Understanding and Countering Gendered Disinformation

A framework for resilience and action

May 2025

CSKA Community Safety Knowledge Alliance

SAPPER LABS

This page is intentionally left blank

This work was prepared by:

www.cskacanada.ca www.sapperlabs.com

With funding from:

Canada

Authors of report

Janos Botschner, PhD
Giovanna Cioffi, CD, PhD
Dave McMahon, MSM, BEng
Julie Ollinger, PhD
Bradley Sylvestre, MA
Ritesh Kotak, JD
Cal Corley, MBA

Additional contributors

Additional support to this project was provided by Actua (www.actua.ca). The following individuals co-authored knowledge products for parents, youth and educators in collaboration with the project team. These resources were produced by Actua in partnership with CSKA:

Janelle Fournier, PhD (ABD)
Mikayla Ellis, BA
Abbey Ramdeo, MT

Suggested report citation:

Botschner, J., Cioffi, G., McMahon, D., Ollinger, J., Sylvestre, B., Kotak, R. & Corley, C. (2025). Understanding and countering gendered disinformation: A framework for resilience and action. Ottawa ON: Community Safety Knowledge Alliance.

Correspondence:

Jbotschner[at]cskacanada.ca

About the Community Safety Knowledge Alliance

The Community Safety Knowledge Alliance is a non-profit applied research organization that supports governments, police, public health and human service leaders in developing, implementing and assessing new approaches to enhancing community safety and well-being service delivery and outcomes.

Over the past decade, CSKA has conducted interdisciplinary research on some of Canada's most pressing social issues, including intimate partner violence, youth radicalization to violence, cybersecurity, food security, drug policy, human rights-based policing, and community reintegration initiatives. CSKA maintains an active posture on issues such as disinformation and artificial intelligence to support adaptive responses to these emerging challenges.

About Sapper Labs Group

Sapper Labs Group conducts research to understand the methods and impacts of disinformation and influence campaigns and networks and as input to the development of processes to support effective countermeasures. SLG is supported by global partners and a comprehensive intelligence sharing network.

The goal of SLG is make the world a better safer place in line with objectives around: countering foreign interference and influence, countering radicalization and extremism, supporting human rights and other activities involving capacity building related to information integrity.

ACKNOWLEDGEMENTS

The authors would like to express sincere appreciation to the following individuals/groups for the guidance they provided to this work. The contents, conclusions and recommendations are those of the authors, alone.

Project Advisory Committee

Michael Doucet, former Executive Director, National Security Intelligence Review Agency
Jennifer Flanagan, Chief Executive Officer, Actua
Carmen Gill, Professor, Department of Sociology, University of New Brunswick
Jennifer Irish, Director, Information Integrity Lab, University of Ottawa
Alan Jones, Executive Advisor, Professional Development Institute, University of Ottawa; former Assistant Director, Canadian Security Intelligence Service
Marcus Kolga, Founder and Director, DisinfoWatch; Fellow, MacDonald-Laurier and Conference of Defence Associations Institutes

Development of Knowledge Products

For parents and educators

Actual National STEM Educator Community of Practice
Actua staff and Actua Network members

For youth

Actua National Black Youth in STEM Program Youth Delegation
Actua Indigenous Youth in STEM Program Youth Delegation
Actua staff and Actua Network members

For police and community groups

Delta Police Department

André Cruz
Communications Assoc.

Cst. Derek Defrane
Domestic Violence Unit

Kim Gramlich
Mgr., Victim Services

Sgt. Alex Quezada
Vulnerable Sector Unit

Greater Sudbury Police Service

Det. Sgt. Adam Demers
Criminal Investigations/ Intimate Partner Violence

Dan Gelinas
Community Mobilization Liaison

Det. Sgt. Lee Rinaldi
Major Sex Crimes

Sudbury YWCA

Marlene Gorman
Executive Director

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Today's interconnected world provides a wealth of opportunities for those wishing to harm women, girls and gender diverse persons individually and at scale. This report describes the mechanisms, impacts, and actors behind technology-enabled gendered disinformation. This is not just a gender issue – it is also a socioeconomic and public safety issue which, in some cases, may also become a national security concern. We illustrate why action is needed now and chart a theory- and evidence-informed path forward.

Technology-enabled gender-based violence – including disinformation – draws from a powerful arsenal of tools. It can be used for illicit surveillance (such as monitoring movement and communications) and to manipulate aspects of the built environment (such as features of "smart" homes and vehicles). It can also be used to pollute the information space with deceptive narratives. Gendered disinformation poses a dual threat: it endangers individuals, especially women and gender-diverse people who are often its direct targets; and it undermines society by eroding trust and cohesion, silencing voices, and weakening democratic norms and processes.

Members of certain populations – notably, marginalized and racialized women, girls and gender diverse persons – may disproportionately encounter greater levels of gendered disinformation. Indigenous women and girls in Canada often face gendered disinformation and related violence due to historical biases, colonial legacies and contemporary social media narratives that can often perpetuate harm.

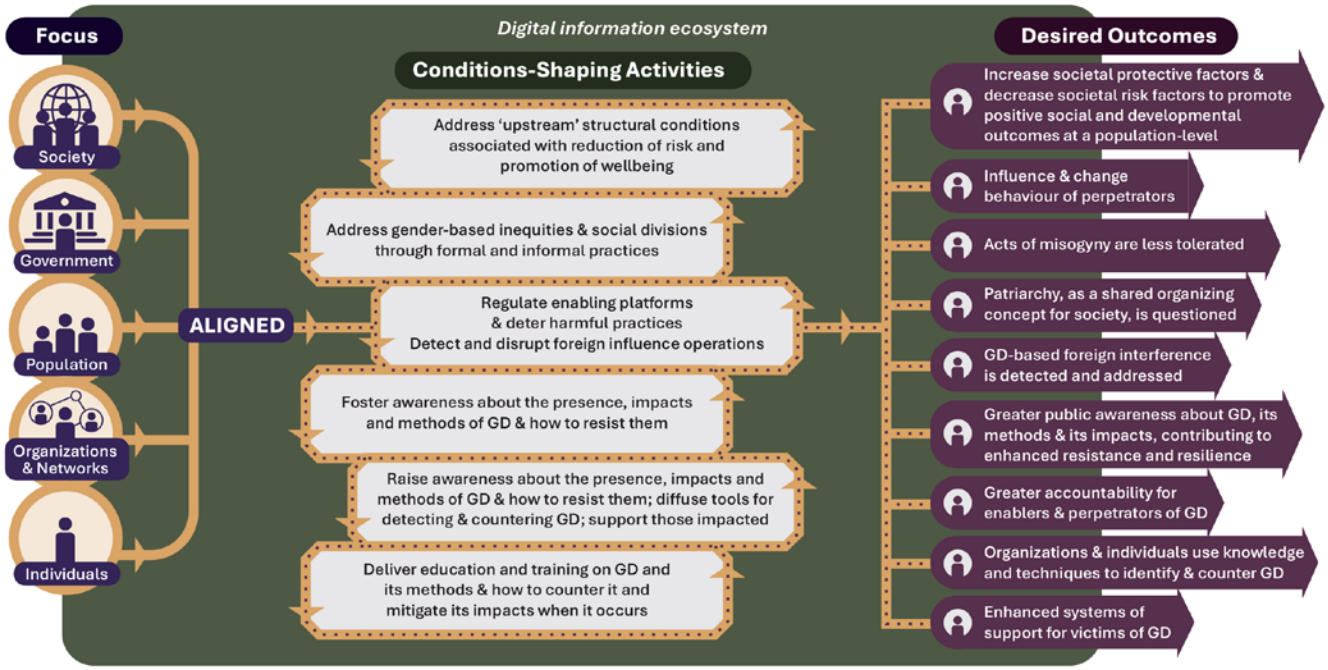
Gendered disinformation is not spread by chance. It can be driven by individual actors, aligned domestic and transnational ideological groups, and even nation states that seek to destabilize democratic societies. When foreign governments are involved, GD becomes an instrument used to sow division, fear, and mistrust across borders – sometimes as a component of broader influence or cyber operations. At a time when online spaces too often amplify misogynist voices and targeted abuse, understanding and countering gendered disinformation has never been more urgent. It is a shared threat across society. Consequently, the resolve and the ability to address gendered disinformation must be a matter of shared responsibility.

The widespread occurrence of gendered disinformation around the world, often leading to violence, underscores the need for international cooperation. This is essential to address the complex, cross-border nature of the issue effectively. By sharing best practices, resources, and intelligence, countries can develop unified strategies to combat disinformation. Domestically, integrating these global insights into national policies and practices will enhance local efforts, ensuring that responses are comprehensive and culturally relevant. Joint initiatives can also strengthen diplomatic relations, promote gender equality, and uphold human rights on a broader scale.

This report provides a novel perspective on gendered disinformation, including a framework for action with a corresponding system of people, processes and technology. Furthermore, it provides a short set of pragmatic recommendations that will have significant impact on combatting gendered disinformation, enhancing human rights protection, and promoting gender equality. These elements are accompanied by a set of information resources for key stakeholder groups seeking to raise awareness and to counter this complex, multi-layered problem.

Building the capacity to counter gendered disinformation will require collaboration. As we navigate geo-political and domestic tensions that threaten the cohesion, unity and sovereignty of Canadian society, our willingness to confront and respond to gendered disinformation will shape the resilience and inclusivity of our digital, democratic and social spaces for years to come.

A preliminary theory of change for addressing gendered disinformation is depicted below. This is discussed in further detail at Figure 9 in this report.



This theory involves multi-level efforts designed to align and create mutually reinforcing conditions, significantly enhancing the likelihood of achieving a range of desired outcomes.

Conclusions

Addressing gendered disinformation is crucial for safeguarding human rights, promoting gender equality, and upholding democratic values. This issue, intertwined with polarization, patriarchy, and misogyny, targets women, girls, and gender-nonconforming individuals, causing harm. A strategic, multi-layered approach is necessary to combat this, focusing on awareness and a coordinated

response. Strengthening resistance to such disinformation requires collaborative efforts to prevent risks, enhance resilience, and align solutions with democratic principles.

Gendered disinformation about Indigenous women and girls in Canada is exacerbated by a colonial history that persists today, reinforcing harmful stereotypes and ignoring ongoing violence. Multi-faceted efforts must be undertaken to break this cycle by challenging false narratives, reforming media practices, and prioritizing Indigenous voices in storytelling. Such measures are vital for transforming the information landscape and supporting reconciliation.

The path forward emphasizes multi-sector collaboration and building broad-based networked capacity to counter gendered disinformation. Increasing awareness and developing new knowledge will be central to this effort. This approach should foster mutual benefits and support collective learning, planning, implementation and further research.

We propose a comprehensive theory of change involving strategically aligned, society-wide interventions grounded in the leading research. This approach includes providing a robust set of knowledge resources and technology examples beneficial to professionals in human services, policy-making, and national security. Furthermore, we recommend creating a cross-sectoral network dedicated to knowledge development and mobilization. This network will support evidence-based, collaborative efforts, ensuring that interventions are informed by the best available evidence and practices. By fostering cooperation across multiple sectors, this critical issue can be tackled effectively and holistically.

Recommendations:

Policy, Legislation and Enforcement

1. That the federal government:

    a. Implement policy and legislative measures to counter gendered disinformation, recognizing that it is a threat that spans community safety and wellbeing, and national security.
        - *The corresponding regulatory framework should ensure platform accountability, transparency, and meaningful financial penalties for non-compliance.*

    b. With targeted investment, initiate cross-departmental, industry, academic and private sector operational coordination and program collaboration to address gendered disinformation within public safety, public health, digital regulation, defence and national security frameworks.

c. Develop a national strategy on gendered disinformation in close partnership with the private sector, research and civil society, integrating public safety, digital governance, and foreign policy approaches.

d. Convene and engage women's advocacy organizations, racial justice groups, security and intelligence professionals, academic researchers, cyber-security experts and relevant community and private sector entities in dialogue on such matters as how to optimize the balance of protection and enforcement with freedom of expression online.

e. Increase data collection and monitoring of gendered disinformation trends and actionable current intelligence.

f. Conduct periodic cross-sector consultations with experts in gender-based violence, cybersecurity, open source intelligence, national security, and digital regulation to understand the evolving landscape of gendered disinformation.

g. Establish gender-responsive online safety laws that hold technology platforms accountable. Options include the re-introduction of Bill C-36 and the applications of relevant elements of a Clean Pipes Strategy.

h. Enhance training for security, intelligence, diplomatic, defence, law-enforcement and policymakers on technology-enabled GD.

i. Invest in digital literacy, research, open source intelligence and enforcement mechanisms to strengthen Canada's resilience against gendered disinformation.

Research and Knowledge Mobilization

2. That the Government of Canada support the creation of a cross-sectoral knowledge mobilization network on gendered disinformation – the Gendered Disinformation Knowledge Network (GenD-Net).

   Such a network would serve as a hub for leadership, information sharing, education and training, research, and policy coordination, program planning, operational coordination and de-confliction ensuring that responses to gendered disinformation are evidence-based, and aligned across sectors.

   The objectives of the network will be to:

   - *Enhance knowledge mobilization and public awareness of gendered disinformation.*
   - *Support curriculum development, stimulate and contribute to education and training.*

- *Strengthen community and cross-sectoral dialogue and collaboration on policy development.*
- *Support defence, intelligence, police and public safety agencies.*
- *Advance research and innovation, including evaluation capacity building.*
- *Bridge gaps in service provision for affected communities.*

Gendered Disinformation as a National Security Issue

3. That the Government of Canada refine and implement options for countering gendered disinformation as a national security issue, including its use as an element of foreign interference. Enhance the capabilities of defensive cyber operations in relation to this threat. More particularly:
    a. Establish a dedicated government funding stream for research and innovation on gendered disinformation that is open to Canadian industry, academia and not-for profit organizations.
    b. Incentivize Canadian industry participation and innovation through public-private partnerships and direct investment.
    c. Develop a national strategy on gendered disinformation as a foreign interference threat, and ensure integration with national defence policy, cyber security and national security strategies.
    d. Fund the creation of a cross-sectoral intelligence-sharing network to combat gendered disinformation, including the creation and maintenance of a national gendered disinformation threat landscape reporting capacity; this would, in-turn, feed into an intelligence "dashboard" (Figure 11) which could be made publicly available as part of building overall awareness an public will to confront this problem (See Annex E4, Attachment B).
    e. Establish legal and policy frameworks to protect women in public life from both foreign and domestic online harm.
    f. Develop a rapid response mechanism to protect individuals facing high-risk disinformation attacks (see Annex E4, Briefing Resources 1 and 4).

Impact of Recommendations

Implementing these recommendations will have significant impacts on combatting gendered disinformation, enhancing human rights protection, and promoting gender equality. By addressing this issue, intertwined with polarization and misogyny, we can safeguard women, girls, and gender-nonconforming individuals from targeted harm. More specific areas impacted are as follows:

Policy and Legislation

By implementing comprehensive policies and legislation, the federal government will strengthen community safety and national security. Establishing regulatory frameworks with platform accountability and penalties for non-compliance will ensure that digital spaces are safer and more transparent. Cross-departmental coordination will enhance efforts to address gendered disinformation within public safety and national security frameworks.

Multi-Sector Collaboration

Creating a national strategy in partnership with the private sector, research institutions and civil society will integrate approaches to enhancing both public safety and social media governance. Engaging diverse organizations in dialogue will balance safety and security with freedom of expression. Furthermore, this approach will help build resilience against gendered disinformation through enhanced data collection, training, and digital literacy investments.

Research and Knowledge Mobilization

A dedicated funding stream for research and innovation, alongside public-private partnerships, will drive industry participation and technological advancements.

Establishing the Gendered Disinformation Knowledge Network (GenD-Net) will enhance public awareness, support curriculum development, and foster cross-sectoral collaboration. By bridging gaps in service provision, it will ensure evidence-based responses aligned across sectors.

National Security

Recognizing gendered disinformation as a national security issue will help refine strategies to counter foreign interference. Developing a rapid response mechanism and legal frameworks will protect individuals from high-risk disinformation attacks.

Overall, when implemented, these measures will help to transform the online information landscape, support reconciliation, and uphold Canadian liberal democratic values by fostering a coordinated, strategic response to gendered disinformation.

COMMON TERMS AND TECHNIQUES

Misinformation is untrue content that is spread by people who believe that it is true – ***untrue information, good or neutral intent*. Disinformation** is untrue content that is spread by people who know that it is untrue. Misinformation could be spread innocently, or to cause harm. Disinformation is always spread knowingly and deliberately to cause harm – ***untrue information, bad intent*. Malinformation** is information that is true, but it's shared in a way that's meant to cause harm – ***true information, bad intent*.**

Fake stories – Fake news articles or social media posts that attack individuals, such as former partners/spouses, or those in public or leadership roles.

Non-consensual image sharing – Can include posting or re-posting intimate images that were meant to be private or exclusive to a partner. It can also involve uploading sexual photos or videos of an ex-partner to social media or pornographic websites without their consent.

Manipulated images & videos – Edited pictures or **"deepfake"** videos that make it look like someone said or did something they never did. Commonly encountered situations include non-consensual, out-of-context, sharing of manipulated or real photos/fake explicit content.

Misinformation about gender roles – Posts or comments claiming that women are naturally bad at leadership, science, or sports.

Harassment & cyberbullying – Online attacks that try to intimidate, humiliate, or silence.

Fake accounts & impersonation – Creating fake online profiles to spread lies, harass someone, or damage their reputation. When this involves creating one or many fake accounts, or taking over existing accounts to make it look like people agree with a fake story, it is called **astroturfing**.

Memes & satire – These are jokes or cartoons that disguise harmful messages about their targets as "just humour."

Doxxing – Broadly sharing private information (like a home address or phone number) online to intimidate or harm a person. Sometimes, this can lead to offline intimidation or violence.

Surveillance and manipulation of "smart" technology – For example, using commercially available tracking devices, to monitor someone's movement; or manipulating home or vehicle systems to intimidate someone.

Cyberflashing is the act of sending someone unsolicited sexual images through digital means, often via text message, social media, dating apps, or file-sharing features like AirDrop or Bluetooth.

Catfishing – a practice in which individuals create fake online identities to deceive others, often for abusive or exploitative purposes.

INTRODUCTION

Gender equality and the safety of women, girls and gender non-conforming persons are under threat. Ideological movements, political forces, and global tensions rooted in patriarchy and misogyny are deepening social and political divides online. In some cases, these divisions are intentionally exploited to harm individuals and destabilize Canadian society.

The aim of this research and development project was to create a framework and a set of corresponding practices to understand and counter online gendered disinformation. This issue occurs against a more general backdrop of: technology-enabled gender-based violence and repression[1], which is a global problem; and foreign interference, which has been flagged as a significant and ongoing threat to Canada and to Canadians (Public Inquiry into Foreign Interference in Federal Electoral Processes and Democratic Institutions, 2025).

Generative AI is driving a surge in disinformation. One specific category – gendered disinformation (GD) – targets women, girls and gender-diverse persons.[2] – through misogynistic harassment and intimate partner violence. Furthermore, it can collectively deny them a voice and undermine their role in society.

Gendered disinformation makes use of distinctions and divisions centred on traditional notions of gender. It targets those who do not conform, such as working women, transgender individuals, or those with non-traditional gender expressions. It targets specific groups to undermine social cohesion and public trust (e.g., by questioning the competence of female leaders or the appropriateness of certain books available in school libraries). GD leverages digital technologies, especially social media, to amplify impact. Finally, GD not only harms individuals, but it

The use of tropes or memes to promote gendered disinformation.

A **trope** is a commonly used theme, idea, or storytelling device that helps people quickly understand a situation or character. They can be visual, verbal, or conceptual, and they often rely on familiar patterns. When used for disinformation, they may reinforce stereotypes.

Example: *"Women are bad drivers."*

A **meme** is a piece of content – often an image, video, or phrase – that spreads quickly online and is shared, adapted, and remixed by different people. They can be humorous, political, or cultural, and they often carry deeper meaning in a short, relatable format. In online disinformation, memes are used to spread false or harmful messages in a way that feels casual and shareable, making them powerful tools for manipulation

Example: *"Real women ['trad wives'] don't chase careers—they support their husbands and raise children the right way."*

[1] Also known as online violence against women.

[2] For consistency, we use the term, "women and gender diverse", to refer to individuals – such as women, gender non-conforming persons, and those with various sexual identities –who are targeted by gendered disinformation or used as instruments for malicious purposes.

deepens existing social divisions, making it more difficult to achieve equality and mutual understanding.

The implications of gendered disinformation are both deeply personal and profoundly political. Targeted individuals may experience psychological distress and reputational damage. These harms can deter affected individuals from participating in politics, activism, or journalism, ultimately silencing important voices in public discourse. At a broader level, the spread of online gendered disinformation erodes public trust in media and institutions – particularly when false content is mistaken for real, or genuine content is dismissed as fabricated. Moreover, disinformation campaigns frequently reinforce harmful narratives, perpetuating stereotypes that disproportionately impact women, girls and LGBTQIA+ communities (e.g., Richardson-Self, 2021; Sobieraj, 2020).

Contested truth and false or misleading content (e.g., text, audio, video and images) scaled through social media is one of the most serious threats to democratic values, civic participation, domestic tranquility, and the ability of people and nations to collaborate in addressing the complex challenges of this century.

Recent examples identified and discussed in print and media publications include: videos of anti-woman influencers denouncing female empowerment, including one posted by "manosphere" influencer Andrew Tate opining that women should not be allowed to drive (Donegan, 2025); the role of video gaming platforms in spreading misogynistic tropes and memes (Stuart, 2025); accounts of how disinformation campaigns can be used to destroy the reputations of political opponents (Ressa, 2022); and the weaponization of the term "woke" to attack various initiatives focused on inclusion (Off, 2024).

In the United States, the National Democratic Institute (2022) described GD as a critical issue for democracies because of its impacts on the participation of women in online political activity, and in recognition of its use by authoritarian and illiberal actors as a tactic of online violence aimed at silencing and undermining the political agency of women and girls. Thus, an increase in GD benefits individual perpetrators of information-based violence and groups seeking to disrupt social harmony and undermine the value of inclusion.

> *A healthy democracy is characterized by a vibrant and diverse range of voices and groups, engaged in a constant process of deliberation, discussion, negotiation and compromise. Because of this characteristic, democracy requires a civic setting in which people can freely express their ideas. To create such a setting, democracy relies on values and principles such as the equality of individuals and respect for others, as well as consideration for the diversity of opinions and beliefs. It requires social and political institutions that encourage the participation of all. By fostering distrust, creating division and preventing compromise, disinformation threatens this fundamental feature of democracy.*
>
> (Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions, 2025).

In recent testimony before the House of Commons Standing Committee on Public Safety and National Security, Marcus Kolga, of DisInfoWatch, argued that "[s]afeguarding Canada's cognitive sovereignty and the integrity of our information environment is essential to defending our democracy and maintaining social cohesion" (Parliament of Canada, 2024).

In the final report of the Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions, Commissioner Hogue opined that disinformation is a pronounced threat to Canadian society.

Technology-Enabled Violence Against Women is a Widespread Problem That Violates Human Rights

The September 2024 SDG Gender Index, published by Equal Measures 2030 (a global coalition of NGOs that use data and evidence to address gender equality), determined that none of the 139 countries assessed has achieved the UN's 2030 SDG benchmarks for gender equality. While Canada is ranked 18th and falls in the 'good' category[3], its progress has stalled over the recent measurement periods (2015-2019, 2019-2022) and is projected to remain unchanged from 2022 to 2030.

Moreover, the report stated that gender-based violence against Indigenous women and girls is a particularly serious issue in Canada. These individuals face greater levels of both intimate and non-intimate partner violence compared to non-Indigenous females. A recent study by the Canadian Women's Foundation found that thirty percent of Indigenous women encounter unwelcome behaviour, including online[4]. The study also found that found one in five Canadian women experiences some form of online harassment.

The increased risk of online abuse by Indigenous women reflects long-standing colonial practices of objectification, sexualization and misrepresentation (Corbett, 2019). These assumptions and biases provide a foundation for contemporary narratives in media and popular culture that perpetuate harm. For example, stereotypes and disinformation about missing and murdered Indigenous women and girls distract from many of the deeper causes of this violence (Corbett, 2019). This can lead to misunderstanding and apathy, undermining public commitment for meaningful change.

[3] Following Belgium, ahead of Spain and the United Kingdom and the United States.
[4] Canadian Women's Foundation (n.d.)

A recent report by the Institute of Global Politics and the Vital Voices Global Partnership (Jankowicz, et al., 2024) on technology-enabled gender-based violence[5] confirms that online abuse of women is a widespread problem across all continents. The associated global statistics[6] are stark: between 2019 and 2020, 85 percent of women had witnessed or experienced online gender-based violence and 38 percent had been personally impacted by it. The Economist Intelligence Unit (EIU, 2020) assessed that these figures likely underestimate the actual prevalence of the issue.

> *The repetitive representation of Indigenous women engaging in "high-risk" lifestyles normalizes the violence against them....*
>
> *The silencing of violence against Indigenous women and girls is made worse in comparison to the media's compassionate framing of white women.*
>
> (Corbett, 2019)

The Economist Intelligence Unit reported a North American prevalence of online gender-based violence of 76 percent[7]. In a University of Maryland study reported by Hess (2014), created a set of fake online accounts with feminine and masculine usernames. They then distributed them across various online chat rooms. Accounts with feminine usernames received an average of 100 sexually explicit or threatening messages per day, whereas those with masculine usernames received only 3.7 such messages.

Sobieraj (2020) suggests that the uneven distribution of identity-based abuse among women is linked to power and inequality. She observes that online attacks are most severe for three groups: women with multiple marginalized identities; those who publicly critique male-dominated spaces; and those perceived as feminist or non-conforming with traditional gender norms. Women at the intersection of all three groups, such as BIPOC[8] feminist members of the LGBTQIA+ community, may be particularly targeted. Researchers studying far-right extremist movements have observed that "persistent anxiety about masculinity" is a core feature of these ideologies (Kesevan, 2024).

According to the EIU (2020) study, nine threat tactics predominated across respondents to its online survey:

- Misinformation and defamation (67 percent);
- Hate speech (65 percent) and violent threats (52 percent);
- Cyber harassment (66 percent), hacking and stalking (63 percent);
- Doxing[9] (55 percent);
- Astroturfing[10] (58) percent;

[5] Known as technology-facilitated violence against women (TF-VAW) in the research literature
[6] Economist Intelligence Unit (2020) data from 2019-2020, reported by Jankowicz, et al. (2024)
[7] While this figure is high, it is the second-lowest across continents, with Europe being the lowest at 74 percent.
[8] i.e., Black, Indigenous, People of Colour
[9] Posting personal information to incite violence
[10] Coordinating the sharing of damaging information across online platforms to give the appearance of

- Impersonation (63 percent); and
- Video- and image-based abuse (57 percent).

The violence stemming from these forms of abuse extends beyond the online environment. A recent New York Times investigation (Mozur, et al., 2024) identified a violence-promoting group hosted on the social media platform Telegram linked to a series of attacks, including a 2022 shooting at an LGBTQIA+ bar in central Europe. The phenomenon, whereby digital platforms facilitate the transition from online rhetoric to offline violence, is called *stochastic terrorism*.[11] It involves the use of mass communication to incite random individuals to commit statistically predictable but individually unpredictable violent acts. The content creators can often assert plausible deniability, claiming they did not directly incite violence. However, their actions contribute to an environment where such acts become more likely.

Case Illustration – Online mysogyny against female political leaders: Canadian example in the news

Threats, harassment and online hate driving women out of politics, MPs warn

Jasmeen Gill – The Canadian Press
March 8, 2025
Source: https://globalnews.ca/news/11073007/threats-harassment-and-online-hate-driving-women-out-of-politics-mps-warn/

Excerpt: **"**As longtime Liberal MP Pam Damoff prepares to leave politics when the next federal election is called, she is wistful but open about what is driving her to leave a career she has had for more than a decade. Vocal about the misogyny and threats she faced during her time in government, she wants public safety officials to take these threats more seriously. 'We've seen a shift in how people treat politicians, and I really worry that at some point, someone will be injured or killed,' Damoff said in an interview."

CBC News (Maimann, 2024) reported on recent research by various NGOs, that female politicians frequently face online abuse. This is attributed to "systemic social media problems" – particularly the lack of enforcement of community guidelines. Sobieraj (2020) highlighted rigorous research showing how online abuse of female officeholders is systematic and persistent. Female politicians and activists are often targeted with online threats, harassment and graphic sexual depictions - tactics designed to undermine their legitimacy, strip individuality, and discourage political engagement (DiMeco, 2019).

In April 2025, British news media reported on a UK parliamentarian who received a series of death and rape threats on social media, which she attributed to followers of social media influencers promoting misogyny (Barker-Singh, 2025). These online attacks began after she criticized a controversial social media owner. This case is noteworthy because it reflects a common feature in foreign information manipulation: certain channels align with actors

popular or grass-roots support

[11] https://en.wikipedia.org/wiki/Stochastic_terrorism

to support their objectives while evading attribution (e.g., Besancenot, 2025). In other words, these channels and followers act as unwitting proxies for other parties. This may also be considered a form of "soft violence", which refers to non-physical actions that, while not criminal, *per se*, are used to undermine social cohesion and assert group dominance, serving as a primary tool for communication, recruitment, and radicalization in violent transnational social movements (Kelshall, 2020).

In her critique of the polluted information ecosystem, Schick (2020) calls for a clear and consistent understanding of the disinformation problem as a crucial first step towards effective action. Heeding this call, we begin by framing online gendered disinformation as part of a broader context of harm and a set of harmful practices.

Conceptualizing Gendered Disinformation

Definitions

In 2022, the UN Women Expert Group sought to develop a common definition for the broad category of technology-facilitated violence against women or gender-based violence (TF-VAW/GBV). They determined that concepts of TF-VAW generally included some or most of the following features (UN Women Expert Group, 2022, pp. 3-4):

- **VAW or GBV:** An implicit reference to existing definitions of violence against women and gender-based violence;
- **Gender dimension/motivation of the act:** A specification that it is an act of gender-based violence, directed towards a woman because she is a woman or that affects women disproportionately. (We advocate for the inclusion of identities and behaviours that do not conform to traditional – particularly, ideologically-driven – formulations of gender within this dimension.)
- **Means:** Naming of ICT or technologies generally and/or specific technologies (e.g. spyware, GPS) as the means through which the violence was perpetrated.
- **Medium or Space:** Referenced as 'online' or 'cyber' or 'digital' spheres.
- **Forms of TF-VAW:** A list of some or several specific forms of TF-VAW, (e.g. sextortion, doxing, trolling).
- **Harm:** Reference to harms generally, or specific forms of harm, that ensue as a result of having experienced TF VAW (e.g physical, sexual, psychological, social, economic, other).
- **Continuum of VAW:** Reference to the fact that TF VAW occurs within a continuum of violence, that can include offline violence, and vice versa. For example, a woman may be stalked online and then the stalker may show up at her place of work, or a partner abusing a woman at home may monitor and control her movements event when they are not home, using GPS enabled technology.

As a result, the UN Women Expert Group (2022) proposed the following common definition for technology-facilitated violence against women, with the proviso that VAW could be replaced by GBV:

> ... any act that is committed, assisted, aggravated, or amplified by the use of information communication technologies or other digital tools, that results in or is likely to result in physical, sexual, psychological, social, political, or economic harm, or other infringements of rights and freedoms (UN Women Expert Group, 2022, p.4).

Within disinformation, a specific subgroup of TF-VAW – **gendered disinformation (GD)** – targets **women and girls *individually* through misogynistic harassment and intimate partner violence, *and collectively*** to subjugate them and deny them voice and participation in democratic society.

The National Democratic Institute defines GD as "the use of false information to confuse or mislead by manipulating gender as a social [wedge] to attack women and/or to sway political outcomes." (Jankowicz, et al., 2021, p.3). When spread online, GD may be viewed as a form of online violence, perpetuating hostile systems against women and posing "a credible threat to democracy" (Sobieraj, 2020, p.152).

Online (or digital) gendered disinformation involves the misuse of information communication technologies (ICT) to:

- Release/propagate false or misleading information about individual females, groups of females, or females, in general; *and/or*
- Overtly and explicitly abuse individual females, groups of females, or females, in general.

Case Illustration – Jess Davies and the long-term harms of online sexual exploitation

'I don't date at all now': one woman's journey into the darkest corners of the manosphere

When Jess Davies was 15, a boy leaked pictures she'd shared with him. At 18, she was a glamour model. A few years later, another man violated her trust. Then she fought back
Anna Moore – The Guardian
April 30, 2025
***Source*:** https://www.theguardian.com/society/2025/apr/30/i-dont-date-at-all-now-one-womans-journey-into-the-darkest-corners-of-the-manosphere?CMP=Share_iOSApp_Other.

Jess Davies, now a women's rights advocate and media professional, was first exposed to online sexual exploitation as a teenager when a private image she had shared with a trusted peer was circulated without her consent. Over the years, she became the target of repeated image-based abuse, including the unauthorized distribution of her photos, cyberflashing*, impersonation, and catfishing**. Her images were misused across pornographic platforms, social media, and anonymous forums, where users engaged in "games" that involved trading, modifying, and humiliating women through manipulated content and explicit commentary. In many cases, images were posted alongside the victim's name and contact information, encouraging coordinated harassment.

Davies' experience demonstrates how the distortion, misuse, or fabrication of content targeting individuals based on gender can intersect with sexual exploitation online. These harms were enabled by weak platform governance, societal stigma, and the absence of clear accountability for perpetrators. Despite years of digital abuse, Davies received no apology from those responsible. Her story also illustrates the long-term psychological, social, and professional impact of such violations, and the urgent need for legal reform, proactive platform responsibility, and survivor-centred support systems to address the growing threat of gendered disinformation and online sexual exploitation.

Case Illustration – Sharing intimate photos without consent: Canadian example in the news

In December, the Winnipeg Police Service were investigating reports of AI-generated nude photos of underage students circulating at Collège Béliveau, a Grade 9-12 high school in Windsor Park
Jen Zoratti – Winnipeg Free Press
February 10, 2024
Source: https://www.winnipegfreepress.com/arts-and-life/2024/02/10/seeing-is-believing-the-real-and-present-danger-of-fake-ai-images
Excerpt: "The speed and ease with which these images can be created and spread is also alarming; one doesn't even need to have a mastery of Photoshop anymore.... And yet, despite this rapid acceleration in technology, it seems as if we're still stuck in 2014 when it comes to the law. ... Manitoba is one of eight provinces that do indeed have intimate image laws, but ours don't refer to altered images. That needs to change, and fast. We cannot afford to have the creation and distribution of sexually explicit AI-generated images dealt with the same way online sexual harassment has traditionally been dealt with, which is to just tell women to 'stay off the internet.'"

Social Contexts and Conditions Enabling Gendered Disinformation

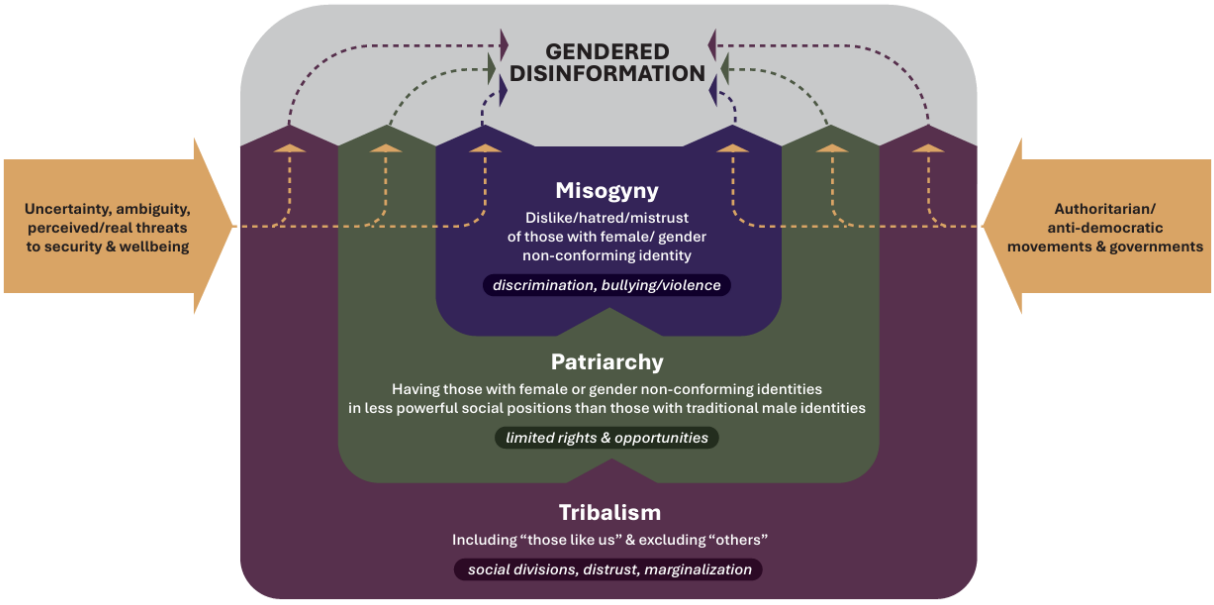
Gendered disinformation occurs within a complex context where cultural and social factors - such as group membership, identity, cognitive biases, beliefs, values, norms, and practices - influence human behavior and societal outcomes. This includes how interactions within cultural and social settings shape attitudes, actions, and development.

This context involves both socio-political and political-economic dimensions:

- How relationships among people, groups, and institutions – shaped by culture, group identity and social norms – influence political processes, policies, and power dynamics;
- How political institutions, processes, and policies – together with political decisions and the distribution of power and resources - affect economic systems and outcomes.

Key features of the broader social context that enable gendered disinformation, including online GD, include misogyny, patriarchy and tribalism (Figure 1).

Figure 1. Broader conditions may constitute fertile ground for the expression and spread of gendered disinformation.



Specific instances of GD may reflect one or multiple forms of systemic marginalization and injustice. These can include active repression – such as intimate partner violence and coercive control (e.g., Gill & Aspinall, 2020) – occurring domestically in what we term "intranational

repression" - or actions carried out at the behest of – or inspired by – foreign actors ("transnational repression") (e.g., Human Rights Watch, 2024).

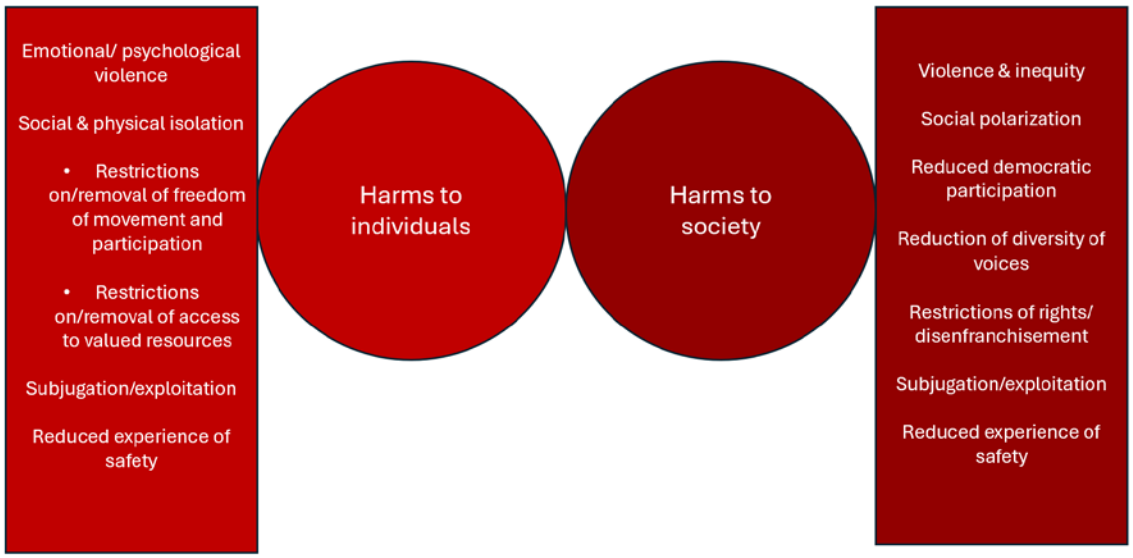
No single element of the broader societal context will, by itself, produce GD or its harms. These elements are akin to changing soil conditions, influencing how negative processes can be initiated, take root and impact their ecosystem. The socio-technical features of our world (such as digital technologies and political movements) create conditions that make either desirable or undesirable outcomes related to GD more or less likely. This means that *no single type of intervention can solve the problem*. However, understanding the layered contexts that can enable harm allow us to develop strategies to shift the environment toward more desirable outcomes.

- **Misogyny:** The hatred, dislike, or mistrust of those identified as being of the female gender. Manifests through discriminatory attitudes, behaviours, and institutional practices that demean, belittle, and oppress females, reinforcing patriarchal structures and limiting their social, economic, and political freedoms and opportunities (e.g., Sobieraj, 2020).
- **Patriarchy:** Male gender holds primary power and dominates in roles of leadership, authority, and control in both public and private spheres (e.g., Richardson-Self, 2021). This system often marginalizes females and limits their opportunities and rights.
- **Tribalism:** Inherited capacity for cooperation arising from cognitive & social practices – including the creation, propagation and promotion/enforcement of narratives – that reinforce shared identities & trust and which may intensify in-group/out-group dynamics (Samson, 2023).

In a mainly intra-national setting, the environment may include individually-focused gender-based violence and politically or ideologically-based harassment and abuse of individuals or groups. This may also involve political discourse or policy positions that reflect patriarchal or misogynistic rhetoric.

In a global setting, it may involve direct foreign interference in domestic life or democratic processes, or indirect foreign influence supporting misogynistic or patriarchal ideological movements (Figure 2).

Figure 2. Types of harms stemming from gendered disinformation.



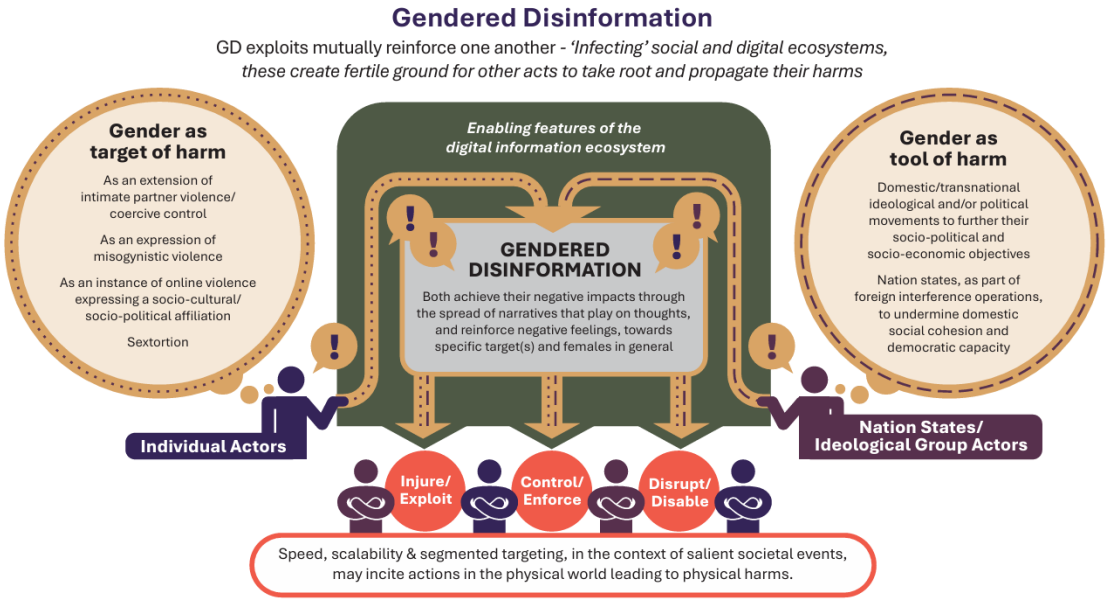
One desired effect is to deter women from participating in civic life and to erase from public policies values related to inclusion and belonging. Another is to exacerbate social divisions in the service of domestic political or ideological objectives, or adversarial geo-political goals.

These social and geographic contexts may also intersect, reflecting the influence of globalized narratives and movements. Perpetrators may be individuals, acting individually against females or groups of females. This positions female identity or gender as a focus of harm.

Group or nation-state perpetrators may target individuals perceived as obstacles to their global ambitions or ideological objectives (foreign interference against individuals), or as a method of disrupting the target nation's social harmony and democratic stability (foreign interference against the nation). These objectives may include the subjugation of women, girls and gender non-conforming persons, positioning female identity or gender as a tool for harm (Figure 3).

The broader socio-technical problems of contested truth and tribalism (of which populist movements are one example) have created a ripe environment for GD to proliferate. These exploits prey on individuals conditioned and motivated to believe false narratives and inaccurate explanations, including conspiracy theories centred on female identity or female/gender non-conforming leaders (e.g., van der Linden, 2023).

Figure 3. Position of females/gender non-conforming persons within gendered disinformation ecosystem.



Enabling Features of the Digital Ecosystem

Gender disinformation exploits are increasingly powered by AI advancements, such as botnets with natural language capabilities, and realistic synthetic media, ("deep fakes") (e.g., Schick, 2020). These emerging technologies complicate detection and countermeasures development.

AI-generated synthetic media, including deepfakes, voice cloning, and image-generation tools, allow the creation of convincing fake audio-visual content with minimal technical skill or cost (Lalonde, et al., 2025). These tools are widely accessible through user-friendly interfaces, democratizing the production of sophisticated disinformation.

The emergence of visual and multimodal disinformation (VMD) marks a significant shift in online abuse, including GD. Rapidly spread online, VMD technologies enable more persuasive, emotionally resonant, and harder-to-detect attacks (Lalonde, et al., 2025).

These capabilities have been weaponized to create non-consensual explicit content, falsely depict women – especially female politicians, activists, and journalists – in compromising scenarios, aiming to impersonate or discredit them. These tactics are not only invasive and damaging, but also serve to intimidate, silence, and discredit women in public life.

Members of racialized, LGBTQIA+ and intersectional communities are particularly vulnerable as potential "targets and tools", however this subgroup has been less well-researched (Thakur & Hankerson, 2021). Disinformation campaigns may draw on pre-existing, culturally potent,

discriminatory narratives related to both race and gender to lend credibility to false information. As a result, intersectional disinformation may weave together multiple harmful tropes and stereotypes, using these layered narratives to make false messages appear more believable and persuasive (Thakur & Hankerson, 2021). The impact of combining stereotypes with manipulated or manufactured audiovisual content, sometimes layered with actual news to enhance credibility, can be significant (Lalonde, et al., 2025).

These technologies, along with the design and business models underlying social media platforms, create opportunities for efficiency in scaling and targeting exploits of every kind. As Ressa (2022), Zuboff (2019) and others have demonstrated: features including the ease with which content may be re-posted within algorithmically optimized networks enables its speed and spread – its potential to amplify content faster than the pace of verification efforts (Lalonde, et al., 2025); the ease of access to networks which may be appropriated or botnets that can mimic popular support for a topic also contribute to the perceived realism of disinformation (Council of Canadian Academies, 2023); and, finally, the business incentive of platforms to attract and hold attention has led to the use of algorithms that arouse emotions and funnel increasingly intense content to users that have been caught-and-held (Bail, 2021). These "recommender" algorithms amplify certain voices and suppress others; they have also been flagged as a critical focus for mitigating online harms; there is a growing dialogue in the EU about regulatory options to disable or modify these algorithms (Ryan, 2025).

By leveraging these features of the digital ecosystem, online GD amplifies its disruptive and harmful effects while creating new avenues for victimization, both directly and indirectly. Indirectly, it fosters repressive elements of culture – hindering the safe and healthy participation in civic life. In some cases, these processes also create conditions conducive to stochastic terrorism, as previously identified.

COUNTERMEASURES: FOSTERING RESILIENCE AND DEVELOPING RESPONSE CAPACITY

The social scientific research base[12] available for developing GD countermeasures is just beginning to emerge. This has happened largely over the past several years, as part of efforts to better understand online identity-based polarization and abuse. Research thus far has varied, spanning

[12] This does not include military/national security research focused on the disruption of foreign interference and/or influence operations.

several disciplines[13] and drawing significantly from analyses of the mechanisms of mis- and disinformation. The focus of this work includes:

- Understanding the forms and mechanisms of disinformation and crafting appropriate interventions;
- Delineating the features and effects of online gendered violence;
- Critiquing the ways that the design, use and practices of social media platforms[14] foster polarization and harm;
- Analyzing the ways that populist disinformation produces and leverages social polarization as part of a broader set of socio-political objectives;
- Exploring and explaining how features of the digital ecosystem intersect with emotion- and identity-based factors to produce polarization; and
- Describing how certain contexts, together with desires for group affiliation and identity-protective cognitive processes, create vulnerabilities for disinformation threats that exploit narratives, symbols and other cultural artifacts.

False narratives[15] are used to harm or dehumanize members of designated out-groups by leveraging psychological biases among in-group members. These threats are designed to short-circuit critical thinking. Images and narratives that instil and amplify perceptions of scarcity and threats to the in-group's social standing blame out-groups for these issues.

Each of these focus areas provide information on the forms, mechanisms and impacts of information and narrative-based harms. They also identify ways to counteract disinformation, online harms, and to prevent or disrupt the social processes that create a fertile environment for online violence.

The Form of Disinformation Operations

Influence operations, including disinformation, often follow a structured format for achieving their objectives. This is known as a "kill chain". Drawing from military terminology, this term refers to a step-by-step outline of the stages of an attack – from initial reconnaissance to final impact. These frameworks help defenders understand and disrupt adversaries at each stage of their operation, rather than only responding to, or after, the attack.

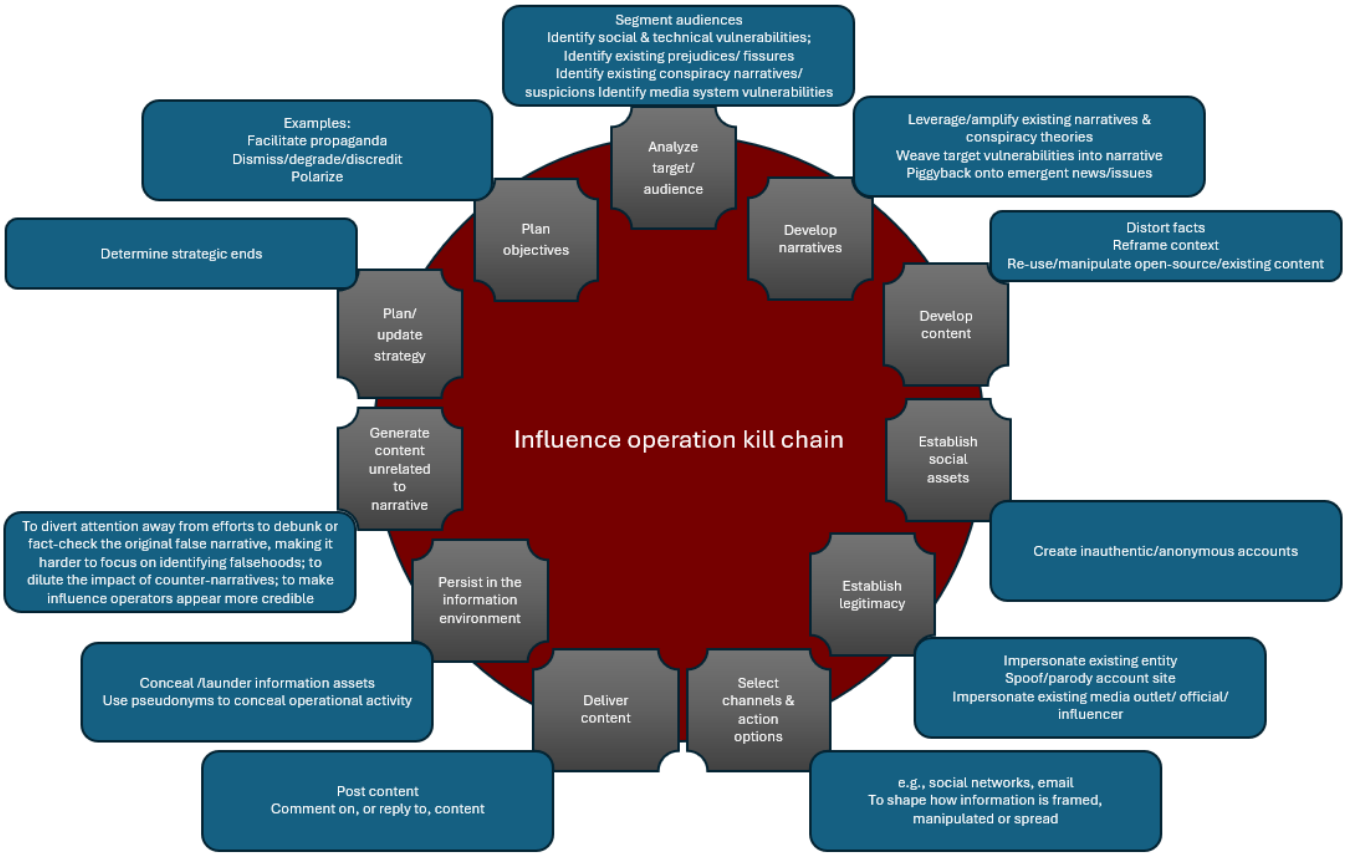
[13] Communications, sociology, social psychology, anthropology, as well as philosophy and history

[14] e.g., policies and practices on moderation and participation, business models, ownership issues, membership

[15] Content may include spoken or written narrative and/or images that invoke and align to particular narratives (e.g., evidence-free allegations that portray a woman as weak, unintelligent or incapable of leadership). As propaganda and influence operations have demonstrated, potent cultural symbolism (including music and other forms of artistic expression) may also be deployed as part-and-parcel of narrative interventions (see Pomerantsev, 2024).

The French Secrétariat général de la défense et de la sécurité nationale (VIGINUM)(2024) developed one such framework to provide insight into the operations used to carry out influence operations (Figure 4). This fairly intuitive structure can help us understand, identify and analyze the tactics, techniques and procedures that can be used as part of disinformation exploits, and determine practical avenues for responding to these where they occur.

Figure 4. General model of an influence operation kill chain (after VINIGUM, 2024).



The 11 components of the VINIGUM model cover areas such as: planning; development; delivery; and sustainability. It also suggests an iterative learning dimension to disinformation operations – based on assessments of responses to the exploit, operators may adjust various features of the campaign to refine targeting opportunities or to address to shifting goals.

However, kill chains are used largely in a responsive fashion, to analyze an attack once it has been detected, since these tools were designed for interdiction, not prevention. The presence of a problem should not be the starting point for security.

Spotting spreading disinformation

Example: The rapid spread of very similar-looking false information by numerous accounts in the immediate aftermath of an actual event – such as a political rally or a demonstration.

This might signal the use of a botnet or camouflaged account activity seeking to discredit or burnish the reputation of an individual or group. Appropriately packaged knowledge of the forms and features of disinformation operations can be an important part of awareness-building.

Understanding the typical format of disinformation operations can help targets, their supporters and responders detect, call out or interdict unfolding campaigns.

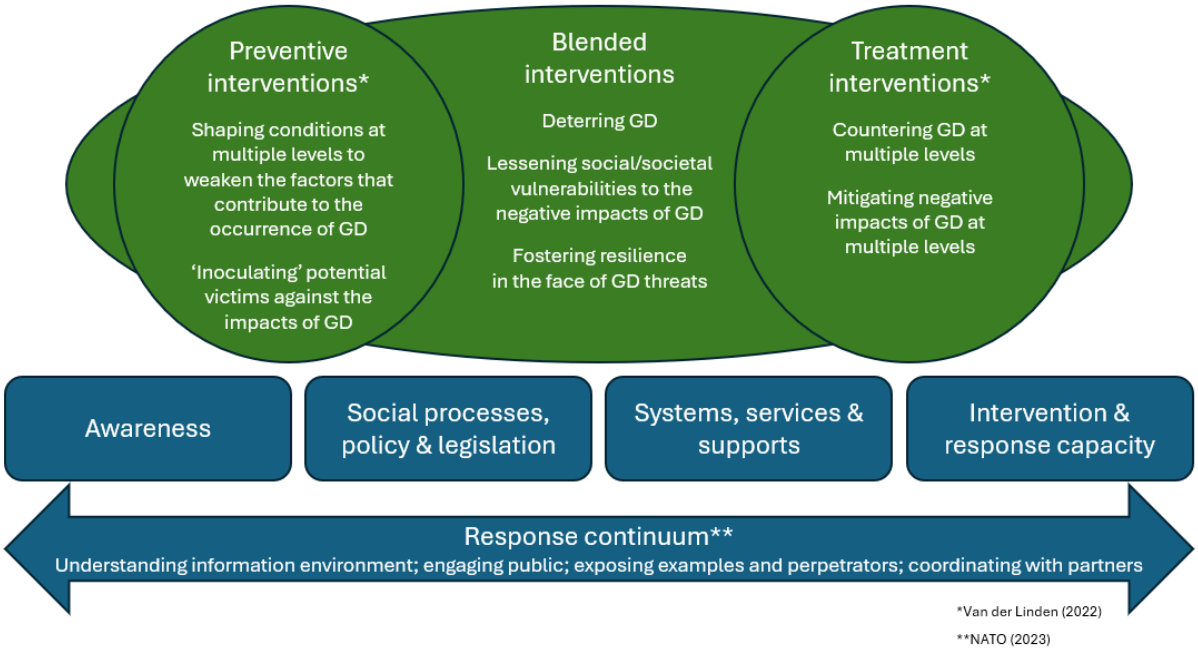
Strategic Interventions to Address Vulnerabilities and Threats

The EIU (2020) study of online violence against women identified that, at that time, efforts to tackle gender-based violence continued to focus mainly on post-experience responses, rather than on prevention. Despite this issue being flagged, the recommendations stemming from a recent report (Jankowicz, et al., 2024) predominantly focused on more 'downstream' or 'midstream' responses, such as platform accountability and victim support or responses to image-based sexual abuse.

van der Linden (2023) suggests that interventions can be helpfully conceptualized along a continuum of upstream to downstream actions categorized as largely preventive or largely "treatment" focused. Similarly, NATO (2023), recognizes that disinformation is simply one element of a wider collection of malicious information activities that range from hostile narratives targeting individuals to foreign information manipulation and interference (FIMI)[16]. At the far end of this continuum, hybrid warfare can make use of military and non-military channels to spread uncertainty among people and weaken the stability and trust within societies (NATO, 2023). Given the complexity of this landscape of digital harms, NATO's approach to countering disinformation emphasizes the importance of partnership and collaboration. Whether confronting GD as an expression of interpersonal violence, or as a geo-political phenomenon, a strategic mix of interventions, including those that fall "mid-stream" will likely be most effective as a framework for action – and as the basis for building greater resilience to the harmful effect is of GD. We propose an expansion of van der Linden's continuum to include mid-stream interventions at the macro and individual levels focusing on reducing vulnerabilities and strengthening resilience (Figure 5).

[16] While not always illegal, FIMI operations are manipulative activities designed to negatively impact "values, procedures and political processes in a target country" (NATO, 2023)

Figure 5. Proposed continuum of interventions for addressing gendered disinformation.



In only a few cases[17] – largely focused on mis- and disinformation not involving GD – has there been research on the effectiveness of specific countermeasures. Consequently, the development of countermeasures against GD, specifically, must be seen as an inferential exercise that ultimately should be tied to further research and evaluation.

An important caveat on any notion that tackling disinformation is a straightforward process is offered by Rid (2020), who observed that:

> Active measures have become more and more active and less measured to such a degree that they are themselves disintegrating – and this disintegration creates a new set of challenges. For the offender, campaigns have become harder to control, harder to contain, harder to steer, harder to manage, and harder to assess, For victims, disinformation campaigns have also become more difficult to manage, more difficult to assess the impact, and more difficult to counter.
>
> ...both open and closed societies... are both overstating and, more rarely, understating the threats and the potential of disinformation campaigns – and thus helping expand and escalate that very threat and its potential. (p. 434)

[17] Namely, work conducted by van der Linden (2023), Hameleers (2022) and Bail (2021) and their associates

This sobering assessment highlights the limitations of single-channel approaches to a complex, issue like GD. A broader vision for change is needed which focuses on shaping the conditions that increase positive outcomes and reduce negative ones.

Change of this magnitude must begin by mobilizing a diversity of experiences, perspectives and allies. It must include focus on individuals as well as groups and populations, and it must seek to counteract factors responsible for cognitive vulnerabilities (e.g., willful disbelief in facts[18]) and affective vulnerabilities (e.g., isolation, deprivation, status-seeking needs and perception of threats to social position[19]). A strategic vision of change should seek to:

- Enable awareness, identification and response capacity;
- Promote safety and pursue accountability;
- Foster a more equitable and inclusive society; and
- Support off-ramps to healthy identities and alternative affiliations for those vulnerable to enrolment in harmful movements and practices[20].

A crucial first step is to raise awareness and to make available concepts and systems for effective action by a variety of stakeholders. While individual action is important, GD is not simply a matter of "personal troubles"[21]. As Sobieraj (2020) and others have suggested, this is a public issue and a shared threat to democracies. Consequently, victims should not shoulder the burden alone – there is a role for everyone. While constructive changes against a problem of this magnitude will take time, systemic and coordinated change grounded in a holistic perspective, and informed by multiple voices, will be more effective than a series of unaligned incremental changes.

At the moment, the best available evidence on what may constitute promising countermeasures focus on points of vulnerability at which harms are either at risk of occurring, or have already begun to be associated with negative outcomes for people[22].

---

[18] e.g., McIntyre (2023), Norman (2021), Samson (2023)

[19] e.g., Bail (2021)

[20] As McIntyre (2023) suggests, in trying to wipe out the sources of the "disease" of untruths, we should also attend to the "sick" – i.e., those who have been deceived into believing and following GD narratives. However, this is difficult work, as van der Linden (2023) has described in relation to belief in conspiracy theories.

[21] Sobieraj (2020, p.138)

[22] Using a public health lens, and drawing from the Institute of Medicine's (IoM) framework of prevention (Pronk, Hernandez & Lawrence, 2013), preventive and blended interventions would correspond approximately to "universal" and "selective" measures, respectively, while treatment focused interventions would correspond to "indicated" measures. The IoM lens may have applicability to the problem of disinformation-polarization as it seeks to support effective action planning tied to an understanding of population-based levels of risks. Accordingly: universal prevention focuses on segments of the population deemed to be low-risk; selective prevention focuses on groups experiencing shared sets of risk factors (and may seek to boost protective factors); and indicated prevention seeks to serve those with emergent, detectable "signs and symptoms" of the problem of interest. The latter may involve individual and small-group delivery of services and supports aimed at preventing the progression of harms (Springer & Phillips,

Understanding and Awareness

Many have described the problem of disinformation as one of a "post-truth crisis" (e.g., McIntyre, 2023) or "infodemic" (van der Linden, 2022). This involves several features, including the creation of contested truths in relation to specific topics (such as climate). It also includes fostering more general conditions promoting contempt – even disgust – for those who have been positioned in some way as 'other', and cynicism towards the truths they share about their experiences (e.g., Bail, 2021; Samson, 2023). In the present case, 'others' are those members of female identifying gender groups that are positioned as objects of: subjugation; humiliation; shame; exclusion; blame; abuse; or any combinations of these harms. Is this light, gendered disinformation can be seen as a form of discourse-based violence.

Undermining the truth and the social position of designated 'others' serves an important aim of authoritarian ideologies within which misogyny is a central feature, and where those who identify as female are both tools and targets. The category of 'other' may include, for example, women, ethnocultural or religious minorities, members of political parties or members of the so-called 'elite' (media, academics, professionals, government and other public institutions, etc.). In some cases, there is overlap among these categories. Rid (2020), Ressa (2022) and McIntyre (2018, 2023) have shown how contested truth, the cultivation of cynicism and distrust, and the tools and platforms of social media, are used by populist authoritarians to undermine social cohesion and to target political adversaries by vilifying them through disinformation exploits that scale and repeat falsehoods. These are used to manipulate public opinion and to serve as a justification for persecution.

Blame and disbelief are key tools in the populist disinformation arsenal (Hameleers, 2022). Among other objectives, these are used to garner allies from among those who may lack healthy social connections and opportunities for secure and meaningful participation in society, and/or who fear the loss of valued, hierarchy-based, identities[23].

Contempt and Control

Sobieraj (2020) has described the ways that online gender-based attacks have the effect of creating a "context of contempt" and a "climate of unsafety" (p.35) that can undermine the willingness and capacity of women to participate in the everyday and democratic life of their communities and society. Corroding the social and political position of women is a key goal of misogynistic ideological movements and is known by its own term, Violence Against Women in Politics (VAW-P) (Jankowicz, Pepera & Middlehurst, 2021).

2021).

[23] These may include forms of supremacy that position specific groups at the top of a socio-political hierarchy which actively subjugates and blames those who are ascribed as being outside of this category.

As Richardson-Self (2021) has observed, recurring discourse focusing on subordinating identity-based groups is often accompanied by a "constellation of other acts" (p. 81), which can include forms of involuntary control and actual or threatened physical violence (Havard & Lefevre, 2023). The range of these tactics, which are used to enforce male dominance, can be referred to as "coercive control" (Stark, 2007).

> *Democracy requires a common understanding of reality, a shared view of what has happened, that informs ordinary citizens' decisions about what should happen, now and in the future. Authoritarians target this shared understanding, seeking to separate us from our own history to destroy our self-understanding and leave us unmoored, resentful, and confused. By setting us against each other, authoritarians represent themselves as the sole solution.*
>
> (Stanley, 2024)

Coercive control involves both physical and non-physical tactics to dominate and isolate the victim (Gill & Aspinall, 2007, 2020). These tactics include threats, monitoring, financial control, and restricting access to loved ones, ultimately eroding the victim's sense of self and freedom (Gill & Aspinall, 2020; Stark, 2007). It is considered an infringement on basic liberty and a form of intimate partner violence when it occurs within these types of relationships – whatever the gender or sexual orientation of the parties.

Common aspects like controlling actions, psychological abuse, sexual jealousy, and stalking can be facilitated by information communication technology (ICT) (Dawson et al., 2019). Douglas, Harris, and Dragiewicz (2019) found that ICT tools, such as smartphones and IoT devices, are used for technology-facilitated violence.

Victim-blaming and disbelief, often seen in coercive relationships and failed responses, align with misogynistic and populist ideologies (Bail, 2021; Cuklanz, 2023; Richardson-Self, 2021; Samson, 2023; Sobieraj, 2020). Similarly, in cyber deception, tactics like uncertainty and misdirection are key (McMahon, 2021).

Foreign Inteference and Manipulation of Information

Like all forms of disinformation, gendered disinformation interferes with the capacity of a society to engage in constructive public dialogue involving a pursuit of common understanding based on shared facts (Richardson-Self, 2021). As a result, gendered disinformation has been used as a component of FIMI operations – efforts by foreign governments or actors to influence public opinion, political decisions, or social stability in another country. This is done by spreading false or misleading information, often through social media, news outlets, or other communication channels. These activities are usually designed to create confusion, distrust, or division among people, and they can target elections, public health responses, or social issues. FIMI is considered a serious threat because it can quietly undermine a country's democracy, security, and public confidence without using traditional weapons or direct attacks.

Work by Bradshaw and Henle (2021) explored how foreign state actors (Russia, Iran and Venezuela) conducted covert influence operations on the Twitter platform to target Western feminist activists and politicians.

Several strategies were used by state-associated operatives to undermine feminist advocates/feminist narratives indirectly or directly. These included narratives designed to:

- **Promote in-group solidarity and out-group divisions** (e.g. around racial and political identities) to amplify negative feelings towards a movement and its supporters – such as that activists were "man-hating" and oppressive;
- **Undermine women's shared sense of a collective identity** by co-opting internal critiques within feminist movements; and
- **Direct online harassment and character attacks against individuals** to delegitimize or discredit them – in some cases, these were combined with threats of physical violence.

The latter were found to be more likely to occur via direct messaging to victims, rather than on public platforms (possibly because Twitter, at that time, was more actively deploying automatic detection measures). This research highlights how digital interference operations involving techniques for promoting social divisions and disrupting collective action are being used to undermine gender equality and weaken democracy by making it harder for women to speak out and mobilize for change (Bradshaw & Henle, 2021).

A related application is gender-based transnational digital repression. This involves the use of TF-VAW by authoritarian regimes to interfere with the exercise of free speech and activism by female diaspora residents of other countries. Research by the Citizen Lab at the Munk School of Global Affairs and Public Policy has shown how invasive monitoring and other forms of surveillance, along with online harassment and various forms of reputational assaults, have been used to extend the control of distant repressive states, or to marginalize victims within their diaspora communities in Canada (Aljizawi, et al., 2024; Michaelsen & Anstis, 2025).

For example, attackers have used mercenary spyware[24] implanted on devices using phishing exploits[25] to collect information and monitor the activities of civil society targets. The spyware is

[24] For example, tools developed by the NSO Group, as described by Deibert (2025), who provides detailed accounts of the use of mercenary spyware against civil society actors.

[25] Phishing exploits are deceptive tactics used by cyber operatives to trick individuals into revealing sensitive information, such as passwords, credit card numbers, or banking credentials. These schemes often involve fraudulent emails, text messages, or websites that closely mimic legitimate sources – like a bank or trusted organization – in order to gain the victim's trust and steal personal data. For example, an email may appear to come from an individual's bank asking them to "verify your account," when, actually, it is a carefully crafted exploit.

used as part of initial reconnaissance activities designed to determine social and technical vulnerabilities as part of the prelude to a disinformation operation. These kinds of exploits follow carefully designed, nested, operational plans, drawing from features of the kill chain used to implement malicious cyber exploits in the context of format of influence operations (Figure 6).

Cyber-enabled exploits are an important feature of contemporary influence operations (including information warfare) because they provide sophisticated hostile actors (individuals, groups or governments) with a number of benefits, such as:

- Securing new sources of private information;
- Diverting attention from the main objectives of an information operations; and
- Interfering with counter disinformation capabilities (Whyte, 2020).

Spyware – a powerful new threat

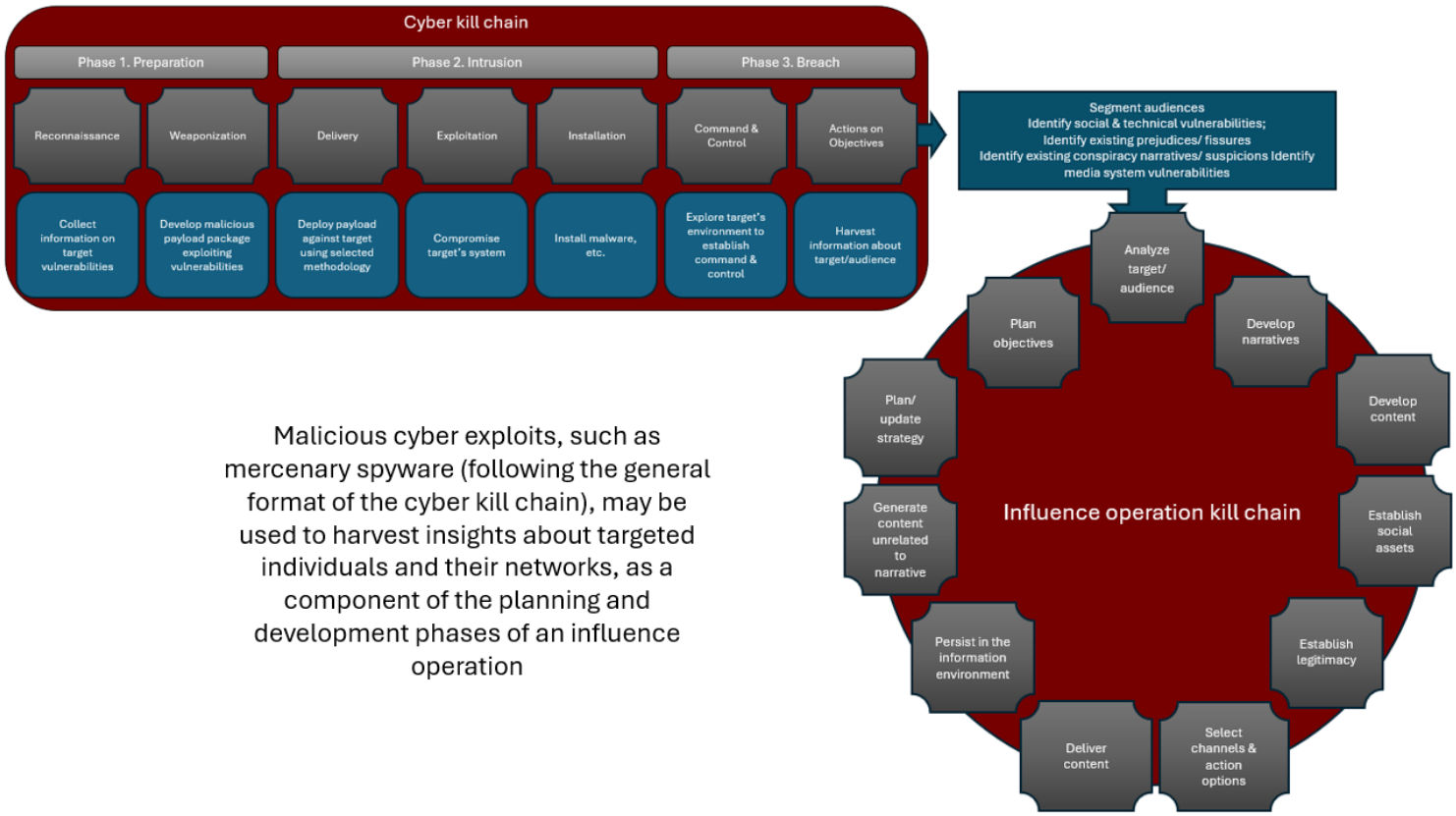
Malware or spyware exploits are tools used by attackers to secretly access or control computers and devices. In influence operations, these malicious programs can be used to steal sensitive or personal information, or monitor activities, as insights used to craft a disinformation campaign. These tools can also be used to manipulate communications, or spread false narratives to achieve political, social, or economic goals.

Imagine someone receives a fake email that looks like it came from their social media site. When they click the link, malware is installed on their device. The attacker uses this malware to steal login details, take control of the account, and then spread disinformation or harmful messages to all the person's contacts, making it look like those messages came from someone they trust.

Michaelsen and Anstis (2025) observed that many of today's authoritarian regimes see traditional gender roles and patriarchal norms as enablers of the social hierarchies and agendas on which the regimes depend for their bases of power and control. They assessed that these structures of domination need to be sustained both domestically and internationally in order to safeguard the stability of the regime. Where misogyny is used as an enforcement tool, polarizing potential adversarial alliances and disrupting in-group cohesion among their opponents helps authoritarian leaders legitimize their agendas in the eyes of their ideologically aligned supporters (Michaelsen & Anstis, 2025).

Among its many uses, spyware is being deployed by authoritarian governments as a tool for supressing dissent. An example of the use of cyber-enabled repression was uncovered by Citizen Lab (Marczak, et al., 2021) which described how two types of mercenary spyware – Cytrox's Predator and NSO Group's Pegasus – were used to compromise the iPhones of an Egyptian journalist and a politician via WhatsApp messages. In one case, both of these forms of spyware were used against the same individual. These compromises were carried out as part of an operation designed to quash dissident voices within civil society.

Figure 6. Example of the way that elements of a typical cyber exploit may connect to those of a targeted influence operation (after Hutchins, Cloppert & Amin, 2010 and VINIGUM, 2024, respectively)



Malicious cyber exploits, such as mercenary spyware (following the general format of the cyber kill chain), may be used to harvest insights about targeted individuals and their networks, as a component of the planning and development phases of an influence operation

Psychological Vulnerabilities Exploited by Disinformation

Education and awareness are important elements of an effective response at the societal level as well as in relation to opportunities for accountability, redress and rehabilitation by perpetrators (e.g., McIntyre, 2023; Cuklanz, 2023).

Understanding how online repression works and, in some cases, using digital tools to uncover and address online abuse are key to combatting gendered violence (e.g., Parrish, et al., 2023; Carty, 2023; Cuklanz, 2023).

One of the ways that fake content works is by showing us things that we anticipate would go together. The truth-likeness of these texts, images, or sounds can fool our systems of perceiving and thinking clearly by capitalizing on everyday expectations, otherwise known as cognitive biases.

Cognitive Biases

Cognitive biases are commonly understood as systematic and widespread mental tendencies that distort how we process information, often leading to outcomes that are inaccurate, flawed, or less than optimal (Korteling & Toet, 2020). They are mental "shortcuts" or patterns of thinking that can lead people to make judgments or decisions that are not fully rational or accurate. They also make us vulnerable to manipulation (McIntyre, 2018).

> ***Cognitive biases are hard to notice***
>
> *Succumbing to cognitive bias can feel a lot like thinking. But especially when we are emotionally invested in a subject, all of the experimental evidence shows that our ability to reason well will probably be affected.*
>
> (McIntyre, 2018)

These biases arise because the brain tries to simplify complex information or make quick decisions under uncertainty. While these shortcuts can be helpful in daily life, they can also cause people to misinterpret information, overlook evidence, or rely too heavily on pre-existing beliefs.

Cognitive biases are one contributor to the believability of disinformation (McIntyre, 2018). Previous research conducted by George, et al. (2021, cited in French, et al., 2025) concluded that confirmation bias – the tendency to favor information that aligns with existing beliefs and dismiss information that contradicts them – was influential in the spread of fake news. McIntyre (2018) suggested that other cognitive biases may be notable for their contributions to a vulnerability to believing false information. These include:

- **Backfire effect:** In which people strengthen their existing beliefs when presented with information or evidence that contradicts them. Instead of changing their views, they double down, often becoming even more committed to their original position. This effect highlights one reason why individuals may resist changing their beliefs, even in the face of clear, corrective information.
- **Dunning-Kruger effect:** The tendency for individuals with limited knowledge in a subject to overestimate their understanding, making them more susceptible to disinformation. This is a version of what is also known as overconfidence bias (see below).

In a recent study, French, et al. (2025) examined social media users' perceptions of how they use and share fake news. They analyzed these results to identify the cognitive biases that appear to shape the believability of fake news when it is being consumed. They found five cognitive biases likely to make fake news believable. These were:

- **Herd mentality:** The tendency of individuals to adopt the thoughts or behaviors of a group, often following peers rather than forming independent judgments. This dynamic can lead people to accept information as true simply because it aligns with the majority view, rather than critically evaluating the evidence themselves.

- **Confirmation bias:** The tendency to favor information that aligns with existing beliefs and dismiss information that contradicts them.
- **Framing cognitive bias:** Involves the ways that people's decisions are influenced by how information is presented, rather than by the facts themselves. Even when the underlying information is identical, different wording or context can lead to different conclusions. In the context of fake news, this bias can cause individuals to judge a story's truth based on how coherent or compelling it sounds, without verifying the information.
- **Overconfidence bias:** This involves overestimating the accuracy or depth of one's knowledge. This can lead people to have excessive confidence in their judgments or decisions, even when their actual understanding is limited.
- **Anchoring bias:** The reliance on the first piece of information encountered, which can shape how new information is interpreted, even if the first source is false.

These biases can cloud judgment and make disinformation seem more believable or harder to challenge. When content aligns with a number of these biases and is congruent with the context in which it is being experienced, it can be difficult to see it as false (van der Linden, 2023).

Despite these difficulties, French, et al. (2025) proposed a set of seven empirically-informed measures that have promise for mitigating the risks posed by these biases. Most of these methods address more than one cognitive bias.

- **Consider-the-opposite strategy:** Encouraging users to actively consider opposing viewpoints—such as linking to articles with contrary perspectives. This can help reduce anchoring bias by broadening the information considered when evaluating truth claims.
- **Analysis of Competing Hypotheses (ACH):** Presenting multiple, competing explanations for a narrative prompts users to evaluate corresponding evidence rather than relying on how information is framed, thereby mitigating confirmation and framing biases.
- **Opt-in obfuscation:** Requiring users to actively choose to view potentially biased or misleading content, by clicking through a warning, can disrupt automatic processing and reduce confirmation and anchoring biases, encouraging exposure to attitude-opposing information.
- **Dynamic flags:** Unlike static warnings, dynamic flags (e.g., blinking alerts or overlays) that require user interaction have been shown to better attract attention and mitigate overconfidence bias, especially when combined with other cues like ACH or evidence ratings.
- **Evidence rating:** Asking users to rate the credibility of content (without enabling content removal) shifts their focus toward evaluating information on its merits. This can reduce framing bias, particularly when users are prompted to think critically about source trustworthiness.

- **Text visualization:** Using visual tools like word clouds disrupts the linear reading of text and reduces the influence of emotionally charged framing. This can help mitigate framing bias by refocusing attention on content rather than narrative style.
- **Hide virality statistics:** Removing public engagement metrics (likes, shares, views) from news content can reduce herd mentality bias by preventing users from using popularity as a proxy for truth.

All of the preceding methods lend themselves to alterations in the user experience design of news and social media platforms. The consider-the-opposite strategy and ACH also suggest opportunities for awareness training and the development of personal reflective habits that can help consumers of online content engage with these platforms with a greater degree of agency. However, some of the core features of disinformation are difficult to resist. This is particularly the case when content is encountered repeatedly.

Repetition is "Sticky" and Contagious: The Truth Illusory Effect and Message Virality

van der Linden (2023) describes a set of studies that demonstrated that, the more a message is repeated, the more true it feels. This is known as the illusory truth effect[26]. Research has shown that even when there is prior knowledge of a particular topic, this does not by itself protect against false truths (van der Linden, 2023; Fazio, et al., 2015). This is more the case when the content is being echoed by what are seen to be credible sources (McIntyre, 2023), such as media outlets that amplify messages without accompanying critical analysis.

When a message is repeated with a high frequency, delivered with fluency, and/or experienced as emotionally charged, and/or received in the midst of felt pressure, it may be harder still to detect as false[27]. For example, the viral nature of conspiracy theories, and their psychological potency – what van der Linden (2023, p. 49) terms an "evidence-resistant worldview" – make them particularly dangerous. He (2015, 2023) reports that even brief exposure to a conspiracy theory can render people less civic-minded. Moreover, a disposition towards a conspiratorial worldview[28] tends to result in people taking one conspiracy theory as evidence of others, however implausible any of these may be (Biddlestone, Azevedo & van der Linden, 2022; van der Linden, 2015, 2023).

The more the message becomes familiar, and the fewer the opportunities for critical reflection, the more likely it is that the message will be perceived as true. Not only might this influence the thoughts and actions of individual information consumers, to the extent that these individuals experience the information as true, they will participate in its amplification by sharing it casually or intentionally among members of their own networks (e.g., McIntyre, 2023; Ressa, 2022).

---

[26] Hasher, Goldstein & Toppino (1977); Fazio, Brashier, Payne & Marsh (2015); Fazio & Sherry (2020)

[27] An important and powerfully negative antecedent of this insight was Hitler's 'Big Lie Rule' of propaganda which involved the assertion that if a big enough lie is told often enough, most people will come to believe it (for additional detail, see van der Linden, 2023 and Pomerantsev, 2023).

[28] Also know as a "monological belief system" (van de Linden, p. 49)

However, knowledge of disinformation – what it looks like and how it works – is thought to be protective (van der Linden, 2023; McIntyre, 2023; Ressa, 2022). A body of research by van der Linden and colleagues summarized in van der Linden (2020) identifies a consistent set of seven features of conspiracy theories that can be used as a short-hand to identify this form of mis- or disinformation. These are summarized by the mnemonic, **CONSPIRE**:

- **Contradictory:** The narrative contains internal contradictions – it doesn't "hang together" logically;
- **Overriding suspicion:** The narrative expresses suspicion about any official positions about the topic;
- **Nefarious intent:** Sinister intentions are attributed to those who are thought to be the conspirators;
- **Something must be wrong**: Believers might let go of aspects of the story but still insist that "something must be wrong";
- **Persecuted victim**: Believers often view themselves as victims of plots created by powerful elites;
- **Immunity to evidence**: Challenges to the conspiracy story are interpreted as evidence of the conspiracy; and
- **Re-interpreting randomness**: Random events that don't seem to have anything to do with the conspiracy story are interpreted as evidence for the conspiracy, even though another cause of the event is more likely.

Spreading knowledge of the typical format of conspiracy theories, and encouraging practice in using this knowledge to notice and analyze false narratives, may help people build resistance to the insidious effects of repetition.

The Role of Identity and Affiliation Needs

Research on tribalism (Samson, 2023) and affective polarization (Bail, 2021) shows how disinformation takes advantage of strong emotions and people's needs related to identity and belonging to make it harder for society to stay united or agree on shared facts.

One of these levers of manipulation is the human tendency to be attracted to social contexts that preserve or build up self-worth among those who feel that their preferred identities and social position are under threat (Bail, 2021; Pomerantsev, 2023).

The "manosphere" has gained attention in recent years. In a 2014 Washington Post article on the perpetrator of misogynist terror attacks near the University of California, Santa Barbara[29], journalist

[29] Wikipedia (n.d.). 2014 Isla Vista killings. https://en.wikipedia.org/wiki/2014_Isla_Vista_killings

Caitlin Dewey described the it as: "that corner of the Internet where boys will be boys, girls will be objects, and critics will be 'feminists,' 'misandrists' or 'enemies.'" It is a vast network of blogs and forums that promote hyper-masculine ideologies and hostility toward women and feminism. While not all components are violent, the core belief is that feminism has corrupted culture and that men should reclaim dominance by embracing traditional gender roles.

A more recent data-driven investigation of these online spaces by Ribeiro, et al. (2021), based on a taxonomy first developed by Lilly (2016), characterized it as a growing and prospering "conglomerate of web-based misogynist movements focused on 'men's issues'" (p. 196). A core shared belief across these online communities is that "masculinity is under siege by feminizing forces; and feminism is hypocritical and oppressive" (Ribeiro, et al., 2021, p.197). Using Lilly's (2016) taxonomy, Ribeiro, et al. (2021, p.196) described four prominent communities within the manosphere:

- **Men's Rights Activities (MRA):** Advocate for men's issues, arguing that social institutions unfairly disadvantage men. This movement is often characterized as misogynistic.
- **Men Going Their Own Way (MGTOW):** Promote the rejection of relationships with women and mainstream society, rooted in the belief that the system is irredeemably biased against men.
- **Pick Up Artists (PUA):** Teaches men manipulative techniques to attract women, frequently involving objectification, harassment, and a belief that modern masculinity is undermined by female dominance.
- **Involuntary Celibates (Incels):** Mostly young men who bond over feelings of sexual rejection and resentment toward women, often expressing violent or self-destructive ideologies linked to real-world acts of violence.

Examining data across a 14-year period, Ribeiro, et al., (2021) found that older sub-groups (MRA, PUA) had declined in popularity and activity, while newer, more extreme – "toxic" – sub-groups (MGTOW, Incels) were "thriving". They concluded that the manosphere is evolving from what was previously a looser conglomerate of related communities towards a cohesive whole, where people are participating in more than one sub-group. This environment also seems to be fertile ground for the emergence and growth of more extreme sub-communities, connected by their adherence to "Red Pill"[30] ideas (Ribeiro, et al., 2021).

[30] Beliefs, often shared in online communities, claiming to expose hidden truths about society, gender, and power, often in opposition to mainstream values. Borrowed from the 1999 film *The Matrix*, the term originally symbolized awakening to reality. Online it is often linked to misogynistic, anti-feminist, and male supremacist ideologies. In these spaces, "taking the red pill" means rejecting feminism, believing men are oppressed, and embracing rigid gender roles. Common in the manosphere (e.g., MRAs, MGTOW, Incels), red pill rhetoric is often as a gateway to extremist content. Ribeiro (2021) found that, by the end of 2018, the */r/TheRedPill* subreddit ranked third in total posts and fourth in monthly active accounts.

Combined with the addictive design of social media platforms (e.g., Lanier, 2018; Zuboff, 2019), online environments that appear credibly to validate and intensify feelings of injustice and outrage create a ripe environment for disinformation. For example, video games have begun to receive attention for the potential role they may play in exposing young men to radicalizing online communities (e.g., Sorell & Kelsall, 2025; Stuart, 2025).

Where individuals are vulnerable to being influenced by the threat of further perceived losses, or additional social exclusion, being simultaneously being welcomed into a fraternity of fellow 'victims', can be a powerful experience (Bail, 2021; Samson, 2023).

Maté (2022; 2024) described in detail the ways that prior histories of trauma, including harsh or abusive early years, may neurologically predispose certain individuals to the malign influences of radicalization. In particular, Maté suggests that experiences of severe trauma lie at the root of risks for enrolment in extreme authoritarian movements. In addition to offering a refuge from feelings of vulnerability, these movements also invite a sense of belonging for those who harbour grievances related to perceived or real experiences of exclusion, dislocation or marginalization (Maté, 2024).

Echoing the role of context in how online messages are received and interpreted, Bail (2021) argues that what he termed the "social media prism" both reflects the broader social landscape back to users, and distorts what is being seen in ways that may create an altered and misguided form of self-worth. He describes how this distorting, but perversely empowering, experience makes it easier to carry out extreme online actions for those who regularly experience dis-empowerment in their off-line lives.

These destructive online behaviours are ways to signal membership in alternative identity-affirming groups. Samson (2023) suggested that identity-protective cognitive processes play a significant role in shoring up disbelief in truth and belief in conspiracy theories. Bail (2021) proposes two relational processes that scaffold increasingly extreme online behavior: the normalization of extremism as a taken-for-granted feature of one's reference group (and a misapprehension that one's reference group is more of the norm than the exception); and an exaggeration of the extremism of opposing sides. Bail argues that these processes, which make a person's own extremism appear reasonable and that of others seem more extreme, creates a feedback loop that intensify extreme thoughts, feelings and behaviour. Where these loops also shore up identity protective processes, and symbolize membership in status affirming groups, they are powerful barriers to change.

Norman (2021) sought to examine ways that toxic ideologies could take hold to inspire a range of harms based on the notion that "bad ideas are mind parasites" with infectious properties (p. 3). Drawing from a range of research and theory, he suggested that, just as physical stress could weaken the body's immune capacity, so could psychological and cultural stress weaken the mental immunity of individuals and groups to cognitive 'pathogens', such as divisive ideologies. He suggests that, not unlike the ways that human viruses propagate by infecting one person and then another, harmful ideologies are spread from person to person and can be amplified through

technologies. They find fertile ground where there exists a tendency towards belief and an active rejection of invitations to disbelief.

Norman (2021) and Samson (2023) suggest that 'mental immunity' can be developed by cultivating people's innate capacity to detect, filtre and remove bad ideas. It is anchored in a practiced capacity to: evaluate ideas presented to us; an openness to constructive doubt about what we are seeing or being told; and a willingness to revise our opinions. Both Norman and Samson suggest that, when individuals, or cultures, fail to nurture these capacities, the result can be a context that promotes, rather than inhibits, harmful ideas. Samson (2023) identifies the rejection of openness to doubt – or willful belief – as a critical vulnerability that disrupts the "linkage between critical thinking and belief revision" (p. 345). Where a key symbol of membership in an identity-affirming group is "willful unreason" (Samson, 2023, p. 345), group members may be more vulnerable to disinformation that is consistent with group belief systems and more resistant to narratives that invite reasoned alternative accounts.

Psychological Propensity to Ideological "Capture"

A new area of research has begun to explore a significant vulnerability to enrolment in disinformation about gender: a psychological propensity towards ideological thinking. Rather than focusing only on the content of belief systems, this work explores the cognitive bases of ideological thinking itself, suggesting how belief formation and susceptibility to disinformation and conspiracies may interact.

Zmigrod and colleagues (Zmigrod, 2022; Zmigrod, et al., 2023) describe an ideological style as being marked by rigid adherence to doctrine, resistance to updating beliefs in the face of new evidence, and strong loyalty to in-groups, often coupled with hostility toward out-groups. They propose that ideological thinking—regardless of its specific content (e.g., political, religious, or conspiratorial) – shares a common psychological structure. This research suggests those who have a strong need for certainty and a low tolerance for ambiguity, for example, are more likely to seek out belief systems that offer clear-cut answers and structured explanations of the world. Psychological, social-emotional and situational factors appear to play a role.

At the psychological level, these individuals appear more likely to exhibit cognitive rigidity – difficulty updating beliefs when presented with new evidence. They may also show a need for cognitive closure – preferring firm conclusions over uncertainty, which can make them especially receptive to dogmatic ideologies that promise order and clarity.

On the social and emotional level, ideological thinking is often reinforced by experiences that promote strong in-group identity and a sense of belonging. Recalling work done by Bail (2021) and Sampson (2023), individuals who feel a deep emotional connection to a group – whether political, religious, or cultural – may be more likely to embrace ideologies that emphasize loyalty and divide the world into "us" versus "them." Zimgrod and colleagues found that this can be accompanied by hostility or distrust toward out-groups, which further entrenches belief systems and resistance to

alternative perspectives. Additionally, individuals with a high need to belong may gravitate toward ideologies that offer not just explanations, but also community, purpose, and meaning.

Situational factors also play a role. In times of instability – such as economic hardship, social unrest, or high levels of misinformation – people may cling to rigid belief systems as a coping mechanism (Zmigrod, et al., 2023). Ambiguous environments, like social media, can further amplify these tendencies by encouraging quick, emotionally driven responses over critical reflection. Taken together, these findings help explain why some people are more vulnerable to ideologically-charged disinformation, including gendered disinformation exploiting identity, fear, and emotional resonance.

As non-egalitarian gender belief systems tend to be grounded in rigid, binary gender roles (presumed to arise from inherent biological differences), these findings map onto factors fostering gendered disinformation, including tribalism, patriarchy and misogyny, and the role that mis- and disinformation and conspiracy theories play in propagating negative gender-based narratives.

The idea of a common psychological propensity towards the adoption of extreme or rigid belief systems, may offer insights into individual susceptibility to disinformation campaigns targeting gender or identity. Because, these ideologies frequently involve hostile or discriminatory treatment of those who deviate from these normative expectations, Zmigrod's research may provide a window into understanding, and intervening in the face of, risks that could escalate to more serious harms.

Another facet of this research is the finding that when individuals are assessing the reliability of incoming information against their prior beliefs, "noisy" or uncertain information environments – like social media can skew this process, making people more likely to accept false information, especially when it aligns with their existing ideological worldviews.

At a broader, societal level, the concept of "rape culture" has increasingly been used to explain how sexual violence is normalized and accepted within digital spaces (Sugiura & Smith, 2020). This social permission structure and social learning environment encompasses a wide range of gendered norms, behaviors, attitudes, beliefs, values, customs, symbols, language, and practices that tolerate, excuse, or even promote or celebrate sexual aggression (Powell and Sugiura, 2018, cited in Sugiura & Smith, 2020). Sugiura and Smith (2020) suggest that, while the concept originally focused on cis-gender[31] women's experiences in heterosexual contexts, it also provides a valuable framework for understanding the power imbalances underlying sexual violence, abuse, and harassment targeting LGBTQIA+ individuals.

These studies suggest that the way people respond to disinformation is shaped not just by political views or media literacy, but by deeper psychological processes or structures, involving elements of ideological rigidity and in-group identity attachments. These widely shared features may have held

[31] The term, cis-gender, refers to a person whose gender identity matches the sex they were assigned at birth.

evolutionary benefits. Today, individuals susceptible to ideological thinking may be particularly vulnerable to gendered disinformation, which often leverages emotionally charged, identity-based narratives. While the research does not focus on gendered disinformation specifically, it offers insights into who may be most at-risk and underscores the need for future work to explore how cognitive style and trust in information shape susceptibility. These insights could help inform more targeted and effective interventions.

Raising awareness of gendered disinformation (GD) helps people see it as a shared public problem. This understanding is crucial for prevention efforts (Jankowicz et al., 2024). It not only mobilizes attention and action but also encourages public discussion, also involving youth, about possible solutions. From a prevention standpoint, these interventions may work to reverse the normalization of these online practices.

Implications for Countermeasures

Fake truths become more real and more 'sticky' the more they are repeated and amplified by credible sources, especially within a context that is consistent with the content of the message. This is the case, even when the targets of GD have prior knowledge about a topic.

Illusory truth is a particularly powerful vector of disinformation. However, its perceived truth-likeness can be reduced when target audience members are aware of disinformation in general, and the falsity of an individual message (or source) in particular.

Abusive narratives that are embdedded within conspiracy theories which have been amplified across social media platforms may be resistent to redress. This is because un-identified conspiracy theories have a high degree of believability – particularly in certain contexts involving sources that are regarded as highly credible.

Belief systems and norms within groups that promote, enable and celebrate misogyny may be part of online-offline feedback loops that make sexual abuse appear more permissible – in both its online and offline forms – for those who are psychologically and situationally susceptible to influence of these kinds.

Under certain circumstances – such as times of significant uncertainty and "noisy" information ecosystems – individuals who are more likely to embrace ideological thinking may represent opportunitistic targets for gendered disinformation – as recruits to/supporters of an agenda and as spreaders of disinformation. Noisy information environments may stem from the multiple channels that characterize today's information ecosystem, or from deliberate tactics to "flood the [information] zone" with a constant barrage of provocations.

Because the misdirected beliefs, lack of empathy/compassion and the blame attached to the targets of online abuse involve elements of social learning, education and awareness can be seen as important elements of an effective response at the societal level as well as in relation to

opportunities for accountability, redress and rehabilitation by perpetrators (e.g., McIntyre, 2023; Cuklanz, 2023).

Public awareness, including campaigns by civil society actors to engage younger audiences and prospective male allies can play an important role in mainstreaming attention to, and action against, TF-GBV.

Social Media Literacy

The National Democratic Institute has proposed providing women with training to reduce online threats. This includes protecting personal information, using social media tools to minimize harassment, and strategies for maintaining mental health and resilience against online abuse (Jankowicz, et al., 2021). Literacy about the hazards of GD is discussed below in the section on mental immunity.

Ressa (2022) provided close-quarter insights into the design features of social media platforms (as described above). Knowledge of the ways that platform characteristic like the ease of re-posting content can contribute to its spread can help users be more prepared to resist these 'virality-by-design' features.

Platform literacy also involves understanding the broader factors at play. Ressa's (2022) experience sheds light on how platform business models can intersect with political and ideological agendas to target and harm women seen as threats to populist movements. This helps us better appreciate the challenges to constructive change. Addressing these issues requires active efforts from journalists, activists, civil society groups, government, and businesses that benefit from societal stability.

Implications for Countermeasures

Knowledge and training on the safe, informed, use of online platforms is advised to lower the risk of inadvertently falling prey to GD. This should be accompanied active messaging that online GD is not simply a personal problem or the result of the actions of those who have been impacted.

At the same time, harms flowing from some of the design features and macro-level dynamics of social media platforms and their business incentives cannot be mitigated by individual behaviour, alone. Addressing the broader risk environment will require multi-level, multi-modal action, with roles for government, civil society and businesses, as well as an informed citizenry.

Debunking: Exposure to Truths and the Viewpoints of Others

A fundamental threat – and objective – of gendered discrimination, and disinformation more generally, is the incitement of contempt for those who are defined as 'other'.

Nearly a century of social psychological research supports the idea that appropriate interpersonal contact[32] between diverse groups can improve relations by reducing perceived differences. However, this is effective only under certain conditions – that groups: share similar status and backgrounds; work cooperatively towards a common goal; and interact in a context that promotes positive cooperation and discourages division[33].

Online gendered disinformation lacks the situational features necessary for positive group interaction. However, research on the contact hypothesis helps us understand the role of situational determinants in behaviour. It suggests that factual corrections can counteract false beliefs arising from disinformation. This process, known as debunking, involves exposing and correcting false or misleading information to clarify the truth.

Debunking can be useful, but its effectiveness is limited. Research shows that misinformation tends to persist even after is corrected or withdrawn. This "continued influence effect" may occur because people tend to avoid the emotional cost of changing previously held beliefs (Susmann & Wegener, 2022).

Lewandowsky, et al. (2020) determined that effective refutations must clearly explain why information is false, and present the truth. Simply refuting a false fact is insufficient. Providing a credible alternative explanation or questioning the source's credibility can also be effective (Lewandowsky & van der Linden, 2021).

The source of the information is important. Debunking messages should come from individuals or organizations perceived as trustworthy by the audience (Lewandowsky, et al., 2020). However, if recipients ignore the source, its characteristics will have negligible effect.

If disinformation leads to misunderstandings about an issue, person or group, one might think that corrective narrative could help. However, this approach is often overly optimistic, as exposing people to contradictory information may actually reinforce their original beliefs.(Ecker, et al., 2020).

In an important study on debunking politically polarizing narratives, Bail, et al. (2018) conducted a field experiment with US Democrats and Republicans on Twitter. Participants were surveyed on policy positions and then exposed to periodic political content from the opposite party via bots. Contrary to expectations, instead of moderating initial views, exposure led to more polarized positions, especially among Republicans. Democrats demonstrated a slight, non-significant increase in liberal leanings. Bail, et al. (2018) suggested that, in light of the strong evidence from

[32] Widely known as the contact hypothesis, these ideas were articulated by Allport in 1954 as well as by Sherif and Sherif in the same year.

[33] Samson (2023); https://en.wikipedia.org/wiki/Contact_hypothesis

previous studies that inter-group contact can foster compromise and mutual understanding, future efforts to reduce political polarization on social media will likely need to focus on identifying the types of content or the positioning of messages that are prone to backfire, and whether other approaches and sources of information might be more effective.

Boukes and Hameleers (2023) explored the effectiveness of satire-based fact-checks as an alternative to traditional methods. They found that regular fact-based content reduced the perceived accuracy and credibility of false information, avoiding a backfire effect. By contrast, satire-based fact checks were found to be effective regardless of prior agreement with the fact-checked information. The researchers suggested that this may mean that refutations based on satire may be less vulnerable to resistance[34] and confirmation biases – possibly because of the cognitive effort required to participate in the narrative invoked by the satirical message.

However, neither approach decreased polarization based on commitment to group membership (in this study, political attitudes) and the gap between in-group like and out-group dislike (affective polarization[35]). Moreover, the use of satire was found to make it more likely that polarization would increase, whereas this was observed for regular debunking messages only when people saw the content of the refutations as confirming their existing views.

Implications for Countermeasures

Truth-restoring narratives and exposure to the views of others, by themselves, may be ineffective in counteracting disinformation. This is especially true – as is the case with GD – where one of the parties to the interaction does not perceive a sufficient degree of similarity with the target(s) of their attacks, where there is pre-existing polarization and when the context of the interaction favours polarization and conflict over harmony and cooperation.

There is emerging research suggesting that attempting to address online polarization simply by providing factual corrections to disinformation may actually exacerbate the problem through the production of backfire effects.

To stand a chance of effectively 'unsticking' disinformation narratives, debunking narratives must clearly articulate the ways that the narrative is incorrect. The counter narratives should lay out the details of the truth about the matter and be delivered by credible, trusted sources. However, the

[34] The case of resistance to belief is an interesting one. Individuals who prioritize self-direction, a human value that encourages independent thinking and actions, generally exhibit a higher need for cognition, whereas those who prioritize conformity, a value focused on maintaining the status quo, typically demonstrate a lower need for cognition (Coelho, Hanel & Wolf, 2020, cited in Kakinohana & Pilati, 2023). Thus, a disposition toward more thinking about the content of a message and its accuracy might be impacted by the cognitive energetic costs of considering the details of a satirical refutation.

[35] Boukes & Hameleers (2023); Bail (2021)

continued influence effect may render this unsuccessful. This is also the case when the content of fact-check messages are perceived as confirming existing views.

Satire-based fact-checks may be more broadly effective in reducing the perceived accuracy of a message and the credibility of the source. However, they may increase, rather than decrease affective polarization, either because it is seen as a critique of one's ideological identity or because of the cognitive load it imposes on people who might otherwise be able to process a message as false.

Upstream measures stand to be more effective than a more downstream response like debunking.

Forewarning and Prebunking: Psychological Inoculation to Disinformation

A promising for of upstream intervention is known as psychological inoculation. This can involve one or both of two components: (1) forewarning, which involves an advance caution to prepare recipients for the threat and motivate resistance; and (2) prebunking, a pre-emptive refutation of the anticipated message.

The continued influence effect causes false information to persist in memory even after convincing correction, as referencing the initial disinformation can reinforce its frame (Lewandowsky, et al., 2020). van der Linden (2021) proposed that instead of post-event debunking, using active-listening and focusing on misinformation techniques can be more effective. This approach leverages people's interest in avoiding manipulation.

A half-century of research on inoculation theory highlights techniques for building resistance to unwanted influence through protective exposure (McGuire, 1970, cited in Lewandowsky & van der Linden, 2021). Lewandowsky and van der Linden (2021) explain that by pre-emptively exposing individuals to a weakened form of manipulation, a cognitive-motivational process, similar to creating "mental antibodies," is triggered, enhancing resistance to future persuasion attempts.

Resistance is thought to be based in a mental default disposition to safeguard existing beliefs in the face of contradictory information. More recent scholarship on inoculation theory – emphasizing the virality of social media content – has expanded attention from narrow-spectrum, issue-specific, arguments toward a broader perspective including general influence and manipulation, as well as both active and passive approaches (Lewandowsky & van der Linden, 2021).

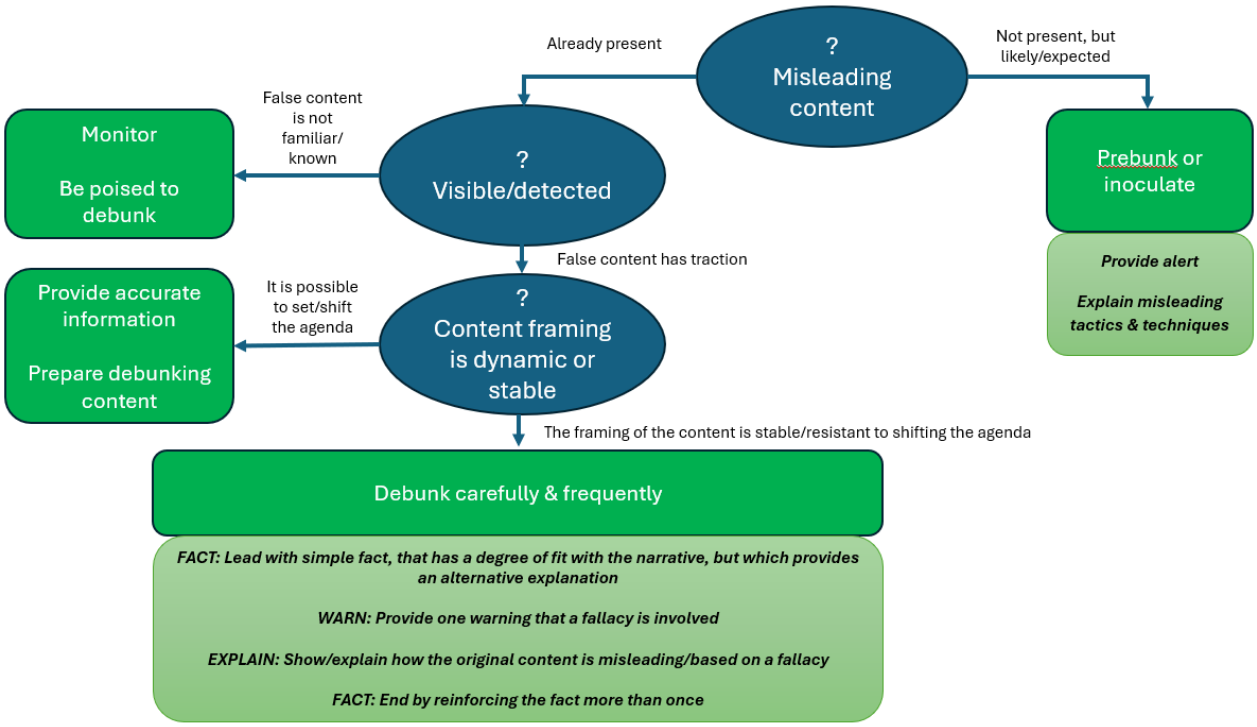
In contrast to the difficulty of counteracting conspiracy theories retrospectively through debunking, research shows that providing people with anti-conspiratorial content – either fact-based or logic-based – that foreshadows conspiracy theorist arguments can be effective (Lewandowsky & van der Linden, 2021). These techniques may also be effective against rhetoric used by online ideological extremists to radicalize prospective adherents to their cause.

A series of studies conducted by van der Linden and colleagues, summarized by van der Linden (2023) and Lewandowsky and van der Linden (2021) demonstrated that forewarning prospective target audiences about techniques of disinformation could neutralize the influence of false messages. This work also conformed to emerging evidence of an association between political ideology and susceptibility to untrue claims.

Research shows that individuals on the populist right may be more susceptible to disinformation and conspiracy theories than those on the ideological left (van der Linden, Panagopoulos, Azevedo, & Jost, 2020) – although both sides are susceptible under certain circumstances. These findings resemble those of Zmigrod's (2022) work on susceptibility to ideological thinking. They also reflect some of Bail's (2021) observations that ideologically conservative individuals may polarize further when exposed to opposing political content. Encouragingly, inoculation interventions can effectively moderate these impacts across target audiences.

Lewandowsky, et al. (2020) proposed a decision tree for determining when to deploy debunking or prebunking countermeasures (Figure 7). While this model focused on *mis*information, it has promise for misleading content more generally, including *dis*information. A more general application of this approach remains to be evaluated.

Figure 7. Decision tree for use of countermeasures (after model proposed by Lewandowsky, et al., 2020).



The key features of this decision tree recognize that prebunking has been found to be more effective than debunking. Consequently, debunking should be used only when necessary – that is, when false content has begun to gain traction and can no longer be ignored. In those cases, it must be done in structured way, and frequently, so that it has the best chance of competing with and displacing the false content through the power of repetition.

Using this formula, a **debunking** message might take the following form:

- **DISINFORMATION CLAIM THAT HAS STARTED TO GAIN TRACTION:** "Women aren't suited for leadership roles because they're too emotional to make rational decisions."
- **FACT:** Women are just as capable as men in leadership roles. Research shows that gender doesn't determine someone's ability to make sound decisions or lead effectively. In fact, many of the world's top-performing leaders are women.
- **WARNING:** Be careful—this kind of claim is based on a harmful stereotype, not facts.
- **EXPLAIN:** The idea that women are "too emotional" to lead is an outdated myth. It plays on old gender stereotypes and ignores real evidence. Emotional intelligence is actually a strength in leadership. Good leaders use both reason and empathy to make smart decisions. Misinformation like this can discourage women from taking on leadership roles and keeps unfair biases alive.
- **FACT (AGAIN):** There's no evidence that women are less effective leaders. Studies show that women perform equally well—or better—than men in leadership positions, across all sectors.

In the case of anticipated disinformation, based on the same false claim, a **prebunk** could include the following:

- **ANTICIPATED DISINFORMATION CLAIM TO GET AHEAD OF:** "Women aren't suited for leadership roles because they're too emotional to make rational decisions."
- **ALERT:** Heads-up! Be careful when you hear people say that women aren't good leaders because they're "too emotional." That's a stereotype designed to discredit women, not a fact.
- **EXPLANATION OF MISLEADING TACTICS:** What's really going on here – This kind of message uses an old stereotype to make women seem less capable. It wrongly suggests that showing emotion is a weakness, when actually, being in touch with emotions can help leaders connect, communicate, and make better decisions. These claims are meant to make people doubt women's abilities and stop them from taking on leadership roles.

To address the growing threat of online mis- and disinformation, Roozenbeek and van der Linden (2019) developed an innovative browser-based game called *Bad News*. The game, which is grounded in inoculation theory, was designed to build cognitive resistance to future disinformation attempts.

Roozenbeek and van der Linden (2019) observed significant improvements in the ability of participants to identify disinformation tactics after playing the game, with the strongest effects among those who were most susceptible to fake news before play started. This improvement was found across a range of demographic categories, including age, education level, political orientation, and gender.

Roozenbeek and van der Linden (2019) concluded that game-based inoculation can offer a "broad-spectrum psychological vaccine" without simply increasing overall skepticism. The latter point is important because one risk of awareness building around disinformation is that people will begin to see disinformation everywhere – leading to a generalized weakening of trust in the information environment and a greater likelihood of "false positives" – assuming that something is false when it is not.

Following the success of *Bad News*, the researchers developed several additional games, focusing on specific disinformation risks. Yet, despite the promise of using gamified approaches to inoculate people against disinformation, a degree of caution is warranted.

Using electronic games to build resistance to disinformation

Bad News is a short, interactive, simulation where players assume the role of a fake news producer. Over approximately 15 minutes, players work to gain followers and credibility by learning and applying six common misinformation tactics: impersonation; emotional manipulation; group polarization; conspiracy theory creation; discrediting sources; and trolling and baiting. Players earn badges for successfully applying these tactics in realistic social media-like scenarios. Ethical behavior is penalized in the game, reinforcing awareness of deceptive practices. The game was launched in partnership with a media outlet and was accessed by tens of thousands of users globally. A subset of over 14,000 participants completed pre- and post-game surveys to assess changes in their ability to recognize misinformation strategies.

Kiili, et al. (2024) conducted a systematic review of research, published between 2019 and 2021, focusing on the use of game-based and gamified learning environments designed to build skills for detecting false information. They identified that most of the 15 studies that matched their inclusion criteria reported positive outcomes for the interventions. However, they discovered considerable variation in what was measured and in the research designs used to assess effectiveness. They also observed that there is currently no standardized framework for describing and comparing across these types of techniques. As a

result, Kiili, et al. (2024) concluded that, not withstanding promising initial results, it is not yet possible to draw general conclusions about the effectiveness of these types of game-based interventions.

Maertens, et al. (2025) conducted a series of longitudinal experiments (with a total of 11,759 participants) to explore the persistence of misinformation resilience over time, following various inoculation interventions. They found that, while text-based and video-based interventions remained effective for up to one month, the effectiveness of game-based interventions – where the acquired skills may be cognitively more demanding to retain – decayed much more quickly. Importantly, they observed that booster interventions that are designed to enhance memory and recall of earlier learnings helped to offset the loss of effectiveness of all forms of interventions. The design of future research and interventions based on this work will likely focus on ways of enabling boosters to be designed and delivered at a pace conducive to sustaining the effective of counter-misinformation narratives. While this would likely be logistically complicated, advances in artificial intelligence – and potentially new features of social media platforms – could make this easier to achieve.

Implications for Countermeasures

Psychological inoculation is one of the most effective countermeasures against disinformation. It also avoids the risk of backfire effects known to be associated with debunking.

Moreover, forewarning about specific techniques of manipulation used in the context of certain topics, along with the refutation of anticipated messages (prebunking) appears to be effective in reducing differences in susceptibility to disinformation tied to divergent political ideologies.

Inoculation techniques may also help to lower the risk of radicalization to extremist ideological movements and so may represent an important upstream public health/public safety opportunity that benefits both prospective victims and prospective perpetrators.

Concepts of “mind parasites” and the mental immune system offer insights into the work that willful belief, distrust and cynicism do in signalling membership in identity-affirming groups. These ideas also underline the importance of broader societal efforts to provide persons vulnerable to radicalization with offramps towards more prosocial identities and their associated symbols and behaviours.

Inoculation measures are a natural fit with awareness and education campaigns and therefore, may be combined as a single package of upstream/prophylactic interventions. Gamified inoculations with booster interventions hold promise as a way to reach populations not always amenable to digital public health programs. However, more research will need to be done, including evaluation, in order to strengthen the empirical base for these efforts.

Contending with these phenomena will be challenging. The creation and propagation of mental immunity will require actions on many fronts to invite new experiences of belonging, and to supplant the symbolic and signalling systems of extremist groups with those of more moderate or prosocial coalitions.

Policy and Regulation: Potential Areas of Focus

While not, strictly speaking, countermeasures, attention to policy and regulatory options may be crucial to re-shaping the broader socio-cultural context towards more inclusive and less polarizing outcomes. Public concern about the prevalence of toxic content on social media has led to growing pressure on platforms to ensure more effective and accountable moderation (Sobieraj, 2020; Richardson-Self, 2021).

In the spyware industry, efforts to hold parties accountable, and to investigate or regulate them, can be challenging, due to complex and shifting ownership structures and corporate relationships (Marczak, et al., 2021). Citizen Lab researchers observed that many of these techniques are similar to those used by arms traffickers and money launderers (Marczak, et al., 2021).

Richardson-Self (2021) suggests enforcing clear standards and guidelines, to define speech identified as hate or abuse, and to require digital platforms to allocate resources to identify harmful content. However, the owners of major platforms such as Telegram and X are alleged to have resisted efforts to increase oversight (Mozur, et al., 2024). Richardson-Self (2021) also suggests user fees to slow the spread of harmful information, but notes this might simply drive users to other sites.

Ermoshina and Musiani (2025) propose implementing measures to ensure safer online spaces by design, analogous to Cavoukian's (2010) concept of privacy by design. Privacy by design encourages the view that privacy ought to be a core component of fair (and, ultimately, more effective) information practices – and essential to the functioning of democratic societies. Embodying seven foundational principles, privacy by design covers three main sets of applications: IT systems; accountable business practices; and physical design (Cavoukian, 2020).

Ermoshina and Musiani's (2025) "federated" model of content moderation offers an alternative to the top-down approaches used by major social media platforms. Instead of a single company setting the rules, this model is built on a network of independently run communities—each with its own moderation policies and user guidelines. Platforms, like Mastodon and Matrix, that use this approach, seek to allow communities to tailor their rules to local values and needs. Users can choose or move between communities that reflect their preferences, giving them more control over their online experience.

Ermoshina and Musiani (2025) suggest that such a decentralized model would support safer online spaces by encouraging moderation that is responsive, community-driven, and transparent. They

argue that it would also reduce the risk of one-size-fits-all policies and give people more say in how harmful content is handled.

While the federated model introduces models for user-centred, ethical, and flexible content moderation, it would also require ongoing investment in technical infrastructure, community participation, and shared responsibility to achieve its vision of supporting safer spaces. However, Ermoshina & Musiani (2025) suggest that, by promoting user choice and rejecting the profit-driven motives of large platforms, the federated model represents a promising path toward more ethical, inclusive, and accountable online environments.

Matthews (2021) assessed four main approaches to online content moderation drawing from private, community and legal models.

- **Legislation or government-led content moderation:** This top-down model involves the government defining what content must be moderated and by whom. It can take various forms, such as holding users accountable, tasking social media platforms with content removal, or establishing independent regulators. Legislation offers clarity and enforceability but may lack the flexibility to keep pace with technological change. Germany's Network Enforcement Act exemplifies this approach, placing the onus on platforms to remove illegal content within 24 hours, though it has faced criticism for encouraging over-censorship and giving too much power to private companies. Canada's proposed Online Harms Act (Bill C-63) draws on similar principles, though it is at a standstill.
- **Social network-led content moderation:** In this approach, social media companies take primary responsibility for moderating content, often driven by legal mandates or public pressure. While companies can tailor moderation to fit their platforms, critics argue that private sector control over speech poses risks to democratic discourse and transparency. There are also concerns about limited investment in moderation, ethics-washing, and monopolistic control due to the dominance of a few major platforms. Without transparency or oversight, users lack recourse when moderation decisions are made.
- **Third-party moderation tools:** This model promotes decentralization by enabling independent developers to build moderation tools that integrate with social media platforms—much like app stores in the tech industry. These tools offer users more choice and control over their online experiences and encourage competition. However, they face challenges such as unclear business models, privacy risks, and potential resistance from platforms. Nonetheless, Matthews (2021) concluded that tools like Block Party (for Twitter/X) demonstrate the potential for user-driven content control, particularly for communities most affected by online harassment.
- **Community-led moderation:** Community-led moderation is a bottom-up, pluralistic, competition-based approach where users set and enforce their own rules, often supported by platform tools and automation. Reddit is a leading example, empowering subreddit communities to self-govern within broad content guidelines. This model increases user

agency and diversity of experience but relies heavily on volunteer labor, raising concerns about sustainability, consistency, and the capacity for effective enforcement. Public responsibility on platforms like Reddit have not proven effective in preventing the development of highly misogynist online communities.

Lalonde, et al. (2025) argue that to improve transparency and accountability, holistic[36] and consistent platform policies should align with a practical regulatory regime than on corporate priorities. However, they are less optimistic about content moderation, as social media business models often hinder effective responses[37] to inappropriate content. Lalonde, et al.'s (2025) analysis of legal and policy responses to VMD – which aligns to the problem of gendered disinformation – is equally concerning. They conclude that, with growing technological sophistication, governments and international bodies are straining with how to regulate it effectively while respecting rights and adapting to evolving technologies.

For example, they point to international efforts by UNESCO and the UN to established non-binding principles to promote ethical AI use and digital platform governance which, however admirable, lack enforcement power.

Legislating against disinformation: A delicate balance

Addressing disinformation through legislation requires a careful balance – ensuring harmful content is identified and limited, while safeguarding forms of expression, such as satire, that play a legitimate and sometimes important role in exposing and challenging false narratives. In some cases, legislation may be used to suppress efforts to expose political agendas and activities which may, themselves, include inaccurate or misleading information.

To illustrate this challenge*, the Texas legislature recently passed a bill that would make it a crime to share altered political media – such as memes, videos, or audio – a unless it includes a government-approved disclaimer. Though originally intended to target AI-generated deepfakes, the legislation (House Bill 366) was expanded to cover any manipulated content that "did not occur in reality," including simple edits and parody. Despite recent amendments, the bill has drawn strong criticism in the US from First Amendment advocates, who argue it is overly broad and vague, potentially chilling political speech and satire. Questions remain about how the law would be applied.

Source: Waltens, B. (2025). Texas house approves former speaker Dade Phelan's meme regulation bill. *Texas Scorecard*, April 30, 2025. https://texasscorecard.com/state/texas-house-approves-former-speaker-dade-phelans-meme-regulation-bill/.

They assess that the European Union has taken the most proactive stance: the *Digital Services Act (DSA)* mandates risk assessments and algorithmic transparency by major platforms, while the *AI*

[36] For example, recognizing the presence and harms of both high-tech deepfakes and lower-tech "cheapfakes", and addressing policies to include a broader level of technological sophistication.

[37] Lalonde, et al. (2025) identify: removal – simple deletion of content; downranking – reducing content visibility by deprioritizing its position in search results and feeds; and demonetization – delinking online content from revenue generation.

Act requires clear labeling of synthetic content. In contrast, U.S. regulation remains fragmented, with states like California enacting laws targeting political deepfakes, but no unified federal framework exists. Domestically, Canada introduced the *Online Harms Act* (Bill C-63) to address AI-generated harms, especially to minors, but the bill stalled in early 2025, leaving a legal gap.

Lalonde, et al. (2025) conclude that, despite progress, there are significant challenges: legal fragmentation across jurisdictions weakens enforcement; the pace of technological change outpaces public policy and regulation; and the need to balance regulatory action with freedom of expression continues to pose dilemmas in democratic societies.

A 2018 report by the Public Policy Forum considered how to contend with the threats to democracy of harmful speech online (Tenove, et al., 2018). A key concern was that current regulations cannot tackle the massive and fast spread of harmful content across social media. One explanation offered is that foreign-owned platforms severely limit Canadians' ability to influence or oversee platform accountability. This creates a pronounced imbalance between the risks and the means available to Canadians to address them (Tenove, et al., 2018).

To help address this problem, the white paper outlined three interconnected public policy recommendations tailored to the Canadian context (Tenove, et al., 2018):

- **Adopt a multi-track framework for harmful speech regulation**
  A coordinated, multi-agency approach is needed to clarify how current laws can better address harmful online speech. This includes establishing a multi-agency task force, requiring social media companies to share data on harmful content with the public and researchers, and launching a multi-stakeholder commission to explore broader social and political impacts—fostering public dialogue on the future of content moderation and oversight.
- **Establish a moderation standards council**
  A new independent council—modeled after the Canadian Broadcast Standards Council—should be created to bring together platforms, civil society, and regulators. The Council would support transparent content moderation, develop and enforce codes of conduct, manage public complaints, and address regulatory conflicts, while contributing to international standards for online content governance.
- **Strengthen civil society and research capacity**
  Canada should significantly invest in research, programs, and civil society initiatives focused on harmful speech. Governments, academic institutions, and tech companies should support this work.

These recommendations are designed to work in a mutually supportive way to foster a healthier, more inclusive, and democratic digital public sphere in Canada. To the extent that they would be able to achieve this vision, each one would need to realize its full potential.

A more community-based approach involves the idea of coordinated acts of 'counterspeech' – speaking back against actions and systems that oppress people (Richardson-Self, 2021). The essence of this approach is to encourage collective action against harmful conditions and behaviours and greater accuracy by confronting biases and false assertions directly. This could take the form of refuting an inaccurate message and/or questioning the credibility of the source. In these ways, counterspeech is similar to notions of debunking discussed previously. Aligned to concerns identified earlier, Richardson-Self cites research by Costello and Hawdon (2020) that suggests that confronting hateful actors may only serve to amplify hateful rhetorics and their narratives.

Surveying the online regulatory landscape, Jankowicz, et al. (2024) assessed that not enough has been done by governments to mitigate online harms – either through incentives and requirements related to oversight, transparency and moderation, or through legislated responses such as criminalizing deepfake image based sexual abuse. Jankowicz, et al. (2024) offered eight recommendations addressing platform accountability and action and to address deepfake image-based sexual abuse. These recommendations covered the following areas:

- Government oversight of platforms to encourage improve duty of care related to women's ability to express themselves safely online;
- Transparency and oversight mechanisms enabling access by journalists and researchers to social media data, in the service of public interest, with appropriate privacy safeguards;
- Explicit provisions within online safety legislation and regulations to address online harms against women;
- Encouraging technology companies to address gender imbalances within their workforces;
- Institution of civil and criminal penalties for the creation and distribution of non-consensual deepfake pornography;
- Measures to interdict the availability and facility by which search engines websites and applications focused on the creation and distributions of deepfake pornography are used to harm women and children;
- Widening the availability of technologies that can be used to challenge deepfakes and protect original images from being misued ("immunizing images", digital "watermarks"); and
- Supporting public awareness campaigns and educational resources aimed at challenging deepfakes and remediating harms that have occurred.

An additional challenge concerns the difficulty of tracing an image back to the original upload. Robust and reliable technologies supporting correct attributions would be useful contibutors to both deterrence and accountability.

Implications for Countermeasures

Improved standards and guidelines are one element of a spectrum of higher-level responses to technology-facilitated harms against women and girls. However, these must be clear and practical, they must be properly resourced and implemented, and they must be transparent and enforceable through effective, accurate, reporting methods.

Legislation and regulations have been proposed to address the harms that flow from irresponsible or inadequately governed platforms. These may include civil and criminal penalties for non-compliance and for harms that stem from a lack of reasonable action by platform owners.

Calls for public funding for awareness and education about social media and gendered violence, including the use of deepfakes as methods of sexual abuse and exploitation, are consistent with the value of fostering awareness identified earlier.

Content moderation is perhaps the most broadly familiar measure in the public mind. Although there is likely a place for improved moderation, foreign ownership of social media platforms, deeper platform design features enabling virality, and business imperatives may lessen the impact of these types of interventions outside of coordinated, global, action by transnational coalitions.

Developing a capacity for reliable and valid attribution would play an important role in deterrence.

Support for Those Affected by Gendered Disinformation

Women involved in politics – especially women of colour – face repeated and ongoing online violence (Sobieraj, 2020). The National Democratic Institute has urged that social media platforms should have specific contacts to whom reports of online abuse could be escalated (Jankowicz, et al., 2021).

Jankowicz, et al. (2024) recommend that, in addition to investing in public awareness campaigns and the development of educational resources about technology-facilitated gender-based violence, governments should also support capacity building within the police and justice communities to enable them to be able to respond more effectively to enforcement situations and community building opportunities. Similarly, they recommend that schools and employers have policies and supports in place for students and staff who experience TF-GBV.

Sobieraj (2020) emphasizes that this is a fundamentally anti-social and anti-democratic phenomenon. Consequently, collective action and a network of support will be necessary to help individuals recover from harms that have occurred and to experience opportunities for resilience in the midst of ongoing threats. More importantly, to create lasting change that inoculates society against the threats of online misogyny, macro-level attention, joined-up action, and more active roles for governments, community-based organizations and digital platforms – in keeping with concerns related to privacy, free-speech and due process – will be necessary.

To bring the focus further upstream, it will also be necessary to address structural, socio-economic and other macro-features of our communities and society that create conditions of risk. These include developmental traumas and experiences impacting boys and young men, which constitute the conditions of exclusion and despair that are grist for the narrative mill of misogynistic authoritarian movements, as suggested earlier.

Implications for Countermeasures

Services and supports for those who have been victimized, or are recovering from, technology-facilitated gender based violence are an important component of a holistic and shared response to this problem.

This can be enhanced by training for the justice, educational and community sectors focusing on strengthening their individual and collective capacities to prevent and intervene in the aftermath of online abuse.

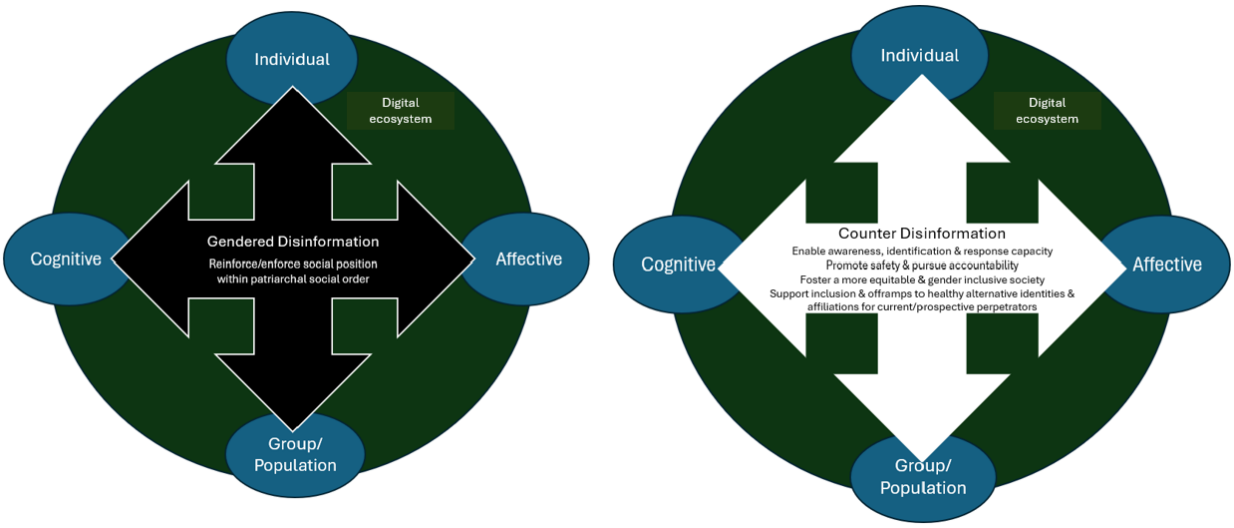
In addition to downstream supports, mid-stream and upstream measure focusing on enhancing conditions that community safety and wellbeing, will serve better developmental outcomes for children, and make communities lower in social determinants of risk and richer in social determinants of wellbeing.

A Strategy for Change

Mindful of the foregoing research, concerns and caveats about the scope and complexity of gendered disinformation, it is important to start with an eye to building momentum and fostering networked capacity for contending with this problem. A strategic mix of countermeasures that span the upstream-midstream-downstream continuum would help shape conditions that are: more resistant to misogyny and disinformation, less conducive and more responsive to technology-facilitated violence against women (whether perpetrated as individual acts of misogyny or as tools of foreign interference), and more supportive of the resilience and recovery of those targeted by GD.

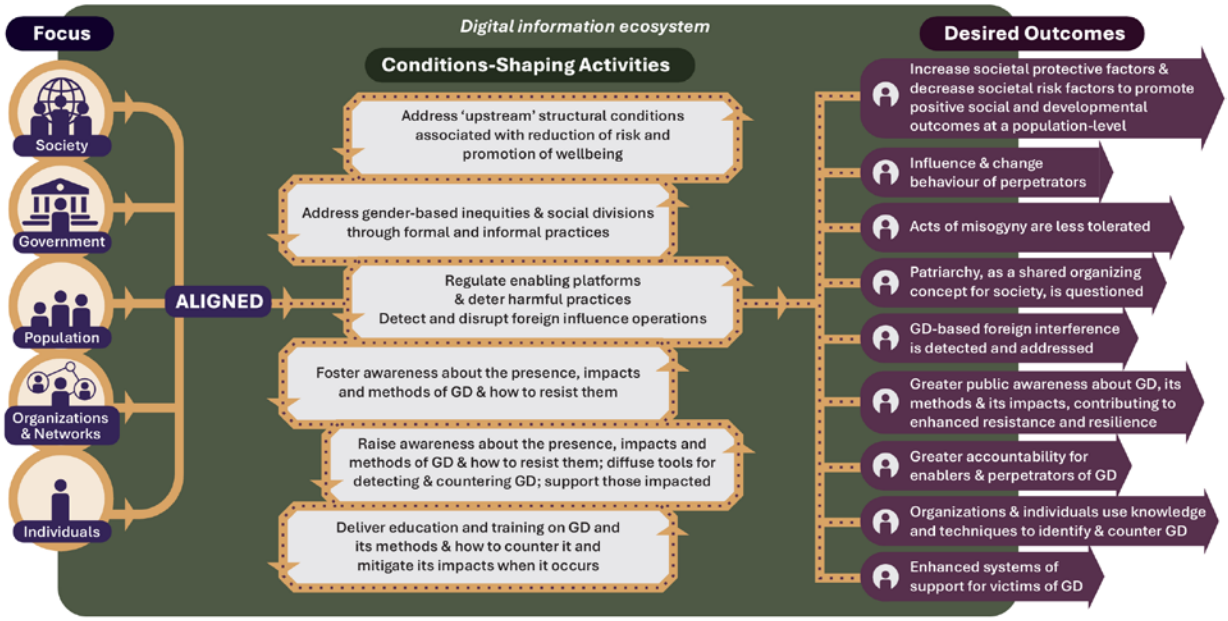
A suite of effective countermeasures should start with attention to awareness, providing skills, tools and opportunities for support to those who have been affected, and work along the length of the intervention continuum (Figure 8).

Figure 8. Focus of gendered disinformation and counter disinformation activities.



Gendered disinformation occurs within a broad and varied socio-cultural context. A corresponding theory of change for addressing GD as a holistic, all of society problem, is shown below (Figure 9). It involves efforts at multiple levels that, if aligned, would create a set of mutually reinforcing conditions that increase the probability of realizing a constellation of desired outcomes.

Figure 9. Preliminary theory of change for addressing gendered disinformation holistically as an all-of-society problem.

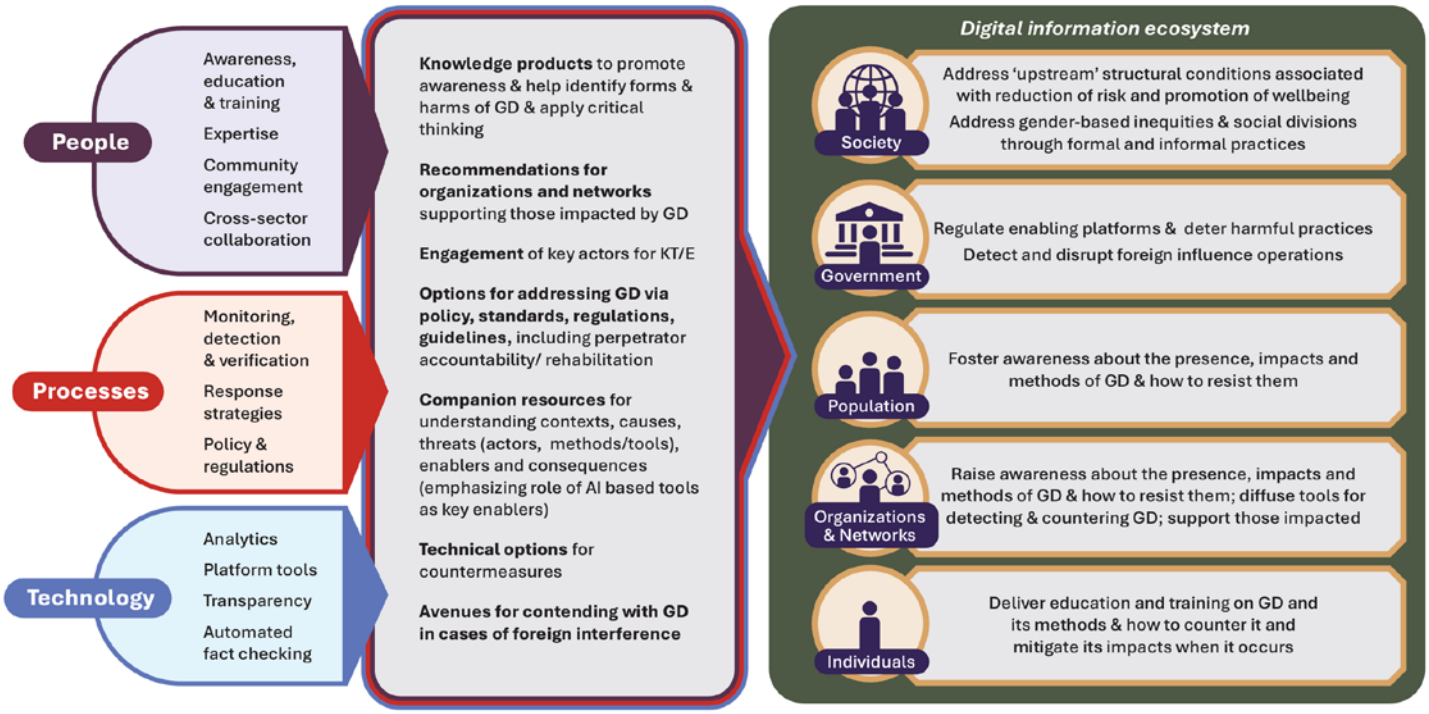


At this juncture, the most promising avenues appear to be those that address: awareness and identification; response capacity; support and empowerment; and policy engagement. The components of this multi-level approach emerge as a basis for achieving initial traction against what has seemed to be an intractable problem. On the basis of the preceding discussion, several considerations stand out:

- Awareness-based interventions are critical for building resistance to GD and resilience enhancing networks of support and accountability.
- Tapping into knowledge-building and training opportunities within the educational and human service systems may help to reduce participation in, and victimization by, malicious information exploits among young people and foster improved response capacity among educators, police, public health and community partners. Awareness building and training involving police and community partners might build on work already underway on the related topics of intimate partner violence and coercive control (e.g., Gill, et al., 2021). Awareness and educational opportunities for school-aged populations might be incorporated into a range of curricular learnings touching on information technology, AI and cybersafety (e.g., social sciences, humanities, computer science) and extracurricular activities focusing on women in science, technology, engineering and math (STEM).
- A focus on gender as a tool of foreign interference and authoritarianism may help the policy, national security and law enforcement communities in their efforts to identify and counter FIMI exploits as well as domestic ideological movements engaging in stochastic terror practices.
- Highlighting how social media platforms help spread GD can increase public understanding and support for policies that balance risk mitigation with freedom of expression.
- Wherever feasible and appropriate, interventions should be tailored to local contexts, considering factors such as literacy, access to technology, and existing gender norms.
- It is not enough to focus on individuals alone as this is not only an individual trouble; it is a collective threat.
- Interventions at the individual level should be combined with organizational, network, technological and policy solutions for maximum effectiveness, as part of a comprehensive approach to combatting gendered disinformation.

A corresponding system for countering gendered disinformation, consisting of people, processes and technology, is summarized below (Figure 10).

Figure 10. Counter GD system of people, processes and technology.



Factors considered in developing this system – which should be considered in its implementation, include the following.

- **Practical utility:** The system should address the widest array of GD threat scenarios.
- **Perceived relevance and value:** The system should address perceived needs across a broad spectrum of users (organizations/agencies, government, individuals) and should serve as a basis for engaging prospective users on needs that are real but, not yet, experienced.
- **Adaptability and maintenance:** The system should allow for flexibility in use and should be able to be updated as new information becomes available (e.g. threat sharing) and/or as threats and the enabling technologies continue to evolve.
- **Capability and cost:** The system should be accessible to a wide range of users, with advanced users able to gain more benefits than those with less technical expertise.

A detailed overview of this system is provided in Appendix C. Appendix D provides a curated set of sample technologies that might be useful to individuals, and human service and educational organizations in identifying and countering potential instances of gendered disinformation.

Appendix E includes a list of the set of accompanying knowledge resources to support awareness and actions among: educators (Annex E1); families and youth (Annex E2a, E2b); a set of additional resources for educators, families and youth (Annex E2c); police and community partner agencies (Annex E3); and government stakeholders (Annex E4). These resources are contained in the companion document, *Understanding and Countering Gendered Disinformation: Knowledge Resources*.

Included within Annex E4 are a set of recommendation to support the development of an expanded national capacity for contending with gendered disinformation. These recommendations follow.

CONCLUSION

Gendered information is a complex issue linked to polarization, patriarchy and misogyny – driven by individuals, groups and nation states. It targets women, girls and gender non-conforming persons, causing harm as victims or tools of repression. No single approach can counter these threats to safety and national security. Therefore, countering gendered disinformation requires a multi-layered, strategic framework that promotes awareness and builds a networked response capacity. It should focus on strengthening resistance to disinformation, and gendered information specifically. Because of the shared nature of these threats, this should involve joined-up coordinated efforts to prevent and mitigate risk, foster resilience, and balance solutions with our democratic values.

Gendered disinformation about Indigenous women and girls is deeply rooted in Canada's colonial history and current realities, with serious consequences. Corbett (2019) observed that inaccurate portrayals in media and culture reinforce negative stereotypes, leading non-Indigenous Canadians to ignore ongoing violence. Corbett recommends breaking this cycle by challenging false narratives, changing harmful media practices, and prioritizing Indigenous voices in storytelling. Together, these measures can help change the harmful information landscape and support reconciliation.

The proposed system emphasizes multi-sectoral collaboration involving people, processes and technology. Knowledge development will be essential to building networked capacity to counter gendered disinformation. This should offer mutual benefits and support shared learning, planning and implementation.

We propose a theory of change involving strategically aligned, society-wide interventions grounded in emerging research. We also offer a set of knowledge resources and technology examples useful to those in human services, policy and national security. Finally, we recommend creating a cross-sectoral knowledge development and mobilization network to support evidence-informed, collaborative efforts on this important issue.

RECOMMENDATIONS

Policy, Legislation and Enforcement

2. That the federal government:

    a. Implement policy and legislative measures to counter gendered disinformation, recognizing that it is a threat that spans community safety and wellbeing, and national security.
        - *The corresponding regulatory framework should ensure platform accountability, transparency, and meaningful financial penalties for non-compliance.*

    c. With targeted investment, initiate cross-departmental, industry, academic and private sector operational coordination and program collaboration to address gendered disinformation within public safety, public health, digital regulation, defence and national security frameworks.

    d. Develop a national strategy on gendered disinformation in close partnership with the private sector, research and civil society, integrating public safety, digital governance, and foreign policy approaches.

    j. Convene and engage women's advocacy organizations, racial justice groups, security and intelligence professionals, academic researchers, cyber-security experts and relevant community and private sector entities in dialogue on such matters as how to optimize the balance of protection and enforcement with freedom of expression online.

    k. Increase data collection and monitoring of gendered disinformation trends and actionable current intelligence.

    l. Conduct periodic cross-sector consultations with experts in gender-based violence, cybersecurity, open source intelligence, national security, and digital regulation to understand the evolving landscape of gendered disinformation.

m. Establish gender-responsive online safety laws that hold technology platforms accountable. Options include the re-introduction of Bill C-36[38] and the applications of relevant elements of a Clean Pipes Strategy[39].

n. Enhance training for security, intelligence, diplomatic, defence, law-enforcement and policymakers on technology-enabled GD.

o. Invest in digital literacy, research, open source intelligence and enforcement mechanisms to strengthen Canada's resilience against gendered disinformation.

Research and Knowledge Mobilization

3. That Canada support the creation of a cross-sectoral knowledge mobilization network on gendered disinformation – the Gendered Disinformation Knowledge Network (GenD-Net).

   Such a network would serve as a hub for leadership, information sharing, education and training, research, and policy coordination, program planning, operational coordination and de-confliction ensuring that responses to gendered disinformation are evidence-based, and aligned across sectors.

   The objectives of the network will be to:

[38] Canada's Bill C-36 (proposed) sought to amend hate speech provisions to better address online harms, including gender-based hate. Canada's Online Harms Act (Bill C-63), officially titled, *An Act to enact the Online Harms Act, to amend the Criminal Code, the Canadian Human Rights Act and An Act respecting the mandatory reporting of Internet child pornography by persons who provide an Internet service and to make consequential and related amendments to other Acts*, aimed to address harmful content on the internet. In particular, issues related to child exploitation, hate speech, and content promoting violence or self-harm. The Bill would establish a Digital Safety Commission to oversee compliance, investigate complaints, and enforce penalties. The Bill also aims to hold platforms accountable for the content that is hosted on their platform. In particular, it creates several Duties on the platform such as a duty to act responsibly, protect children and keep all the records. If the Bill were to receive Royal Assent, then the legislation would increase penalties for hate crime, expand the definition of hate crime and amend elements of the Criminal Code of Canada. It needs to be noted that the Bill was not passed prior to the 2025 election, hence, it is currently not codified in law.

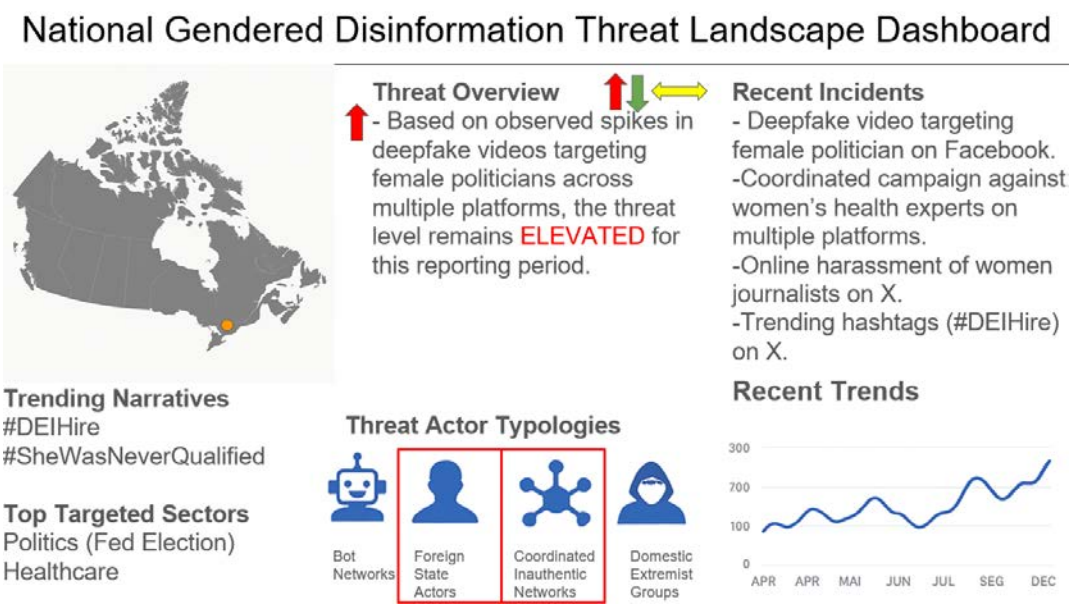
[39] A "clean pipes" strategy is a cybersecurity approach where internet service providers filter out malicious traffic—such as malware, phishing, and botnet activity—before it reaches end users. By blocking known threats at the network level, it helps create a safer online environment and reduces the burden on individuals and organizations to defend themselves. This strategy is part of national cybersecurity efforts in several countries, including the United Kingdom, Australia, and Singapore, which have partnered with internet service providers (ISPs) to implement network-level threat filtering to protect citizens and critical infrastructure.

- *Enhance knowledge mobilization and public awareness of gendered disinformation.*
- *Support curriculum development, stimulate and contribute to education and training.*
- *Strengthen community and cross-sectoral dialogue and collaboration on policy development.*
- *Support defence, intelligence, police and public safety agencies.*
- *Advance research and innovation, including evaluation capacity building.*
- *Bridge gaps in service provision for affected communities.*

Gendered Disinformation as a National Security Issue

4. That the Government of Canada refine and implement options for countering gendered disinformation as a national security issue, including its use as an element of foreign interference. Enhance the capabilities of defensive cyber operations in relation to this threat. More particularly:
    g. Establish a dedicated government funding stream for research and innovation on gendered disinformation that is open to Canadian industry, academia and not-for profit organizations.
    h. Incentivize Canadian industry participation and innovation through public-private partnerships and direct investment.
    i. Develop a national strategy on gendered disinformation as a foreign interference threat, and ensure integration with national defence policy, cyber security and national security strategies.
    j. Fund the creation of a cross-sectoral intelligence-sharing network to combat gendered disinformation, including the creation and maintenance of a national gendered disinformation threat landscape reporting capacity; this would, in-turn, feed into an intelligence "dashboard" (Figure 11) which could be made publicly available as part of building overall awareness an public will to confront this problem (See Annex E4, Attachment B).
    k. Establish legal and policy frameworks to protect women in public life from both foreign and domestic online harm.
    l. Develop a rapid response mechanism to protect individuals facing high-risk disinformation attacks (see Annex E4, Briefing Resources 1 and 4).

Figure 11. Sample depiction of a proposed gendered disinformation dashboard.



Impact of Recommendations

Implementing these recommendations will have significant impacts on combatting gendered disinformation, enhancing human rights protection, and promoting gender equality. By addressing this issue, intertwined with polarization and misogyny, we can safeguard women, girls, and gender-nonconforming individuals from targeted harm. More specific areas impacted are as follows:

Policy and Legislation

By implementing comprehensive policies and legislation, the federal government will strengthen community safety and national security. Establishing regulatory frameworks with platform accountability and penalties for non-compliance will ensure that digital spaces are safer and more transparent. Cross-departmental coordination will enhance efforts to address gendered disinformation within public safety and national security frameworks.

Multi-Sector Collaboration

Creating a national strategy in partnership with the private sector, research institutions and civil society will integrate approaches to enhancing both public safety and social media governance. Engaging diverse organizations in dialogue will balance safety and security with freedom of expression. Furthermore, this approach will help build resilience against gendered disinformation through enhanced data collection, training, and digital literacy investments.

Research and Knowledge Mobilization

A dedicated funding stream for research and innovation, alongside public-private partnerships, will drive industry participation and technological advancements.

Establishing the Gendered Disinformation Knowledge Network (GenD-Net) will enhance public awareness, support curriculum development, and foster cross-sectoral collaboration. By bridging gaps in service provision, it will ensure evidence-based responses aligned across sectors.

National Security

Recognizing gendered disinformation as a national security issue will help refine strategies to counter foreign interference. Developing a rapid response mechanism and legal frameworks will protect individuals from high-risk disinformation attacks.

Overall, when implemented, these measures will help to transform the online information landscape, support reconciliation, and uphold Canadian liberal democratic values by fostering a coordinated, strategic response to gendered disinformation.

REFERENCES

Ai Ramiah, A. & Hewstone, M. (2013). Intergroup contact as a tool for reducing, resolving, and preventing intergroup conflict evidence, limitations, and potential. *American Psychologist*, 68(7):527-542. DOI: 10.1037/a0032603.

Aljizawi, N., Anstis, S., Michaelsen, M., Arroyo, V., Baran, S., Bikbulatova, M., Böcü, G., Franco, C., Geybulla, A., Iliqud, M., Lawford, N., LaFlèche, E., Lim, G., Meletti, L., Mirza, M., Panday, Z., Posno, C., Reichert, Z., Taye, B., & Yang, A. (2024). *No escape: The weaponization of gender for the purposes of digital transnational repression* (Citizen Lab Report No. 180). University of Toronto. https://citizenlab.ca/2024/12/the-weaponization-of-gender-for-the-purposes-of-digital-transnational-repression/.

Bail, C.A. (2021). *Breaking the social media prism: How to make our platforms less polarizing*. Princeton, NJ: Princeton University Press.

Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Fallen Hunzaker, M.B., Lee, J., Mann, M., Merhout, F. & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Procedings of the National Academy of Science*, 115(37), 9216-9221. https://www.pnas.org/doi/epdf/10.1073/pnas.1804840115.

Barker-Singh, S. (2025). MP tells Sky News she was targeted online by Tate brothers after Commons contribution. *Sky News*, April 3, 2025. https://news.sky.com/story/mp-tells-sky-news-she-was-attacked-online-by-tate-brothers-after-commons-contribution-13340655.

Besancenot, M-D. (2025). Next time you hear someone say “it’s just coms”, pull out the striking visuals provided by the European External Action Service (EEAS) in their last report on information threats. LinkedIn post, March 26, 2025. https://www.linkedin.com/feed/update/urn:li:activity:7310592850847559680/.

Biddlestone, M., Azevedo, F. & van der Linden, S. (2022). Climate of conspiracy: A meta-analysis of the consequences of belief in conspiracy theories about climate change. *Current Opinion in Psychology*, 46, 101390. https://doi.org/10.1016/j.copsyc.2022.101390.

Bijlsma, A. M. E., van der Put, C. E., Vial, A., van Horn, J., Overbeek, G., & Assink, M. (2022). Gender differences between domestic violent men and women: Criminogenic risk factors and their association with treatment dropout. *Journal of Interpersonal Violence, 37*(23-24), NP21875-NP21901. https://doi.org/10.1177/08862605211072704

Boukes, M. & Hameleers, M. (2023) Fighting lies with facts or humor: Comparing the effectiveness of satirical and regular fact-checks in response to misinformation and disinformation. *Communication Monographs*, 90(1), 69-91, DOI:10.1080/03637751.2022.2097284.

Bradshaw, S. & Henle, A. (2021). The gender dimensions of foreign influence operations. *International Journal of Communication*, 15(2021), 4596-4618. https://ijoc.org/index.php/ijoc/article/view/16332/3584.

Canadian Women's Foundation (n.d.). *The facts about gendered digital hate, harassment, and violence*. https://canadianwomen.org/the-facts/online-hate-and-cyberviolence/.

Cavoukian, A. (2010). *Privacy by design: the definitive workshop*. A foreword by Ann Cavoukian, Ph.D. *IDIS, 3*, 247–251. https://doi.org/10.1007/s12394-010-0062-y.

CCA (Council of Canadian Academies). (2023). *Vulnerable Connections*. Ottawa (ON): Expert Panel on Public Safety in the Digital Age, CCA. https://cca-reports.ca/wp-content/uploads/2023/04/Vulnerable-Connections_FINAL_DIGITAL_EN_UPDATED.pdf.

Coelho, G.L.H., Hanel, P.H.P. & Wolf, L.J. (2020). The very efficient assessment of need for cognition: developing a six-item version. *Assessment*. 27(8):1870–1885. doi: 10.1177/1073191118793208.

Corbett, E. (2019). When disinformation turns deadly: The case of missing ad murdered Indigenous women and girls in Canadian media. In J. McQuade, T. Kwok & J. Cho (Eds.), *Disinformation and digital democracies in the 21st century*. Toronto, ON: The NATO Association of Canada, 19-23.

Costello, M. & Hawdon, J. (2020). Hate speech in online spaces. In T. Holt & A. Bossler (Eds.), *The Palgrave handbook of cybercrime and cyberdeviance*. London: Springer Nature, 1397-1416.

Dawson, M., Sutton, D., Carrigan, M., Grand'Maison, V., Bader, D., Zecha, A., & Boyd, C. (2019). *#CallItFemicide: Understanding gender-related killings of women and girls in Canada 2019*. Canadian Femicide Observatory for Justice and Accountability. https://femicideincanada.ca/callitfemicide2019/pdf.

Deibert, R. (2025). *Chasing shadows: Cyber espionage, subversion and the global fight for democracy*. New York: Simon & Schuster.

Dewey, C. (2014). Inside the 'manosphere' that inspired Santa Barbara shooter Elliot Rodger. *The Washington Post*, May 27, 2014. https://www.washingtonpost.com/news/the-intersect/wp/2014/05/27/inside-the-manosphere-that-inspired-santa-barbara-shooter-elliot-rodger/.

DiMeco, L. (2019). Gendered disinformation, fake news, and women in politics. *Council on Foreign Relations*, December 6, 2019. https://www.cfr.org/blog/gendered-disinformation-fake-news-and-women-politics.

Douglas, H., Harris, B. & Dragiewicz, M. (2019). Technology-facilitated domestic and family violence: Women's experiences. *British Journal of Criminology*, 59, 551-570. doi:10.1093/bjc/azy068.

Ecker, U. K. H., Lewandowsky, S., & Chadwick. M. (2020). Can corrections spread misinformation to new audiences? Testing for the elusive familiarity backfire effect. *Cognitive Research: Principles and Implications*, 5, 41. https://doi.org/10.1186/s41235-020-00241-6.

Economist Intelligence Unit (2020). *Measuring the prevalence of online violence against women*. https://onlineviolencewomen.eiu.com/.

Equal Measures 2030 (2024). *A gender equal future in crisis? Findings from the 2024 SDG Gender Index*. Seattle, WA: Equal Measures 2030.

Ermoshina, K. & Musiani, F. (2025). Safer spaces by design? Federated socio-technical architectures in content moderation. *Internet Policy Review*, *14*(1). https://doi.org/10.14763/2025.1.1827.

Fazio, L.K, Brashier, N.M., Payne, B.K. & Marsh, E.J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 144(5), 993-1002.

Fazio, L.K. & Sherry, C.L. (2020). The effect of repetition on truth judgments across development. *Psychological Science*, 31(9), 1150-1160.

French, A.M., Storey, V.C. & Wallace, L. (2025). The impact of cognitive biases on the believability of fake news. *European Journal of Information Systems, 34*(1), 72-93, DOI: 10.1080/0960085X.2023.2272608.

Gill, C. & Aspinall, M. (2020). *Understanding coercive control in the context of intimate partner violence in Canada*. Research paper for the Office of the Federal Ombudsman for Victims of Crime, Department of Justice Canada. Fredericton, NB: University of New Brunswick.

Gill, C., Campbell, M.A. & Ballucci, D. (2021). Police officers' definitions and understandings of intimate partner violence in New Brunswick, Canada/ *The Police Journal, 94*(1), 20-39. https://doi.org/10.1177/0032258X19876974.

George, J., Gerhart, N., & Torres, R. (2021). Uncovering the truth about fake news: A research model grounded in multi-disciplinary literature. *Journal of Management Information Systems, 38*(4), 1067–1094. https://doi.org/10.1080/07421222.2021.1990608.

Hameleers, M. (2022). *Populist disinformation in fragmented information settings: Understanding the nature and persuasiveness of populist and post-factual communication*. London: Routledge.

Hasher, L., Goldstein, D. & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, 16(1), 107-112.

Hess, A. (2014). Why women aren't welcome on the Internet. *Pacific Standard*, January 6, 2014. https://psmag.com/social-justice/women-arent-welcome-internet-72170/.

Human Rights Watch. (2024). *We will find you: A global look at how governments repress nationals abroad*.

https://www.hrw.org/sites/default/files/media_2024/02/global_transnationalrepression0224web_0.pdf

Hutchins, E., Cloppert, M.J. & Amin, R.M. (2010). *Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains*. Unpublished report. Lockheed Martin. https://www.lockheedmartin.com/content/dam/lockheed-martin/rms/documents/cyber/LM-White-Paper-Intel-Driven-Defense.pdf

Jankowicz, N., Pepera, S. & Middlehurst, M. (2021). *Addressing online misogyny and gendered disinformation: A how-to guide*. National Democratic Institute. https://www.ndi.org/sites/default/files/Addressing%20Gender%20%26%20Disinformation%202%20%281%29.pdf

Jankowicz, N., Gomez-O'Keefe, I., Hoffman, L. & Vidal Becker, A. (2024). *It's everyone's problem: Mainstreaming responses to technology-facilitated gender-based violence*. New York, NY: Columbia University SIPA Institute of Global Politics and the Vital Voices Global Partnership. https://igp.sipa.columbia.edu/sites/igp/files/2024-09/IGP_TFGBV_Its_Everyones_Problem_090524.

Kakinohana, R.K. & Pilati, R. (2023). Differences in decisions affected by cognitive biases: examining human values, need for cognition, and numeracy. *Psicologia Reflexao e Critica*., *36*(1):26. doi: 10.1186/s41155-023-00265-z.

Kelshall, C. (2020). Soft violence, social radicalisation and violent transnational social movements (VTSMs). Paper presented at the November 25, 2020 meeting of the CASIS West Coast Security Conference, Vancouver, BC. *Journal of Intelligence, Conflict and Warfare, 3*(3). https://doi.org/10.21810/jicw.v3i3.2800.

Kesivan, M. (2024). India is witnessing the slow-motion rise of fascism. *The Guardian*, September 8, 2024. https://www.theguardian.com/commentisfree/article/2024/sep/08/india-slow-motion-rise-of-fascism?CMP=Share_iOSApp_Other.

Kiili, K., Siuko, J. & Ninaus, M. (2024). Tackling misinformation with games: a systematic literature review. *Interactive Learning Environments, 32*(10), 7086-7101, DOI: 10.1080/10494820.2023.2299999.

Kolga, M. (2024, October 16). *Testimony before the Canadian House of Commons Standing Committee on Public Safety and National Security, October 1, 2024*. Macdonald-Laurier Institute. https://macdonaldlaurier.ca/marcus-kolga-warns-against-threat-of-russian-cognitive-warfare-mli-in-parliament/.

Korteling, J.E. & Toet, A. (2020). Cognitive biases. In S. Della Sala (Ed.), *Reference Module in Neuroscience and Biobehavioral Psychology*. Amsterdam-Edinburgh: Elsevier ScienceDirect. https://doi.org/10.1016/B978-0-12-809324-5.24105-9.

Lalonde, M., Boulianne, G., Rutherford, N., Beaulieu, M., Ghodrati, H. & Dahmane, M. (2025). *Visual and multi-modal disinformation: Analysis, challenges, solutions*. Montreal, QC: Computer Research Institute of Montreal (CRIM) & Ottawa, ON: Information Integrity Lab, University of Ottawa.

Lanier, J. (2018). *Ten arguments for deleting your social media accounts*. New York, NY Holt.

Lewandowsky, S., Cook, J., Ecker, U., Albarracín, D., Kendeou, P., Newman, E.J., Pennycook, G., Porter, E., Rand, D. G., Rapp, D. N., Reifler, J., Roozenbeek, J., Schmid, P., Seifert, C. M., Sinatra, G., Swire-Thompson, B., van der Linden, S., Wood, T.J., & Zaragoza, M. S. (2020). *The debunking handbook 2020*. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1247&context=scholcom.

Lewandowsky, S. & van der Linden, S. (2021) Countering misinformation and fake news through inoculation and prebunking. *European Review of Social Psychology*, DOI: 10.1080/10463283.2021.1876983.

Lilly, M. 2016. *The world is not a safe place for men: The representational politics of the manosphere*. Unpublished masters thesis, University of Ottawa. https://ruor.uottawa.ca/server/api/core/bitstreams/1eee5112-7f22-4ffc-a49d-a978a56bed05/content.

Maertens, R., Roozenbeek, J., Simons, J.S., Lewandowsky, S., Maturo, V.,, Goldberg, B., Xu, R. & van der Linden, S. (2025). Psychological booster shots targeting memory increase long-term resistance against misinformation. *Nature Communications, 16*, 2062 (2025). https://doi.org/10.1038/s41467-025-57205-x.

Maimann, K. (2024). Instagram ignored 93% of abusive comments toward female politicians: Report. *CBC Online News*, August 19, 2024. https://www.cbc.ca/news/women-politicians-online-abuse-1.7298168.

Marczak, B., Scott-Railton, J., Razzak, B.A., Al-Jizawi, N., Anstis, S., Berdan, K. & Deibert, R. (2021). *Pegasus vs. Predator: Dissident's doubly-infected iPhone reveals Cytox mercenary spyware*. The Citizen Lab Research Report No. 147, University of Toronto, December 2021. https://citizenlab.ca/2021/12/pegasus-vs-predator-dissidents-doubly-infected-iphone-reveals-cytrox-mercenary-spyware/

Maté, G. (2022). *The myth of normal: Trauma, illness and healing in a toxic culture*. Toronto, ON: Knopf Canada.

Maté, G. (2024). We each have a Nazi in us. We need to understand the psychological roots of authoritarianism. *The Guardian*, September 6, 2024. https://www.theguardian.com/commentisfree/article/2024/sep/06/authoritarianism-roots-origin?CMP=Share_iOSApp_Other.

Matthews, M. (2021). *Four approaches to content moderation and their risks and benefits*. Information and Communications Technology Council (ICTC), October 20, 2021. https://ictc-ctic.ca/articles/four-approaches-to-content-moderation-and-their-risks-and-benefits.

McIntyre, L. (2023). *Post-truth*. Cambridge, MA: MIT Press.

McIntyre, L. (2023). *On disinformation: How to fight for truth and protect democracy*. Cambridge, MA: MIT Press.

McMahon, D. (2021). *Cyber deception: The art of camouflage, stealth and misdirection*. Unpublished paper. Ottawa, ON: Clairvoyance Cyber Corp.

Michaelsen, M. & Anstis, S. (2025): Gender-based digital transnational repression and the authoritarian targeting of women in the diaspora, *Democratization*, DOI: 10.1080/13510347.2025.2476178.

Mozur, P., Satariano, A., Krolik, A. & Myers, S.L. (2024). How Telegram became a playground for criminals, extremists and terorists. *New York Times*, September 7, 2024. https://www.nytimes.com/2024/09/07/technology/telegram-crime-terrorism.html?smid=nytcore-ios-share&referringSource=articleShare&sgrp=c-cb.

National Democratic Institute (2022). *Interventions for ending online violence against women and girls*. https://www.ohchr.org/sites/default/files/documents/issues/expression/cfis/gender-justice/subm-a78288-gendered-disinformation-cso-ndi-annex-3.pdf.

Norman, A. (2021). *Mental immunity*. New York: Harper.

North Atlantic Treaty Organization. (2023). *NATO's approach to countering disinformation*. https://www.nato.int/cps/en/natohq/topics_219728.htm.

Off, C. (2024). *At a loss for words: Conversation in the age of rage*. Toronto: Random House Canada.

Pain, P. (2023). "Suddenly we were the story": Women journalists, the #MeToo movement and online misogyny in India. In L.M. Cuklanz (Ed.), *Gender violence, social media and online environments* (pp. 113-129). London: Routledge.

Parliament of Canada, House of Commons Standing Committee on Public Safety and National Security. (2024). *Minutes of Proceedings.* 44th Parliament, 1st session, meeting no. 121. Retrieved from the Parliament of Canada website: https://www.ourcommons.ca/documentviewer/en/44-1/SECU/meeting-121/evidence.

Pomerantzev, P. (2024). *How to win an information war*. London: Faber.

Powell, A., & Sugiura, L. (2018). Resisting Rape Culture in Digital Society. In W. S. DeKeseredy, C. M. Rennison, & A. K. Hall-Sanchez (Eds.), *The Routledge International Handbook of Violence Studies* (pp. 469–479). Milton: Routledge.

Pronk, N.P., Hernandez, L.M., Lawrence, R.S. (2013). An integrated framework for assessing the value of community-based prevention: A report of the Institute of Medicine. *Prevention of Chronic Disease*,10:120323. DOI: http://dx.doi.org/10.5888/pcd10.120323.

Public Inquiry Into Foreign Interference in Federal Electoral Processes and Democratic Institutions (2025). *Final report. Volume 1: Report summary*. His Majesty the King in Right of Canada. https://foreigninterferencecommission.ca/fileadmin/report_volume_1.pdf.

Ressa, M. (2022). *How to stand up to a dictator*. New York, NY: Harper Collins.

Ribeiro, M. H., Blackburn, J., Bradlyn, B., De Cristofaro, E., Stringhini, G., Long, S., Greenberg, S., & Zannettou, S. (2021). The evolution of the manosphere across the web. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media* (pp. 196–207). AAAI Press. https://ojs.aaai.org/index.php/ICWSM/article/view/18053/17856.

Richardson-Self, L. (2021). *Hate speech against women online: Concepts and countermeasures*. Lanham, MD: Rowman and Littlefield.

Rid, T. (2020). *Active measures: The secret history of disinformation and political warfare*. New York: Picador.

Roozenbeek, J., van der Linden, S. (2019). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications, 5*(65). https://doi.org/10.1057/s41599-019-0279-9.

Ryan, J. (2025). Europe's race to arm is pointless if its adversaries are waging war online. *The Guardian*, April 15, 2025. https://www.theguardian.com/commentisfree/2025/apr/15/us-europe-military-spending-trump-ireland?CMP=Share_iOSApp_Other.

Samson, D.R. (2023). Out tribal future: How to channel our foundational human instincts into a force for good. New York, NY: St. Martin's Press.

Schick, N. (2020). *Deepfakes: The coming infocalypse*. New York, NY: Twelve.

Secrétariat général de la défense et de la sécurité nationale (VIGINUM) (2024). Matryoshka: A pro-Russian campaign targeting media and the fact-checking community. https://www.sgdsn.gouv.fr/files/files/20240611_NP_SGDSN_VIGINUM_Matriochka_EN_VF.pdf.

Sugiura, L. & Smith, A. (2020). Victim Blaming, Responsibilization and Resilience in Online Sexual Abuse and Harassment. In: Tapley, J., Davies, P. (eds) *Victimology*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-42288-2_3.

Sobieraj, S. (2020). *Credible threat: Attacks against women online and the future of democracy*. Oxford, UK: Oxford University Press.

Sorell, T. & Kelsall, J. (2025). Violent video games, recruitment and extremism. *Criminal Justice Ethics*. DOI: 10.1080/0731129X.2025.2484974.

Springer. F. & Phillips, J. (n.d.). *The institute of medicine framework and its implication for the advancement of prevention policy, programs and practice*. Center for Substance Abuse Prevention. http://ca-sdfsc.org/docs/resources/SDFSC_IOM_Policy.pdf

Stanley, J. (2024). *Erasing history: How fascists re-write the past to control the future*. New York: Simon and Schuster.

Stark, E. (2007). *Coercive control: How men entrap women in personal life*. Oxford University Press.

Stuart, K. (2025). Video games can't escape their role in the radicalisation of young men. *The Guardian*, March 24, 2025. https://www.theguardian.com/games/2025/mar/24/video-games-cant-escape-their-role-in-the-radicalisation-of-young-men?CMP=Share_iOSApp_Other.

Susmann, M.W. & Wegener, D.T. (2022). The role of discomfort in the continued influence effect of misinformation. *Memory and Cognition*, 50, 435-448. https://doi.org/10.3758/s13421-021-01232-8.

Tenove, C., Tworek, H.J.S., & McKelvey, F. (2018). *Poisoning Democracy: How Canada Can Address Harmful Speech Online*. Public Policy Forum. https://ppforum.ca/wp-content/uploads/2018/11/PoisoningDemocracy-PPF-1.pdf.

Thakur, D. & Hankerson, D.L. (2021). *Facts and their discontents: A research agenda for online disinformation, race, and gender.* Center for Democracy & Technology. https://osf.io/preprints/osf/3e8s5.

UN Women Expert Group (2022). *Technology-facilitated violence against women: Towards a common definition. Report of the meeting of the Expert Group*. World Health Organization. https://www.unwomen.org/sites/default/files/2023-03/Expert-Group-Meeting-report-Technology-facilitated-violence-against-women-en.pdf

van der Linden, S. (2013). 'What a hoa'. *Scientific American Mind*, 24(4), 40-43.

van der Linden, S. (2015). The conspiracy effect: Exposure to conspiracy theories (about global warming) decreases pro-social behavior and science acceptance. *Personality and Individual Differences, 87*, 171-173.

van der Linden, S. (2021). The best way to deal with Covid myths this Christmas? Pre-bunk rather than debunk. *The Guardian*, December 23, 2021. https://www.theguardian.com/commentisfree/2021/dec/23/covid-myths-christmas-vaccines-virus-missinformation.

van der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 28 (March), 460-467. DOI: 10.1038/s41591-022-01713-6.

van der Linden, S. (2023). *Foolproof: Why misinformation infects oru minds and how to build immunity*. London, UK: Norton.

van der Linden, S., Panagopoulos, C., & Roozenbeek, J. (2020). You are fake news: Political bias in perceptions of fake news. *Media, Culture & Society, 42*(3), 460–470. https://doi.org/10.1177/0163443720906992.

Whyte, C. (2020). Cyber conflict or democracy "hacked"? How cyber operations enhance information warfare, *Journal of Cybersecurity*, *6(1)* https://doi.org/10.1093/cybsec/tyaa013.

Zmigrod, L. (2022). A psychology of ideology: Unpacking the psychological structure of ideological thinking. *Perspectives on Psychological Science, 17*(4), 1072-1092. DOI: 10.1177/17456916211044140.

Zmigrod, L., Burnell, R. & Hemeleers, M. (2023). The misinformation receptivity framework. *European Psychologist, 28*(3), 173-188. https://doi.org/10.1027/1016-9040/a000498.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. New Yok, NY: Public Affairs.

ANNEXES

Annex A: Project Team

Community Safety Knowledge Alliance

Dr. Janos Botschner, PhD – *Project Lead.* Janos is a social scientist with deep experience in applied research and evaluation and strategic consulting across a range of contexts. He holds a joint doctorate in applied social and developmental psychology. Janos has held a number of adjunct faculty appointments and administrative positions during a lengthy career in the broader public sector. Janos' professional work covers the spectrum of issues related to collaborative public safety and community well-being, with a focus on understanding, and responding adaptively to, the complex issues and emerging opportunities of today's world.

Cal Corley, MBA Cal is CEO of the Community Safety Knowledge Alliance and a former Assistant Commissioner of the RCMP. Over the course of his career, Cal gained extensive experience in both operations and executive management, serving in such areas as national security, criminal intelligence, drug enforcement, human resources, and leading reform initiatives. He also served on secondments at the Privy Council Office and at Public Safety Canada.

Ritesh Kotak, JD, MBA is a Technology and Cybersecurity analyst and a licensed lawyer in Ontario. Ritesh started his career in public safety working for two police organizations focusing on cybercrime investigations and innovation. He left policing to pursue an MBA and then worked in Big Tech for two years focusing on innovation and smart cities. He left the Tech sector to attend law school and received a JD with a Law and Technology Option. Ritesh is a frequent contributor on mainstream media and is an international public speaker. Ritesh has also appeared twice as a witness in House of Commons Committees.

Sapper Labs Group

Dave McMahon, Hon. B.Eng., M.S.M. – *Project Co-Lead.* Chief Intelligence Officer at SLG, Dave is a deep generalist and expert with 40 years of experience in intelligence operations, cyber and cognitive warfare. He has an honours degree in Computer Engineering from the Royal Military College of Canada. Dave served with the Canadian Armed Forces, the Canadian Security Intelligence Service (CSIS), the Communications Security Establishment (CSE), the Security Intelligence Review Committee (SIRC), and the Office of the Communications Security Establishment Commissioner (OCSEC). He was a principal architect of a number of national offensive cyber and foreign intelligence programs for Canada. Dave co-chaired the interdepartmental committee on Information Warfare and psychological operations.

Dr. Giovanna Cioffi, CD, Hon. BA, MDEM, MES, PhD, served as an Army intelligence analyst r and expert in Cyber warfare and Psychological Operations. She was Deputy Chief of Targets/Ground Force Analyst/Special Purpose Reconnaissance Analyst (Operation IMPACT), a Captured Equipment and Material Analyst (Digital Forensics) (Op IMPACT), and a National Security Team Open Source Intelligence Analyst with a multinational joint intelligence task force covering global extremism.. She also worked as an Intelligence Operator and Intelligence Analyst at CANSOFCOM as well as Civil Military Cooperation and Psychological Operations Analyst/Tactical Operator with the CAF.

Dr. Juliane Ollinger, PhD is a research scientist with a PhD in Microbiology (Cornell 2008) with a focus on infectious diseases. Julie brings her strong research background and critical analysis skills to the Sapper team and has contributed to intelligence investigations focused on Due diligence, National Security, Foreign Interference, Disinformation and Support to Defence Operations.

Bradley Sylvestre, MA is an analyst with Sapper Labs Group focused on open-source intelligence (OSINT) and strategic analysis. His research interests broadly encompass strategic competition, foreign interference, espionage, disinformation and deep fake research. Prior to joining Sapper Labs Group, Bradley worked as a strategic analyst with the Canadian Armed Forces and Department of National Defence. Within the force development enterprise, his efforts supported work to identify the necessary capabilities to enable and sustain the Canadian Armed Forces and missions through current intelligence. Bradley holds a MA in International Affairs from the Norman Paterson School of International Affairs (NPSIA), also located in Ottawa.

Actua

Actua and CSKA collaborated to produce resources tailored to parents, youth and educators, based on knowledge synthesized by CSKA and SLG. The following staff members led Actua's involvement in this work.

Mikayla Ellis, BA, Senior Manager, Outreach.

Janelle Fournier, PhD (ABD), Senior Manager, Education.

Abbey Ramdeo, MT, Manager, National Educator Learning Program.

Annex B: Advisory Committee

Michael (Mike) Doucet is a senior leader of portfolios focusing on public safety and technology. He served as executive director of the Security Intelligence Review Committee, now known as National Security and Intelligence Review Agency. He currently serves as Executive Director, Office of the CISO, at OPTIV, a cyber advisory and solutions company, providing strategic advice on cyber programs, technology and risk.

Jennifer Flanagan is the President and CEO of Actua, which has become Canada's largest STEM outreach organization. It represents a national network of 43 universities and colleges that engage youth, ages 6-26, in STEM learning experiences, and advancing equity, diversity and inclusion in STEM. Actua's activities annually reach 350,000 young people. In 2021, Jennifer was awarded in the Manulife Science and Technology category, which recognizes women in STEM roles who are challenging the status quo for knowledge and female empowerment.

Dr. Carmen Gill is a professor in the Department of Sociology at the University of New Brunswick. She works in partnership with police agencies in Canada. Her research focuses on police intervention in intimate partner violence (IPC), domestic homicide and treatment of perpetrators and victims through the criminal justice system. Carmen is currently leading a three-year national research project entitled: Coercive control, risk assessment and evidence of intimate partner violence: Police response in partnership with the Canadian Association of Chiefs of Police (CACP), the Canadian Police Knowledge Network (CPKN) and l'École nationale de police du Québec. Carmen was previously the leader of the Canadian observatory on the justice system response to intimate partner violence (2006-2016). She led the development of the national framework for collaborative police action on IPV with CACP.

Jennifer Irish is the Director of the Information Integrity Lab at the University of Ottawa, advancing understanding, analysis and knowledge transfer related to disinformation and misinformation. She serves concurrently as an Associate at uOttawa's Telfer Centre for Executive Leadership, as Program Director of its Executive Security and Intelligence Leadership Certificate. She brings to these positions previous extensive leadership experience in Canada's foreign service and government in national security and international affairs. Ms. Irish's 35-year career in the foreign service and federal government included 5 diplomatic postings abroad, and leadership experience in international security, global environment issues, human rights, and humanitarian affairs. In Canada's Security and Intelligence Community she led in intelligence assessment and global threats and trends analysis, including as Director General at the Integrated Terrorism Assessment Centre which assesses threats to Canada related to terrorism and extremism. She also served as the Director of Operations of the Intelligence Assessment Secretariat of Canada' Privy Council Office, which provides strategic global intelligence assessments for high-level government decision-makers. Ms. Irish also provides professional training services in leadership, management,

business and decision-making processes, strategic analysis, briefings, and engagement strategies, to national security. law enforcement and other clients.

Alan Jones is executive adviser to the University of Ottawa Professional Development Institute and a retired CSIS officer who served in numerous operational and policy positions, including assistant director of CSIS. Alan's CSIS career included being the Chair of the G8 working Committee on Terrorism, Senior Policy Advisor in the Privy Council Office, Security and Intelligence Secretariat, Director General of the Counter Terrorism Branch and Director General of the International Terrorism Branch. In 2008 Alan became the Assistant Director for Operations, responsible for all operational programs and in 2010 he became the Assistant Director for Technology which included both corporate and operational technology.

Marcus Kolga is an international award-winning documentary filmmaker, journalist, digital communications strategist, and a leading Canadian expert on Russian and Central and Eastern European issues. Marcus has a focus on communications and media strategies as tools of foreign policy and defence, and continues to write commentary for national and international media including the Globe and Mail and Toronto Star. He is the co-founder and publisher of UpNorth.eu, an online magazine that features analysis and political and cultural news from the Nordic and Baltic region. Marcus is involved with international human rights organizations and national political organizations. In 2015, Marcus was awarded the Estonian Order of the White Star by President Toomas Hendrik Ilves.

Annex C: System of People, Processes and Technology Aligned to Theory of Change

Levels and outcomes from theory of change addressed by the present project

Focus of Interventions	Ecological Levels			
	Individual	**Microsystem**	**Exosystem**	**Macrosystem/ Chronosystem**
	Physical, mental & social development & wellbeing	Family, peers, schools, religious groups, health system	Neighbours, legal & social welfare services, community-based services, mass & social media	Attitudes & ideologies of broader culture/ Major global & environmental events occurring over time
Society			Acts of misogyny are less tolerated GD-based foreign interference is detected and identified	
Government			Greater accountability for enablers & perpetrators of GD GD-based foreign interference is detected and identified Foreign interference/transnational oppression is proactively targeted	
Population			Public awareness of GD-based foreign interference	
Organizations & Networks		Organizations & individuals use knowledge and techniques to identify & counter GD	Enhanced systems of support for victims of GD GD-based foreign interference is publicly identified	Greater public awareness about GD, its methods & its impacts
Individuals	Organizations & individuals use knowledge and techniques to identify & counter GD			

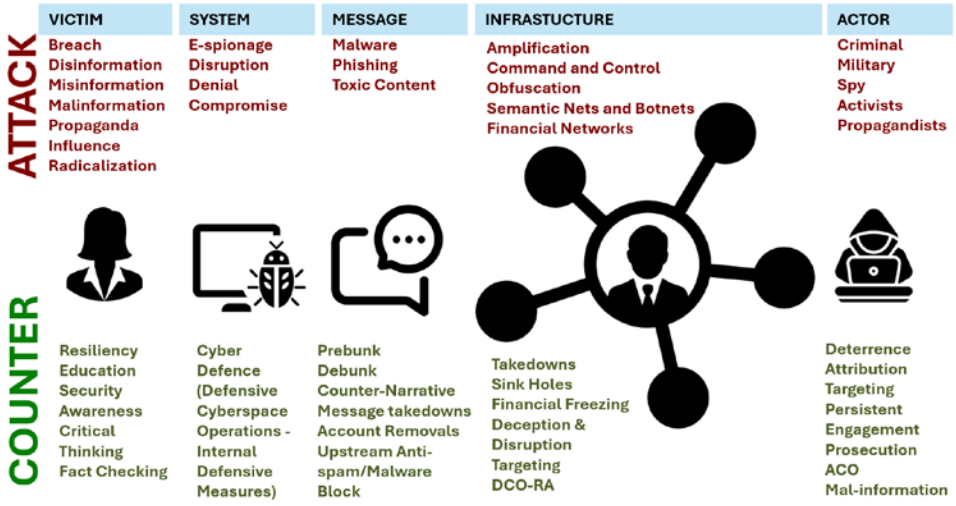
The main focus of the present project

Project outputs, corresponding ecological levels and areas of primary focus

Focus of Interventions	Ecological Levels			
	Individual	**Microsystem**	**Exosystem**	**Macrosystem/ Chronosystem**
	Physical, mental & social development & wellbeing	Family, peers, schools, religious groups, health system	Neighbours, legal & social welfare services, community-based services, mass & social media	Attitudes & ideologies of broader culture Major global & environmental events occurring over time
Society				
Government				**Knowledge products** Information sheet for government re GD as a national security threat/avenues for contending with GD in cases of foreign interference/transnat'l oppr'n **Options for addressing GD via policy, standards, regulations** Information sheet for organizations & government (consistent with theory of change)
Population		**Awareness & action focused knowledge products** Information sheets for individuals		
Organizations & Networks		**Awareness & action focused knowledge products** Information sheets for individuals & organizations **Additional resources** Literature/research synthesis	**Awareness & action focused knowledge products** Information sheets **Solution system** Technical options and framework for application of countermeasures – people, processes & technologies/tools) to counter GD and enhance system capacity **Recommendations for organizations and networks supporting those impacted by GD** **Engagement of key actors for KT/E**	**Options for addressing GD via policy, standards, regulations** Information sheet for organizations & government (consistent with theory of change) **Engagement of key actors** For KT/E and to enhance capacity and resilience
Individuals	**Awareness & action focused knowledge products** Information sheets for individuals			

Proposed system of people, processes and technology for countering gendered disinformation, aligned to theory of change

Cyber disinformation ecosystem and targeted countermeasures



VICTIM/AUDIENCE

- Building **resiliency** within the target audience and users starts with **security awareness** education, **critical thinking** and promoting access to **authoritative sources** of information.
- **Fact-checking**, media **literacy** programs and increased **transparency** in social media advertising can help the audience make informed decisions.
- **Cyber safe** programs and access to trusted end-device apps, platform and upstream security services.
- **Counter conspiracy beliefs** without challenging a **person's identity** may therefore be an effective strategy.

MESSAGE/PAYLOAD

- We can tackle **toxic messaging** and **malware** with content-based spam and malware filters supported by artificial intelligence (AI), **debunking/pre-bunking** of posts and **suspending accounts** of maligned influencers.
- **Counter-narratives** are highly-effective but should always be **truth-based** as part of an **information peacekeeping** strategy (IPK) or global peace and stabilization operations.

INFRASTRUCTURE

- Disinformation campaigns rely on cyberspace to propagate and amplify their malicious content or message with **botnets**. Cyberspace also offers an effective means to hide through **obfuscation** and non-attribution networks.
- **Enumerating** foreign global disinformation infrastructure requires effective **open-source intelligence** and targeting resources.
- Protective DNS services like **Canadian Shield** help.
- Ultimately **taking down** or **sink-holing** malevolent infrastructures has been a more **effective strategy** than chasing billions of toxic messages consumed by a target audience.

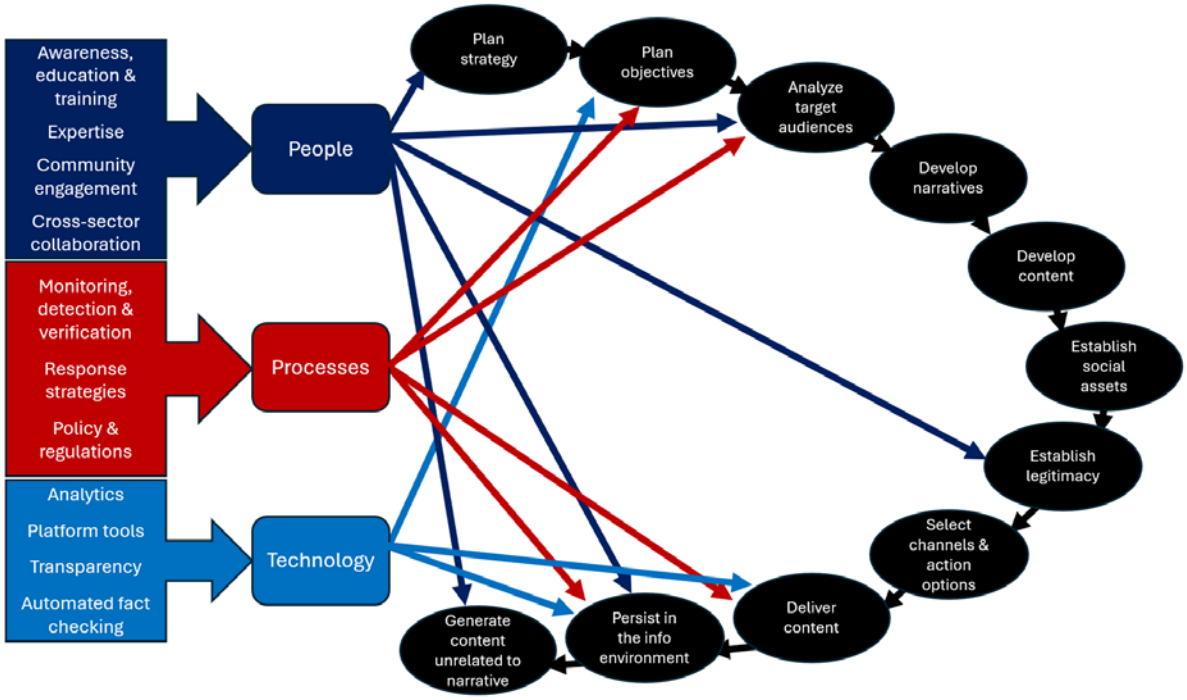
ACTOR

- The threat **actor** sits at the top of the food chain. Whether that is a **hostile intelligence service** (HoIS) or **paramilitary**or **transnational criminal organization**, **troll** farms or person-of-influence.
- Targeting the threat actor requires **strong attribution** substantiated with sophisticated **intelligence**, but is worth the effort.
- Cut the head off the troll and the disinformation campaign goes silent. Effects can include **sanctioning** companies and individuals, **freezing assets**, dismantling **financial networks**, **disrupting command and control**, maintaining **persistent engagement**, or following through with **indictment** and **prosecution**.
- **Industry** has been actively disrupting and **dismantling adversary networks**, **exposing** and **prosecuting actors** effectively for quite some time. **Public-private** collaboration, where appropriate, will be important from the perspectives of: effectiveness, accountability and trust.

PERSISTENT ENGAGEMENT

- The importance of **persistent engagement** at its core is to preserve our advantages and defend national interests in, through and from cyberspace by **contesting adversaries'** malicious cyber and influence activity during **day-to-day competition**.
- Strategic advantage is achieved through operations that **hunt** down the threat, **close the attribute chain**, **defend forward**, **contest** and **counter** the adversary **in real-time**.

Influence operation kill chain components addressed by proposed system of people, processses and technology



System features		Influence operation kill chain components targeted by proposed system of PPT										
		Plan strategy	Plan objectives	Analyze target audience	Develop narratives	Develop content	Establish social assets	Establish legitimacy	Select channels & action options	Deliver content	Persist in the information environment	Generate content unrelated to narrative
People	Awareness, education & training	■		■				■			■	■
	Expertise	■		■				■			■	■
	Community engagement	■		■				■			■	■
	Cross-sector collaboration	■		■				■			■	■
Processes	Monitoring, detection & verification		■	■						■	■	
	Response strategies		■	■						■	■	
	Policy & regulations		■	■						■	■	
Technology	Analytics		■							■	■	
	Platform tools		■							■	■	
	Transparency		■							■	■	
	Automated fact checking		■							■	■	

System features		System focus re social ecosystem (engagement & knowledge mobilization)				
		Individuals	Orgzn's & Netwks	Population	Government	Society
People	Awareness, education & training	Knowledge products to promote awareness & help identify forms & harms of GD & apply critical thinking				
	Expertise		Engagement of key actors for KT/E Recommendations for organizations and networks supporting those impacted by GD		Engagement of key actors for KT/E Avenues for contending with GD in cases of foreign interference	
	Community engagement					
	Cross-sector collaboration					
Processes	Monitoring, detection & verification		Options for addressing GD via policy, standards, regulations, guidelines, including perpetrator accountability/ rehabilitation		Options for addressing GD via policy, standards, regulations, guidelines, including perpetrator accountability/ rehabilitation Avenues for contending with GD in cases of foreign interference	
	Response strategies					
	Policy & regulations					
Technology	Analytics	Companion resources for understanding contexts, causes, threats (actors, methods/tools), enablers and consequences (emphasizing role of AI based tools as key enablers)	Companion resources for understanding contexts, causes, threats (actors, methods/tools), enablers and consequences (emphasizing role of AI based tools as key enablers) Technical options for countermeasures		Companion resources for understanding contexts, causes, threats (actors, methods/tools), enablers and consequences (emphasizing role of AI based tools as key enablers) Avenues for contending with GD in cases of foreign interference	
	Platform tools					
	Transparency					
	Automated fact checking					

System features		Influence operation kill chain components targeted by proposed system of PPT				
		Plan strategy	Analyze target audience	Establish legitimacy	Persist in the information environment	Generate content unrelated to narrative
People	Awareness, education & training (Create and diffuse knowledge to support ...)	Awareness about broader agendas and uses of online platforms to promote GD/MDM	Understanding of the ways that MDM/GD campaigns seek to: exploit and exacerbate social divisions; build on existing conspiracy theories; & leverage social media vulnerabilities/ properties to deepen and amplify impacts	Critical thinking about the ways that media platforms, accounts and influencers may be compromised and/or inauthentic (fake) purveyors of GD/MDM	Tools (concepts, systems) to detect information assets, operational activities and other TTPs used by those who are conducting influence operations	Individual and collective ability to observe and discern attempts to obscure the presence of malicious exploits within the information ecosystem
	Expertise (Translate & share knowledge about...)	Translate knowledge about broader context of misogyny and the ways it interacts to enable and benefit from GD/MDM			Awareness of the TTPs used by malicious actors to conceal information assets and operational activity	
	Community engagement (Engage relevant stakeholders to understand, amplify knowledge & collaborate)	Foster awareness about broader reasons online platforms are used to promote GD/MDM				Individual and collective capacity to distinguish GD/MDM content from distracting 'noise' and direct counter measures toward the GD/MDM within the information ecosystem
	Cross-sector collaboration (Support coordinated, joined-up action to...)	Diffusion of knowledge of concepts and harms to enable shared understanding and joined-up action to prevent and address harms	Promote awareness of, pre-bunk/de-bunk: attempts at manipulating social divisions & conspiracy theories Provide inputs on regulatory/ voluntary options to moderate vulnerabilities of tech. platforms		Foster networked capacity to detect relevant signals from contrived noise & to collaborate on measures to locate, identify and counteract GD/MDM content within operations that include aspects of concealment, diversion and distraction	

System features		Influence operation kill chain components targeted by proposed system of PPT			
		Plan objectives	Analyze target audience	Deliver content	Persist in the information environment
Processes	Monitoring, detection & verification (Describe/ explore approaches to...)		Examine and promote awareness of platform vulnerabilities related to hosting and propagating malicious information content Promote awareness of the general features of conspiracy theories to enable the discovery of information 'viruses' Promote awareness of the general features of common forms of GD/MDM	Document verified threat intelligence on GD/MDM Explore options for sharing threat intelligence in support of networked capacity to identify and respond to information operations involving GD/MDM	
	Response strategies (Identify the features of promising counter-narratives/ counter-measures)		Build knowledge of the features of counter-narratives that may counteract conspiracy theories and other forms of GD/MDM	Explore options for developing and enabling networked capacity to identify and respond to information operations involving GD/MDM, based on shared threat intelligence, and updated policy and regulation (where relevant & appropriate)	
	Policy & regulations (Examine & consider objectives, principles and tools...)		Engage policy advocates and policy professionals on dialogue about the role of policies and regulations, and technical options for achieving an effective balance between control versus freedom of speech in policies and regulations targeting GD/MDM, including the development and promotion of false content including conspiracy theories, the use of concealment to evade detection, and the use of GD/MDM as part of foreign interference		
		Consider when, how, by whom and under which powers and authorities, identified information operations involving GD warrant the use of state resources to disrupt/degrade/defend forward against foreign adversaries			

System features		Influence operation kill chain components targeted by proposed system of PPT		
		Plan objectives	Deliver content	Persist in the information environment
Technology	Analytics (Detect, identify, document & track...)	Identify contexts in which GD/MDM operations commonly take place, and the harms that they seek to perpetrate (e.g., harms against individuals, groups, society); consider objectives and the probability of GD/MDM operation being underway	Identify situations where it is likely that GD/MDM is being delivered, and the formats in which it this delivery might be occurring (e.g., deepfakes)	Distinguish GD/MDM content from distracting 'noise' and direct counter measures toward the GD/MDM
	Platform tools (Identify & limit the harms of...)	Consider actors likely to be responsible for using technology platform features and operational environments (e.g., telecos) to conduct certain GD/MDM activities (where evidence exists and can be collected, document links to specific actors)	Document and disseminate information about the features (design & business) of technology platforms (news media, social media, others) that enable malicious content to achieve depth, and scale of penetration against target audiences Document and disseminate verified information about actors using GD/MDM, including foreign interference (threat intelligence sharing)	
	Transparency (Reveal and promote awareness of...)			
	Automated fact checking (Identify and support accuracy...)	Identify and share information about frequent targets of malicious information exploits and the characteristics of the exploits	Use pre-bunking (when possible) and/or de-bunking (when necessary)	

Focus re awareness and prevention of victimization

Stakeholder	Awareness
Individuals (Directly-targeted individuals, parents, partners, allies)	• If this has happened to you, or someone you know, you/they are not alone • It is not "rare" or "isolated" problem; it is prevalent in Canada and beyond (include statistics) • It is not acceptable; in some cases, it may be illegal • Harms of GD/MDM for individuals and society (e.g., coercive control, emotional/psychological abuse, discrimination, polarization, intimidation/fear, disenfranchisement) • Forms of GD/MDM and their mechanisms of action (e.g., conspiracy theories, manipulated media/deepfakes) • How GD/MDM is produced, distributed and consumed using the features and exploiting the vulnerabilities of technology platforms (e.g., design and structures of platforms that create psychological rewards, along with ease and speed, of 'likes' or re-posting • Identifying the necessary skills at each stage to counter GD/MDM
Organizations & networks (e.g., community-based agencies/NGOs, police, school boards/threat risk teams)	• GD/MDM is a form of technology facilitated violence against women and girls • It is not "rare" or "isolated" problem; it is prevalent in Canada and beyond (include statistics) • It is not acceptable; in some cases, it may be illegal • It can be part of coercive control or intimate partner violence designed to harass, threaten or intimidate those who are targets of this behaviour • It is part of a broader context of harm that oppresses and subjugates females to a male-dominated social hierarchy • It can contribute to online and physical environments that are unwelcome and/or unsafe for women and girls • In extreme cases, it may be understood as an attempt to incite similar behaviour by others, and/or physical harms against women • It can discourage the participation of women and girls in the life of Canadian society, including in positions of influence and leadership • How GD/MDM is propagated using the features and exploiting the vulnerabilities of technology platforms • Incorporate media literacy programming: Develop curricula to teach critical thinking skills and how to identify manipulated media, especially targeting women and marginalized groups (e.g., IREX's Learn to Discern (L2D) initiative builds communities' resilience against disinformation and hate speech) • Digital safety training: Provide guidance on protecting personal data and images online to prevent their use in deepfake creation. This includes establishing institutional protocols for reporting and responding to online attacks.

Stakeholder	Awareness
Organizations & networks (e.g., community-based agencies/NGOs, police, school boards/threat risk teams)	• Propose: review of current regulatory and legislative conditions that may help or hinder GD/MDM; and study of options for legal frameworks that would support enhanced action (e.g., content moderation, technical properties) by platforms, deterrence of would-be perpetrators, and accountability for both platforms and individuals • User empowerment: Engage platforms (or alternative avenues) such that users are provided with tools to control their online experience and report harmful content • Engage civil society, levels of government and political parties on creation of codes of conduct or declarations of principles for electoral periods that address gendered disinformation • Establish complaints referral and adjudication processes for gendered disinformation cases • Coordinate with social media platforms to enhance dissemination of credible information and restrict problematic content • Community based rumour management has been used by the Sentinel Project to lessen the risk of mass atrocities – attention to the role of various communities in amplifying or dampening GD may be an important component of a holistic response involving multi-stakeholder dialogue and collaboration, along with developing coordinated response networks where the risk of GD/MDM may be elevated
Government (Policy, legislation, direct action through authorized agencies)	

Focus re response capacity

Stakeholder	Capacity
Individuals (Directly-targeted individuals, parents, partners, allies)	• Identification: Critical thinking; safe use practices • Resistance: Critical thinking; fact checking; counter narrative skills, focusing on the message (ignore, evade, address – pre-bunk/de-bunk); Cambridge University's gamified training tools, such as Bad News, can help adults and youth develop skills for identifying MDM • Reporting: Organizations; law enforcement • Resilience: Dialogue and information sharing; peer support with resistance, managing impacts; formal services (e.g., women's support organizations, shelters)
Organizations & networks (e.g., community-based agencies/NGOs, police, school boards/threat risk teams)	• Prevention/risk reduction: Awareness training for employees and the people they serve; security policies and procedures (e.g., for employees, service users, students) • Identification: Technological and non-technological tools, depending on technical maturity of the organization and its mandate and authorities • Response (e.g., counter measures): Counter narrative skills, focusing on the message (ignore, evade, address – pre-bunk/de-bunk) - e.g., develop and deploy counter speech campaigns (e.g., correct, de-emphasize false gendered content) to combat gendered disinformation • Reporting: Dialogue, information sharing, collaboration, MOUs/service protocols • Explore multi-sector partnering with civil society organizations to build coalitions to enhance monitoring capabilities

Stakeholder	Capacity

Government (Policy, legislation, direct action through authorized agencies)	Regulatory and legislative options, implications, opportunities and challenges
	
	• Convening, enabling and leadership:
	  ○ Establish a national task force to study and address gendered disinformation.
	  ○ Fund creation of a consortium of NGO and academic partners to conduct research, and serve as a clearing house for information, on: the nature, prevalence and impacts of GD/MDM (independent monitoring); inoculation strategies; technological innovations; and citizen literacy programs focusing on women and girls; public policy and regulatory options; technology governance; and legal/freedom of speech issues. Use national task force to govern creation and implementation of strategic information agenda and accompanying KT/E plan.
	  ○ Create a cross-departmental committee having external representation from national task force and consortium to prioritize attention to gendered disinformation: in foreign policy; and in processes shaping technology and disinformation policy.
	  ○ Explore opportunities to enhance coordination between platforms, fact-checkers, and election authorities
	
	Legal, enforcement options
	
	• Explore and assess legal protections that may be enacted to criminalize the creation and distribution of non-consensual pornography and malicious deepfake content.

Stakeholder	Capacity

Government (Intelligence and direct action supporting national security - e.g., through prevention/response to foreign interference and transnational oppression)
 	
(Potential for collaboration with private sector organizations, where appropriate and consistent with relevant legislative authorities)
	Cyber defence capacity (strategies, actions & techniques) taken by state of organizations to protect information ecosystems from cyber threats
	
	• Defensive cyberspace operations (DCO) – broad, strategic approach:
	  ○ Proactive measures: Prevention; proactive deterrence/disruption at-source; threat identification and intelligence sharing; cross-sectoral cooperation; international cooperations and cyber diplomacy to identify, track and respond to cyber threats
	  ▪ Target actors and their infrastructures through threat intelligence, targeting, takedowns, disruption, deception, prosecution etc – by integrating the F3EAD targeting framework (Find, Fix, Finish, Exploit, Analyze and Disseminate) and the DISARM framework (Detect, Interpret, Segment, Analyze, Respond, and Mitigate) and applying these to the Influence Operation Kill Chain (Annex A)
	  ○ Reactive measures (e.g., incident response) to protect the cyber domain from a range of cyber aggressions
	
	• Internal defensive measures (IDM) within organizations or networks
	  ○ Attention to reducing vulnerabilities and strengthening systems against influence operations involving GD/MDM

Focus re organizations and networks supporting those impacted by gendered disinformation

Stakeholder	Awareness	Capacity
Organizations & networks service providers (e.g., community-based agencies/NGOs, police, school boards/threat risk teams)	Information sheet supporting awareness of GD/MDM for stimulating awareness of GD/MDM & outlining	• Prevention/risk reduction: Awareness training for employees and the people they serve; security policies and procedures (e.g., for employees, service users, students); information on potential impacts for victims of GD/MDM • Identification: Awareness training focusing on GD/MDM as technology facilitated violence against women which may, in some cases, be an instance of coercive control/intimate partner violence • Response (e.g., counter measures): Identification and availability of legal and support resources for supporting victims of violence against women towards those impacted by GD/MDM. Develop and implement institutional protocols to support those attacked and address reports of attacks. Consider and assess opportunities for collaborative approaches to providing support services for targets of broadly focused gendered disinformation campaigns. • Reporting: Dialogue, information sharing, collaboration, MOUs/service protocols, exploration of policy and legal options with police, crowns, policy actors, women's groups and others

Annex D: Curated Sample Technology Options for Individuals, Human Service (Including Police) and Educational Organizations

ANALYTICS/AUTOMATED DETECTION

Note: Weblinks have not been provided as these may change, over time. Subscription fees, where indicated are current, as of April, 2025.

1. **Sentinel Deepfake Detection System**
   - **What it does**: AI detection platform that works with governments, media, & defence agencies to protect democracies from disinformation campaigns, synthetic media & information operations.
   - **How to use it**: Users can report gendered deepfakes for review.
   - **Subscription**: No public access; used by governments, media, & defence agencies.
   - **Example**: A deepfake targeting a female journalist is flagged & removed before going viral.

2. **WeVerify, DuckDuckGoose, DeepfakeProof**
   - **What they do**: Content verification, tracking, & debunking (WeVerify); AI powered deepfake detection for images, videos, & audio (DuckDuckGoose); Helps users identify deepfakes while browsing the web (DeepfakeProof).
   - **How to use them**: Chrome Plugin (WeVerify); Upload files via a regular browser to DuckDuck Goose; As a real-time deepfake detection plugin for Chrome (DeepfakeProof).
   - **Subscription**: Free/Open source platform (WeVerify); Subscription Required (DuckDuckGoose); Free Chrome Plug-in (DeepfakeProof).
   - **Example**: A fake nude image of a female politician is detected & debunked.

3. **Reality Defender**
   - **What it does**: Equips enterprises, governments, & platforms with the tools to detect AI generated or manipulated content in real time.
   - **How to use it**: Upload content to the software for real-time video identity, image & text authentication.
   - **Subscription**: Subscription required.
   - **Example**: A fake video targeting a women's rights activist is debunked before being used in a smear campaign.

4. **MeVer: Verification, Media Analysis, & Retrieval**
   - **What it does**: Developing technologies & services for understanding, searching, & verifying media content
   - **How to use it**: Journalists & researchers analyze disinformation content & networks.

- **Subscription**: Offers resources (tools, software, & datasets) via GitHub & other repositories.
- **Example**: A smear campaign against female journalists is traced to coordinated disinformation actors.

5. **RAND's Countering Truth Decay Initiative**
   - **What it does**: RAND researchers are studying the causes, consequences, & means of countering truth decay.
   - **How to use it**: Free resource.
   - **Subscription**: Research available on RAND's website for free.
   - **Example**: A journalist or researcher may explore Truth Decay research & commentary to understand the drivers, trends, & consequences of Truth Decay as a System.

PLATFORM/CONTENT GENERATOR TOOLS

1. **SynthID (Digital Watermarking)**
   - **What it does**: Watermarks & identifies AI generated content by embedding digital watermarks directly into AI generated images, audio, text, or video.
   - **How to use it**: Integrated into AI-generated media, detected by compatible tools.
   - **Subscription**: Available via Google Cloud's AI tools (Google DeepMind).
   - **Example**: A fake image of a female CEO is debunked using SynthID detection.

TRANSPARENCY

1. **Hoaxy – Tracking Gendered Disinformation**
   - **What it can do**: Hoaxly visualizes the spread of information online using the X/Twitter & Bluesky APIs.
   - **How to use it**: An API is used to retrieve recent posts matching your search query.
   - **Subscription**: Free until Hoaxly reaches its monthly post limit, then live search is only available to users with Basic ($100/month), Pro ($5000/month), or Enterprise (price available upon request) access.
   - **Example**: Hoaxy reveals bot activity pushing a false claim against a female official.

2. **Systematic Data Collection & Reporting**
   - **What it can do**: Track trends in gendered disinformation & AI-generated attacks.
   - **How it can be used**: Governments, researchers, journalists, & civil society can utilise reports for situational awareness, policy development, & advocacy.
   - **Subscription**: Varies - there is a wide variety of open source reporting available.
   - **Example**: A media watchdog report documents rising deepfake attacks on female politicians, which provides a situational awareness on deepfake trends.

3. **Gender-Sensitive Monitoring**
   - **What it can do**: One can utilise AI tools (e.g. Reality Defender), social network analysis (Hoaxy, Never), & qualitative methods to track gendered disinformation.
   - **How it can be used**: Quantitative / qualitative research to identify gendered attacks online.
   - **Subscription**: Varies - some tools are free, others require paid access.
   - **Example**: Through gender-sensitive monitoring, a researcher is able to show that women candidates face twice as many disinformation attacks as men.

4. **Enhanced User Reporting for Harmful Content**
   - **What it can do**: Improve response time & categorization of gendered disinformation reports.
   - **How it can be used**: As a mass-reporting campaign.
   - **Subscription**: Unknown, would depend on platform implementation.
   - **Example**: A journalist targeted by deepfakes reports it to an enhanced moderation system.

5. **Global Coalition for Digital Safety (World Economic Forum)**
   - **What it can do**: Develop politics & global coordination on digital safety, including gendered disinformation.
   - **How it can be used**: Advocacy groups can engage with the coalition to push for stronger policies.
   - **Subscription**: Dependent on how the coalition is set up.
   - **Example**: An NGO joins the coalition to push for stricter deepfake detection on social media.

Annex E: List of Accompanying Knowledge Resources

E1: Tackling Online Gendered Disinformation: Educator Guide

E2a: Tackling Online Gendered Disinformation: A Family Resource

E2b: Tackling Online Gendered Disinformation: Youth Guide

E2c: Tackling Online Gendered Disinformation: Additional Resources for Educators, Families & Youth

E3: Gendered Disinformation: A Resource for Police and Human Service Agencies

E4: Knowledge Resources for Government