RESEARCH REPORT

Reference: PVE-2026-RPT-001

Date: 12 February 2026

To: Department of the Prime Minister and Cabinet

From: Dr Simon McCallum, Victoria University of Wellington

Subject: Game Design Principles and Violent Extremism: Research Findings, Industry Tools, and Educational Content

For: Information, Noting, and Consideration

---

1. Executive Summary

This research considered the relationship between online gaming, recruitment from that environment into violent extremism and options to manage those relationships. It is based on a literature review, a review of the Charlie Kirk assassination attempt and engagement with New Zealand's gaming industry.

The research found that one-third of gamers surveyed internationally reported exposure to extremist content within gaming spaces. One-quarter reported encountering active recruitment attempts. More than 50 percent of female gamers reported experiencing gender-based harassment. Nearly one-third of gamers perceived toxicity as a normalised feature of the communities in which they participate.

Young New Zealanders encounter extreme and objectionable content online with little guidance or support, curiosity is a primary driver of engagement with such content. It is not merely content, but the systems of engagement that create risk for young people.

The NZ gaming sector has identified a clear gap in the tools available for small-to-medium game developers to manage the intersection of community safety and violent extremism prevention.

An AI bot has been developed to fill this gap. The AI Bot is called PVE Bot, it is designed with seven rule-based detection analysers that will operate without any dependence on generative artificial intelligence. This ensures that the tool is accessible to small studios with limited resources and that no community data will need to be sent to overseas AI providers for analysis.

It provides a graduated response system comprising seven levels (Level 0 through Level 6), reflecting the research finding that punitive bans alone are insufficient and that interventions should be proportionate to the severity of the behaviour detected.

There is alignment of interest with the gaming industry to manage this threat using such a bot as 60 percent of gamers report spending less money when exposed to toxicity, 60 percent report quitting communities, and 70 percent report avoiding interactions entirely where toxic behaviour has undermined trust in an online community.

There is also a broader economic argument for intervention as the NZ gaming sector generated NZD 759.57 million in revenue in 2024/2025 and revenue is projected to exceed NZD 1 billion in 2026; currently employing1,418 people

This report recommends consideration of development and trialling of a bot such as PVE bot with a group of SME game developers in conjunction with the New Zealand Game Developers Association. (NZGDA)

There is also a gap in educational material for schools.

Existing resources address online safety as a content problem (harmful material to be avoided) or a behaviour problem (bullying to be addressed). They do not equip students to understand the structural mechanics by which engagement is manufactured and manipulated.

Additional age specific support is needed which equips students to understand the structural mechanics by which engagement is manufactured and manipulated. This content can start by explaining game mechanics and then transfer that understanding to immunise teens to some of the manipulative systems.

This report recommends the introduction of age specific education that build on the existing material. An example of such material is included in as Appendix E which could be trialled in schools in 2026.

Both of these resources are developed based on working with the NZ game development industry and in reference to the international research conducted by the Royal United Services Institute's Extremism and Gaming Research Network (RUSI EGRN). Their mixed-methods study of over 2,200 gamers across seven countries found that one-third of gamers have been exposed to extremist content within gaming spaces, and one-quarter have encountered active recruitment attempts. More than 50 percent of female gamers have experienced gender-based harassment. Nearly one-third of gamers perceive toxicity as normalised within their communities.

The content is based on Self-Determination Theory (SDT) and Te Ao Māori values. SDT identifies three core psychological needs (Autonomy, Competence, Relatedness) whose exploitation is central to radicalisation. Te Ao Māori provides a New Zealand framework (Mana

Motuhake, Pūmanawa, Manaaki, Whānaungatanga, Kaitiakitanga, Kotahitanga, Pakiki, Auahatanga) in which these dimensions are interconnected and inseparable. This insight emerged from Dr McCallum's work with Kevin Shedlock and the Āwhina Māori support programme at Victoria University of Wellington.

Both deliverables align with all three pillars of the DPMC Counter-Terrorism and Violent Extremism Strategy (December 2025): identifying and understanding threats; preventing and reducing radicalisation risk; and protecting people, places, and infrastructure. They also align with the Preventing and Countering Violent Extremism (PCVE) Strategic Framework's whole-of-society approach.

We recommend that DPMC note these research findings and consider their integration with existing government platforms and education initiatives. We also recommend that DPMC consider supporting the further development, testing, and deployment of PvEbot for Discord, subject to ministry approval and integration planning with existing gaming communities.

---

2. Purpose and Scope

This report addresses the following research question: How do game design principles parallel extremist recruitment mechanics, and what practical tools can be developed to address the exploitation of gaming communities for radicalisation?

The scope of the research encompasses gaming communities (with a focus on Discord as the predominant platform for game developer community interaction), the New Zealand game development sector, and the New Zealand education system. The research draws on international literature and case analysis but is directed toward New Zealand-specific applications and policy alignment.

The methodology comprises four components:

a. A literature review of international research on the gaming-extremism nexus, drawing on publications from the Royal United Services Institute Extremism and Gaming Research Network (RUSI EGRN), the Global Project Against Hate and Extremism (GPAHE), the International Centre for Counter-Terrorism (ICCT), and the Frontiers journal series on identity fusion in gaming cultures;

b. Case analysis of the Charlie Kirk assassination attempt (September 2025) and the emerging pattern of Mixed, Unclear, Unstable (MUU) ideology in acts of targeted violence, with specific attention to gaming references and subcultural signifiers;

c. Tool design: the design of PvEbot, a Discord moderation tool grounded in Preventing Violent Extremism (PVE) research, Self-Determination Theory, and Te Ao Māori values;

d. Curriculum design: the development of age-differentiated educational content for New Zealand schools (Years 1 to 13) aligned to the Te Mataiaho curriculum refresh.

This report has been prepared for the Department of the Prime Minister and Cabinet following presentations by Dr McCallum to DPMC (September 2025) and to the New Zealand Game Developers Conference (NZGDC 2025).

---

3. Background: The Gaming-Extremism Nexus

3.1 International Research Landscape

International research on gaming communities and violent extremism has grown since 2019. The largest empirical study was conducted by the Royal United Services Institute's Extremism and Gaming Research Network (EGRN), combining quantitative survey data from over 2,200 gamers across seven countries with qualitative interviews and platform analysis.

The RUSI EGRN research identifies three primary mechanisms by which extremist actors exploit gaming communities:

a. Direct use of games for propaganda and recruitment, including the creation of custom modifications for mainstream titles (Arma 3, GTA V, Roblox user-generated content) that embed extremist messaging within gameplay;

b. Gaming-adjacent platforms, particularly Discord, used as encrypted communication and recruitment infrastructure. Discord servers associated with gaming communities provide a trust-rich environment where recruitment can occur under the cover of shared gaming interests;

c. Gamification of real-world violence, where attacks are framed using game mechanics, aesthetics, and language. This includes the use of scoreboards, achievement framing, and live-streaming in the presentation of violent acts.

The RUSI quantitative findings confirm the scale of the problem. One-third of gamers surveyed reported exposure to extremist content within gaming spaces. One-quarter reported encountering active recruitment attempts. More than 50 percent of female gamers reported experiencing gender-based harassment. Nearly one-third of gamers perceived toxicity as a normalised feature of the communities in which they participate.

These statistics must be read alongside Professor Robert Axelrod's foundational work on the evolution of cooperation. Axelrod demonstrated that when the proportion of defectors (in this context, toxic or hostile participants) exceeds approximately 10 percent of a community,

cooperative behaviour collapses. Community members cease to extend trust, and interactions become guarded or cease altogether. The RUSI data quantifies the economic and social cost of this collapse: 60 percent of gamers report spending less money when exposed to toxicity, 60 percent report quitting communities, and 70 percent report avoiding interactions entirely.

The economic consequences are direct. When Axelrod's cooperation threshold is breached and toxicity becomes normalised, the effects cascade through the community. The 60 percent reduction in spending harms game studios whose revenue depends on community engagement. The 60 percent attrition rate means studios lose their most engaged members: the players who provide feedback, generate content, moderate discussions, and serve as advocates. The 70 percent interaction avoidance rate means even members who remain cease to contribute. For NZ game studios, which are typically small and community-dependent, these effects can be existential.

The GamerGate controversy (2014 onward) demonstrated how gaming communities could be mobilised for coordinated harassment and, subsequently, political radicalisation. What began as a targeted campaign against women in the game industry became, as Bezio (2018) documents, a rehearsal ground for alt-right political mobilisation. Massanari (2017) showed how platform design (algorithms, governance, moderation) actively facilitated this toxic mobilisation. Braithwaite (2016) analysed how the movement weaponised "geek masculinity" to enforce exclusion. Wells et al. (2024) describe GamerGate as "the official debut of organised right-wing extremism on the contemporary centre stage of youth culture." The trajectory from GamerGate to the alt-right demonstrates that gaming community toxicity is not merely a social nuisance; it is a radicalisation pathway.

Other relevant international research includes:

a. The Global Project Against Hate and Extremism (GPAHE) report on extremism in gaming, which documents the use of mainstream gaming platforms as recruitment environments;

b. The International Centre for Counter-Terrorism (ICCT) publication on design principles for prevention and counter-violent extremism (P/CVE) in gaming platforms, which provides an architectural framework for interventions;

c. The Frontiers journal article "Not just a game: Identity fusion and extremism in gaming cultures" (Kowert, Martel, & Swann, 2022), which describes the process by which group identity in gaming communities can override personal identity and create willingness to engage in extreme acts on behalf of the group;

d. The Global Network on Extremism and Technology (GNET) Building Tech Capacity project, which supports the development of technical tools for counter-extremism;

e. Schlegel (2020), who analyses both top-down gamification (extremist organisations using points systems, apps, and radicalisation metrics) and bottom-up gamification (organic emergence of game-like dynamics in extremist online communities, including scoreboards and competitive challenges for violent acts). Valentini (2020), discuss Onlife Extremism, where life is a mix of online and offline events. Discussing only one side of the experience will not capture the pathways.

3.2 The New Zealand Context

New Zealand's direct experience of online radicalisation is defined by the 15 March 2019 Christchurch mosque attacks, in which 51 people were killed and 40 injured. The perpetrator's attack was internet-native in its conception and execution. The manifesto was structured as online content. The attack was live-streamed on a platform adjacent to gaming communities. The perpetrator's engagement with extremist content occurred substantially in online spaces. Battersby and Ball (2019) situate this attack within a longer history of right-wing extremism in New Zealand, demonstrating that the country was not immune to such violence prior to 15 March. Lakhani and Wiedlitzka (2023) provide empirical analysis of how the attacker deliberately gamified the violence: livestreaming in first-person-shooter format, using kill-count comparisons to prior attacks, and framing the event on 8chan with game-like mechanics. Cunningham, La Rooij, and Spoonley (2023) edited a volume where the chapters provide the most comprehensive account of the radical right in Aotearoa New Zealand, tracing the ideological ecosystem that connects to online radicalisation.

The Royal Commission of Inquiry into the Christchurch mosque attacks (2020) made recommendations spanning intelligence, firearms, social cohesion, and online regulation. In response, the New Zealand Government established He Whenua Taurikura as the National Centre of Research Excellence for Preventing and Countering Violent Extremism (which has since ended), and developed the PCVE Strategic Framework, which adopts a whole-of-society approach to prevention.

The DPMC Counter-Terrorism and Violent Extremism Strategy, released in December 2025, provides the current strategic framework. It is built on three pillars: (1) Identify, Understand, and Disrupt Threats; (2) Prevent and Reduce Radicalisation Risk; and (3) Protect People, Places, and Infrastructure. The deliverables described in this report align with all three pillars, as detailed in Section 7.

The NZ Classification Office has published research on extremism and online harms, including *Content that Crosses the Line* (2023), a qualitative study finding that young New Zealanders encounter extreme and objectionable content online with little guidance or support, and that curiosity is a primary driver of engagement with such content. The Films, Videos, and Publications Classification Act 1993 provides a legislative framework for objectionable material, and the Harmful Digital Communications Act 2015 addresses online harassment. Enforcement in gaming-adjacent platforms is difficult: content is ephemeral, encoded in subcultural language, and distributed across multiple jurisdictions. These spaces require complementary community-level tools that operate at the speed and granularity of the platforms themselves.

The New Zealand Game Developers Association represents over 3,000 individual members and more than 50 studio members. NZ game developers build communities around their products, predominantly on Discord. These communities are the social infrastructure of the sector. When toxicity or extremist exploitation damages them, the cost flows directly to engagement, revenue, and reputation.

The Australian eSafety Commissioner has documented extremist exploitation of user-generated content platforms popular with New Zealand users. On Roblox (a platform hosting user-generated games, comparable to Steam or Xbox), Australian police have identified recreations of Nazi concentration camps, Chinese Communist re-education camps for Muslims, and Islamic State-style conflict zones embedded within user-created content. Minecraft has been similarly exploited. These platforms are among the most popular with New Zealand children and young people, and the user-generated content model creates vulnerabilities that platform-level moderation struggles to address at scale.

The sector has identified a clear gap in the tools available to it. No purpose-built, research-informed moderation tool exists for small-to-medium game developers to manage the intersection of community safety and violent extremism prevention. Commercial moderation tools focus on content removal after the fact; they do not address the underlying psychological drivers of radicalisation, do not provide cultural context, and do not offer pathways back into a supportive community for individuals who have been excluded. This gap is the direct motivation for the design of PvEbot. PVE can stand for Prevention of Violent Extremism, or in gaming terminology PvE is used for Player vs Enemy, as opposed to PvP Player vs Player. PvE environments are where players defeat an enemy rather than fight each other. PvE environments are considered safer that PvP.

3.3 Post-Ironic Violence: A New Pattern

The assassination of Charlie Kirk in the United States in September 2025 illustrates a new pattern in targeted violence. The case did not occur in New Zealand, but it demonstrates an emerging pattern that traditional analytical frameworks cannot address.

The perpetrator, Tyler Robinson, was described in reporting and subsequent analysis as "chronically online," with a social life centred substantially on Discord and gaming communities. His digital footprint was extensive and deeply embedded in internet culture. His attack was saturated with references that were simultaneously sincere and ironic, meaning that traditional motive analysis misses the motivations for the killing.

Bullet casings recovered from the scene were inscribed with references to the game Helldivers 2 (in which players "defend democracy" against aliens and robots under the direction of a satirically fascist government), the Italian anti-fascist song "Bella Ciao" (which also features prominently in the video game Far Cry 6), and internet meme language including trolling phrases directed at whoever would read them. The inscriptions combined anti-fascist sentiment, gaming references, trolling humour, and lethal violence in a single violent act.

This pattern has been classified by researchers and intelligence analysts as MUU ideology: Mixed, Unclear, and Unstable. Unlike traditional single-issue extremism (whether Islamist, ethno-nationalist, or single-issue), MUU ideology is characterised by a "salad bar" approach in which the perpetrator selects elements from multiple, often contradictory, ideological sources. The unifying thread is not ideological coherence but rather the internet-native culture of irony, escalation, and transgression that permeates certain gaming and online communities.

The implications for prevention are direct. Content moderation systems designed to detect specific ideological keywords (white supremacist terminology, Islamist recruitment language) will fail to identify MUU-pattern actors whose language is coded in irony and gaming references. Detection requires subcultural literacy: an understanding of the memetic language, gaming references, and layered irony that characterise these communities. This requirement informed the design of PvEbot's post-ironic violence analyser, described in Section 6.

The Charlie Kirk case reinforces a distinction central to this report: gaming culture and political violence intersect not because games cause violence, but because gaming communities and their adjacent platforms provide the social infrastructure through which vulnerable individuals encounter and act upon extremist ideas. The research on video game violence is clear that violent content is not correlated with an increase in social violence. It claims that gaming communities, like any community, can be exploited, and that game design mechanics provide a useful analytical lens for understanding how that exploitation works.

3.4 Game Design Mechanics as Recruitment Parallels

There are structural parallels between game design mechanics and extremist recruitment tactics. These parallels are grounded in established research on persuasive design. Zagal, Bjork, and Lewis (2013) define "dark design patterns" in games: design elements whose purpose is to advance the designer's interests at the player's expense, including temporal manipulation, social pressure exploitation, and obscured costs. Hamari, Koivisto, and Sarsa (2014) demonstrate through a systematic review of 24 empirical studies that gamification (points, badges, leaderboards, progression systems) reliably shapes behaviour. These same mechanics appear in extremist recruitment contexts. The parallels are not metaphorical. They reflect the same underlying psychological principles applied in different contexts.

The following table summarises the key parallels:

Game Design Mechanic	**Recruitment Tactic**	**Psychological Function**
Creating belonging and community	Welcoming vulnerable individuals into in-group	Need for Relatedness
Creating clear opponents (us-vs-them)	Defining out-groups and enemies	Enemy framing and identity consolidation

Game Design Mechanic	Recruitment Tactic	Psychological Function
Engagement and commitment cycles	Escalating involvement and commitment	Sunk-cost exploitation and escalation commitment
Secret knowledge and hidden content	"Redpilling": revealing "hidden truths"	Epistemic manipulation and in-group distinction
Respect and acknowledgement (internal)	Status and recognition within the group	In-group status and competence fulfilment
Progression mechanics and levelling	Radicalisation pipeline stages	Structured escalation from casual to committed

These parallels are not coincidental. Przybylski, Rigby, and Ryan (2010) demonstrate that games engage players by satisfying three basic psychological needs (autonomy, competence, and relatedness), producing powerful motivational pull. Both game designers and extremist recruiters are, in effect, solving the same problem: how to engage a person, hold their attention, create emotional investment, and motivate increasingly committed action. Game designers do this to create entertainment. Extremist recruiters do this to create soldiers. The mechanics are transferable because the underlying psychological needs are the same. Players seeking to satisfy these needs could do so in a game or via a radicalisation group.

Understanding these parallels is valuable for two reasons. First, it provides a framework for detection: the stages of recruitment can be mapped to the stages of game engagement, allowing community managers to identify when engagement patterns shift from healthy to exploitative. Second, it provides a framework for prevention: by ensuring that community members' genuine psychological needs are met through healthy engagement, the appeal of extremist recruitment is diminished.

---

4. Theoretical Framework

4.1 Self-Determination Theory (SDT)

Self-Determination Theory, as developed by Deci and Ryan (2012), identifies three core psychological needs that are essential for human wellbeing and intrinsic motivation:

> **a. Autonomy**: the need for self-determination, choice, and agency. Individuals need to feel that their actions are self-endorsed and volitional, not coerced or externally controlled;

b. Competence: the need for mastery, meaningful contribution, and the experience of effectiveness. Individuals need to feel that they can achieve outcomes and that their contributions matter;

c. Relatedness: the need for belonging, genuine connection, and caring relationships. Individuals need to feel that they are part of a community and that they are valued by others.

The relevance of SDT to violent extremism is direct and well-supported. Kruglanski et al. (2014) introduce Significance Quest Theory, demonstrating that the desire to matter (closely paralleling SDT's competence and relatedness needs) is a central motivational driver of radicalisation. Their 3N model (Need, Narrative, Network) formalised in Kruglanski et al. (2018) shows how unfulfilled significance needs lead to extremist action when coupled with a justifying narrative and an enabling network. Vansteenkiste, Ryan, and Soenens (2020) articulate the "dark pathway" of need frustration: the active thwarting of autonomy, competence, and relatedness, which is distinct from mere absence of satisfaction. Need frustration creates vulnerability. An individual who lacks genuine belonging (Relatedness) is susceptible to groups offering false community. An individual who lacks purpose or self-direction (Autonomy) is susceptible to groups offering coercive ideology framed as a mission. An individual who lacks recognition or mastery (Competence) is susceptible to groups offering in-group status and the experience of significance.

Each stage of the radicalisation pipeline can be mapped to a specific SDT deficit:

Radicalisation Stage	**SDT Need Exploited**	**Mechanism**
Initial contact and welcome	Relatedness	Vulnerable individual receives warmth, acceptance, belonging
Ideological framing	Autonomy	Individual is offered a worldview that provides clarity and purpose
Skill development and contribution	Competence	Individual is given tasks, roles, and recognition within the group
Escalating commitment	All three	Sunk-cost effects, identity investment, and social pressure compound

Radicalisation Stage	SDT Need Exploited	Mechanism
Action	Identity fusion	Group identity overrides personal identity; action is taken on behalf of group

SDT also identifies negative drivers: actions and dynamics that undermine psychological needs and increase vulnerability:

a. Coercion (anti-Autonomy): controlling behaviour, ultimatums, removal of choice;

b. Imposter syndrome amplification (anti-Competence): persistent messaging that an individual is inadequate, unworthy, or unable to contribute meaningfully;

c. Othering (anti-Relatedness): exclusion, dehumanisation, and the denial of belonging to individuals or groups.

These negative drivers are detectable in community discourse and serve as early indicators of a community that is in trouble. PvEbot is designed to monitor for these patterns as part of its behavioural analysis capabilities.

4.2 Te Ao Māori Values

The integration of Te Ao Māori values into this research is not a checkbox ticking exercise like some Vision Mātauranga statements. Instead, the dimensions identified by SDT are, in Te Ao Māori, understood as fundamentally interconnected and incapable of existing in isolation.

The following table maps the Te Ao Māori values to their SDT correspondences and their application in the research:

Te Ao Māori Value	SDT Correspondence	Application
Mana Motuhake (self-determination)	Autonomy	Every person has inherent mana. Moderation actions protect mana; they do not diminish it
Pūmanawa (inherent potential)	Competence	Everyone has unrealised potential. Rehabilitation is framed as unlocking potential, not as punishment served

Te Ao Māori Value	SDT Correspondence	Application
Manaaki / Manaakitanga (caring, hospitality)	Relatedness	Care for others is a communal responsibility. The tool models manaaki in all responses
Whānaungatanga (relationships, belonging)	Relatedness	Genuine belonging is built through relationships, not transactions
Kaitiakitanga (guardianship)	Community responsibility	Community members are guardians of each other's wellbeing
Kotahitanga (unity, collective action)	Collective resilience	Working together across differences to maintain community health
Pakiki (critical questioning)	Critical thinking	Genuine inquiry is valued and protected; bad-faith "just asking questions" (a known recruitment tactic) is distinguished from authentic curiosity
Auahatanga (creativity)	Creative expression	Creative expression, including game development, is valued as positive community engagement

The key insight from Te Ao Māori is that for collectivist cultures, including Māori and Pasifika communities, the three SDT dimensions cannot be addressed independently. Autonomy without Relatedness is isolation. Competence without community recognition is meaningless. Relatedness without respect for Mana Motuhake is assimilation, not belonging. Effective approaches to motivation, resilience, and community safety must engage all three dimensions simultaneously.

This insight has practical implications for both the PvEbot design and the educational content. Intervention messages that address only one dimension (for example, a warning that appeals solely to individual choice without acknowledging the individual's place in the community) will be less effective for people from collectivist-cultures than messages that integrate all three. The educational content is designed with this principle throughout, not as a separate cultural module but as an integrated framework.

4.3 Radicalisation Pipeline Model

The research adopts a four-stage radicalisation pipeline model, drawing on established frameworks in the literature. Moghaddam (2005) proposes a six-floor "staircase to terrorism" in which individuals progressively narrow their perceived options until violence appears to be the only course. Borum (2003) identifies a four-stage cognitive escalation from perceived deprivation through injustice framing to dehumanisation. Horgan (2008) argues against static profiles in favour of understanding radicalisation as a dynamic process with distinct phases of becoming involved, being involved, and disengaging. McCauley and Moskalenko (2017) separate radicalisation of opinion from radicalisation of action, noting that most individuals with radical views never act. The model adopted here synthesises these frameworks and is refined through case analysis:

a. Stage 1, Desensitisation: Hateful rhetoric is normalised within the community. Toxic language, dehumanising "jokes," and transgressive humour create an environment in which extremist content does not register as exceptional. The Axelrod threshold (approximately 10 percent toxic participants) is relevant here: once toxicity is normalised, the community's capacity for self-correction is compromised;

b. Stage 2, In-group identity formation: Community members develop a shared identity that is defined substantially by opposition to out-groups. "Us-versus-them" framing becomes the dominant mode of discourse. Loyalty to the in-group is tested through increasingly extreme statements and actions;

c. Stage 3, Ideological framing: Extremist actors frame their ideology as the logical defence of the in-group identity established in Stage 2. The ideology is presented not as an external imposition but as an organic extension of the group's existing values and grievances;

d. Stage 4, Identity fusion and action: The individual's personal identity becomes inseparable from the group identity. The Frontiers research on identity fusion in gaming cultures describes this as the point at which an individual is willing to take extreme action on behalf of the group, not because they have been ordered to, but because the group's cause has become their own. Pre-existing social bonds form the basis of this fusion.

This pipeline model is not deterministic. Most individuals who enter Stage 1 will **not** progress to Stage 4. However, each stage creates the conditions for the next, and without intervention, the pipeline operates as a ratchet: it is easier to move forward along the pipeline than to move back. This observation informs the design of PvEbot's graduated response system, which is intended to intervene at the earliest possible stage.

---

5. Key Deliverable 1: Educational Content for Schools

5.1 The Education Gap

New Zealand has an established landscape of digital safety education. Netsafe is the primary agency, providing the Cyberbullying Prevention Toolkit (a progressive learning pathway for Years 5 to 13 aligned with Social Sciences, Health, and Digital Technologies), the Headspace Invaders gamified platform (which explores misinformation and identity-based harm), and the Sticks 'n Stones peer-led anti-bullying programme for secondary schools. Keep It Real Online, operated by the Department of Internal Affairs, provides parent and educator resources on online harms including violent extremism. The Think: Protect: Connect toolkit, developed through the United Nations, is available via Keep It Real Online and addresses safeguarding against online radicalisation. Bullying-Free NZ provides a nine-element school framework for systemic bullying prevention, and the Ministry of Education's "Ata" and "Oho" collections offer Social and Emotional Learning modules. The Ministry's Digital Citizenship Modules provide further curriculum-aligned resources.

These resources are valuable and widely used. They represent significant government and sector investment in online safety education, and the deliverables described in this report are designed to complement and extend them, not to replace them. The educational content directs teachers to use established Netsafe and Ministry of Education resources alongside our modules. However, there are a couple of specific gaps that the existing resources do not currently address:

> **a.** Game-design-informed understanding of manipulation mechanics. Existing resources address online safety as a content problem (harmful material to be avoided) or a behaviour problem (bullying to be addressed). They do not equip students to understand the structural mechanics by which engagement is manufactured and manipulated: the progression systems, commitment escalation, and in-group/out-group dynamics that operate beneath the level of specific content. This understanding requires game design literacy, which is the specific expertise of the game development sector;
>
> **b.** Age-differentiated activity-based learning that connects bullying behaviour to radicalisation pathways. Bullying is an entry-point behaviour. The dynamics of exclusion, in-group formation, status competition, and escalation that characterise school bullying are the same dynamics that, in online contexts, serve as the early stages of radicalisation. Existing resources tend to address bullying and extremism as separate topics. The educational content developed through this research explicitly connects them, showing students (at age-appropriate levels) how the same patterns operate across contexts.

5.2 Curriculum Alignment: Te Mataiaho

The educational content is designed for alignment with Te Mataiaho, the Ministry of Education's curriculum refresh for New Zealand schools. Te Mataiaho employs an Understand-Know-Do framework that emphasises conceptual understanding, knowledge, and practical capability. The draft for Years 0 to 10 was released in Term 4 2025, with implementation planned for Term 1 2027.

The primary curriculum alignment is with Te ao tangata (the Social Sciences learning area), particularly the new Civics and Society strand, which addresses how individuals and communities participate in democratic life and maintain social cohesion. The educational content also connects to the Digital Technologies learning area where modules address online platforms and digital communication.

The five Key Competencies of the New Zealand Curriculum are addressed across the modules:

Key Competency	**How It Appears in the Modules**
Thinking	Critical analysis of manipulation mechanics, SDT analysis of case studies, systems thinking in senior modules
Using Language, Symbols, and Texts	Media literacy, understanding coded language and symbols, deconstructing memes and ironic content
Managing Self	Emotional regulation, recognising personal vulnerability to manipulation, confidence self-assessment
Relating to Others	Upstander behaviour, community guardianship, manaaki in online spaces
Participating and Contributing	Community design activities, policy analysis, active contribution to safe communities

5.3 Pedagogical Approach

The educational content follows a pedagogical approach grounded in social constructivism: students build knowledge through experience and dialogue socially. The teacher's role is "guide on the side" rather than "sage on the stage." Teachers are positioned as lead learners, exploring these ideas alongside students.

Every module follows the same pattern: activity first, debrief second. Students experience something through a game, simulation, or structured exercise. The teacher then leads a

structured discussion about what was discovered during the activity. The Te Ao Māori values and SDT connections are introduced as part of the debrief, not as pre-loaded vocabulary.

Didactic content (material that must be directly taught) is kept to approximately 10 to 15 percent of each module. The remainder is experiential learning and facilitated discussion. This ratio reflects both the pedagogical evidence (students retain and apply knowledge more effectively when they construct it through experience) and the practical reality that didactic content about extremism risks being perceived as patronising ("cringe") by older students, undermining its effectiveness.

The Six Animal Avatars model developed by Dr McCallum is introduced from Years 7 to 8 onward as a structured group work tool. Post covid many students still have limited ability to work effectively in groups. The system is designed to teach Self Determination Theory as part of group dynamics and so can be used to introduce the psychological context linked to student behaviour. Each avatar represents a functional role within a team (Leader, Manager, Risk Manager, Enthusiast, Process Master, Facilitator), mapped to SDT dimensions and animals to make it easier to remember. The cognitive separation between the student and the role provides a buffer from criticism: comments can be directed at the role ("The Cat should have flagged that risk") rather than the person. This lightens the mood and makes group dynamics discussable.

> **a.** The Six Animal Avatars model has a specific function within the PVE education context. Each role maps to a combination of SDT needs. When students rotate through different animals across modules, they gain practical experience of each psychological dimension. This builds the fluency with SDT that the Years 9 to 13 modules require for analysing recruitment mechanics, identity fusion, and radicalisation pathways. The connection to prevention is direct: a student who has practised identifying risks in a classroom group (the Cat role) is better equipped to identify the "secret knowledge" recruitment technique; a student who has practised maintaining group inclusion (the Wolf role) is more likely to act as an upstander when they encounter exclusion online. The model teaches students to recognise and practise the same psychological competencies that extremist recruiters exploit, with the goal of being able to identify and defuse these tactics.
>
> **b.** The Six Animal Avatars model is IP developed by Dr McCallum. It can be licenced for use in New Zealand state and state-integrated schools and kura. Commercial use in paid products or professional development requires written permission. Schools may adapt the model for their context (for example, using animals that are more aligned with their kura) provided the SDT mapping and functional role structure are maintained. Licensing details are set out in the Teacher Guide.

Confidence self-assessment is introduced from Years 9 to 10 onward. Students rate their confidence in different aspects of their work, building meta-knowledge: knowing what you know and what you do not know. This is a transferable skill that also serves as a protective factor against "redpilling" (the extremist tactic of claiming to reveal hidden truths to those who "dare to

see"), because students who have practiced assessing what they know are better equipped to evaluate claims about secret knowledge.

5.4 Age-Differentiated Content

The educational content is differentiated across five age bands, comprising 14 modules in total:

Years 1 to 3 (ages 5 to 7): Two modules.

At this level, the approach is very light touch. Modules focus on kindness (online and offline), fair play, and identifying trusted adults. Te Ao Māori values are introduced as simple kupu (words), specifically Manaaki (caring) and Whānaungatanga (relationships), within activity debriefs. SDT is implicit: activities naturally support Autonomy, Competence, and Relatedness without naming the theory. There is no mention of extremism. The foundation being laid is the understanding that communities have norms, that kindness matters, and that unfairness should be identified and addressed.

Module 1 (Kindness Online and Offline) uses simple cooperative games followed by debrief questions such as "How did it feel when someone was kind to you in the game?" and "What could we do if someone is not being kind?" Module 2 (Good Friends, Good Games) explores fair play through games with unfair rules, with the debrief drawing out the concept of Kaitiakitanga (guardianship) by asking "Who looks after us when things are not fair?"

Years 4 to 6 (ages 8 to 10): Three modules.

Modules address fair play and exclusion, standing up for others (bystander-to-upstander behaviour), and digital footprints. Bullying is explicitly addressed as an entry-point behaviour: the dynamics of exclusion, in-group formation, and escalation are explored through activities and then connected, in the debrief, to broader patterns. Kotahitanga (unity) and Pakiki (critical questioning) are introduced alongside the values from Years 1 to 3. SDT remains implicit but activities are deliberately designed to address all three needs.

The Standing Up, Not Standing By module is particularly significant for the broader research objectives. It introduces the concept of the bystander effect in age-appropriate language and then equips students with concrete strategies for becoming upstanders: individuals who act when they witness harm. The debrief connects this to Whānaungatanga (we are all in relationship with each other) and Manaaki (we have a responsibility to care). This module lays the groundwork for the more sophisticated treatment of community guardianship in later years.

Years 7 to 8 (ages 11 to 12): Three modules.

Modules address online communities (design, trust, moderation), pressure and persuasion techniques, and bystander-to-upstander behaviour in online contexts. The six animal avatars are introduced for structured group work. SDT is named in plain language (choice, mastery, belonging) and connected explicitly to Te Ao Māori values. Students design their own online

community rules and moderation approaches, experiencing the trade-offs involved. The full range of Te Ao Māori values is available for debrief discussions.

Years 9 to 10 (ages 13 to 14): Three modules.

Modules address game design mechanics as manipulation tools, identity and belonging (including identity fusion), and critical media literacy (including post-ironic content and algorithmic amplification). SDT is taught explicitly and used as an analytical framework. Students analyse how game mechanics exploit each SDT need and apply the same analysis to recruitment tactics. Confidence self-assessment is introduced. Animal avatars are used with full intentionality: students choose roles based on the task and reflect on how the role affected the discussion.

The Game Design and Manipulation module is the centrepiece of the Years 9 to 10 content. Students examine real game design mechanics (engagement loops, variable reward schedules, progression systems, social pressure mechanics) and then analyse how the same mechanics appear in recruitment tactics. This is not presented as a condemnation of games; rather, it equips students with design literacy, the ability to see the mechanics operating beneath the surface of any designed experience, whether a game, a social media platform, or a recruitment campaign. Students who understand how engagement is manufactured are better equipped to make conscious choices about where they invest their attention and identity.

The Identity and Belonging module addresses identity fusion directly, using the Frontiers research as a foundation. Students explore the difference between healthy belonging (where group membership enriches personal identity) and identity fusion (where group identity replaces personal identity). Activities include scenario analysis in which students identify at what point a fictional character's relationship with a group shifts from healthy to fused. The debrief connects this to Mana Motuhake: genuine belonging does not require the surrender of personal identity.

Years 11 to 13 (ages 15 to 16+): Three modules.

Modules address radicalisation pathways (the full pipeline model with SDT analysis), designing safe communities (moderation design, rehabilitation pathways, privacy considerations), and systems thinking applied to the New Zealand counter-terrorism ecosystem (stakeholder analysis, policy analysis). SDT is a full analytical framework. Students engage in rubric negotiation for assessment (choosing the weighting of assessment criteria within defined ranges), building Autonomy into the assessment process. The final module includes a policy brief assignment in which students analyse a real aspect of the NZ PCVE landscape.

The Designing Safe Communities module asks students to design a moderation system for a fictional online community. This exercise requires students to grapple with the same trade-offs that face real community managers: the tension between safety and free expression, the challenge of proportionate response, the question of rehabilitation versus exclusion, and the privacy implications of monitoring. Students present their designs using the animal avatar roles

and evaluate each other's designs against SDT criteria. This module draws directly on the PvEbot design process and provides students with genuine design experience.

The Systems Thinking and Extremism module positions students as policy analysts examining the New Zealand counter-terrorism ecosystem. Students map the stakeholders (DPMC, NZ Police, DIA, Netsafe, He Whenua Taurikura, the Classification Office, community organisations, technology companies, educators), identify the relationships between them, and analyse gaps. The policy brief assignment requires students to write a formal recommendation on one aspect of the system, using the same format and conventions employed in this report. This is the most sophisticated module in the package and is designed for Year 12 and 13 students.

The Years 9 to 13 modules have been mapped to specific NCEA Social Sciences assessment standards. Three standards align directly with the module activities: AS 92050 (Level 1, demonstrating understanding of decisions on a social issue), AS 92051 and AS 91599 (Level 1 and Level 3 social action standards), and Unit Standard 30910 (Level 1, identifying strategies to respond to online bullying). Module activities generate assessment evidence directly, so students can meet standard requirements through the learning activities rather than separate assessment tasks. The educational content includes Achieved, Merit, and Excellence grading examples for each standard, enabling teachers to use the modules as contexts for NCEA internal assessment.

5.5 Te Ao Māori Integration

Te Ao Māori values are not a separate module or an add-on section within the educational content. They are woven throughout every activity and every debrief at every age level. Values are introduced in context. When an activity raises a question about fairness, the debrief names Kotahitanga. When an activity explores caring for others, the debrief names Manaaki. They are not presented as vocabulary lists to be memorised.

This approach is aligned with the Tātaiako framework for cultural competencies for teachers of Māori learners, which emphasises the integration of Māori perspectives throughout teaching practice rather than their confinement to separate cultural modules.

The key insight from the theoretical framework is maintained throughout: SDT dimensions are interconnected, not independent. Activities and debriefs consistently model this interconnection. A discussion about belonging (Relatedness/Whānaungatanga) leads naturally to a discussion about respect for individual choice (Autonomy/Mana Motuhake) and the recognition of potential (Competence/Pūmanawa). This is not a pedagogical convenience. It is a reflection of the Te Ao Māori understanding that these dimensions cannot be separated.

For Māori and Pasifika students, this integrated approach is pedagogically essential. Motivation frameworks that treat Autonomy, Competence, and Relatedness as independent variables will be less effective for students from collectivist cultural backgrounds, for whom these dimensions are experienced as a unified whole. The educational content is designed to work for all students, but it is particularly attentive to the needs of Māori and Pasifika learners, who are

disproportionately represented in both the at-risk populations for online radicalisation and the populations underserved by individualistic pedagogical approaches.

---

6. Key Deliverable 2: PvEbot, Community Safety Tool Design

6.1 The Gap in Current Tools

Existing community moderation tools have a significant gap. Commercial tools (Discord's built-in AutoMod, third-party bots such as MEE6 and Dyno, and platform-level content moderation systems) focus on content removal. They detect prohibited words or phrases, delete messages, and impose sanctions (mutes, kicks, bans). They do not address the psychological drivers that make individuals vulnerable to radicalisation. They do not provide pathways back into community for individuals who have been banned.

This content-removal approach has a specific and well-documented failure mode: when an individual is banned from a community, they migrate to an unmoderated or less-moderated platform. In these spaces, radicalisation accelerates without oversight, community norms, or the possibility of intervention. The individual's grievance is compounded by the experience of exclusion, further undermining Relatedness and increasing vulnerability to groups that offer belonging. The ban, intended as a protective measure, becomes an accelerant.

No research-informed tool existed for small-to-medium New Zealand game developers to manage the intersection of community safety and violent extremism prevention. Game Developers at the workshop as part of NZGDC identified this gap, and PvEbot was designed to address it.

The economic case for intervention is supported by the RUSI data. The loss of 60 percent of spending, 60 percent of community membership, and 70 percent of social interaction when toxicity is present represents a material commercial risk to a sector that generated NZD 759.57 million in revenue in 2024/2025 and is projected to exceed NZD 1 billion in 2026. The majority of NZ game studios are small enterprises whose commercial model depends on community engagement for revenue and growth. With 95 percent of sector revenue derived from international exports, damage to community health has a direct impact on New Zealand's export earnings and the 1,418 jobs the sector now supports.

6.2 Detection Architecture

PvEbot is designed with seven rule-based detection analysers that will operate without any dependence on generative artificial intelligence. This design decision will ensure that the tool is accessible to small studios with limited resources and that no community data will need to be sent to overseas AI providers for analysis. This was identified by the Game Developers as a

critical requirement for their game communities which are accurately suspicious of International AI tools.

The seven analysers and their intended functions are as follows:

Analyser	Function	Research Basis
Keyword	Designed to detect extremist terminology with resistance to common obfuscation techniques (character substitution, spacing, Unicode manipulation)	RUSI pattern databases and established extremist lexicons
Pattern	Designed to identify coded numbers (1488, 88, 14), symbols, and platform migration links (invitations to move conversation to less-moderated platforms)	Dr McCallum's modus operandi analysis of extremist signalling
Behavioural	Designed to monitor message frequency anomalies, channel-hopping patterns, and the toxic fraction of community discourse	Axelrod's Evolution of Cooperation (toxic fraction threshold)
Contextual	Designed to distinguish legitimate game discussion from genuine threats, reducing false positives in gaming communities where violent language is common in game context	Post-ironic violence research and the challenge of context-dependent content
Post-ironic	Designed to detect violence normalised through irony, memes, and gaming language, identifying the layered sincerity characteristic of MUU-pattern content	Charlie Kirk case analysis and MUU ideology research
Recruitment	Designed to identify us-versus-them framing, platform migration invitations, "redpill" language, and escalating commitment patterns	RUSI finding that one-quarter of gamers encounter recruitment attempts
Identity fusion	Designed to detect dehumanisation of out-groups, willingness-to-fight rhetoric, martyrdom glorification, and language indicating personal-group identity merger	Frontiers research on identity fusion and extremism in gaming cultures

Each analyser is designed to operate independently and produce a confidence score. The scores will be aggregated using a weighted system that accounts for the co-occurrence of multiple indicators (for example, recruitment language combined with platform migration

invitations and identity fusion markers). This aggregation will reduce false positives while maintaining sensitivity to genuine risk patterns.

The contextual analyser is designed to address a challenge unique to gaming communities. In a gaming Discord server, language that would be alarming in other contexts is routine. Players discuss "killing," "destroying," "wiping out the enemy," and "headshots" as normal gameplay vocabulary. Without contextual understanding, a keyword-based system would generate overwhelming false positives. The contextual analyser will maintain an understanding of the community's gaming context and distinguish between language used in game discussion and the same language used in non-game contexts. This distinction is not merely lexical; it will involve analysis of conversational flow, channel context (a message in a game-discussion channel will be treated differently from the same message in a general chat channel), and temporal patterns (a burst of violent language during a game session will be treated differently from the same language in an unprompted conversation).

The post-ironic violence analyser is designed to translate MUU ideology research into operational tooling. This analyser will identify content in which violence is normalised through layers of irony and humour, such that the speaker maintains plausible deniability ("I was just joking") while the content functions as genuine desensitisation or signalling. The analyser will look for patterns including escalating transgression within conversations, the use of gaming aesthetics to frame real-world violence, the combination of conflicting ideological references (characteristic of MUU ideology), and the specific meme formats and language patterns associated with post-ironic violence communities.

The design allows studios to optionally enable generative AI enhancement using Anthropic Claude, OpenAI, or a locally hosted open-source model via Ollama. This will be strictly opt-in. Studios will provide their own API keys or local infrastructure. The tool will fall back to rule-based detection automatically if AI is unavailable. The expected path for studios that choose to enable AI-enhanced detection is locally hosted open-source models (such as Llama 3.2 via Ollama), which can run on New Zealand-based infrastructure, ensuring that AI inference data also remains within Aotearoa.

6.3 Graduated Response System

The design includes a graduated response system comprising seven levels (Level 0 through Level 6), reflecting the research finding that punitive bans alone are insufficient and that interventions should be proportionate to the severity of the behaviour detected.

Level	Response	Automatic?	Description
L0	Monitoring	Yes	Passive observation. No user-facing action. Pattern data will be logged for trend analysis

Level	Response	Automatic?	Description
L1	Nudge	Yes	Subtle positive redirection. The bot will introduce prosocial content or redirect conversation without identifying the trigger
L2	Warning	Yes	Private message to the user referencing community guidelines. Framed as a choice (supporting Autonomy) and acknowledging past positive contributions (supporting Competence)
L3	Restriction	No, moderator approval	Reduced channel access. The user will retain community membership but with limited scope
L4	Mute	No, moderator approval	Temporary communication restriction. The user will be able to read but not post
L5	Kick	No, moderator approval	Removal from the server with the option to rejoin after a cooldown period
L6	Ban	No, moderator approval	Categorised ban (toxic, abusive, extremist, or illegal) with a structured rehabilitation pathway (except for illegal activity)

Levels 0 through 2 will be applied automatically by the bot. Levels 3 through 6 will require human moderator approval before any action is taken. This will ensure proportionality and prevent over-reliance on automated systems. For Levels 3 through 6, the tool will generate triage summaries for moderators that include the detected patterns, the recommended response level, the rationale grounded in SDT and Te Ao Māori values, and suggested message templates.

Every user-facing response generated by PvEbot will be grounded in Self-Determination Theory and Te Ao Māori values. Warning messages will be framed as choices, not commands (supporting Autonomy and Mana Motuhake). Past positive contributions will be acknowledged when issuing interventions (supporting Competence and Pūmanawa). Connection to the community will be maintained even during restrictions (supporting Relatedness and Whānaungatanga). Ban messages will preserve mana while being firm about behaviour.

6.4 Rehabilitation Pathways

Every ban category except "illegal" will have a structured rehabilitation pathway. This feature addresses a question: "Whose job is it to rehabilitate banned players?" The answer embedded in PvEbot's design is that rehabilitation is the community's responsibility, exercised through structured manaaki.

The four planned rehabilitation pathways are as follows:

Ban Category	Pathway	Approximate Duration	Key Features
Toxic behaviour	Cooldown	10-21 days	Three phases: (1) waiting period, (2) acknowledgement of harm, (3) probationary return with monitoring
Abusive behaviour	Guided Return	30-65 days	Four phases: (1) reflection period with access to educational resources, (2) limited access to low-risk channels, (3) contribution to community (positive action), (4) full return with continued monitoring
Extremist content	Mentored Rehabilitation	60-180 days	Five phases: (1) mentor assignment from trained community volunteers, (2) guided one-on-one conversations addressing the unmet SDT needs that made extremist content appealing, (3) supervised participation in community activities, (4) graduated return to full access, (5) ongoing check-ins
Illegal activity	No rehabilitation	N/A	Report to NZ Police (direct threats of violence), Department of Internal Affairs (objectionable material), or Netsafe (online harassment) as appropriate

The mentored rehabilitation pathway for extremist content will be the most intensive and reflects a core research finding: radicalisation happens through relationships, and de-radicalisation must also happen through relationships. An algorithmic ban cannot undo the social processes that led to radicalisation. A sustained human relationship, grounded in genuine care (manaaki), can.

The mentored rehabilitation pathway will not require community moderators to act as counsellors or social workers. Mentors will be trained community volunteers who provide connection and structured conversation. The pathway will include clear escalation points at which moderators will be directed to refer the individual to professional support services, including He Whenua Taurikura, Netsafe, or NZ Police, depending on the nature and severity of the concern.

The design of the rehabilitation pathways draws on the distinction between punitive and restorative approaches to community harm. Punitive approaches (ban and exclude) protect the

immediate community but displace the problem to unmoderated spaces. Restorative approaches (structured return with accountability and support) address both the immediate community harm and the underlying vulnerability that generated the harmful behaviour. The PVEbot rehabilitation pathways are restorative in design, reflecting the Manaakitanga principle that care for others, including those who have caused harm, is a communal responsibility.

Each rehabilitation pathway will include measurable milestones and clear criteria for progression between phases. The cooldown pathway will require a written acknowledgement of the specific community guideline that was violated. The guided return pathway will require a reflective statement and a positive contribution to the community (for example, helping a new member or contributing constructively to a discussion). The mentored rehabilitation pathway will require sustained engagement with a mentor and demonstrated understanding of the impact of the individual's behaviour on the community. These milestones are designed to address the specific SDT deficits associated with each ban category.

The question of who bears the cost of rehabilitation is a policy question that extends beyond the scope of this design. Community moderators are a mix of Game Development company employees working on building the games community and/or volunteers. Mentors will be volunteers. The time and emotional labour required for rehabilitation, particularly at the mentored level, may be significant. The tool will provide the structure and guidance, but the human resources must come from the community. This is, in Dr McCallum's analysis, both the strength and the limitation of the approach: it locates responsibility where relationships exist, but it depends on communities having the capacity and willingness to invest in restoration.

6.5 Deployment and Data Sovereignty

PvEbot is designed for deployment on New Zealand-based cloud infrastructure to maintain data sovereignty over community moderation data. For example Catalyst Cloud could be the hosting provider for the shared community instance. Catalyst Cloud is a NZ Government-approved cloud provider operating from Wellington datacentres. All community moderation data remains within Aotearoa New Zealand, subject to NZ privacy law, including the Privacy Act 2020.

Three deployment models are planned:

a. Shared instance: The NZGDA would host a single PvEbot instance on Catalyst Cloud. Any NZ game developer Discord server would be able to invite the bot. The estimated cost is approximately $45 per month (NZD plus GST) for the entire sector, with $300 free credit available for new Catalyst Cloud accounts covering the first approximately six months;

b. Individual studio instance: Studios with specific requirements (large communities, custom detection patterns, data sovereignty preferences) can deploy their own instance on Catalyst Cloud, Voyager, SiteHost, or other NZ-based providers. Cost varies from approximately $10 to $45 per month per studio;

c. Hybrid: The shared bot with per-guild configuration. Studios can enable their own AI provider, customise detection sensitivity, and manage their own moderation data while sharing the underlying infrastructure.

The tool is designed to store message hashes (SHA-256), never raw message content. Detection events will record what was detected (category, severity, indicators) but not the original text. Configurable data retention schedules will allow automatic purging of old data. This approach will respect user privacy while maintaining moderation capability and comply with the Privacy Act 2020.

GenAI analysis will default to disabled (rule-based detection only). When studios opt in, locally hosted open-source models (such as Llama 3.2 via Ollama) will be the recommended path, ensuring that AI inference data also remains within New Zealand infrastructure. Cloud AI providers (Anthropic Claude, OpenAI) will remain available as a secondary option but will not be the default or recommended path, as they involve sending community data to overseas servers.

6.6 Applicability Beyond Gaming

While PvEbot was designed for the NZ game development sector, its architecture is intentionally adaptable. The detection analysers, graduated response system, and rehabilitation pathways are not specific to gaming. Any Discord-based community, or with adaptation any online community platform, could benefit from the same approach.

Potential applications beyond gaming include:

a. School Discord servers and online learning environments;
b. Community sports clubs and recreational organisations that maintain online communication channels;
c. Youth organisations and community groups;
d. Professional associations and industry bodies.

The detection patterns would require customisation for each context (the contextual analyser, for example, would need different baseline expectations for a sports community than for a gaming community), but the underlying architecture, response system, and rehabilitation pathways are transferable. The open-source nature of the design means that adaptation will be possible without licensing costs or vendor dependency. Community organisations would be able to deploy customised instances at the same cost structure described for the gaming sector.

This broader applicability is noted here because it has implications for the DPMC Strategy's whole-of-society approach. The research and tool development described in this report began with the gaming sector because that is where the research evidence and the developer expertise converge, but the underlying problem (exploitation of online community dynamics for radicalisation) is not confined to gaming. Any community that builds belonging, identity, and engagement online faces the same risks and could benefit from the same approach.

6.7 Implementation Status and Next Steps

PvEbot exists as a research-informed design. The architecture, detection analysers, graduated response system, and rehabilitation pathways described in this section represent a detailed design specification grounded in the research evidence. Moving from design to operational deployment will require further work in three areas: (a) development and testing of the detection analysers against real community data to validate detection accuracy and false-positive rates; (b) integration planning with existing New Zealand gaming communities, including the NZGDA member studios whose Discord servers will be the initial deployment environment; and (c) ministry approval for deployment, given the sensitivity of the data and the counter-terrorism context in which the tool operates.

---

7. Alignment with DPMC Counter-Terrorism and Violent Extremism Strategy

7.1 Pillar 1: Identify, Understand, and Disrupt Threats

The DPMC Counter-Terrorism and Violent Extremism Strategy (December 2025) identifies the need to build and maintain an understanding of the evolving threat landscape, including the exploitation of online platforms and emerging patterns of radicalisation.

The research and deliverables described in this report contribute to Pillar 1 in the following ways:

a. PvEbot's seven detection analysers are designed to provide operational capability to identify extremist activity within gaming communities across the full spectrum from keyword-level indicators to identity fusion markers. The post-ironic violence analyser and the identity fusion analyser represent novel contributions that address gaps in existing detection approaches;
b. The post-ironic violence research, grounded in the Charlie Kirk case analysis and the MUU ideology framework, contributes to the understanding of an emerging threat pattern that traditional analytical frameworks are not equipped to address. This research has been shared with DPMC and is available to inform intelligence assessment and policy development;
c. The behavioural analyser's monitoring of toxic fraction within communities will provide an early warning system based on the Axelrod threshold (approximately 10 percent), alerting community managers to community health decline before specific extremist content appears;
d. The research (mapping game design mechanics to recruitment parallels, SDT analysis of vulnerability, Te Ao Māori framework for community resilience) contributes to the body of knowledge available to New Zealand's counter-terrorism ecosystem;
e. The rapid analysis capability demonstrated in the Charlie Kirk case (research shared with DPMC within four days of the event) illustrates the value of maintaining standing

expertise in the gaming-extremism nexus. Incidents involving gaming references and subcultural signifiers will recur, and the ability to provide rapid, contextually informed analysis to government depends on the sustained development of the research base described in this report.

7.2 Pillar 2: Prevent and Reduce Radicalisation Risk

Pillar 2 of the Strategy focuses on addressing the drivers and conditions that lead to radicalisation, building resilience in communities and individuals, and providing pathways away from extremism.

The deliverables align with Pillar 2:

a. The educational content for schools builds resilience in young New Zealanders aged 5 to 16 by equipping them to understand the mechanics of manipulation and recruitment, to recognise when their psychological needs are being exploited, and to develop the critical thinking skills (Pakiki) needed to evaluate extremist claims. This is prevention at its most fundamental: building the capacity of individuals to resist recruitment before it begins;
b. The SDT-based approach addresses underlying psychological drivers of vulnerability, not merely the symptoms. By providing a framework for understanding why individuals become vulnerable to recruitment (unmet needs for belonging, purpose, and recognition), the research enables interventions that address root causes rather than surface behaviours;
c. PVEbot's rehabilitation pathways will provide structured routes back into community for individuals who have exhibited toxic, abusive, or extremist behaviour. The mentored rehabilitation pathway for extremist content is specifically designed to address the unmet SDT needs that made the individual vulnerable, through sustained human relationship rather than punitive exclusion;
d. The graduated response system is designed to provide positive alternatives to recruitment at each intervention level. Nudges will redirect toward prosocial engagement. Warnings will acknowledge positive contributions. Even restrictions will maintain community connection. This approach will ensure that individuals at risk encounter genuine SDT need fulfilment through legitimate community participation, reducing the appeal of extremist alternatives.

7.3 Pillar 3: Protect People, Places, and Infrastructure

Pillar 3 of the Strategy addresses the protection of New Zealand communities, physical spaces, and critical infrastructure from the consequences of violent extremism.

The deliverables contribute to Pillar 3 in the following ways:

a. Community-level protection through the graduated response system. PvEbot is designed to provide real-time protection for gaming communities, intervening before toxic or extremist behaviour escalates to the point of causing significant harm to community members;

b. Data sovereignty. The planned deployment of PvEbot on Catalyst Cloud (NZ Government-approved, Wellington datacentres) will ensure that all community moderation data remains within Aotearoa New Zealand, subject to New Zealand law. This will protect the data infrastructure of community moderation from overseas jurisdiction and surveillance risks;
c. Sector-wide coverage at minimal cost. The shared hosting model will allow the entire NZ game development sector to be covered for approximately $45 per month. This will ensure that protection is available to small studios that could not afford individual enterprise moderation solutions;
d. Guidance for referral. The tool's "illegal" ban category will include clear guidance for moderators on when and how to refer matters to NZ Police (direct threats of violence), the Department of Internal Affairs (objectionable material under the Films, Videos, and Publications Classification Act 1993), or Netsafe (online harassment under the Harmful Digital Communications Act 2015). This will ensure that serious threats are escalated to the appropriate authorities rather than being managed solely at the community level;
e. Scalable, low-cost protection infrastructure. The shared hosting model will mean that protection is not dependent on individual studios having the resources or expertise to build their own moderation systems. A single investment of approximately $45 per month will cover the entire NZ game development sector. This will be infrastructure-level protection at a cost that is negligible relative to the harm it addresses.

---

8. Recommendations

Based on these research findings, we recommend that DPMC: **a. note** the research findings on the intersection of game design principles and violent extremism, including the RUSI Extremism and Gaming Research Network findings on the scale of extremist exploitation in gaming spaces (one-third of gamers exposed to extremist content, one-quarter encountering recruitment attempts, more than 50 percent of female gamers experiencing harassment); **b. note** the design of PVEbot as a community safety tool for the New Zealand game development sector, grounded in Preventing Violent Extremism research, Self-Determination Theory, and Te Ao Māori values, with seven rule-based detection analysers, a graduated response system, and structured rehabilitation pathways. Full implementation will require further development, testing with gaming communities, and ministry approval; **c. note** the development of age-differentiated educational content for New Zealand schools (Years 1 to 13, ages 5 to 16+) aligned to Te Mataiaho, designed to build resilience to manipulation and radicalisation through activity-based learning, grounded in Self-Determination Theory and Te Ao Māori values; **d. consider** integration of the educational content with the Keep It Real Online platform and Netsafe's existing educational resources (including the Cyberbullying Prevention Toolkit and Headspace Invaders) to extend its reach to New Zealand schools and educators. The educational content has been designed to complement these resources and includes NCEA assessment alignment (AS 92050, AS 92051, AS 91599, and Unit Standard 30910) with grading examples, enabling teachers to use the modules as contexts for internal assessment; **e. consider** a pilot

programme with selected New Zealand schools to test and refine the educational content, in partnership with the Ministry of Education and potentially He Whenua Taurikura, prior to wider distribution; **f. consider**, subject to successful implementation and testing, extending the PVEbot architecture to non-gaming community platforms (for example, community sports clubs, youth organisations, school communication platforms) where similar vulnerability to the exploitation of online communities exists; **g. note** that these deliverables align with all three pillars of the Counter-Terrorism and Violent Extremism Strategy (December 2025): Pillar 1 (Identify, Understand, and Disrupt Threats), Pillar 2 (Prevent and Reduce Radicalisation Risk), and Pillar 3 (Protect People, Places, and Infrastructure), and with the PCVE Strategic Framework's whole-of-society approach; **h. note** that the New Zealand Game Developers Association and Dr McCallum welcome engagement with DPMC and He Whenua Taurikura on further research, the implementation and testing of PVEbot with existing gaming communities, and the integration of these deliverables into the broader counter-terrorism and violent extremism prevention ecosystem.

---

9. References

Axelrod, R. (1984). *The Evolution of Cooperation*. Basic Books.

Battersby, J., & Ball, R. (2019). Christchurch in the context of New Zealand terrorism and right wing extremism. *Journal of Policing, Intelligence and Counter Terrorism*, 14(3), 191-207. https://doi.org/10.1080/18335330.2019.1662077

Bezio, K. M. S. (2018). Ctrl-Alt-Del: GamerGate as a precursor to the rise of the alt-right. *Leadership*, 14(5), 556-566. https://doi.org/10.1177/1742715018793744

Borum, R. (2003). Understanding the terrorist mind-set. *FBI Law Enforcement Bulletin*, 72(7), 7-10. https://www.ojp.gov/ncjrs/virtual-library/abstracts/understanding-terrorist-mind-set

Braithwaite, A. (2016). It's about ethics in games journalism? Gamergaters and geek masculinity. *Social Media + Society*, 2(4), 1-10. https://doi.org/10.1177/2056305116672484

Cunningham, M., La Rooij, M., & Spoonley, P. (Eds.). (2023). *Histories of Hate: The Radical Right in Aotearoa New Zealand*. Otago University Press. ISBN: 9781990048401.

Deci, E. L., & Ryan, R. M. (2012). Self-determination theory. In P. A. M. Van Lange, A. W. Kruglanski, & E. T. Higgins (Eds.), *Handbook of Theories of Social Psychology* (Vol. 1, pp. 416-436). SAGE Publications. https://doi.org/10.4135/9781446249215.n21

Department of the Prime Minister and Cabinet. (2025). *New Zealand's Counter-Terrorism and Violent Extremism Strategy*. New Zealand Government. https://www.dpmc.govt.nz/our-programmes/national-security/counter-terrorism

Department of the Prime Minister and Cabinet. (n.d.). *Preventing and Countering Violent Extremism Strategic Framework*. New Zealand Government. https://www.dpmc.govt.nz/our-programmes/national-security/counter-terrorism/preventing-and-countering-violent-extremism

Films, Videos, and Publications Classification Act 1993 (New Zealand).

Global Project Against Hate and Extremism (GPAHE). (2024). *How the Far-Right Spreads Hate through Gaming*. GPAHE. https://globalextremism.org/post/how-the-far-right-spreads-hate-through-gaming/

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025-3034). IEEE. https://doi.org/10.1109/HICSS.2014.377

Harmful Digital Communications Act 2015 (New Zealand).

Hartgers, M., & Leidig, E. (2023). *Fighting Extremism in Gaming Platforms: A Set of Design Principles to Develop Comprehensive P/CVE Strategies*. International Centre for Counter-Terrorism. https://icct.nl/publication/fighting-extremism-gaming-platforms-set-design-principles-develop-comprehensive-pcve

He Whenua Taurikura. (2022). *National Centre of Research Excellence for Preventing and Countering Violent Extremism*. https://hwt.ac.nz/

Horgan, J. (2008). From profiles to pathways and roots to routes: Perspectives from psychology on radicalization into terrorism. *The ANNALS of the American Academy of Political and Social Science*, 618(1), 80-94. https://doi.org/10.1177/0002716208317539

Kowert, R., Martel, A., & Swann, W. B. (2022). Not just a game: Identity fusion and extremism in gaming cultures. *Frontiers in Communication*, 7, 1007128. https://doi.org/10.3389/fcomm.2022.1007128

Kruglanski, A. W., Gelfand, M. J., Bélanger, J. J., Sheveland, A., Hetiarachchi, M., & Gunaratna, R. (2014). The psychology of radicalization and deradicalization: How significance quest impacts violent extremism. *Political Psychology*, 35(S1), 69-93. https://doi.org/10.1111/pops.12163

Kruglanski, A. W., Jasko, K., Webber, D., Chernikova, M., & Molinario, E. (2018). The making of violent extremists. *Review of General Psychology*, 22(1), 107-120. https://doi.org/10.1037/gpr0000144

Lakhani, S., & Wiedlitzka, S. (2023). "Press F to Pay Respects": An empirical exploration of the mechanics of gamification in relation to the Christchurch attack. *Terrorism and Political Violence*, 35(7), 1586-1603. https://doi.org/10.1080/09546553.2022.2064746

Massanari, A. (2017). #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3), 329-346. https://doi.org/10.1177/1461444815608807

McCallum, S. (2025). *Research supporting rapid analysis of Charlie Kirk and links to gaming*. Presented to the Department of the Prime Minister and Cabinet.

McCallum, S. (2025). *Extremism and Gaming*. Presented at the New Zealand Game Developers Conference (NZGDC) 2025.

McCauley, C., & Moskalenko, S. (2017). Understanding political radicalization: The two-pyramids model. *American Psychologist*, 72(3), 205-216. https://doi.org/10.1037/amp0000062

Ministry of Education. (2023). *Te Mataiaho: The New Zealand Curriculum*. New Zealand Government. https://www.education.govt.nz/our-work/changes-in-education/curriculum-changes/te-mataiaho/

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

New Zealand Classification Office. (n.d.). *Extremism and online harms research*. https://www.classificationoffice.govt.nz/

New Zealand Game Developers Association. (2025). *NZGDA 2025 Annual Survey: New Zealand Game Development Industry Breaks Records with $759M in Revenue and 1,418 Jobs*. NZGDA.

Privacy Act 2020 (New Zealand).

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, 14(2), 154-166. https://doi.org/10.1037/a0019440

Royal Commission of Inquiry into the terrorist attack on Christchurch masjidain on 15 March 2019. (2020). *Ko tō tātou kāinga tēnei: Report of the Royal Commission of Inquiry*. New Zealand Government. https://christchurchattack.royalcommission.nz/

Schlegel, L. (2020). Jumanji extremism? How games and gamification could facilitate radicalization processes. *Journal for Deradicalization*, 23, 1-44. https://journals.sfu.ca/jd/index.php/jd/article/view/359

Teaching Council of Aotearoa New Zealand. (n.d.). *Tātaiako: Cultural Competencies for Teachers of Māori Learners*. https://teachingcouncil.nz/resource-centre/tataiako-cultural-competencies-for-teachers-of-maori-learners/

Te Mana Whakaatu — Classification Office. (2023). *Content that Crosses the Line: Conversations with young people about extremely harmful content online*. Wellington. https://www.classificationoffice.govt.nz/resources/content-that-crosses-the-line/

Vansteenkiste, M., Ryan, R. M., & Soenens, B. (2020). Basic psychological need theory: Advancements, critical themes, and future directions. *Motivation and Emotion*, 44(1), 1-31. https://doi.org/10.1007/s11031-019-09818-1

Valentini, D., Lorusso, A.M. and Stephan, A. (2020). Onlife Extremism: Dynamic Integration of Digital and Physical Spaces in Radicalization. *Frontiers in psychology,* 11, https://doi.org/10.3389/fpsyg.2020.00524

Wallner, C., White, J., & Regeni, N. (2025). *Extremism in Gaming Spaces: Policy for Prevention and Moderation*. RUSI Policy Brief. Royal United Services Institute.

Wells, G., Romhanyi, A., Reitman, J. G., Gardner, R., Squire, K., & Steinkuehler, C. (2024). Right-wing extremism in mainstream games: A review of the literature. *Games and Culture*, 19(4), 469-492. https://doi.org/10.1177/15554120231167214

White, J., Sherwood, H., Sherwood-Sheridan, I., Campbell-Obaid, J., & Sherwood-Sheridan, S. (2024). *Radicalisation through Gaming: The Role of Gendered Social Identity*. Whitehall Report 2-24. Royal United Services Institute. https://static.rusi.org/radicalisation-through-gaming-role-of-gendered-social-identity-whr-december-2024.pdf

Zagal, J. P., Bjork, S., & Lewis, C. (2013). Dark patterns in the design of games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games* (pp. 39-46).

Appendices

The following appendices provide detailed technical specifications and supporting data referenced throughout this report:

- **Appendix A: RUSI Extremism and Gaming Research Network, Key Statistics.** Detailed analysis of the RUSI EGRN mixed-methods study findings, economic impact data, NZ sector impact analysis, Australian eSafety Commissioner findings, and the three mechanisms of exploitation.
- **Appendix B: PVEbot Detection Architecture, Technical Specification.** Full technical specification of the seven rule-based detection analysers (keyword, pattern, behavioural,

contextual, post-ironic violence, recruitment, and identity fusion), including detection patterns, scoring algorithms, confidence calculations, and the optional generative AI enhancement layer.

- **Appendix C: Graduated Response System and Rehabilitation Pathways, Detailed Specification.** Detailed specification of the seven-level graduated response system (L0 to L6), the SDT and Te Ao Māori response frameworks with message templates, the four rehabilitation pathways (cooldown, guided return, mentored rehabilitation, and illegal activity referral), and the restorative design philosophy.
- **Appendix D: Deployment Model, Data Sovereignty, and Cost Analysis.** Data sovereignty principles, the three deployment models (shared community instance, individual studio instance, and hybrid), GenAI configuration options, NZ hosting provider comparison, cost analysis, security architecture, and broader applicability beyond gaming.
- **Appendix E: Educational Modules.** Educational content covering multiple years with progressively complex understanding of the motivation structures and pathways to extremist beliefs.

Prepared by:

Dr Simon McCallum Senior Lecturer, Victoria University of Wellington

Contact: simon.mccallum@vuw.ac.nz