25. Generative Artificial Intelligence

Stéphane J. BAELE

https://orcid.org/0000-0002-3632-0888

Abstract

The end of the 2010s and the beginning of the 2020s have witnessed exponential progress in Artificial Intelligence (AI). As with other fields of human activity, the technology has started to percolate into extremism and terrorism. This chapter appraises the significance of extremist use of AI (or 'AI extremism'), proceeding in three steps. First, an introduction situates generative AI at the core of AI extremism and mentions illustrative cases. Second, two ideal-typical diagnoses – and therefore prognoses – of AI extremism are presented, which stand in opposition. On the one hand, an *alarming* diagnosis/prognosis casts AI as an overwhelmingly negative technological evolution bolstering online extremism and repression in critical and irreversible ways. On the other hand, a *reassuring* scenario stresses the technology's significant opportunities for counter-extremism and violence prevention, and highlights human societies' capacity to prevent nefarious dual-use through targeted policies, regulations, and interventions. However, an argument is made in a third step that both pessimistic and optimistic diagnoses/prognoses rest on the same simplistic narrative of an AI "revolution", interpreted either in a modern-liberal way emphasising progress or in a reactionary way emphasising regress, which is typical of reactions to important new technologies. Theories of historical state development indeed change the perspective through which we appraise AI, towards a third diagnosis/prognosis dialectically merging the two others, whereby AI simultaneously strengthens extremists and (strong) states, raising profound ethical questions.

Keywords: extremism, Artificial Intelligence, AI, deepfake, ChatGPT, terrorism, radicalization

The Reality of "Extremist AI"

The past two decades have witnessed a breakthrough in a technology that was essentially dormant since its earliest forms dating back to the 1950s and 1960s: Artificial Intelligence (AI), understood as the type of computer "systems endowed with the intellectual processes characteristic of humans, such as the ability to reason, discover meaning, generalize, or learn from past experience" (Encyclopaedia Britannica 2024). In the 10 years spanning 2013-2023, key metrics characterising the development of this technology have shown genuinely exponential trajectories in the amount of computation involved in AI models, size of the datasets used to train them, and the accuracy and power of their results (Giattino, Mathieu, Samborska & Roser 2025; see also OECD 2023). Consequently, AI models – that is, computational frameworks trained to perform specific tasks by learning patterns from data in supervised or unsupervised ways, using mathematical algorithms to process inputs and generate outputs – are now present in a diverse range of domains such as medicine, communications, and the military. They are also making their ways into political extremism. This is no surprise: extremists have always sought to harness new technologies to strengthen their strategies and tactics (Veilleux-Lepage 2020; Rassler & Veilleux-Lepage 2024). Even the most fundamentalist and traditionalist groups have played determinant roles in importing new technologies into extremism, as the case of the Islamic State (IS) illustrates (e.g., Baele, Boyd & Coan 2019). As Baele and Brace observe, "like every new technological breakthrough in the past […] AI also unlocks new paths to harm. Just as they learned to use the printing

press, guns, audio and video tapes, the Internet, or photo-editing software, violent and non-violent extremists now assimilate AI into their projects and modus operandi". There is therefore no doubt that the use of AI models by extremists – what Baele and Brace (2023) call "AI extremism" – will not only happen but will also normalize, in the same way the Internet has become a central feature of extremist communications and socialization.

Specifically, *generative AI* – that is, AI models designed to learn patterns in text, image, sound, or video content (input) to create similar content (output) –[1] has the most immediate implications for extremism compared to other types of models (such as strategic decision-making), not only because its implementation for extremist objectives requires less skills and infrastructure than other types of AI models, but also because it naturally plugs into the steady growth of extremist online communications and content over the past 30 years. As such, generative AI – which produces deepfake imagery and videos, synthetic text, cloned voices and sounds – is a new technology that feeds from, and enhances, existing online extremism. The aim of this chapter is to appraise the severity and nature of this relationship between extremism and generative AI.

Extremist use of generative AI had been foreseen by a handful of observers. Notably, Chesney and Citron (2019a, 2019b) published two articles in 2019 arguing that deepfakes would soon become a key tool in great powers' grey zone operations, chiefly through the instrumentalization of ideological fault-lines and local extremisms in rival states. Commissioned by OpenAI, McGuffie and Newhouse (2020) later offered a first demonstration that extremist content could easily be generated by a Large Language Model (LLM – an AI model trained on vast text data to identify linguistic patterns and generate on that basis human-like language). Such a feat had already occurred back in 2016, when Microsoft released on Twitter an AI chatbot named Tay, which was immediately "trained" through its discussions to engage in racist, antisemitic, and misogynist rants. Commenting on this episode, Wolf, Miller and Grodzinsky (2017) warned that this technology was inherently vulnerable to extremist abuse. More recently, LLMs' ability to produce all sorts of sometimes niche extremist content has been further evidenced by Baele and colleagues (Baele 2023; Baele, Naserian & Katz 2024) and researchers at the AI company Palisade (Gade, Lermen, Rogers-Smith & Ladish 2023; Lermen, Rogers-Smith & Ladish 2023). Lakomy (2023) has broadened the warning towards the potential use of generative AI for know-how and information purposes (e.g., bomb-making, cyber operations).

The early 2020s witnessed the materialisation of these warnings, providing the first instances of extremist actors adopting – and adapting – the incoming technology to pursue their goals, chiefly using generative models to enhance their information operations and propaganda. The circulation of fake Neo-Nazi images on far-right Telegram channels (Tech Against Terrorism 2023), the multiplication across social media of fake pictures of Israeli soldiers sporting diapers during the Gaza war following Hamas spokesperson Abu Obaida's allegations that they wore pampers (Siegel 2024), the creation of synthetic racist memes on 4chan/pol (Belanger 2023), the dissemination of fake posters by al-Qaeda (Tech Against Terrorism 2023), or the dissemination of several types of images by IS digital enthusiasts (flags, abstract art, mujahideen, etc.) (Criezis 2024), are examples of early extremist use deepfake imagery. The customization by 4chan/pol users of the LlaMA language model into an anti-Semitic chatbot (Siegel 2023) and their interactions with a racist version of the ChatGPT (Baele & Brace 2024) are two cases of the type of AI-powered extremist text generation foreseen by McGuffie and Newhouse, Baele and colleagues, Lakomy, and the Palisade team.

The proliferation of synthetic jihadist nasheed (Siegel 2024),[2] the use by IS supporters of AI transcription and translation services to extend the reach of statements and messages initially made in Arabic (Tech Against Terrorism 2023), and al-Qaeda's propagandists and digital supporters' use of deepfake news videos (Katz 2024), demonstrate uses of generative AI going beyond the most intuitive tools.

Reflecting on these early instances, Baele and Brace (2024) sought to comprehensively map all affordances opened up by generative AI and other types of models in order to chart the potential of this new technology for extremist endeavour. The array of nefarious uses is very large. Generative AI can be

used by extremists for: propaganda creation, blackmail, harassment campaigns, impersonation tactics, artificial online platform boosting, adversarial online platform flooding, radicalizing chatbots, pollution of information environment, creation of misperception-inducing content, assistance in offensive cyber operations, violence advice, speech recognition and translation, and advice on online privacy. To this day however, extremist dual-uses of generative AI remain sporadic and disjointed and have yet to be fully integrated into a coherent and forward-looking strategy. The question raised by these developments is that of the magnitude – and nature – of the change they induce in (online) extremism. The next section spells out, in an ideal-typical manner (whereby positions are artificially exaggerated to their purest expression to highlight their differences), the two major appraisals of the problem: one claims that the technology tips the balance firmly in favour of extremists vis-à-vis mainstream political worldviews as well as law-enforcement, the other that, if anything, it impedes their efforts.

AI Extremism: Two Diagnoses

The alarming diagnosis

A first diagnosis stresses the many dangers of generative AI when it comes to online extremism, thereby offering an ominous prognosis. Adopting a pessimistic stance, sometimes rooted in the longstanding technophobe tradition, proponents of this view characterize AI as a game-changing technology offering multiple new possibilities to extremists. Similar to the deployment of the Internet in the 1990s and 2000s, AI's generalization is said to enable extremist projects in ways that decisively shift the balance in their favour against policy, civil society, and law-enforcement efforts to keep extremism at bay. AI is thought to turbocharge an online extremism problem which is already out of control, with sprawling radical digital ecosystems gaining more and more traction. Criezis (2024), for instance, talks about "a pivotal shift in digital warfare", while Siegel and Doty (2023) argue that AI tools are, "unlike the propaganda products of the past, […] much more effective for radicalisation, recruitment, and retention". This alarming diagnosis ideal-type, which is not far from the AI "doomsday" or "existential risk" discourses regularly voiced in the media or by philosophers like Nick Bostrom (e.g., 2014; for critical discussions see Samuels 2025; Swoboda et al. 2025), rests on five mutually reinforcing benefits of generative AI..

First, the technology offers the possibility to mass-produce very large volumes of content. Once a model is fine-tuned to generate the desired type of extremist visuals/text (which is not a prohibitively long or complex endeavour, as shown by the abovementioned Palisade and Baele studies), it can automatically produce a theoretically endless flow of content. This was done in 2022, when AI developer and YouTube tech influencer Yannic Kilcher trained the GPT3 LLM on three years of posts from the abovementioned 4chan/pol board and used a bot to inject tens of thousands of new synthetic posts created by this "GPT4-Chan" into the board. Authentic, human users took a long time to understand the nature of the sudden popularity of their platform, evidencing the highly credible character of the contributions made by the LLM-fed chatbot.

Even when not credible, synthetic content can still participate in the simplification, acceleration, and amplification of extremist narratives and tropes – a phenomenon called the "Synthetic Narrative Amplification Phenomenon" by Siegel (2024), who explains that "even overtly artificial images can streamline complex narratives into easily digestible, AI-generated memes […] that resonate widely, enhancing virality". The dissemination by far-right social media groups, during the 2024 riots in Dublin, of numerous fake images of stereotypically white, muscular, tall Irishmen protesting together is a good example; these images were obviously inauthentic, but nonetheless reproduced and conveyed motivating ingroup archetypes and narratives. The same goes for the relentless production of AI-generated images of beautiful white women and families in certain far-right online spaces, fuelling the popularity of what Tebaldi (e.g., 2023) calls "granola Nazism".

Second, generative AI coupled with social media bots can create persuasive dynamics of polarization/radicalization, by both creating artificially large extremist spaces and engaging in sustained

personal dialogues. As Siegel (2023) explains, "bigoted chatbots exacerbate the problem of echo chambers, where individuals are primarily exposed to information confirming their existing beliefs, […] continuously [reinforcing] extremist views, making it increasingly difficult for users of manufactured problematic chatbots to escape insular cycles of hate". The "GPT4chan" case mentioned above illustrates this problem particularly well: when Kilcher controversially flooded the discussion board with the automated injection of tens of thousands of inauthentic posts produced by his racist LLM, human participants obliviously continued to chat, albeit with many more responses reinforcing their views. Mantello, Ho and Podoletz (2023) further argue that chatbots "will play a dominant role in online radicalization" because they get better in their "ability to read and respond to human emotions and, in turn, increase their anthropomorphic tendencies." The case of Jaswant Singh Chail – whose intense discussions with AI friend-bot *Replika* played a role in his resolve to attempt to kill Queen Elisabeth II on Christmas Day 2021 (for a report, see for example Singleton, Gerken & McMahon 2023) – illustrates the risk of AI bots engaging in deeply personal conversations leading to extremism and violence, even when not designed to do so (but, crucially, designed to align with, and further encourage, the user's preferences, style, and projects).

Third, generative AI expands the reach of online propaganda. Besides the augmentation of volume, the technology enables extremists to easily create content in a multitude of languages, thereby engaging audiences well beyond their normal reach. Instant translation models and multilingual models make it easy to offer a given content (e.g., a magazine) in several linguistic variants, without relying on translators or affiliates with language proficiency (as ISIS did in 2014-2015). Voice cloning enables this new content to be directly voiced by leading figures of extremist movements in a targeted dissemination strategy, akin to New York City mayor Eric Adams' "robocalling" practice, whereby he called New Yorkers directly in their native language (such as Mandarin or Yiddish), with almost null time-lag, despite his inability to do so without AI enhancement (for report, see for example Fitzsimmons & Mays 2023).

The fourth effect stems from the previous: generative models unlock significant economies of scale that enable extremists to create more content with less resources – and faster. As Baele (2022) noted, "the real power of language models is not so much that it could automatically produce large amounts of problematic content in one click […], but rather that they enable significant economies of scale. In other words, the cost of creating such content is about to plummet." As such, generative AI critically offsets the main limiting factors in the production of extremist and terrorist propaganda: time, material resources, and skilled human staffing.

Finally, fifth, extremist content creation is only one affordance of generative AI; beyond propaganda, the array of nefarious uses of generative AI is vast. The technology opens up many avenues for illegal action that extremists will no doubt find appealing. From blackmailing government officials with compromising deepfakes, to influencing their work, or gaining access to information or funds, to large-scale harassment campaigns against "enemy" individuals, the possibilities are as numerous as the tactics and strategies followed by extremist political actors.

While these severe issues call for a swift, targeted, and genuinely effective response, none currently exist. Safeguards are easily bypassed by people with technical skills (Gade, Lermen, Rogers-Smith & Ladish 2023; Lermen, Rogers-Smith & Ladish 2023), and quick solutions provided by tech companies (such as watermarks on fake images) are dodged and unlikely to deter (Criezis 2024). Powerful open-source models proliferate that can be fine-tuned, with no possible control, in whatever ways the user deems useful (such as *DeepFaceLab* for videos or *TurtoiseTTS* for voice). Moreover, extremist models have even started to emerge, such as the *Wizard-Vicun* built on Meta's *Llama*, directly generating extremist content (Siegel 2023). Law-making is too slow, generic, and geographically inadequate to sufficiently address the issue and keep pace with fast technological developments and the decentralized nature of digital criminality; as both the history of extremist and terrorist innovation and new "ecosystem" approaches to the field teach us (e.g., Hutchinson et al. 2022), extremist actors inevitably

adapt to changes in the legislative environment (for example by migrating to another service, relocating activity in another region, etc.).

Moreover, liberal democracies' commitment to the rule of law and their extensive bureaucratic practices and rules (data protection, procurement, etc.) make the adoption of AI – and adversarial use of cyber tools more broadly – by the security services too slow and inadequate (see Kello 2024 for example). In other words, intelligence and law-enforcement services lack the same level of swift, extensive, and agile access to, and mobilization of, generative AI enjoyed by the extremist actors they are in charge of countering. Even if they were more reactive, gain-seeking companies – or more frequently individuals – with little ethical concern will always be available to monetize their skills.[3]

The reassuring diagnosis

Opposed to this first rather alarming diagnosis, the second one offered in commentaries on the emerging confluence of AI and extremism is more optimistic, imbued by a liberal, modern, technophilic understanding of scientific and technological evolution as progress. In this second ideal-typical stance, new technologies solve problems and largely benefit humanity; while they do bring about negative externalities, especially in their early days, these are eventually tamed as states establish rational regulations and cooperate internationally. In this view, the prognosis of AI extremism is not dire for four reasons: the emergence of the problem is gradually tackled by policies and law; the market adjusts to avoid negative effects impacting or putting off consumers; AI technology is used in positive ways to combat its malign effects; and extremists themselves are thought to be unlikely to make significant uses of AI. Major tech companies, Silicon Valley libertarians, and economic actors seeking productivity gains, regularly voice discourses close to this ideal-type – indeed Bourne (2019) calls them the "AI cheerleaders".

An impressive array of policy initiatives endeavouring to limit the negative externalities of AI are indeed underway. National AI strategies have already emerged that seek to maximise the benefits of the technology while minimising its risks (e.g., Galindo, Perset & Sheeka 2021; Radu 2021). More specific instruments, such as the US *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* (2023) or the UK Government's decision to introduce into the Online Safety Act the Law Commission's 2022 recommendations to criminalize deepfake porn (and the comparable criminalization introduced in the US in May 2025), start to address nefarious dual-use. Beyond these national frameworks, international initiatives have quickly mushroomed to tackle the risks of generative AI, with a strong focus on combatting extremist and manipulative information operations. Among them, the G7 has launched the Hiroshima Process on Generative Artificial Intelligence in partnership with the Organisation for Economic Co-operation and Development (OECD), which seeks to asphyxiate "the capacity of generative AI to exacerbate the challenges of disinformation and manipulation of opinions, […] considered by G7 members as one of the major threats stemming from generative AI" (OECD 2023). The threat is no less than "at the top of the agenda for G7 governments" (OECD 2023).

Also in 2023, the UK convened the AI Safety Summit, which resulted in the "Bletchley Declaration" whose signatories – including, crucially in times of increasing geopolitical tensions, China alongside the European Union and the US – acknowledged the "necessity and urgency of addressing" the problem of generative AI misuse and pledged to combat "the potential for unforeseen risks stemming from the capability to manipulate content or generate deceptive content" (Bletchley Declaration 2023). The EU AI Act, which exited the complex European legislative process in early 2024, moved beyond declarations to introduce several targeted provisions, while international organisations like Europol have sought to coordinate national understandings of, and responses to, the threat of generative AI (e.g., Europol 2023). Considering the sudden rise and incredibly fast pace of progress in AI, the swiftness of these reactions is remarkable; two years of development for an ambitious and complex instrument like the EU AI Act is remarkable , especially compared to the organisation's usual timeline.

Warned by the *Tay* incident mentioned above, and similar episodes, tech companies are working to ensure their products are of little use to extremists. If anything, they enforce excessively stringent safeguards when releasing their models to the public; generative models such as *Bing* or *ChatGPT* regularly generate complaints and mockery on social media from users when they refuse to produce content remotely connected to an ever-growing list of trigger words or themes such as violence or sex. The production or commission by major AI companies of countless studies and papers on the risks of generative AI is a testimony of their evident interest in keeping the technology safe. Even open-source models are said to carry, overall, more benefits than risks. According to Eiras and colleagues (2024), for instance, risks associated with open-source are "not a substantial concern for existing models given their limited capabilities"; openness, they argue, "propels advancements in safety technologies" and helps "developing safeguards against misuse". Open-source models should therefore, the argument goes, be protected against policies seeking to control or stifle them.

Concurrently, these models are now routinely harnessed to detect extremist content, offering an efficient set of tools for large-scale action against sprawling extremist ecosystems metastasizing onto social media. AI-led detection software still comes with challenges but is getting ever more powerful (Fernandez & Alani 2021; Gunton 2022; Govers et al. 2023; Mussiraliyeva, Bagitova & Sultan 2023). Stokes (2021) observes that beyond digital/social media providers, intelligence services start to successfully use machine learning in counterterrorism efforts, allowing them to "more quickly identify and prioritize suspects, analyse connections between suspects, and even utilize facial and voice recognition technologies to track persons of interest". As Baele, Brace and Naserian (2025) have shown, integrating several types of AI-powered computer vision models into a digital extremism detection and assessment workflow can allow law-enforcement to efficiently analyse the growing scale of online extremist visual content. Recognizing the defects of ambitious projects such as predictive radicalization or violence algorithms, Schröter (2020) nonetheless explains that AI systems will strengthen automated content moderation, and steer recommendation systems and search engine results away from extreme content, helping to construct a healthier online environment.

But perhaps more fundamentally, it is argued that generative AI is unlikely to be used by extremists in a blanket way. Baele and Brace (2024) conceptualize how groups' characteristics (e.g., its particular ideology, its specific situation in the broader radical milieu, the practices it values with symbolic and social capital, etc.) make their use of generative AI more or less likely. Indeed, using the technology comes with legitimacy risks for the extremist user keen to appear authentic or sacred (for example a respected cleric writing theology, or a jihadist leader penning poetry), runs against some religious or ideological norms and rules (the issue of what AI tools are *haram* is for example debated in Islamist forums), and is ill-suited for propaganda outlets aiming for "intellectual" impact rather than mass reach. In sum, it is argued, generative AI will make its way among the technologies used by extremists, but will only be used in limited ways, not by everyone, and only for specific tasks.

Beyond the AI Revolution Narrative: A Third Diagnosis

Presented in such an ideal-typical way, these two diagnoses – and the prognoses they entail – seem (by design) fundamentally opposed and contradictory. While the reassuring stance encourages unchecked open-source AI science, stresses the potential of the new technology to manage the problem of online extremism, and underlines the improbable character of large-scale "AI extremism", the more alarming view calls (without much hope) for determined multi-level action limiting the damages inevitably unleashed by this new episode of reckless technological innovation generating societal harms. This ideal-typical abstract opposition between the two stances is not only useful to artificially underline major thinking lines about the issue, it is also, perhaps paradoxically, a useful starting point to foresee what is ahead. It can be suggested that the two diagnoses actually participate, together, in a single master narrative about Artificial Intelligence, which ought to be overcome in a dialectical way to reach a more accurate diagnosis and prognosis.

Like every significant technological progress, AI is understood and appraised by societies through established narratives and metaphors (Carbonell, Sánchez-Esguevillas & Carro 2016), and both the alarming and reassuring understandings rest on what could be called the "AI revolution" narrative according to which AI is a radically new technology holding unprecedented capabilities – a "game-changer" transforming society in profound ways. Köstler and Ossewaarde's (2022) study of AI framing in the German media, for example, showed that the technology is usually presented as possessing "a disruptive potential tantamount to the invention of electricity". As Coeckelbergh (2020, 2021) explains, such a "technological singularity" narrative – characterized by a very short timeline comprising a "before", a seismic technological change of almost unprecedented proportions, and an "after" – is fuelled by frequent "breakthrough" announcements that naturally generate opposing "hype and fears" positions – our two ideal-types. Schwartz (2019) similarly observes how both "incredible miracle" and "incredible horror" discourses rest on – and together construct – a single, exaggerated perception. In that sense, the alarming and reassuring views correspond to the long-established technophobe and technophile discourses on new technological "revolutions". Social narratives have profound political implications: by offering a particular understanding of a given development, they influence our responses to it. As such, AI narratives "link past, present, and future in particular ways that are ethically and politically significant" (Coeckelbergh 2021), and "shape the ways in which we develop, regulate, and use technologies" (Westerstrand, Westerstrand & Koskinen 2024). Both ideal-typical diagnoses summarized above entail similarly naïve prognoses resting on a simplistic revolution master narrative – whether blindly trusting progress and technology or calling to intensely fight the existential risk of AI dual-use, they are equivalent in their one-sidedness and demonstrate an incomplete appraisal of what is at stakes with AI extremism. As abstract constructs, they are nonetheless useful to dialectically build another analysis of AI (and "AI extremism"), which still acknowledges the newness and significance of the technology but simultaneously situates it in the longer term and comprehends it in the lineage of past technological progress.

This third view is informed by Charles Tilly's influential work on the historical development of powerful states in Europe (Tilly 1992). In his account, today's great powers have gradually emerged and consolidated from an environment of relentless security competition and power accumulation thanks to their enduring resilience to, and conduct of, wars. A key factor in remaining successful in this cyclical process of war / consolidation has been states' ability to integrate new technology (first and most crucially the heavy cannon) into their military, and to set up an efficient tax system to finance cutting-edge war capabilities. More recently, Peter Andreas has adapted Tilly's famous line "war made the State and the State made war" to characterize the relationship between state consolidation and transnational crime such as smuggling (e.g., Andreas 2013). Among the many mechanisms constitutive of an intimate relationship between transnational crime and states highlighted by Andreas, one is of crucial importance for our analysis of AI extremism: he argued that state powers have expanded through a ratchet mechanism whereby new threats emerge[4] (e.g., drug trafficking, smuggling, terrorism) that lead states to criminalize them and subsequently accumulate new capabilities to fight the problems and the "enemy" behind it. Applying this line of thought to terrorism and extremism, Hegghammer (2021) insightfully argued that the "war on terror" is best interpreted through this lens: following a particularly acute instance of a threat (Islamist terrorism on 9/11), the US state broadened and deepened its criminalization of "terrorism", subsequently enhancing in significant ways its power through legal instruments (like the Patriot Act [2001] or the Intelligence Reform and Terrorism Prevention Act of [2004]), new IT tools expanding the reach of intelligence agencies (like the National Security Agency - NSA), and a flurry of hardline security practices at home and abroad (such as extraordinary renditions).

This conceptual framework unlocks a third diagnosis and prognosis of extremist AI, which captures more comprehensively the nature and consequences of this phenomenon. This third interpretation rests on a longer-term perspective where scientific progress proceeds with occasional technological accelerations, some of them holding significant implications for security: AI is merely the most recent.[5] These implications for security are twofold. First, challengers to states (non-state extremist, terrorist, or

insurgent organisations, rival states, etc.), will seek to harness the new technology to enhance their capabilities and thereby undermine state power. Second, states will fear such an adversarial use of the new technology and will thus be pressed not only to limit its use among challengers and more broadly within society through an expanding criminalization net, but also to harness the technology himself. This is what happened worldwide with older technologies like firearms (with the notable exception of the USA), and it is exactly what is beginning to happen with AI. These two mechanisms therefore entail a "ratchet" scenario akin to an arms race: state challengers (may) gain power by adopting a new technology, the state responds by using the technology to augment its own capabilities, by legally restricting or criminalizing certain uses of the technology, and requesting increased traditional means of power to counter the threat (coercive security practices, expanded surveillance, counter-measures, punishment structures, etc.), but this reaction puts challengers in a disadvantaged situation that pushes them to adapt and further enhance their own capabilities – and so on. Ultimately, as Hegghammer (2021) warned, the result is a more intrusive security state navigating a worsened security environment.

In a way, it is not scientific progress in AI per se that triggers this chain-reaction; it is the emergence of, and belief in, the two main "revolution" discourses and associated diagnoses pertaining to the AI revolution master narrative, which *together* contribute to the simultaneous augmentation of the threat and the reinforcement of states' powers. At the time of writing, this ratchet mechanism is worryingly starting to eventuate, with all the early uses of generative AI by extremist actors described in the first section of the chapter, followed by governments' legislative efforts to criminalize "dangerous" generative AI uses (as the EU AI Act extensively does) and concomitant development of their own AI systems to counter the threat (for example, DARPA launched in 2018 a $2 billion investment in a campaign called "AI Next"). As McKendrick already noted in 2019, "developments in AI have amplified the ability to conduct surveillance without being constrained by resources" (McKendrick 2019); Burton (2022) went further, claiming that AI has been *securitized* by the state, with adverse effects including even more radical views (against a state perceived to be oppressive). In sum, extremists' use of generative AI is not merely a narrow, technical problem; rather, it raises profound ethical and policy questions gravitating around difficult quandaries like the contours of just securitization, the right limits to state power, or the possibility of a "liberty-security balance".

Conclusions

AI models designed to create fake text, images, audio, and videos – usually grouped under the label "generative AI" – have quickly entered the digital world; users produce evermore credible synthetic content to maximize efficiency, cut costs, gain resources, bolster creativity, reach new publics, and gather new know-how. Extremists are no exception: a diverse array of extremist actors, from the far-right to the far-left and from Salafi jihadists to Christian fundamentalists, have already harnessed the technology to pursue the very same goals. Facing this particular "dual-use", two ideal-typical diagnoses can be abstracted. The first, alarming one, warns of the irreparable damage generative AI will cause as soon as it starts to be deployed in a consistent and coherent manner within more or less elaborate extremist strategies. The second, more reassuring diagnosis views, in the liberal-modern tradition, "AI extremism" as a minor and temporary negative externality of an otherwise empowering new technology, stressing the proliferation of tech solutions, national initiatives, and international cooperation developed to tame the problem.

This chapter argues, however, that both diagnoses are incomplete; while they appear to be incompatible, they actually participate together in an "AI singularity" or "AI revolution" master narrative that prevents us from considering how the technology generates the same kind of ratchet mechanism seen with several technological innovations in the past. On the one hand, the alarming diagnosis is right in stressing the force multiplier character of generative AI for extremists, yet neglects the potential held by the technology for state power (and, as has always been the case in the history of technology-based power, hybrid private-public structures). On the other hand, the reassuring diagnosis

is right in highlighting this potential, but wrongly assumes that states and tech companies will solve the problem in the first round (which in turn naïvely implies that their interests converge) and, worse, blissfully ignores the dark side of states' and tech companies' increased capabilities. The latter issue is particularly worrying in an age of authoritarian resurgence and democratic backsliding, with authoritarian states enthusiastically embracing AI to enhance repression and surveillance as well as information operations in foreign countries, and liberal democracies' security services too often paralysed by the regulatory and bureaucratic quandaries opened up by models resting on deep learning.

References

Andreas P. (2013) *Smuggler Nation: How Illicit Trade Made America*. New York: Oxford University Press.

Baele S. (2022) Artificial Intelligence and Extremism: The Threat of Language Models for Propaganda Purposes. *CREST Guide*.

Baele S., Boyd K., Coan T. (2019) *ISIS Propaganda: A Full-Spectrum Extremist Message*. Oxford: Oxford University Press.

Baele S., Brace L. (2024) *AI Extremism. Technologies, Tactics, Actors*. Dublin: VOX-Pol.

Baele S., Brace L., Naserian E. (2025) More is More: Scaling up Online Extremism and Terrorism Research with Computer Vision. *Perspectives on Terrorism* 29(1): 34-64.

Baele S., Naserian E., Katz G. (2024) Is AI-Generated Extremism Credible? Experimental Evidence from an Expert Survey. *Terrorism and Political Violence*, online before print.

Belanger A. (2023) 4Chan Users Manipulate AI Tools to Unleash Torrent of Racist Images. *Ars Technica*, 5 October 2023.

Bletchley Declaration (2023) *Declaration Agreed by Countries Attending the AI Safety Summit 2023 at Bletchley Park, Buckinghamshire*.

Bostrom N. (2014) *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bourne C. (2019) AI cheerleaders: Public Relations, Neoliberalism and Artificial Intelligence. *Public Relations Inquiry* 8(2): 109-125.

Burton J. (2023) Algorithmic Extremism? The Securitization of Artificial Intelligence (AI) and its Impact on Radicalism, Polarization and Political Violence. *Technology in Society* 75.

Carbonell J., Sánchez-Esguevillas A., Carro B. (2016) The Role of Metaphors in the Development of Technologies. The Case of the Artificial Intelligence. Futures 84(B): 145-153.

Chesney R., Citron D. (2019a) Deep Fakes: Looming Challenge for Privacy, Democracy, and National Security. *California Law Review* 107(6): 1753-1820.

Chesney R., Citron D. (2019b) Deepfakes and the New Disinformation War: The Coming Age of Posttruth Geopolitics. *Foreign Affairs* 98(1): 147-155.

Coeckelbergh M. (2020) *AI Ethics*. Cambridge, Mass.: MIT Press.

Coeckelbergh M. (2021) Time Machines: Artificial Intelligence, Process, and Narrative. *Philosophy & Technology* 34: 1623-1628.

Criezis M. (2024) AI Caliphate: The Creation of Pro-Islamic State Propaganda Using Generative AI. *GNET Insight*.

Eiras F., et al. (2024) Risks and Opportunities of Open-Source Generative AI. *arXiv:2405.08597*.

Encyclopaedia Britannica (2024) *Artificial Intelligence*, https://www.britannica.com/technology/artificial-intelligence.

Europol (2023) ChatGPT: The Impact of Large Language Models on Law Enforcement. *Europol Tech Watch Flash*.

Fernandez M., Alani H. (2021) Artificial Intelligence and Online Extremism. In McDaniel J., Pease K. (eds.) *Predictive Policing and Artificial Intelligence*. London: Routledge.

Fitzsimmons E., Mays J. (2023) Since When Does Eric Adams Speak Spanish, Yiddish and Mandarin? He Doesn't. *New York Times*, 20 October 2023, https://www.nytimes.com/2023/10/20/nyregion/ai-robocalls-eric-adams.html.

Gade P., Lermen S., Rogers-Smith C., Ladish J. (2023) BadLlama: Cheaply Removing Safety Fine-tuning from Llama 2-Chat 13B. *arXiv:2311.00117*.

Galindo L., Perset K., Sheeka F. (2021) An Overview of National AI Strategies and Policies. *OECD Going Digital Toolkit Notes* 14.

Giattino C., Mathieu E., Samborska V., Roser M. (2023) Artificial Intelligence, *OurWorldInData.org*.

Govers J., Feldman P., Dant A., Patros P. (2023) Down the Rabbit Hole: Detecting Online Extremism, Radicalisation, and Politicised Hate Speech. *ACM Computing Surveys* 55(14): 1-35.

Gunton K. (2022) The Use of Artificial Intelligence in Content Moderation in Countering Violent Extremism on Social Media Platforms. In Montasari R. (ed.) *Artificial Intelligence and National Security*. Cham: Springer.

Hegghammer T. (2021) Resistance Is Futile. The War on Terror Supercharged State Power. Foreign Affairs 100(5): pp.44-53.
Hutchinson J., Droogan J., Waldek L., Ballsun-Stanton B. (2022) Violent Extremist & REMVE Online Ecosystems: Ecological Characteristics for Future Research & Conceptualization. *RESOLVE Research Briefs*.
IBM (2023) What is Generative AI? *IBM Explainers*, 20 April 2023, https://research.ibm.com/blog/what-is-generative-AI.
Kello L. (2024) Digital Diplomacy and Cyber Defence. In Bjola C., Manor I. (eds.) *The Oxford Handbook of Digital Diplomacy*. Oxford: Oxford University Press
Köstler L., Ossewaarde R. (2022) The Making of AI Society: AI Futures Frames in German Political and Media Discourses. *AI & Society* 37: 249-263.
Katz R. (2024) Extremist Movements are Thriving as AI Tech Proliferates. *SITE Special Reports*, 16 May 2024, https://ent.siteintelgroup.com/Articles-and-Analysis/extremist-movements-are-thriving-as-ai-tech-proliferates.html.
Lakomy M. (2023) Artificial Intelligence as a Terrorism Enabler? Understanding the Potential Impact of Chatbots and Image Generators on Online Terrorist Activities. *Studies in Conflict & Terrorism*, online before print.
Lermen S., Rogers-Smith C., Ladish J. (2023) LoRA Fine-tuning Efficiently Undoes Safety Training in Llama 2-Chat 70B. *arXiv:2310.20624*.
Mantello P., Ho T., Podoletz L. (2023) Automating Extremism: Mapping the Affective Roles of Artificial Agents in Online Radicalization. In Pashentsev E. (eds) *The Palgrave Handbook of Malicious Use of AI and Psychological Security*. Cham: Palgrave Macmillan.
McGuffie & Newhouse (2020) The Radicalization Risks of GPT-3 and Advanced Neural Language Models. *arXiv:2009.06807*.
McKendrick K. (2019) Artificial Intelligence Prediction and Counterterrorism. *Chatham House Research Papers*.
Mussiraliyeva S., Bagitova K., Sultan D. (2023) Social Media Mining to Detect Online Violent Extremism using Machine Learning Techniques. *International Journal of Advanced Computer Science and Applications* 14(6): 1384-1393.
Nieweglowska M., Stellato C., & Sloman S.A. (2023) Deepfakes: Vehicles for Radicalization, Not Persuasion. *Current Directions in Psychological Science* 32(3): 236–241.
OECD (2023) *G7 Hiroshima Process on Generative Artificial Intelligence (AI)*. OECD Publishing.
Papachristos A. (2022) The Promises and Perils of Crime Prediction. *Nature Human Behaviour* 6: 1038–1039.
Payne K. (2021) *I Warbot: The Dawn of Artificially Intelligent Conflict*. London: Hurst.
Radu R. (2021) Steering the Governance of Artificial Intelligence: National Strategies in Perspective. *Policy and Society* 40(2): 178-193.
Rassler D., Veilleux-Lepage Y. (2024) The Paradox of Progress: How 'Disruptive,' 'Dual-use,' 'Democratized,' and 'Diffused' Technologies Shape Terrorist Innovation. *Zeitschrift für Technikfolgenabschätzung in Theorie und Praxis* 33(2): 22-28.
Samuels R. (2025) *The Global Solution to AI* (Chapter 1 - The Existential Threats of AI). Cham: Palgrave Macmillan.
Schröter M. (2020) Artificial Intelligence and Countering Violent Extremism: A Primer. *GNET Reports*.
Schwarz E. (2019) Günther Anders in Silicon Valley: Artificial Intelligence and Moral Atrophy. *Thesis Eleven* 153(1): 94-112.
Siegel D. (2023) 'RedPilled AI': A New Weapon for Online Radicalisation on 4chan. *GNET Insights*.
Siegel D. (2024) AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State's Generative AI Digital Arsenal. *GNET Insights*.
Siegel D., Chandra B. (2023) "Deepfake Doomsday": The Role of Artificial Intelligence in Amplifying Apocalyptic Islamist Propaganda. *GNET Insights*.
Siegel D., Doty M. (2023) Weapons of Mass Disruption: Artificial Intelligence and the Production of Extremist Propaganda. *GNET Insights*.
Singleton T., Gerken T., McMahon L. (2023) How a Chatbot Encouraged a Man who Wanted to Kill the Queen. *BBC News*, 6 October 2023, https://www.bbc.com/news/technology-67012224.
Stokes S. (2021) *The AI Revolution and Its Implications on Domestic Counterterrorism*, https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/ai-revolution-and-its-implications-domestic#_ftnref6.
Swoboda T., Uuk R., Lauwaert L., Rebera A., Oimann A-K., Chomanski B., Prunkl C. (2025) Examining Popular Arguments Against AI Existential Risk: A Philosophical Analysis. *arXiv:2501.04064*.Tebaldi C. (2023) Granola Nazis and the Great Reset. Enregistering, Circulating and Regimenting Nature on the Far Right. *Language, Culture & Society* 5(1): 9-42.
Tech Against Terrorism (2023) Early Terrorist Experimentation with Generative Artificial Intelligence Services. *Tech Against Terrorism Briefings*.
Tilly C. (1990) *Coercion, Capital, and European States, AD 990-1990*. Cambridge, Mass.: Blackwell.

Tilly C. (2017) War Making and State Making as Organized Crime. In Castañeda E., Schneider C. (eds.) *Collective Violence, Contentious Politics, and Social Change*. London: Routledge.

Veilleux-Lepage Y. (2020) *How Terror Evolves. The Emergence and Spread of Terrorist Techniques*. Lanham: Rowman & Littlefield.

Westerstrand S., Westerstrand R., Koskinen J. (2024) Talking Existential Risk into Being: A Habermasian Critical Discourse Perspective to AI Hype. *AI & Ethics*, online before print.

Wolf M., Miller K., Grodzinsky F. (2017) Why We Should Have Seen That Coming: Comments on Microsoft's Tay 'experiment', and Wider Implications. *ACM SIGCAS Computers & Society* 47(3): 54-64.

[1] IBM (2023) defines generative AI as the family of "deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on".

[2] i.e. Islamic a cappella songs.

[3] This is already the case in the porn industry, where money can buy well-designed deepfakes with specific demands (who appears in the video, what happens in it, etc.), no questions asked.

[4] Or are *claimed to* emerge, as Securitization Theory and Andreas himself would rather argue.

[5] And quantum computing probably the next one.