Professor Xu

4 April 2026

CSDS 395

How Platform Design Shapes Extremism in Anonymous Online Spaces

In recent years, it has become increasingly difficult to ignore the influence of online communities on real-world beliefs and behavior. Platforms that were once considered niche or fringe now play a visible role in shaping political discourse, spreading conspiracy theories, and, in some cases, contributing to acts of real-world violence. Much of the public conversation has focused on algorithm-driven platforms like social media feeds, where recommendation systems guide users toward increasingly tailored content. However, less attention has been given to platforms that operate without these algorithms altogether.

One of the most prominent examples of these platforms is 4chan, an anonymous imageboard whose design fosters a unique and often extreme online culture. Created in 2003, 4chan allows users to post content anonymously without creating accounts (Thorleifsson 289). Threads are short-lived and quickly replaced by newer content, which creates an environment that prioritizes immediacy over accountability (Thorleifsson 289). Unlike platforms such as Facebook, TikTok, or Twitter, 4chan does not rely on algorithmic recommendation systems. Instead, visibility is determined by user interaction: threads that receive replies rise to the top, while those that do not quickly disappear (Elley 2).

At first glance, this lack of algorithmic curation might seem to reduce the risk of reinforcing extreme viewpoints. In fact, algorithms used in platforms such as TikTok have been widely criticized for amplifying political extremism (Shin and Jitkajornwanich 1020). Yet the reality is more complicated. Even without algorithms, 4chan's design creates its own mechanisms for amplifying extreme content.

I argue that the structure of anonymous, low-moderation platforms like 4chan plays a significant role in facilitating online radicalization. Through its combination of

anonymity, ephemerality, and user-driven visibility, the platform produces feedback loops that reinforce extreme beliefs and behaviors. By examining 4chan's technical design, the psychological mechanisms of radicalization, and real-world case studies, this paper demonstrates how social networking platforms can contribute to harmful social outcomes. It also considers potential mitigation strategies, including both technical interventions and educational approaches.

Before examining how radicalization occurs on 4chan, it is important to understand the platform's technical structure in more detail. Unlike most modern social media platforms, which are built around user profiles and long-term engagement, 4chan is designed to be temporary and anonymous by default. Posts are organized into threads, and each board has a limited number of active threads at any given time. When new threads are created, older ones are pushed off the board and eventually deleted. This system creates a cycle of constant turnover, where visibility is short-lived, and users must continually post in order to keep a discussion active (Elley 2).

This structure has important consequences for how users interact. Because threads disappear quickly, there is little incentive to build a long-term reputation or maintain consistent beliefs over time. Instead, users are encouraged to focus on immediate reactions to gain attention before the thread disappears. In this sense, the platform rewards intensity over consistency, which can contribute to increasingly extreme content as users compete for visibility and validation (Thorleifsson 290).

Another key feature of 4chan is its moderation model, which is relatively "hands-off" compared to mainstream platforms. While some illegal content is removed, enforcement is inconsistent and often relies on volunteer moderators (Kasimov et al. 134). This creates an environment where harmful or extremist content can remain visible for long enough to spread and gain traction. More importantly, it allows these communities to develop their own internal norms with little external oversight. On boards like /pol/, these norms often include the normalization of racist and misogynistic content, which can make such views seem more acceptable over time (Thorleifsson 291).

The absence of algorithmic curation is often assumed to make platforms like 4chan less influential in shaping user beliefs. For example, Shin and Jitkajornwanich

claim that “algorithms that personalize user content have raised significant social concerns because they have the potential to intensify and reinforce spirals that can result in users’ adoption of radical and extremist beliefs” (1020). While true, this assumption overlooks the ways in which user-driven systems can produce similar effects. Instead of being guided by recommendation algorithms, users are exposed to content that has already gained attention from others on the board. This creates a kind of crowd-driven visibility, where popular or provocative ideas are amplified simply because they generate more replies. Thorliefsson claims that “4chan is non-hierarchical, yet there is a struggle to obtain symbolic status and recognition by fellow users by producing content others will engage with” (289). In practice, this can function as a feedback loop, where extreme content attracts engagement, which in turn increases its visibility, encouraging even more extreme contributions.

To understand why this environment can lead to radicalization, it is necessary to examine the underlying social mechanisms at play. Radicalization does not typically occur all at once, rather, it is a gradual process in which individuals adopt increasingly extreme beliefs over time (Berjawi et al. 5–6). Online spaces like 4chan accelerate this process by combining social reinforcement with constant exposure to provocative content. Users are not only consuming information but actively engaging with a community that rewards extremist behavior.

One of the most important mechanisms in this process is the use of irony and humor. On 4chan, extremist ideas are often presented in a joking or exaggerated manner, making it difficult to distinguish between what is meant seriously and what is not (Thorleifsson 290). This ambiguity allows users to engage with controversial ideas without fully committing to them, creating a kind of “plausible deniability.” Over time, however, repeated exposure to these ideas, even in a joking context, can lead to their normalization. For example, one study found that “exposure to neo-Nazi discourse online increased radical opinion extremism even among those individuals who were regularly exposed to opposing (anti-far-right) views in their offline social circles” (Kasimov et al. 130). What begins as ironic participation can gradually shift into genuine belief, especially when these ideas are reinforced by other users.

In addition to humor, memes play a central role in how ideas are communicated and spread on the platform. Memes allow highly complex political or ideological messages to be condensed into simple, highly shareable formats. Because they are so easy to reproduce and modify, memes can spread quickly across threads and other platforms (Kasimov et al. 131). This makes them an effective tool for both expressing and reinforcing group identity, as well as for introducing new users to specific narratives and ideologies. Thorliefsson writes that "the mastering of memes functions as a gatekeeper, marking communal belonging to fellow anons and distance to normies, a slang pejorative label for individuals deemed to be conventional or mainstream" (290). Over time, repeated exposure to these memetic messages can shape how users understand the world around them.

Building on these mechanisms, 4chan can be more specifically understood not just as a platform where radical ideas appear, but as an environment that actively facilitates their development and spread. This is especially visible on the site's "Politically Incorrect" board, also known as /pol/, which has become one of the most studied spaces for online extremism (Phillips and Campion 2). While not all users on /pol/ hold extreme views, the board's culture tends to reward posts that are provocative and offensive, creating an atmosphere where more moderate perspectives are often ignored or dismissed.

On /pol/, political discussion is frequently framed through a mix of irony, cynicism, and hostility toward mainstream institutions. Users often present extremist viewpoints alongside humor or self-improvement rhetoric, blending personal identity with broader ideological narratives (Elley 2). As Elley notes, posts may frame participation in these communities as part of a larger mission tied to cultural or civilizational renewal, encouraging users to see themselves as active participants in a broader ideological struggle (3). This framing can make engagement feel purposeful rather than casual, which may deepen users' commitment over time.

Furthermore, anonymity plays a central role in reinforcing this environment. Because users are not tied to stable identities, they can experiment with increasingly extreme positions without facing long-term consequences. This freedom is not just

structural but also psychological. Many users draw a clear distinction between their online activity and their real-world identity, treating what they post as separate from who they "actually" are (Thorleifsson 290). As a result, the belief that something is "just online" does little to limit its real psychological or social impact.

Over time, this dynamic becomes a self-reinforcing cycle. Users post increasingly extreme content under the assumption that it does not reflect their real identity, receive validation from others, and then continue escalating their behavior. What may begin as experimentation or ironic participation can gradually become routine, blurring the line between performance and genuine belief (Elley 2). As Whittaker observes, users in these spaces often "role-play an idealized version of themselves online which is more zealous and supportive of violence" (28).

This process is further intensified by inconsistent moderation. Without strong or consistent enforcement of content standards, harmful ideas are not only allowed to persist but are repeatedly reinforced through exposure and engagement. When extremist or conspiratorial content becomes common, it can begin to feel normal within that space, even if it would be widely rejected elsewhere (Whittaker 28). In this way, radicalization tends to occur gradually rather than through any single moment of transformation.

Several real-world case studies illustrate how these dynamics can extend beyond the platform itself. One early case is Gamergate, a loosely organized harassment campaign that emerged in part from discussions on 4chan. What initially appeared as a dispute around video game journalism quickly escalated into widespread harassment, particularly targeting women in the gaming industry (Massanari 334). This case demonstrates how 4chan can facilitate coordinated action with harmful consequences.

A chilling, more extreme example can be seen in the Christchurch mosque shooting, which resulted in 51 deaths. Shortly before the attack, the perpetrator, Brenton Tarrant, posted a 74 page manifesto on 8chan, another loosely moderated imageboard modelled after 4chan (Thorleifsson 292). However, his involvement with these platforms would actually begin much earlier. Tarrant had reportedly followed 4chan since his early teenage years and later described it as an important space for

sharing and spreading his ideas (Phillips and Campion 2). He also claimed to have uploaded his manifesto to 4chan in addition to 8chan, suggesting that he viewed these platforms not just as communities but as effective channels for ideological communication and recruitment (Phillips and Campion 3).

In the aftermath of the attack, Tarrant's manifesto circulated widely on 4chan's /pol/ board, where some users expressed admiration and incorporated its themes into ongoing discussions (Thorleifsson 293). This response shows how real-world violence can be absorbed into online culture. More broadly, it highlights the ongoing relationship between platforms like 4chan and the process of radicalization, showing that these platforms can play a role not only in shaping beliefs but in sustaining and amplifying them over time (Phillips and Campion 3).

A similar pattern can be seen in the emergence of QAnon. The conspiracy theory first appeared in the form of anonymous posts on 4chan before gradually moving to platforms such as 8chan. What began as a series of cryptic messages did not stay confined to those spaces for long. Over time, it evolved into a far-reaching movement with tangible political consequences, culminating in the 6 Jan. 2021 insurrection on the US Capitol Building (Kasimov et al. 129). The trajectory of QAnon shows how ideas that originate in small, anonymous communities can gain momentum and circulate far beyond their original context.

The effects described throughout in this paper do not remain contained within online spaces. Instead, they carry over into everyday life, shaping both individual perspectives and broader social dynamics. Even though platforms like 4chan may seem fringe or isolated, the ideas and behaviors that emerge there can have wide-reaching consequences.

For individual users, repeated exposure to highly polarized or extremist content can gradually reshape how they interpret the world. Over time, ideas that can initially seem fringe can begin to feel familiar, and eventually self-evident. On /pol/, for instance, engagement is often driven by outrage and conspiracy-oriented narratives that position far-right perspectives as uniquely credible (Elley 2). Users frequently share so-called "red pills," a term used to describe supposedly hidden truths that

fundamentally reframe the way a person understands the world (Elley 2). These "truths" often rely on misleading statistics, selective evidence, or outright falsehoods, yet they are presented in such a way that encourages anger and alienation (Elley 2).

With enough repetition, this kind of content can become difficult to disengage from. Users may begin to see themselves as part of a small group that understands what others cannot, which may feel both empowering and isolating at the same time (Elley 2). As this perspective deepens, outside sources of information may seem less trustworthy or less relevant, making users increasingly reliant on the platform for both information and community (Elley 2). In this way, the normalization of extremist narratives is not just about belief, but about reshaping how individuals interpret reality itself.

On a broader scale, the circulation of extremist content and conspiracy theories contributes to ongoing patterns of polarization and misinformation. Ideas that take shape on platforms like 4chan do not remain contained there; instead, they often spread to more mainstream platforms, where they reach larger and more diverse audiences. To navigate the moderation and censorship of mainstream sites, some users have adopted strategies to "hide in plain sight." One common strategy involves the use of coded language, or "dog whistles," which signal extremist ideas to those familiar with them while remaining ambiguous to others (Kasimov et al. 131). For example, the use of triple parentheses, often in reference to a nebulous "(((they)))," has been used to reference antisemitic conspiracies without stating them explicitly (Tuters and Hagen 2219). Such memes not only enable the spread of harmful ideas to an "in-the-know" audience but also spark curiosity among others, who may seek to understand the meaning behind the symbols, inadvertently exposing themselves to these ideologies.

Memes play a similar role in this process. Far-right groups like the Proud Boys use them to frame their worldview in ways that appear more approachable or mainstream. By drawing false equivalencies, like comparing Black Lives Matter to groups like the KKK or Nazis, these messages reshape how issues are perceived while maintaining a veneer of neutrality (Kasimov et al. 139). In both cases, memes are used to circumvent moderation and gradually introduce harmful ideas to wider audiences, demonstrating how online extremism can move from niche communities into broader public discourse.

Alongside these broader ideological effects, there are also more direct forms of harm. Platforms like 4chan have been repeatedly linked to harassment and targeted abuse. The combination of anonymity and minimal accountability makes coordinated campaigns easier to organize and carry out. Incidents such as Gamergate demonstrate how these efforts can extend beyond the platform, affecting individuals' personal and professional lives in lasting ways (Massanari 334). In this way, the harms associated with 4chan are not limited to abstract ideological concerns but include tangible, real-world consequences for those targeted.

While these effects are significant, they make more sense when placed within the broader context of online radicalization. Much of the public conversation around this issue has centered on algorithm-driven platforms like TikTok, YouTube, and Facebook, where recommendation systems place a central role in shaping the user experience. These algorithms are built to maximize engagement, often by surfacing content that aligns with a user's existing interests. Over time, this can reinforce increasingly narrow or extreme viewpoints over time (Shin and Jitkajornwanich 1021). Because of this, algorithmic amplification is often treated as the central explanation for online radicalization.

However, the case of 4chan complicates this narrative. Even without recommendation algorithms, the platform still produces many of the same outcomes associated with algorithm-driven systems. Instead of relying on personalized recommendations, 4chan amplifies content through collective user behavior. Posts that provoke strong reactions, whether positive or negative, tend to receive more replies, which keeps them visible longer and draws in more engagement. This process is still only part of the issue. As users respond to and build on each other's posts, they also reinforce shared ways of interpreting events and ideas.

This is especially noticeable in the use of insider language, memes, and symbols. Elements like greentext formatting or triple parentheses require insider knowledge to fully interpret (Ludemann 2732; Kasimov et al. 131). Knowing how to use these memes signals belonging. Thorliefsson describes this phenomenon, writing that "memes work to reinforce the bond of the community and to mark in-group members from new or

inexperienced users and to mere lurkers, members who observe, but do not participate" (290). Additionally, users who understand and replicate these signals are rewarded with recognition, while those who do not are often dismissed or told to "lurk moar [sic]," reinforcing a kind of gatekeeping dynamic (Ludemann 2732). Over time, these interactions help shape a shared worldview that feels internally consistent, even if it conflicts with perspectives outside the platform (Thorleifsson 290).

What emerges from this process is a kind of informal filtering system. Content that resonates with the community tends to stick around and gain traction, while posts that do not fit are ignored or pushed aside. This suggests that the reinforcing "echo chamber" effects often attributed to algorithms can also emerge organically through patterns of user interaction and community structure.

Given these dynamics, it is not surprising that researchers and policymakers have explored a range of strategies to mitigate the harms associated with online radicalization. These approaches can be divided into two broad categories: technical interventions and educational or social strategies. One survey of the field describes them as "hard" and "soft" approaches, where hard approaches involve platform-level changes such as content moderation, while soft approaches focus on education and awareness (Berjawi et al. 21–22). Both play an important role, though neither is without limitations.

One of the more direct responses has been to improve content moderation. This can involve removing extremist material, banning users who repeatedly violate guidelines, or limiting the visibility of harmful posts (Berjawi et al. 21). On larger mainstream platforms, these approaches are frequently supported by machine learning and deep learning systems designed to analyze the content of posts themselves. As Berjawi et al. explain, researchers widely use "[machine learning] and [deep learning] techniques as a content-based approach to detect extremist and hate speech content" (2). These systems focus on identifying problematic language without requiring user reports, allowing platforms to remove harmful material automatically.

In recent years, more advanced models have become central to this work. Neural network-based approaches can process large amounts of text and identify complex

linguistic patterns associated with extremism. Techniques such as “Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM),… Bidirectional Encoder Representations from Transformers (BERT), and Robustly Optimized BERT Approach (RoBERTa) are used to classify content as extremist or non-extremist” (Berjawi et al. 13). Their ability to capture context, tone, and subtle semantic cues makes them particularly useful for identifying coded or implicit forms of extremist language, which are common on platforms like 4chan.

Some approaches look beyond individual posts and instead examine patterns of interaction between users. Graph-based methods, for instance, map relationships between users and posts to identify clusters of coordinated extremist activity. For example, a system called INSiGHT (Investigative Search for Graph-Trajectories) applies graphing algorithms to “match patterns of behavior over time to find groupings connected to extremist organizations” (Berjawi et al. 23). Other systems, such as the TTDF framework, combine real-time data collection with machine learning classification to identify terrorist-related content on Twitter, using a platform that includes “crawling…, pre-processing…, training…, and classification” (Berjawi et al. 23–24). Tools like NewsGuard and Perspective API take a different approach, attempting to mitigate harmful content by assigning credibility or toxicity scores based on attributes such as “toxicity,” “insult,” and “profanity” (Berjawi et al. 24).

Even so, many of the datasets studied by Berjawi et al. were sourced from platforms like Twitter, Facebook, Telegram, and Whatsapp, where each user has a persistent identity (8). On platforms like 4chan, where users are anonymous and content is highly ephemeral, it may be difficult to construct stable networks or track long-term behavioral patterns. As a result, graph-based detection methods and reputation-based systems may be less effective, as there is limited data linking users across posts or tracking them over time.

In contrast to these technical approaches, many researchers place greater emphasis on the importance of education and media literacy. Recent studies consistently point to what are often called “soft” strategies, which are aimed towards helping users recognize misinformation, understand how online communities shape

beliefs, and more critically evaluate the content they encounter (Berjawi et al. 23). Rather than focusing only on removing harmful material after it appears, these approaches try to address the conditions that make individuals more susceptible to it in the first place.

A central component of this strategy involves educational programs that raise awareness about the risks of online radicalization. Governments and research organizations have developed training initiatives designed to show how extremist content is created and spread. For instance, the Radicalization Awareness Network (RAN) has introduced programs that focus on helping young people identify manipulative or misleading content online (Berjawi et al. 23). Other studies similarly highlight the value of structured training in encouraging more critical engagement with online material (Berjawi et al. 23).

Some approaches move beyond formal instruction and instead try to change the kinds of content users encounter. One proposed strategy is to reduce exposure to homogenous viewpoints while increasing interaction with diverse opinions with the goal of disrupting echo chambers and reducing ideological polarization (Berjawi et al. 23). There has also been work on designing discussion platforms to facilitate dialogue between individuals with differing perspectives, encouraging users to engage with opposing viewpoints in a more constructive way (Berjawi et al. 23). These efforts can be seen, in part, as an attempt to counteract the insular nature of communities like 4chan, where shared beliefs are often reinforced rather than challenged.

Looking across these examples, it becomes clear that online radicalization cannot be explained by algorithms alone. While much of the public discourse focuses on recommendation systems as the primary driver of extremism, I argue that platform design plays an equally important role. Features such as anonymity, rapid content turnover, and minimal moderation create an environment where extreme ideas can emerge, spread, and become normalized through user interaction alone. Even without algorithmic curation, 4chan produces many of the same reinforcing dynamics seen on mainstream platforms.

At the level of individual experience, these environments can gradually influence how users make sense of the world. Repeated exposure to extreme or conspiratorial content, even when framed ironically, can shift what feels reasonable or believable. On a wider scale, the effects of this process do not remain contained to niche anonymous messageboards. Ideas developed within these spaces spread outward through memes, coded language, and cross-platform migration, contributing to polarization and misinformation. The case studies discussed in this paper, including Gamergate, the Christchurch mosque attack, and the spread of QAnon, illustrate how online spaces can influence both discourse and real-world outcomes.

Attempts to respond to these issues are complicated. Technical approaches, such as machine learning and graph-based detection algorithms, can help identify harmful content, but they tend to be less effective in fast-moving anonymous environments like 4chan. Educational and media literacy efforts offer a more preventative path, although they require sustained investment and may not reach the individuals most at risk. Any response to these problems must also navigate the tension between limiting harm and preserving free expression.

Ultimately, the case of 4chan shows that the risks associated with social media technologies cannot be reduced to recommendation algorithms alone. The way these platforms are structured and how people use them also plays an important role. Addressing these challenges will likely require not just better technical solutions, but a more nuanced understanding of how online spaces shape identity and belief formation.

Works Cited

Berjawi, Omran, et al. “A Comprehensive Survey of Detection and Prevention Approaches for Online Radicalization: Identifying Gaps and Future Directions.” *IEEE Access* [Piscataway], vol. 11, 2023, pp. 1–1. *ohiolink-cwru.primo.exlibrisgroup.com*, https://doi.org/10.1109/ACCESS.2023.3326995.

Elley, Ben. “‘The Rebirth of the West Begins with You!’—Self-Improvement as Radicalisation on 4chan.” *Humanities & Social Sciences Communications* [London], vol. 8, no. 1, 2021, pp. 1–10. *ohiolink-cwru.primo.exlibrisgroup.com*, https://doi.org/10.1057/s41599-021-00732-x.

Kasimov, Andrey, et al. “‘Pepe the Frog, the Greedy Merchant and #stopthesteal’: A Comparative Study of Discursive and Memetic Communication on Twitter and 4chan/Pol during the Insurrection on the US Capitol.” *New Media & Society* [London, England], vol. 27, no. 1, 2025, pp. 127–50. *ohiolink-cwru.primo.exlibrisgroup.com*, https://doi.org/10.1177/14614448231172963.

Ludemann, Dillon. “Digital Semaphore: Political Discourse and Identity Negotiation through 4chan’s /Pol/.” *New Media & Society*, vol. 25, no. 10, Oct. 2023, pp. 2724–43. *SAGE Journals*, https://doi.org/10.1177/14614448211034848.

Massanari, Adrienne. “Gamergate and The Fappening: How Reddit’s Algorithm, Governance, and Culture Support Toxic Technocultures.” *New Media & Society* [London, England], vol. 19, no. 3, 2017, pp. 329–46. *ohiolink-cwru.primo.exlibrisgroup.com*, https://doi.org/10.1177/1461444815608807.

Phillips, Justin Bonest, and Kristy Campion. “Is He /Ourguy/, a False Flag, or Something Else? Debating Breivik’s and Tarrant’s Terrorism on 4chan’s /Pol/ Board.” *Studies in Conflict & Terrorism*, vol. 0, no. 0, Aug. 2024, pp. 1–23. *Taylor and Francis+NEJM*, https://doi.org/10.1080/1057610X.2024.2388337.

Shin, Donghee, and Kulsawasd Jitkajornwanich. “How Algorithms Promote Self-Radicalization:

Audit of TikTok's Algorithm Using a Reverse Engineering Method." *Social Science Computer Review*, vol. 42, no. 4, Aug. 2024, pp. 1020–40. *SAGE Journals*, https://doi.org/10.1177/08944393231225547.

Thorleifsson, Cathrine. "From Cyberfascism to Terrorism: On 4chan/Pol/ Culture and the Transnational Production of Memetic Violence." *Nations and Nationalism*, vol. 28, no. 1, 2022, pp. 286–301. *Wiley Online Library*, https://doi.org/10.1111/nana.12780.

Tuters, Marc, and Sal Hagen. "(((They))) Rule: Memetic Antagonism and Nebulous Othering on 4chan." *New Media & Society* [London, England], vol. 22, no. 12, 2020, pp. 2218–37. *ohiolink-cwru.primo.exlibrisgroup.com*, https://doi.org/10.1177/1461444819888746.

Whittaker, Joe. "Rethinking Online Radicalization." *Perspectives on Terrorism*, vol. 16, no. 4, Aug. 2022, 159426615, pp. 27–40. *EBSCOhost*, https://doi.org/10.2307/27158150.