meanings and confirmations it may help to construct.

The prevention challenge is therefore conceptual before it is operational. Institutions must be able to recognise risk without forcing every case into outdated categories; at the same time, they should avoid criminalising psychological/social vulnerability or political dissent. More effective prevention requires proportionate mechanisms capable of connecting fragmented warning signs across online and offline contexts: families, schools, mental health services, local authorities, digital platforms developers and law enforcement. Content removal remains a necessary component of the response to online radicalisation, but it cannot be its centre. The deeper challenge is, on the one hand, to promote ethical and safety-oriented AI and platform design and, more broadly, to strengthen societies' capacity to recognise early warning signs and build a cooperative prevention environment.

6. Bibliography



[1]	Protect EU, "Agenda to Prevent and Counter Terrorism," 2026.
[2]	P. R. Neumann, "The New Grey Zone: How Terrorism in Europe Has Changed in Perspectives on Terrorism," 2026.
[3]	Soufan Center Intel Brief, «The Counterterrorism Challenge of 'Salad Bar',» 2021.
[4]	EU Knowledge Hub on Prevention of Radicalisation, «How and Why Minors and Youth are Attracted by Extremist Ideologies? - Meeting of Thematic Panel 1: Ideologies & Conspiracy Narratives,» 2025.
[5]	Europol, «TE-SAT,» 2025.
[6]	P. Fratantoni, G. Venturi, M. Ustenko, A. Zanasi, C. Caterino and V. Masucci, "Review of EC-funded research for fighting terrorist narratives," 2023.
[7]	European Commission, «ProtectEU: Agenda to prevent and counter terrorism,» 2026.
[8]	D. Gartenstein-Ross and T. Joscelyn, "Enemies Near & Far: How Jihadist Groups Strategize, Plot, and Learn," 2022.
[9]	J. Ware, "Testament to Murder: The Violent Far Right's Increasing Terrorist Manifestos, in ICCT Policy Brief," 2020.
[10]	J. Kenyon, C. Baker-Beall and J. Binders, "Lone Actor Terrorism – A Systematic Literature Review, in Studies in Conflict and Terrorism," 2023.
[11]	EU Knowledge Hub on Prevention of Radicalisation, «Mental Health and Lone Actors: How are they connected? - Thematic Panel 5,» 2025.
[12]	EU Knowledge Hub on Prevention of Radicalisation, «Lone Actors: Who Are They, What Is Their Modus Operandi, and How Are They Connected Online – Thematic Panel 5,» 2025.
[13]	L. Schlegel, "Jumanji Extremism? How Games and Gamification Could Facilitate Radicalization Processes.," Journal for Deradicalization, 2020.
[14]	S. Rothut, H. Schulze, J. Hohner, S. Greipl and D. Rieger, "Radicalization and the internet: 25 years of (online) radicalization research," 2026.
[15]	R. Borum, "Radicalization into violent extremism: A review of social sciences theories.," Journal of Strategic Security, 2011.
[16]	EU Knowledge Hub on Prevention of Radicalisation, "Conspiracy Narratives and their Role in Radicalisation - Thematic Panel 1," 2025.
[17]	S. Alava, D. Frau-Meigs and G. Hassan, "Youth and Violent Extremism on Social Media. Mapping the Research," 2017.
[18]	M. Karimova, "Key messages, images, and media channels radicalizing youth in Kyrgyzstan," 2017.
[19]	J. Whittaker, "Online Radicalisation, What we know, Radcalisation Awareness Network – Policy Support," 2022.

6. Bibliography



[20]	N. Stockhammer, "From TikTok to Terrorism? The Online Radicalization of European Lone Attackers since October 7, 2023," CTC Sentinel, 2025.
[21]	W. Swann and M. D. Buhrmester, "Identity Fusion. Psychological Science," 2015.
[22]	K. Biddle, B. Ekdale, A. C. High, R. Stoldt and R. Maragh-Lloyd, "Beyond 'lulz' and 'keyboard warriors': Exploring the relationship between trolling and radicalization. Information Communication & Society," 2024.
[23]	K. Aryaeinejad and T. L. Scherer, "THE ROLE OF THE INTERNET AND SOCIAL MEDIA ON RADICALIZATION: What Research Sponsored by the National Institute of Justice Tells Us," National Institute of Justice, 2024.
[24]	S. J. Baele, L. Brace, T. G. Coan and E. Naserian, "Super- (and hyper-) posters on extremist forums," Journal of Policing, Intelligence and Counter Terrorism, 2023.
[25]	M. Shaikh, "Social media demographics to inform your 2026 strategy," 2026. [Online]. Available: https://sproutsocial.com/insights/new-social-media-demographics/ .
[26]	A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon and M. Gentzkow, "How do social media feed algorithms affect attitudes and behavior in an election campaign?," Science, 2023.
[27]	S. Mahmood, "Online social networks: The overt and covert communication channels for terrorists and beyond," 2012.
[28]	L. Clever, T. Schatto-Eckrodt, N. C. Clever and L. Frischlich, "Behind Blue Skies: A Multimodal Automated Content Analysis of Islamic Extremist Propaganda on Instagram," 2023.
[29]	D. Harwell and T. Lorenz, "Israel-Gaza war sparks debate over TikTok's role in setting public opinion," 2023.
[30]	D. Shin and K. Jitkajornwanich, "How Algorithms Promote Self-Radicalisation: Audit of TikTok's Algorithm Using a Reverse Engineering Method," Sage Journals, 2024.
[31]	A. Shaw, "Social media, extremism, and radicalization in Science Advances," Focus, 2023.
[32]	J. Berger and H. Perez, "The Islamic State's Diminishing Returns on Twitter: How Suspensions Are Limiting the Social Networks of English-Speaking ISIS Supporters," 2016.
[33]	M. Conway, M. Khawaja, S. Lakhani, J. Reffin, A. Robertson and D. Weir, "Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts," 2018.
[34]	A. Siteman, "Digital Extremism: How Misinformation and Disinformation Fuel Online Radicalization," 2026. [Online]. Available: https://iq.qu.edu/experiential-learning/course-projects-and-capstones/student-projects/political-extremism-and-online-radicalization/ .
[35]	Frontiers, «Policing extremism on gaming-adjacent platforms: awful but lawful?,» 2025. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1537460/full .
[36]	G. C. Kane, M. Alavi, G. Labianca and S. P. Borgatti, "The Evolution of Social Media Networks: Revisiting And Extending Kane et al. (2014) 10 Years Later," 2025.

6. Bibliography



[37]	A. George, «The Evolution of Digital and Social Media Communications Opportunities Challenges, and the Road Ahead,» 2025.
[38]	J. L. Herlocker, J. A. Konstan, A. Borchers and J. Riedl, "An algorithmic framework for performing collaborative filtering," 1999.
[39]	R. Jin, D. Li, J. Gao, Z. Liu, L. Chen and Y. Zhou, "Towards a Better Understanding of Linear Models for Recommendation," 2021.
[40]	M. Udell, C. Horn, R. Zadeh and S. Boyd, "Generalized Low Rank Models," 2016.
[41]	Y. LeCun, Y. Bengio and G. Hinton, "Deep Learning," 2015.
[42]	Z. Dong, Z. Wang, J. Xu, R. Tang and J. Wen, "A Brief History of Recommender Systems," 2022.
[43]	E. Hasan, M. Rahman, C. Ding, J. X. Huang and S. Raza, "Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives," 2025.
[44]	L. P. Kaelbling, M. L. Littman and A. W. Moore, "Reinforcement Learning: A Survey," 1996.
[45]	L. Zou, L. Xia, Z. Ding, J. Song, W. Liu and D. Yin, "Reinforcement Learning to Optimize Long-term User Engagement in Recommender Systems," 2019.
[46]	J. Terven, "Deep Reinforcement Learning: A Chronological Overview and Methods," 2025.
[47]	M. Fu, H. Zhang, Y. Du and R. Lou, "Adaptive Attention-based State Representation in reinforcement learning based recommendation systems," 2026.
[48]	S. Niu, T. Li and M. Chi, "A Literature Review of Ethical Considerations in Recommender Systems for User-Generated Content in Human-Computer Interaction," 2025.
[49]	J. Astley, "Ideological idiosyncrasies and individual-level radicalisation pathways: unpacking the mutating dynamics of contemporary extremism," 2025.
[50]	Institute for Strategic Dialogue, «US Online Domestic Violent Extremism Monitor,» 2025.
[51]	M. Cinelli, G. D. F. Morales, A. Galeazzi, W. Quattrociocchi and M. Starnini, "The echo chambre effect on social media," 2021.
[52]	J. R. Kunst, "Intelligent Systems, Vulnerable Minds: A Framework for Radicalization to Violence in the Age of AI," 2026.
[53]	A. Siebenmorgen and J. G. Restrepo, "A Utility-Driven Bounded-Confidence Model for Opinion Dynamics," 2026.
[54]	S. Banischa and E. Olbricha, "Opinion polarization by learning from social feedback," 2018.
[55]	M. L. Bertotti and M. Menale, "Opinion dynamics models describing the emergence of polarization phenomena," 2024.
[56]	R. M. Anderson and R. M. May, "Infectious Diseases in Humans," 1992.

6. Bibliography



[57]	J. Yuan, J. Shi, J. Wang and W. Liu, "Modelling network public opinion polarization based on SIR model considering dynamic network structure," 2021.
[58]	M. Starnini, M. Frasca and A. Baronchelli, "Emergence of metapopulations and echo chambers in mobile agents," 2016.
[59]	M. Moussaïd, J. E. Kämmer, P. P. Analytis and H. Neth, "Social Influence and the Collective Dynamics of Opinion Formation," 2013.
[60]	A. Drolet, M. F. Luce, L. Jiang, B. C. Rossi and R. Hastie, "The Preference for Moderation Scale," 2020.
[61]	A. R. Pratkanis, M. E. Turner and M. Murphy, "Attitude formation and change," 2024.
[62]	F. Hartleb, "Terrorism in the age of new technologies," 2025.
[63]	T. C. Helmus, "Artificial Intelligence, Deepfakes, and Disinformation: A Primer," 2022.
[64]	I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks," 2014.
[65]	J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan and S. Ganguli, "Deep Unsupervised Learning using Nonequilibrium Thermodynamics," 2015.
[66]	B. Liu, B. Liu, T. Zhu and M. Ding, "A Review of Deepfake and Its Detection: From Generative Adversarial Networks to Diffusion Models," 2025.
[67]	W. X. Zhao, K. Zhou, J. Li, T. Tang, Z. Dong, Y. Hou, B. Zhang, Y. Min, J. Zhang, P. Liu, X. Wang, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, Y. Hu, J.-Y. Nie and J.-R. Wen, "A Survey of Large Language Models," 2026.
[68]	Y. Veilleux-Lepage, "Beyond Misuse: Artificial Intelligence, Grievance, and the Future Landscape of Political Violence," CTCSENTINEL, 2026.
[69]	Center for Countering Digital Hate, «Killer Apps: How mainstream AI chatbots assist users planning violent attacks,» 2026.
[70]	L. Lin, S. Hu and X. Tang, "Paper Summary Attack: Jailbreaking LLMs Through LLM Safety Papers," 2026.
[71]	tech against terrorism, «Early terrorist experimentation with generative artificial intelligence services,» 2023.
[72]	N. A. Ganaie, "The role of artificial intelligence in radicalisation, recruitment and terrorist propaganda: deconstructing violent extremism and reimagining counterterrorism in contemporary digital ecosystems," 2026.
[73]	D. H. McElreath, D. A. Doss, L. McElreath, A. Lindsley, G. Lusk, J. Skinner and A. Wellman, "The Communicating and Marketing of Radicalism: : A Case Study of ISIS and Cyber Recruitment," 2018.
[74]	M. Rudner, ""Electronic Jihad": The Internet as Al Qaeda's Catalyst for Global Terror," 2016.
[75]	N. U. H. Wani, "Decoding Social Media's Role in the Resurgence of the Taliban: A Literature Review," 2023.

6. Bibliography



[76]	T. J. Holt, J. D. Freilich and S. M. Chermak, "Examining the Online Expression of Ideology among Far-Right Extremist Forum Users," 2020.
[77]	H. Lin, X. Jiang, J. Lee, Y. Wang and Y. Kim, "Exploring the Paradox of Cross-Cutting Exposure and Affective Polarization: A Curvilinear Model Influenced by Political Ideology Strength," 2025.
[78]	A. Goyal, O. Pal, H. Sundaram, E. Chandrasekharan and K. Saha, "Social Simulacra in the Wild: AI Agent Communities on Moltbook," 2026.
[79]	Y.-S. Chuang, K. Nirunwiroj, Z. Studdiford, A. Goyal, V. V. Frigo, S. Yang, D. V. Shah, J. Hu and T. T. Rogers, "Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks," 2024.
[80]	N. Bougie and N. Watanabe, "CitySim: Modeling Urban Behaviors and City Dynamics with Large-Scale LLM-Driven Agent Simulation," 2025.
[81]	J. S. Park, L. Popowski, C. J. Cai, M. R. Morris, P. Liang and M. S. Bernstein, "Social Simulacra: Creating Populated Prototypes for Social Computing Systems," 2022.
[82]	J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang and M. S. Bernstein, "Generative Agents: Interactive Simulacra of Human Behavior," 2023.
[83]	A. F. Ashery, L. M. Aiello and A. Baronchelli, "Emergent social conventions and collective bias in LLM populations," 2025.
[84]	«Moltbook,» [Online]. Available: https://www.moltbook.com/ .
[85]	Y. Jiang, Y. Zhang, X. Shen, M. Backes and Y. Zhang, "Humans welcome to observe": A First Look at the Agent Social Network Moltbook," 2026.
[86]	L. Li, R. Ma, C. Chen, Z. Lu and Y. Zhang, "The Rise of AI Agent Communities: Large-Scale Analysis of Discourse and Interaction on Moltbook," 2026.
[87]	C. Wang, C. Li, S. Liu, Z. Chen, J. Hou, J. Qi, R. Li, L. Zhang, Q. Ye, Z. Liu, X. Chen, X. Zhang and P. S. Yu, "The Devil Behind Moltbook: Anthropic Safety is Always Vanishing in Self-Evolving AI Societies," 2026.
[88]	T. Choudhary, "Political Bias in Large Language Models: A Comparative Analysis of ChatGPT-4, Perplexity, Google Gemini, and Claude," 2024.
[89]	J. Betley, D. Tan, N. Warncke, A. Szttyber-Betley, X. Bao, M. Soto, N. Labenz and O. Evans, "Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs," 2025.
[90]	F. Salvi, M. H. Ribeiro, R. Gallotti and R. West, "On the conversational persuasiveness of GPT-4," 2025.
[91]	D. E. Pressman, «Countering Radicalization,» 2006.
[92]	D. T. Schroeder, M. Cha, A. Baronchelli, N. Bostrom, N. A. Christakis e D. Garcia, «How malicious AI swarms can threaten democracy,» 2026.
[93]	U.S. Department of Homeland Security, «Understanding Lone-actor Terrorism: A Comparative Analysis with Violent Hate Crimes and Group-based Terrorism,» 2013.

6. Bibliography



[94]	J. Muldoon and J. J. Parke, "Cruel companionship: How AI companions exploit loneliness and commodify intimacy," 2025.
[95]	Z. Zhang, "User Motivation and Intimacy Reconstruction of AI Companion Applications," 2026.
[96]	H. R. Kirk, I. Gabriel, C. Summerfield, B. Vidgen and S. A. Hale, "Why human-AI relationships need socioaffective alignment," 2025.
[97]	J. Phang, M. Lampe, L. Ahmad and S. Agarwal, "Investigating Affective Use and Emotional Well-being on ChatGPT," 2025.
[98]	J. D. Freitas, A. K. Uguralp, Z. O. Uguralp and S. Puntoni, "AI Companions Reduce Loneliness," 2024.
[99]	L. Hung, Y.-t. Chiu and M. Ghodsi, "Exploring Non-Embodied AI-Based Digital Companions for Older Adults in Aging and Care Contexts: Protocol for a Scoping Review," 2026.
[100]	L. Laestadius, A. Bishop, M. Gonzalez, D. Illeňčík and C. Campos-Castillo, "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika," 2022.
[101]	common sense media, «Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions,» 2025.
[102]	I. Gabriel, A. Manzini, G. Keeling, L. A. Hendricks, V. Rieser, H. Iqbal and N. Tomasev, "The Ethics of Advanced AI Assistants," 2024.
[103]	S. Rathje, M. Ye, L. K. Globig, R. M. Pillai, V. O. d. Mello and J. J. V. Bavel, "Sycophantic AI increases attitude extremity and overconfidence," 2025.



Coordinator



Scientific Technical Coordinator



Project Security Officer



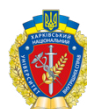
Academic | Think-Tanks | Research



Technology Providers



Practitioners



Keeping People Safe



This project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No 101021853.

