From: s 47F @wfanet.org>
Sent: Tuesday, 21 June 2022 8:22 PM
To: Julie Inman Grant; s 22 ; s 22 ; s 22 ; s 22
Cc: s 47F
Subject: GARM: Key updates at Y3
Attachments: GARM_3 Years of Progress_June 2022[41][90].pdf; GARM Adjacency Framework v1 17Jun22.pdf; GARM Brand Safety Floor Suitability Framework 2Jun22.pdf

Hi Julie + Team –

I wanted to reach out to you with some key updates from GARM that you should find relevant.
We're reaching our 3 year anniversary since our launch, and we've released a comprehensive update on where we've been, what's new today, what's next tomorrow.

The big news we are sharing today is:

1/ We've aligned on definition for Misinformation to integrate into the GARM Brand Safety Floor + Suitability Framework. This is a significant step and is aligned with the work for Digi and more recently with our direct involvement with the EC on their CoP on Disinformation.

2/ We've also just released the GARM Adjacency Standards Framework by which to assess the placement of ads next to safe but sensitive content which will allow for ad placement in a more consistent way

3/ We also recently released our GARM Aggregated Measurement Report in May which saw more metrics but most importantly that YouTube have had their monetization safety metrics accredited by the leading media auditor, MRC – which is a significant undertaking and achievement

We're inspired by your work and continue to reach higher based on your impact.
We're looking forward to reconnecting with you in the coming weeks to discuss local measurement – which you will see is a priority area for our next steps.

Best,

s 47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

GARM: Adjacency Standards Framework

Global Alliance for Responsible Media

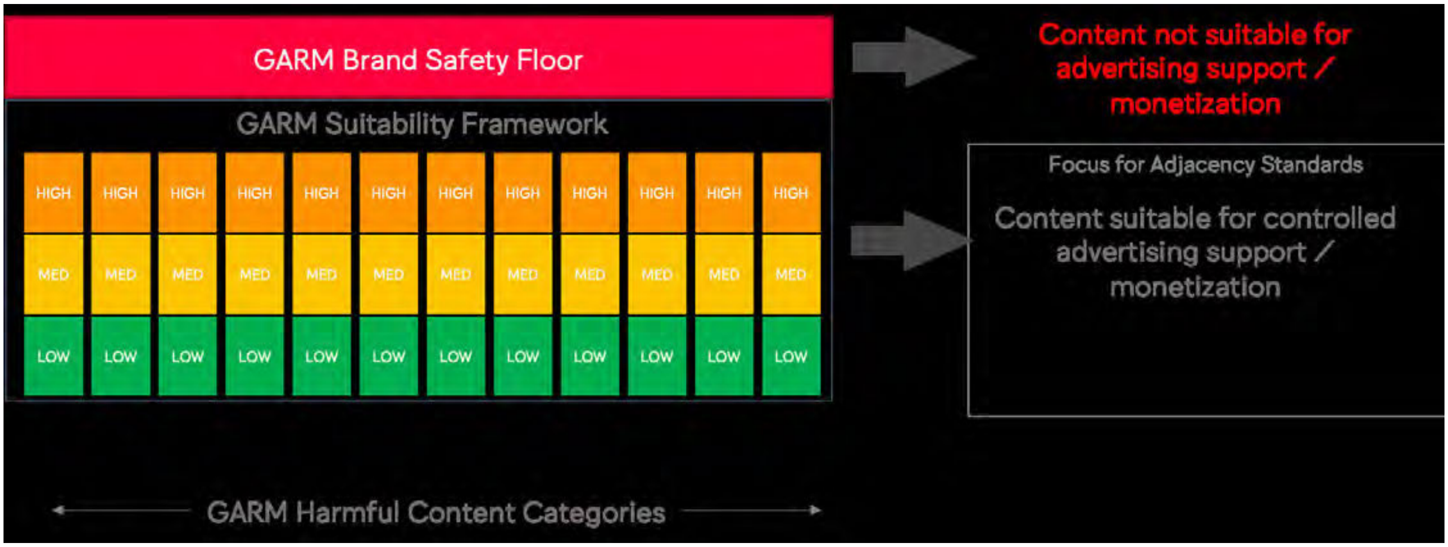
CONTEXT FOR THIS SOLUTION

The Global Alliance for Responsible Media (GARM) is an industry-first effort that unites marketers, media agencies, media platforms, and industry associations to safeguard the potential of digital media by reducing the availability and monetization of harmful content online. These steps are essential to create a safer digital media environment that enriches society through content, communications, and commerce.

In September 2020 we took our first significant step and created a solution in a common framework of shared definitions, known as the GARM Brand Safety Floor + Suitability Framework. That foundational framework set out an agreed set of sensitive content categories with different risk levels, each with monetization guidelines that range from content that is not suitable for advertising (The Brand Safety Floor) to content that is suitable to be eligible for monetization but may present varying degrees of sensitivity to the advertiser (The Suitability Framework).

The Adjacency Standards Framework is designed to serve as a companion to the GARM Suitability Framework, providing advertising industry participants with a common structure for evaluating the brand safety and suitability of an ad placement relative to an ad's position to nearby content (i.e., "adjacency") within specific media formats. The Adjacency Standards Framework works within the confines of the GARM Suitability Framework – where sensitive content may be supported by advertising.

Our belief is that greater transparency and common frameworks will allow for advertising buyers to support content more aligned with corporate, brand and campaign beliefs via paid media insertion.


GARM Brand Safety Floor
GARM Suitability Framework
HIGH HIGH HIGH HIGH HIGH HIGH HIGH HIGH HIGH HIGH HIGH HIGH
MED MED MED MED MED MED MED MED MED MED MED MED
LOW LOW LOW LOW LOW LOW LOW LOW LOW LOW LOW LOW
GARM Harmful Content Categories
Content not suitable for advertising support / monetization
Focus for Adjacency Standards
Content suitable for controlled advertising support / monetization

These standards in this framework were developed by a dedicated GARM Working Group consisting of advertisers, media agencies, media platforms, and industry associations. This Working Group formed in January 2021 and made its formal recommendation in December 2021 to the GARM Steer Team and GARM Community. The Adjacency Standards Working Group was opened to members wanting to join and help define and design the solution.

The standards are informed by a research process that considered studies spanning both observed and claimed consumer research provided by Edelman, BMW/Cheq, General Motors, Johnson & Johnson, Meta, Twitter, TikTok, and OMG. Additionally, GARM commissioned dedicated community-focused research around advertiser and media agency needs relative to adjacency controls. The following is a synopsis of our findings:

1. **Research findings on consumer impact had a wide range based on content severity:** Our review of consumer research ranged from perception-based research and behavioral impact research. The current research set available also compared harmful content which should not be monetized and sensitive and suitable content which could be monetized. Our assessment of the multiple studies' findings concluded that harmful content unfit for advertising support required greater adjacency, whereas suitable content required lesser adjacency standards. We have therefore focused our adjacency recommendations on a minimum standard for adjacency on suitable content.

2. **Advertisers and Agencies desire comparable thresholds for formats across platforms:** In our research within the GARM Community, nearly 9-in-10 media buyers expressed a need for cross-platform format consistency[1]; advertisers and agencies wanted a single standard. When asked for priority ordering of formats, the Community prioritized all formats with the following: Audio, Video, Livestreaming, and Feed[2].
3. **Advertisers and agencies are increasingly seeing adjacency and content targeting essential to their brand safety and suitability strategies:** Nearly 9-in-10 buyers say that suitability and adjacency controls are Very or Extremely Important to their operations, and nearly 2-in-3 advertising buyers would invest in platforms where the controls are more readily available.

GOALS FOR SOLUTION

This shared framework will provide individual GARM participants with:

Consistent Units of Measure	Ensuring that there's a common approach to evaluating an ad placement's position relative to content above the floor within the Suitability Framework categories
Standardized Expectations on Ad Placements	Creating industry-standard, format-centric specifications to inform the development of relevant ad solutions across media environments
Improved Transparency	Establishing a deeper framework by which to report accuracy of ad placements for advertiser, agency, platform and ad tech solutions providers

HOW THIS SOLUTION WILL BE USED

- GARM Community Member organizations will endorse this adjacency standards framework as a minimum for placement, and a starting point for post-campaign measurement
- Platforms and ad tech providers will adopt and operationalize these standards through practices and solutions as they become available
- GARM Working Group Leaders and the GARM Steer Team will work with GARM member platforms and providers to track the adoption of adjacency standards via solutions in a shared framework
- GARM will work with industry auditing bodies like the MRC to incorporate adjacency standards into existing accreditation processes where appropriate
- GARM will work via the Solutions Developers Working Group to integrate Adjacency Standards into existing post-campaign verification services

[1] GARM Community Adjacency Needs Research Study, July 2021

[2] Ibid

GARM: Adjacency Standards Framework

The following table is the approved adjacency standards for GARM. These standards will then be utilized in solutions by platforms (first party tools) and via independent providers who have integrations with media platforms and publishers (third party tools)

Some key terms and notes on this framework:

1. FORMATS COVERED: The formats identified and included here are based on current media format availabilities and investment levels: Feed, Stories, Video In-Stream, Audio In-Stream
   - NOTE: Livestream Audio and Livestream video are currently omitted from the current version of the framework. The Working Group will require further exploration into the technical and operational complexities of this format, given high-profile incidents, before developing formalized adjacency standards in a future update. This update will address how the Safety Floor is upheld and the technical implementation of the Suitability Framework.
2. STANDARD: Denotes how ad adjacency is evaluated within respective Format environments, and at current is based on "spatial" evaluation (e.g. units of space between ad and content on a screen) or temporal evaluation (e.g. units of sequence in which ads and content appear on-screen)"
3. ADJACENCY UNIT: This is the actual "unit of measurement" proposed to identify what content should be evaluated and categorized based relation to an ad's placement. This is treated an industry standard and a minimum upon which providers and platforms can provide additional spacing/separation should they need to.

FORMAT		ADJACENCY STANDARD	MINIMUM ADJACENCY UNIT	NOTES
Feed	This covers content (text, video, image, audio) that is featured in a newsfeed or timeline environment irrespective of the screen being mobile or computer.	Spatial	+/- 1	• Adjacency controls should apply to individual and group feed and timeline (e.g., lists and groups that are public and/or private) • Comments on videos are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Stories	This covers sequenced content from a single creator in a carousel environment, where ads may appear within or between such segments.	Spatial	+/- 1	• Comments on stories are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Video: In-stream	This is prerecorded video content that is uploaded to a website or platform that features ads before, in between or after specified video content	Temporal	+/- 0 Directly Adjacent	• Adjacency solutions and controls should apply to Pre/Mid/Post and Parallel ad units • Comments on videos are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Audio: In-stream	This is pre-recorded audio content that is uploaded to a website or platform that features ads before, in between or after specified video content.	Temporal	+/- n Same as ad unit length (n = ad length)	• Adjacency controls should apply to Pre/Mid/Post and Parallel ad units • Comments on content are not a focus for suitability controls – but platform must be able to uphold Floor on audio content

GARM: Brand Safety Floor + Suitability Framework

CONTEXT FOR THIS SOLUTION

The Global Alliance for Responsible Media (GARM) is an industry first effort that unites marketers, media agencies, media platforms, industry associations, and advertising technology solutions providers to safeguard the potential of digital media by reducing the availability and monetization of harmful content online. These steps are essential to create a safer digital media environment that enriches society through content, communications, and commerce. Harmful content and its creators threaten the potential for digital media and disrupt the connections everyone seeks. Our first step in safeguarding the positive potential for digital is to provide platforms, agencies, and marketers with the framework with which to define safe and harmful content online.

Our position is that you cannot address the challenge of harmful online content if you are unable to describe it using consistent and understandable language.

The GARM has developed and will adopt common definitions to ensure that the advertising industry is categorizing harmful content in the same way across the board. These eleven key categories have been identified in consultation with experts from GARM's NGO Consultative Group. Establishing these standards is the essential foundation needed to stop harmful content from being monetised through advertising. Individual GARM members will adopt these shared principles in their operations, whether they are a marketer, agency, or media platform.

We fundamentally believe that, together, these definitions are the cornerstone for us to find balance between supporting responsible speech, bolstering public safety, and providing for responsible marketing practices. With this framework of consistent categories in place, we will be able to improve transparency in the availability, monetization, and inclusion of content within advertising campaigns. This is essential to help platforms, agencies, and advertisers make decisions essential to the advertising industry.

In November 2019, the GARM initiated work towards this challenge under a working group focused on advancing shared language and standards for advertising & media (as seen in our GARM Charter here). The output of this work is the following:

1. A common understanding of what harmful and sensitive content is via content categories
2. A common understanding of where ads should not appear, as expressed in a Brand Safety Floor
3. A common way of delineating different risk levels for sensitive content, as expressed in a Brand Suitability Framework

The output of the work is a framework of Shared Definitions that sets the limits for monetization of harmful content in agreed upon categories. This work, the GARM Brand Safety Floor + Suitability Framework was first published in September 2020.

In June 2021, we began work to update the framework to include Misinformation as an additional harmful content category. This important addition builds upon individual GARM member work, GARM member collaboration with regulatory and NGO bodies, and more recently GARM collaboration with the European Commission on the Code of Practice on Misinformation.

GOALS FOR SOLUTION

This shared framework, which is activated by the IAB TechLab's industry-wide taxonomy, will provide individual GARM participants with:

A Consistent Categorization	Ensuring that there's a common way to categorize sensitive content
Transparency	Creating transparency for industry participants on where sensitive content may be present in the interest of consumer safety and responsible marketing
Clarity in Exceptions	Establishing a method for platforms to report on special exception cases in the interest of responsible speech and public interest

HOW THIS SOLUTION WILL BE USED

- Platforms will adopt, operationalize and continue to enforce monetization policies with a clear mapping to GARM brand suitability framework

- Platforms will leverage their community standards and monetization policies to uphold the GARM brand safety floor
- Advertising technology providers will adopt and integrate GARM definitions into targeting and reporting services via clear mapping or overt integration
- Agencies will leverage the framework to guide how they invest with platforms at the agency-wide level and at the individual campaign level
- Marketers will use the definitions to set brand risk and suitability standards for corporate, brand and campaign levels

GARM: Brand Safety Floor + Suitability Framework

GARM Global Alliance for Responsible Media

CONTENT CATEGORY	BRAND SAFETY FLOOR – Content not appropriate for any advertising support
Adult & Explicit Sexual Content	• Illegal sale, distribution, and consumption of child pornography • Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated
Arms & Ammunition	• Promotion and advocacy of Sales of illegal arms, rifles, and handguns • Instructive content on how to obtain, make, distribute, or use illegal arms • Glamorization of illegal arms for the purpose of harm to others • Use of illegal arms in unregulated environments
Crime & Harmful acts to individuals and Society, Human Right Violations	• Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity – Explicit violations/demeaning offenses of Human Rights (e.g. human trafficking, slavery, self-harm, animal cruelty etc.), • Harassment or bullying of individuals and groups
Death, Injury or Military Conflict	• Promotion, incitement or advocacy of violence, death or injury • Murder or Willful bodily harm to others • Graphic depictions of willful harm to others • Incendiary content provoking, enticing, or evoking military aggression • Live action footage/photos of military actions & genocide or other war crimes
Online piracy	• Pirating, Copyright infringement, & Counterfeiting
Hate speech & acts of aggression	• Behavior or content that incites hatred, promotes violence, vilifies, or dehumanizes groups or individuals based on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status, or serious disease sufferers.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	• Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	• Promotion or sale of illegal drug use – including abuse of prescription drugs. Federal jurisdiction applies, but allowable where legal local jurisdiction can be effectively managed • Promotion and advocacy of Tobacco and e-cigarette (Vaping) & Alcohol use to minors
Spam or Harmful Content	• Malware/Phishing
Terrorism	• Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society
Debated Sensitive Social Issue	• Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite greater conflict;
Misinformation	• Misinformation is defined as the presence of verifiably false or willfully misleading content that is directly connected to user or societal harm

Brand Suitability Framework: Sensitive content appropriate for advertising supported by enhanced advertiser controls

CONTENT CATEGORY	High Risk	Medium Risk	Low Risk
Adult & Explicit Sexual Content	• Suggestive sexual situations requiring adult supervision/approval or warnings • Full or liberal Nudity	• Dramatic depiction of sexual acts or Sexuality issues presented in the context of entertainment • Artistic Nudity	• Educational, Informative, Scientific treatment of sexual subjects or sexual relationships or sexuality
Arms & Ammunition	• Glamorization /Gratuitous depiction of illegal sale or possession of Arms • Depictions of sale/use/distribution of illegal arms for inappropriate uses//harmful acts	• Dramatic depiction of weapons use presented in the context of entertainment • Breaking News or Op-Ed coverage of arms and ammunition	• Educational, Informative, Scientific treatment of Arms use, possession or illegal sale • News feature stories on the subject
Crime & Harmful acts to individuals and Society, Human Right Violations	• Depictions of criminal/harmful acts or violation of human rights	• Dramatic depiction of criminal activity or human rights violations presented in the context of entertainment • Breaking News or Op-Ed coverage of criminal activity or human rights violations	• Educational, Informative, Scientific treatment of crime or criminal acts or human rights violations • News feature stories on the subject
Death, Injury or Military Conflict	• Depiction of death or Injury • Insensitive and irresponsible treatment of military conflict, genocide, war crimes, or harm resulting in Death or Injury • Depictions of military actions that glamorize harmful acts to others or society	• Dramatic depiction of death, injury, or military conflict presented in the context of entertainment • Breaking News or Op-Ed coverage of death, injury or military conflict	• Educational, Informative, Scientific treatment of death or injury, or military conflict • News feature stories on the subject
Online piracy	• Glamorization /Gratuitous depiction of Online Piracy	• Dramatic depiction of Online Piracy presented in the context of entertainment • Breaking News or Op-Ed coverage of Online Piracy	• Educational, Informative, Scientific treatment of Online Piracy • News feature stories on the subject
Hate speech & acts of aggression	• Depiction or portrayal of hateful, denigrating, or inciting content focused on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status or serious disease sufferers, in a non-educational, informational, or scientific context	• Dramatic depiction of hate speech/acts presented in the context of entertainment • Breaking News or Op-Ed coverage of hate speech/acts	• Educational, Informative, Scientific treatment of Hate Speech • News features on the subject
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	• Glamorization /Gratuitous depiction of profanity and obscenity	• Dramatic depiction of profanity and obscenities presented in the context of entertainment by genre • Breaking News or Op-Ed coverage of profanity and obscenities Genre based use of profanity, gestures, and other actions that may be strong, but might be expected as generally accepted language and behavior	• Educational or Informative, treatment of Obscenity or Profanity • News feature stories on the subject
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	• Glamorization /Gratuitous depictions of illegal drugs/abuse of prescription drugs • Insensitive and irresponsible content/treatment that encourages minors to use tobacco and vaping products & Alcohol	• Dramatic depiction of illegal drug use/prescription abuse, tobacco, vaping or alcohol use presented in the context of entertainment • Breaking News or Op-Ed coverage of illegal drug use/prescription abuse, tobacco, vaping or alcohol use	• Educational, Informative, Scientific treatment of illegal drug use/prescription abuse, tobacco, vaping or alcohol • News feature stories on the subject
Spam or Harmful Content	• Glamorization /Gratuitous depiction of Online Piracy	• Dramatic depiction of Spam or Malware presented in the context of entertainment • Breaking News or Op-Ed coverage of Spam or Malware	• Educational, Informative, Scientific treatment of Spam or Malware • News feature stories on the subject

CONTENT CATEGORY	High Risk	Medium Risk	Low Risk
Terrorism	• Depiction of terrorist actions that are disturbing, agitating or promotes harmful acts to others or society • Terrorist content requiring a viewer advisory • Insensitive and irresponsible treatment of terrorism/ related crimes	• Dramatic depiction of terrorism presented in the context of entertainment • Breaking News or Op-Ed coverage of acts of terrorism	• Educational, Informative, Scientific treatment of terrorism • News feature stories on the subject
Debated Sensitive Social Issue	• Depiction or discussion of debated social issues and related acts in negative or partisan context	• Dramatic depiction of debated social issues presented in the context of entertainment • Breaking News or Op-Ed coverage of partisan advocacy of a position on debated sensitive social issues	• Educational, Informative, Scientific treatment of debated sensitive social issues and related acts including misinformation • News feature stories on the subject
Misinformation	• Glamorization/Gratuitous depiction of misinformation	• Dramatic depiction of misinformation presented in the context of entertainment • Breaking News or Op-Ed coverage of misinformation	• Educational, Informative, Scientific treatment of misinformation. • News features describing various misinformation campaigns as such

GARM Global Alliance for Responsible Media

GARM: 3 Years of Progress

Uncommon Collaboration and its Impact on Brand Safety

A look back and reflecting on our journey

Three years ago, the WFA led 16 organizations in the launch of the Global Alliance for Responsible Media (GARM) at the Cannes Lions Festival of Creativity. GARM is a cross-industry initiative, launched by brands, to remove harmful content from ad-supported digital and social media. That challenge is big, however GARM has been able to develop a unique structure with focused areas for collaboration that have driven meaningful impact.

GARM's launch was propelled forward by *uncommon collaboration,* a unique way of working recognizing that all sectors of the advertising industry and companies benefit from partnering to create new brand safety standards and solutions that could be accepted industry-wide, where there had been no established protocols.

From an initial set of 16 launch companies, we've grown to 122 members (61 advertisers, six agency holding companies, 11 media platforms, nine ad tech companies, and 35 industry associations). GARM has six active working groups, powered by 80 media leaders from a larger GARM Community. The Community meets monthly to review the progress of Working Groups and share best practices and thought leadership.

Along with our launch focus of advertisers, agencies, platforms and industry associations, we've set focused areas for engagement:

- Bringing advertising technology companies into a Solutions Developers Working Group to help us drive consistency in implementation and faster speed to market
- Broadening membership to National Advertiser Associations to help us embed GARM work at the local level
- Formalizing our ways of working with NGOs via a consult group to ensure our work isn't insular to the advertising industry

A view on digital media safety: our right to play & win

Through the work of the GARM Steer Team, we've brought into focus how digital media safety should be viewed holistically. The GARM Steer Team's view on digital media safety is framed by two broad questions:

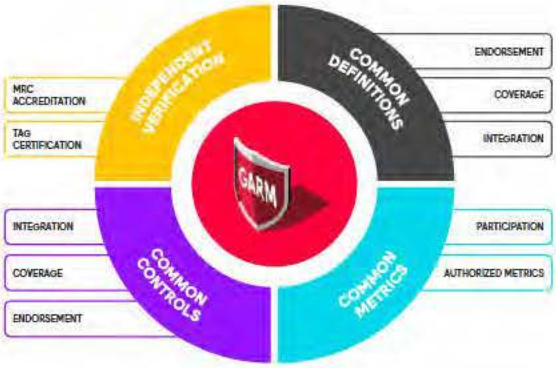
Is the platform safe for consumers?	Does the service cause harm? Is engagement positive? Is support universal? Is moderation fair? Is there choice and control?	Place human safety at the core to eliminate harm from the attention flywheel	PLATFORM SAFETY
Is the platform suitable for advertising?	Is content categorised? Are there placement controls? Is there transparency in where ads appeared?	Make digital media safety practices analogous to other media	BRAND SAFETY

The first area of consideration is **Platform Safety**, which focuses on the product, essential technology design and oversight. These are individual media platform responsibilities that sit well outside advertiser control and competency. These also raise questions that are fundamentally local and regulatory in nature. Areas outside of advertiser control are content legality, freedom of expression and algorithmic oversight. These are areas where we need regulators to step in and define the right balance between consumer protection and freedom of expression. Further, advertisers don't have full competency in areas like technology design and algorithms where we need technical regulation as seen in markets like the EU via the Digital Services and Digital Markets Acts and in Australia with the eSafety Commissioner.

Advertisers have had a clear stake in **Brand Safety**, and with GARM we are able to overcome fragmentation within the advertiser, agency and platform communities to set a direction forward. This has enabled us to effectively establish safety standards for where ads show up. Advertiser standards on content safety work best for our industry applied at a global level to drive impact, and GARM's global scope of influence positions us to continue leading here.

GARM's focus remains on Brand Safety, which is centered on monetization safety – how advertising investment is steered away from harmful content and behaviors. Since our start, we've been able to drive positive, forward momentum around four core areas we set out in the GARM Charter, launched at the WEF Davos Summit in January 2020. They are:

- Common Definitions
- Common Metrics
- Common Tools
- Independent Verification



With that understanding of where we have direct impact on Brand Safety, and indirect influence on Platform Safety – our hypothesis is that transparency, control and accountability will better allow the advertising industry to reward positive content and engagement.

All our solutions are meant to improve upon that –brands should be able to invest in content that aligns with their values and purpose, and with proper tools and partnerships in place there should be no surprises.

Common Definitions: Setting the Limits for Advertising Support

What we've delivered so far – The GARM Brand Safety Floor + Suitability Framework

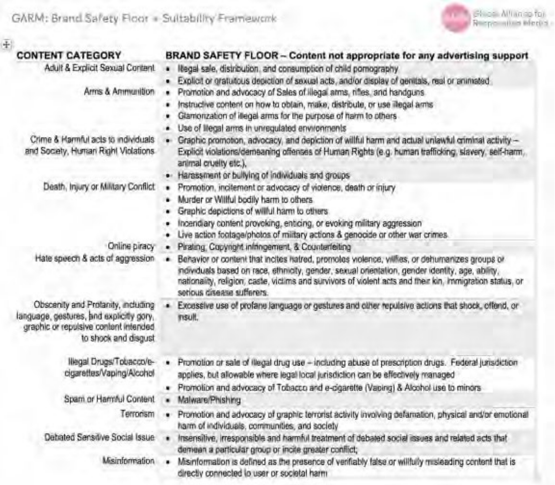
Eliminating harmful content from advertising campaigns requires a shared understanding of what are sensitive topics, and what are the limits for advertising support. Prior to our work in this area, platforms, advertisers and agencies had separate views and vocabulary on harmful content. In September 2020, GARM started with the 4A's APB initial proposals and enhanced them through a multistakeholder process and in collaboration with GARM's NGO Consult Group. From there, GARM drove an agreement across advertisers, agencies and platforms to set a framework that limits advertising support for harmful content, through the Brand Safety Floor, while providing for a Suitability Framework to manage advertising placement in sensitive content categories, while acknowledging critical nuances across platforms and formats.

What's new today – GARM-inspired ad tech solutions and a new category for Misinformation

Two years from the launch of the GARM Brand Safety Floor + Suitability Framework, we are proud to show that this solution is impacting how brands set strategies, how media agencies build media buys, and how platforms and ad tech partners structure their tools. Because of this work, advertisers have more tangible control and transparency over monetized content. We are also proud to announce that we have added Misinformation into this framework, in coordination with our work with the European Commission and several of our NGO partners. This important work helps to solidify individual GARM member work and multistakeholder collaboration with regulators. The new standard is designed to provide a structure for demonetizing harmful misinformation and to build on the success that the framework has already delivered.

What's next tomorrow – Democratizing application in local markets

As noted, much of our work is global in nature. However, we recognize that this framework in the hands of more advertising buyers and sellers will fundamentally further the demonetization of harmful content through simple demand and supply marketplace dynamics. To that end, we are working through WFA's National Associations Council to help translate, calibrate and embed it via local educational efforts with national advertiser associations.



GARM: Brand Safety Floor + Suitability Framework

Global Alliance for Responsible Media

CONTENT CATEGORY	BRAND SAFETY FLOOR – Content not appropriate for any advertising support
Adult & Explicit Sexual Content	• Illegal sale, distribution, and consumption of child pornography • Explicit or gratuitous depiction of sexual acts, and/or display of genitals, real or animated
Arms & Ammunition	• Promotion and advocacy of Sales of illegal arms, rifles, and handguns • Instructive content on how to obtain, make, distribute, or use illegal arms • Glamorization of illegal arms for the purpose of harm to others • Use of illegal arms in unregulated environments
Crime & Harmful acts to individuals and Society, Human Right Violations	• Graphic promotion, advocacy, and depiction of willful harm and actual unlawful criminal activity – Explicit violations/demeaning offenses of Human Rights (e.g. human trafficking, slavery, self-harm, animal cruelty etc.), • Harassment or bullying of individuals and groups
Death, Injury or Military Conflict	• Promotion, incitement or advocacy of violence, death or injury • Murder or Willful bodily harm to others • Graphic depictions of willful harm to others • Incendiary content provoking, enticing, or evoking military aggression • Live action footage/photos of military actions & genocide or other war crimes
Online piracy	• Pirating, Copyright infringement, & Counterfeiting
Hate speech & acts of aggression	• Behavior or content that incites hatred, promotes violence, vilifies, or dehumanizes groups or individuals based on race, ethnicity, gender, sexual orientation, gender identity, age, ability, nationality, religion, caste, victims and survivors of violent acts and their kin, immigration status, or serious disease sufferers.
Obscenity and Profanity, including language, gestures, and explicitly gory, graphic or repulsive content intended to shock and disgust	• Excessive use of profane language or gestures and other repulsive actions that shock, offend, or insult.
Illegal Drugs/Tobacco/e-cigarettes/Vaping/Alcohol	• Promotion or sale of illegal drug use – including abuse of prescription drugs. Federal jurisdiction applies, but allowable where legal local jurisdiction can be effectively managed • Promotion and advocacy of Tobacco and e-cigarette (Vaping) & Alcohol use to minors
Spam or Harmful Content	• Malware/Phishing
Terrorism	• Promotion and advocacy of graphic terrorist activity involving defamation, physical and/or emotional harm of individuals, communities, and society
Debated Sensitive Social Issue	• Insensitive, irresponsible and harmful treatment of debated social issues and related acts that demean a particular group or incite greater conflict;
Misinformation	• Misinformation is defined as the presence of verifiably false or willfully misleading content that is directly connected to user or societal harm

Common Metrics: Tracking Industry Efforts

What we've delivered so far – The GARM Aggregated Measurement Report

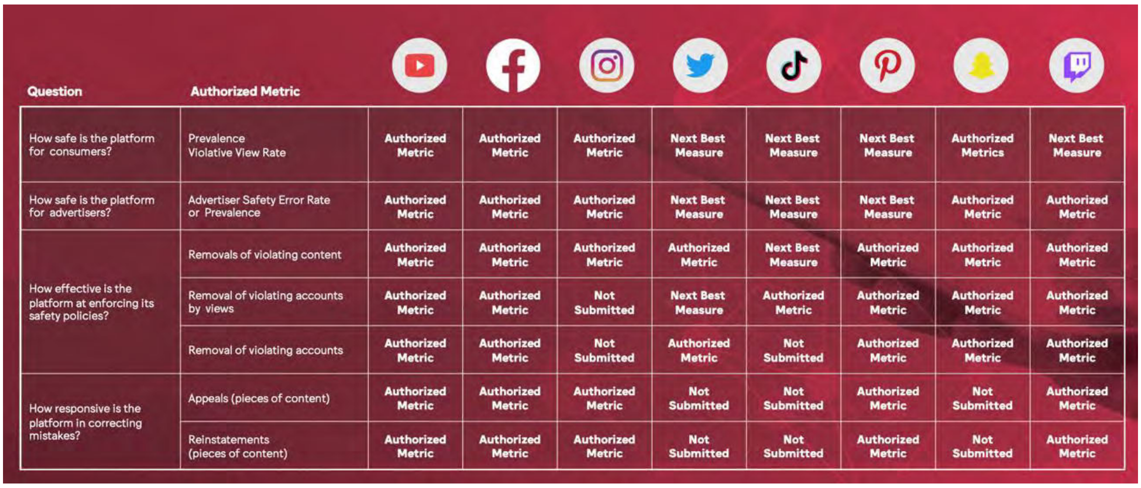
Following our agreements on Definitions, our focus turned to driving transparency through tracking industry progress in removing harmful content from advertising. In April 2021, GARM developed a common framework to assess industry progress in removing harmful content from advertising-supported media. In partnership with seven member platforms we defined four core questions and eight authorized metrics to drive transparency for the advertising industry. As a result of the GARM Aggregated Measurement Report, new metrics have been shared that have never been available before ranging from YouTube's Violative View Rate and Advertising Safety Error Rate to Meta's reporting on Prevalence of Hate Speech to Snap's Violative View Rate to Pinterest's reporting on removal of Misinformation content by views and to Twitter's sharing of Violative Impressions Rate.

What's new today – Increased participation and first accreditations

Since the launch of this effort, we've had more platforms join, and the measurement best practices and authorized metrics are shaping new work from new platforms. We've also been consistent in our call to have independent verification of metrics for monetization and transparency reporting. We're pleased to share that platforms are responding to our calls to have their monetization safety metrics audited. YouTube is the only platform at present to have their monetization safety metrics accredited by the MRC – this is a significant independent verification of the platform's safety for advertising and should help improve the confidence in the safety of YouTube's monetized content. We urge other platforms to follow. We also note that Meta have completed a first-party audit via EY of their transparency reporting that helps attest to the accuracy of their own processes.

What's next tomorrow – Increased disclosure for local markets

What's next in this area is taking the global metrics we have and gaining regional and language-level insights. We must help the industry go beyond global understandings to local trust-building transparency. Leading platforms are already exploring how to provide more specific metrics and an overview of global sampling methodologies, and we anticipate being able to report out a roadmap with Volume 5 of the report in less than a year from now. We also want to leverage work already done by platforms with regulators and NGOs to help understand safety incidents with a pertinent lens for advertisers: how many people were reached by the harmful content and was it supported by advertising?



Question	Authorized Metric								
How safe is the platform for consumers?	Prevalence Violative View Rate	Authorized Metric	Authorized Metric	Authorized Metric	Next Best Measure	Next Best Measure	Next Best Measure	Authorized Metrics	Next Best Measure
How safe is the platform for advertisers?	Advertiser Safety Error Rate or Prevalence	Authorized Metric	Authorized Metric	Authorized Metric	Next Best Measure	Next Best Measure	Next Best Measure	Authorized Metric	Authorized Metric
How effective is the platform at enforcing its safety policies?	Removals of violating content	Authorized Metric	Authorized Metric	Authorized Metric	Authorized Metric	Next Best Measure	Authorized Metric	Authorized Metric	Authorized Metric
	Removal of violating accounts by views	Authorized Metric	Authorized Metric	Not Submitted	Next Best Measure	Authorized Metric	Authorized Metric	Authorized Metric	Authorized Metric
	Removal of violating accounts	Authorized Metric	Authorized Metric	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric	Authorized Metric	Authorized Metric
How responsive is the platform in correcting mistakes?	Appeals (pieces of content)	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric
	Reinstatements (pieces of content)	Authorized Metric	Authorized Metric	Authorized Metric	Not Submitted	Not Submitted	Authorized Metric	Not Submitted	Authorized Metric

Common Tools: Driving widescale safety for ad placement

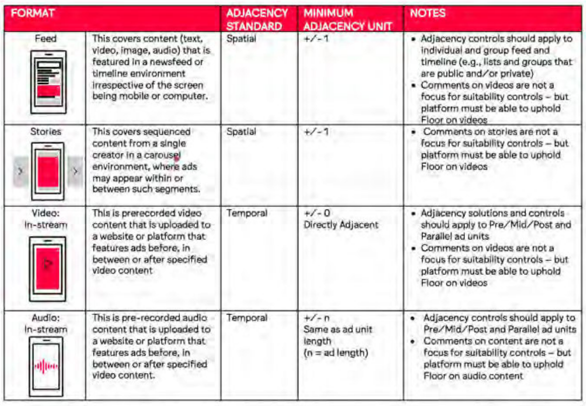
What's new today – The GARM Adjacency Standards Framework

In January 2021 we started work on the GARM Adjacency Standards Framework via a dedicated Working Group. Designed to serve as a companion to the GARM Suitability Framework, these standards provide advertising industry participants with a common methodology for evaluating the brand suitability of an ad placement relative to an ad's position to nearby content (i.e. "adjacency") within specific media formats. The Adjacency Standards Framework works within the confines of the GARM Suitability Framework – where sensitive content can be supported by advertising, and effectively allows advertising buyers and advertising sellers more consistency and control over placements relative to sensitive content.

This Working Group assessed a series of existing tools and consumer research, and also conducted research within the GARM Community to assess needs across advertisers and agencies. We're happy to share that we have now defined adjacency standards for News Feed, Stories, In-stream Video, In-stream Audio, and Display overlay. These are minimum standards and we are eager to see the industry take these up as a means of managing brand suitability. Given the recent incidents involving Livestream formats, we are working internally to ensure that there is a robust Safety Floor in place, before advancing those formats into a Suitability Framework.

What's next – Moving from Adjacency Standards and building toward Controls

With the standards for many formats defined, and livestream as a fast follower, this GARM Working Group will then pivot to working together to educate the marketplace on current and planned development. This Working Group will be collaborating with platforms and ad tech partners take these standards and putting them into practice. Once they are available, GARM's goal is to have them assessed as part of existing independent audits.



FORMAT		ADJACENCY STANDARD	MINIMUM ADJACENCY UNIT	NOTES
Feed	This covers content (text, video, image, audio) that is featured in a newsfeed or timeline environment irrespective of the screen being mobile or computer.	Spatial	+/-1	• Adjacency controls should apply to individual and group feed and timeline (e.g., lists and groups that are public and/or private) • Comments on videos are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Stories	This covers sequenced content from a single creator in a carousel environment, where ads may appear within or between such segments.	Spatial	+/-1	• Comments on stories are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Video: In-stream	This is prerecorded video content that is uploaded to a website or platform that features ads before, in between or after specified video content	Temporal	+/-0 Directly Adjacent	• Adjacency solutions and controls should apply to Pre/Mid/Post and Parallel ad units • Comments on videos are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Audio: In-stream	This is pre-recorded audio content that is uploaded to a website or platform that features ads before, in between or after specified video content.	Temporal	+/-n Same as ad unit length (n = ad length)	• Adjacency controls should apply to Pre/Mid/Post and Parallel ad units • Comments on content are not a focus for suitability controls – but platform must be able to uphold Floor on audio content

Independent Verification: Building trust in process and operations

What's new today – More platforms with more accreditations

Digital platforms have become the cornerstones of advertising. We must have trust and transparency at the core of the business, especially in safety. We've aligned two audit standards – TAG Brand Safety Certification which looks at process, and MRC Content Level Brand Safety Accreditations which looks at definitions, implementation, and reporting. We're pleased to share that all GARM platforms are TAG-certified. We are happy to see that YouTube has led the way in being the only platform at present to have earned MRC's Content Level Brand Safety Controls certification – inclusive of operations and reporting on monetization. We are encouraged by public commitments and continued steps made by other platforms like Meta and Twitter.

What's next – Getting Local on Independent Verification Results

We know that brand safety is rooted in culture, which is why it is important to take the results of global audits and understand them at a local and language level. Again, this is something we will do with WFA's National Associations Council. This will allow us to better line up advertising growth at a local market-level and safety for users. We believe that as markets develop for advertising revenue, there should be commensurate investments into safety.

GARM's go-forward

Demonetizing harmful content online is a big challenge, and content continually shifts as the culture evolves. We are starting to see impact in our work, and we do believe that GARM is an effective forum to address brand safety. And we need to acknowledge two provocations:

Provocation 1: Brand Safety isn't a substitute for Platform Safety

Platforms have become a mainstay in consumers' lives and in the industry. However, platforms have increasingly been forced to make hard decisions on freedom of expression, consumer safety and technology transparency. We are supportive of regulation that works both in the interest of society and in the interest of industry to set respective floors and the right duty of care that all users deserve, creating robust and consistent thresholds for platform safety. We have been, and continue to be, supportive of progressive, comprehensive regulation that creates a duty of care around design, resourcing, oversight, and moderation. Freedom of expression and safety are not mutually exclusive. Similarly, while we recognize that regulators want to avoid stifling development of new platforms for safety requirements, every user's safety should be protected. We believe this helps platforms have common ground on safety requirements, recognizing they operate in a competitive arena.

Provocation 2: We must get the future right by being proactive

GARM was created in reaction to a lack of effective and holistic safeguards that inadvertently had advertising funding harmful content. With GARM's work underway and impact being seen, we must now help the industry understand safety requirements before commercialization begins in the metaverse. We're being asked by our members to start on this journey, and we're here in Cannes to recommit to our mission and renew it and scale it to new spaces. We must ensure that advertising is aligned with sustainable and responsible growth models. We anticipate sharing a plan for this in conjunction with our work with regulators, our NGO Consult Group, and partners like the World Economic Forum. We are challenging ourselves to share progress towards a framework in January 2023.

While we are proud to share our progress three years on, we are clear; our impact shouldn't distract us from the challenge and opportunities at hand. It is our intent to build upon our accomplishments to date and this truly uncommon coalition to scale our efforts to engage our partners locally and define proactive practices that will future proof our industry for the good of the digital media ecosystem and our society.

To our members and our supporters – we are humbled by your commitment and contributions to our work.

To those interested in joining GARM, we invite you to join a community of the committed and like-minded.

Contact information

s 47F

[redacted]@wfanet.org

About the World Federation of Advertisers

The World Federation of Advertisers (WFA) is the voice of marketers worldwide, representing 90% of global marketing communications spend – roughly US$900 billion per annum – through a unique, global network of the world's biggest brand owners and national advertiser associations in more than 60 markets. WFA champions more effective and sustainable marketing communications. More information at wfanet.org.

About the Global Alliance for Responsible Media

The Global Alliance for Responsible Media (GARM) was formed to identify specific collaborative actions, processes and protocols for protecting consumers and brands from safety issues. Alliance members will work collaboratively to identify actions that will better protect consumers online, working towards a media environment where hate speech, bullying and disinformation is challenged, where personal data is protected, and used responsibly when given, and where everyone is, especially children, better protected online. Alliance members acknowledge their collective power to significantly improve the health of the media ecosystem. More information at wfanet.org/GARM.

Note: All WFA benchmarks, survey results, agendas and minutes are reviewed by King & Spalding, our competition lawyers.

WFA Competition law compliance policy

The purpose of the WFA is to represent the interests of advertisers and to act as a forum for legitimate contacts between members of the advertising industry. It is obviously the policy of the WFA that it will not be used by any company to further any anti-competitive or collusive conduct, or to engage in other activities that could violate any antitrust or competition law, regulation, rule or directives of any country or otherwise impair full and fair competition. The WFA carries out regular checks to make sure that this policy is being strictly adhered to. As a condition of membership, members of the WFA acknowledge that their membership of the WFA is subject to the competition law rules and they agree to comply fully with those laws. Members agree that they will not use the WFA, directly or indirectly, (a) to reach or attempt to reach agreements or understandings with one or more of their competitors, (b) to obtain or attempt to obtain, or exchange or attempt to exchange, confidential or proprietary information regarding any other company other than in the context of a bona fide business or (c) to further any anti-competitive or collusive conduct, or to engage in other activities that could violate any antitrust or competition law, regulation, rule or directives of any country or otherwise impair full and fair competition.

From: s 47F @wfanet.org>
Sent: Thursday, 23 June 2022 1:35 PM
To: Julie Inman Grant; s 22 ; s 22 ; s 22 ; s 22
Cc: s 47F ; s 47F s 47F s 47F s 22
Subject: Re: GARM: Key updates at Y3 [SEC=OFFICIAL]

Categories: EM Actioning

Hi there –

Really great to hear from you so quickly. s 22 and I are searching for a time for us to reconvene.

1/ On Misinfo – yes – Digi was actually a starting point for negotiations 😀. We also ran some sessions with some platforms on Algo Transparency and I know Christchurch are too. I will assemble my notes for cross-comparison once we are thru with all platforms as we have more to hear from, and will be happy to share insights with you.

2/ on Local – yes – this will be a big priority for us in 2H 2022 and we'd like to get a good consult with you on your work and needs and make sure we take other key market needs into a 'best of breed' local reporting request and drive that from the global center in service of the markets – and having local regulatory backing would help force negotiations over the line.

3/ on Cascading – wow I am very humbled – that would be great and perhaps something we get s involved in too! We are developing training for November – so it's work planned already!

4/ on Notices – yes – this is a good heads up. Let's figure out how we can stay close to this as it will help some of the leaders in GARM better understand and have a structured view on 'implementation' versus just headlines on incidents.

Best as always,

s
47F
s 47F
Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant
Date: Wednesday, June 22, 2022 at 10:49
To: s 47F , s 22 , s 22 , s 22 s 22
Cc: s 47F , s 47F , s 47F , s 47F , s 22
Subject: RE: GARM: Key updates at Y3 [SEC=OFFICIAL]

s 47F

Thanks very much for these updates, and all of your pioneering work through GARM. Thanks also for your kind comments on LinkedIn! I know the team had a helpful call with you a few weeks ago on Safety by Design and the Basic Online Safety Expectations, both of which we are really keen to continue working with you on. A couple of quick comments on your helpful updates:

1. Great to hear about consistent definitions on mis/disinformation, which is an issue we face across all online harms. You may know that mis/disinfo and the voluntary code in Australia overseen by our colleagues in the Australian Media and Communications Agency, who I would be happy to put you in touch with if of interest. Of course many of the issues in terms of algorithmic promotion, incentives, and recidivism are common across other online harms that eSafety regulates though.

2. It's very good to see the GARM framework continuing to deepen and broaden. Your advocacy for local metrics is also welcome – another common issue. We'd also love to discuss how the Safety by Design initiative and risk assessment tools might be integrated into your reporting framework in the future, and become (even more of) a common benchmark for industry.

3. I think we were first brought together via the work you had done with WEF. We are co-chairing a stream with a tech leader (we think, AWS) on Digital Safety and are seeking to develop a toolkit that could be used with a range of players across the ecosystem (including the education, investment and VC communities). Wondering if we might be able to toolkit leverages or adapting some of the incredible tools you've built – I view this as a broader distribution mechanism for a wider but important set of audiences.

4. Finally, as I think you're aware, we will also be issuing our first notices to industry in the next few months under the Basic Online Safety Expectations. We hope that process generates some unique insights into company safety interventions that will help us to collectively drive accountability.

I know the team are keen to have a follow-up discussion on all of this, as well as some of the future looking pieces around the metaverse and immersive tech. I'll leave to them to follow-up, but thank you again for all the work in this area.

All the best,

Julie

From: s 47F
Sent: Tuesday, 21 June 2022 8:22 PM
To: Julie Inman Grant ; s 22 ; s 22 ; s 22 ; s 22
Cc: s 47F s 47F ; s 47F
Subject: GARM: Key updates at Y3

Hi Julie + Team –

I wanted to reach out to you with some key updates from GARM that you should find relevant.
We're reaching our 3 year anniversary since our launch, and we've released a comprehensive update on where we've been, what's new today, what's next tomorrow.

The big news we are sharing today is:

1/ We've aligned on definition for Misinformation to integrate into the GARM Brand Safety Floor + Suitability Framework. This is a significant step and is aligned with the work for Digi and more recently with our direct involvement with the EC on their CoP on Disinformation.

2/ We've also just released the GARM Adjacency Standards Framework by which to assess the placement of ads next to safe but sensitive content which will allow for ad placement in a more consistent way

3/ We also recently released our GARM Aggregated Measurement Report in May which saw more metrics but most importantly that YouTube have had their monetization safety metrics accredited by the leading media auditor, MRC – which is a significant undertaking and achievement

We're inspired by your work and continue to reach higher based on your impact.
We're looking forward to reconnecting with you in the coming weeks to discuss local measurement – which you will see is a priority area for our next steps.

Best,

s 47F

s 47F
Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

Meeting Brief – GARM

To:	s 22
From:	
Date:	28 June 2022
Subject:	Follow up meeting with GARM
Meeting date, time and platform:	Thursday 7 July 2022, 11-11:45am AEST via MS Teams [rescheduled]
Company representatives:	s 47F
eSafety representatives:	s 22 [redacted]

Purpose

To meet with representatives of GARM to provide follow-up information on the BOSE and Safety by Design, and learn more about GARM's key updates in terms of brand safety and measurement.

Recent engagements

- JIG and representatives of eSafety IAE team engaged with GARM in November 2021 as part of an introductory catch-up to provide an overview of eSafety and GARM and discuss opportunities for collaboration.
- Representatives of eSafety IAE and StratPol team re-engaged with GARM in May 2022 to provide an update on the BOSE and Safety by Design activities.
- Most recently, s 47F from GARM reached out to JIG via email to provide an update on GARM's 2022 activities. Representatives from StratPol liaised with GARM to set up a follow-up meeting in June 2022.

Agenda

1. Introductions (if required)
2. Updates from GARM – local reporting metrics
3. Updates from eSafety – Safety by Design and how they could be used in GARMs reporting framework (primarily) and BOSE (briefly); flagging AANA children's code consultation (time permitting)
4. Questions/AOB

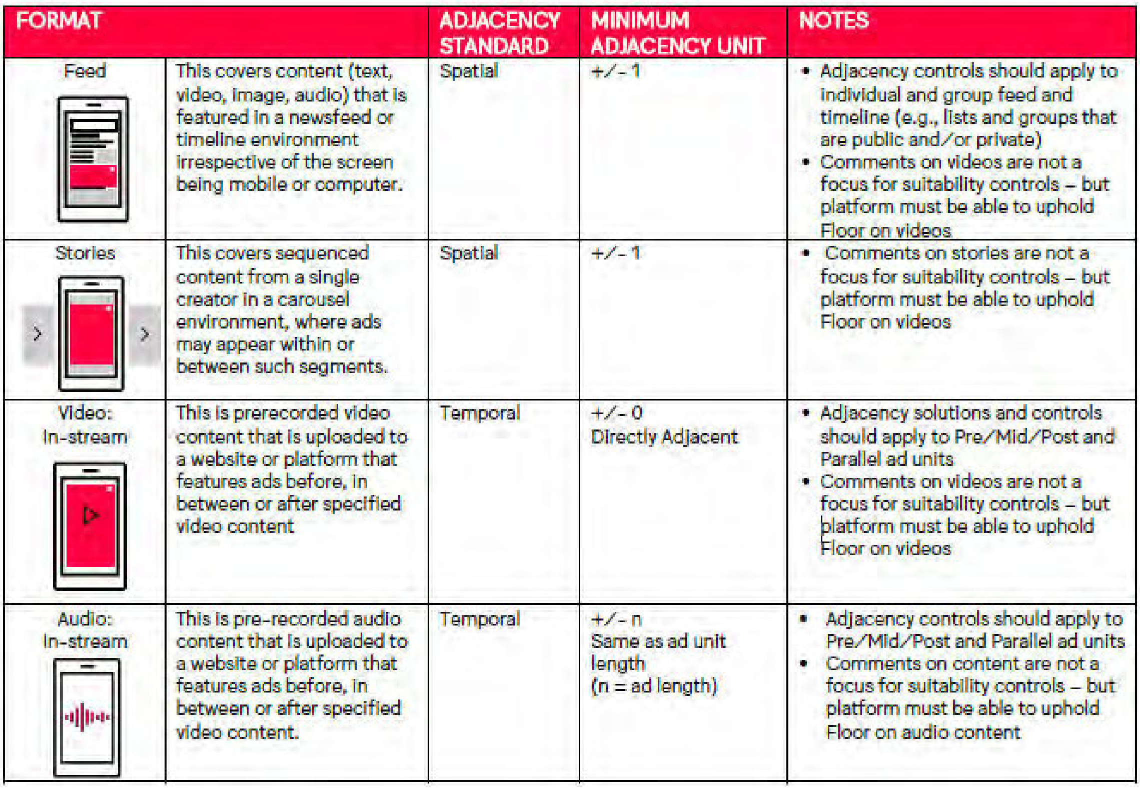
Background

GARM Update

- In email correspondence with JIG, GARM noted three important updates:
  - 1. We've aligned on definition for Misinformation to integrate into the GARM Brand Safety Floor + Suitability Framework. This is a significant step and is aligned with the work for DIGI and more recently with our direct involvement with the EC on their CoP on Disinformation.
  - 2. We've also just released the GARM Adjacency Standards Framework by which to assess the placement of ads next to safe but sensitive content which will allow for ad placement in a more consistent way.
  - 3. We also recently released our GARM Aggregated Measurement Report in May which saw more metrics but most importantly that YouTube have had their monetization safety metrics accredited by the leading media auditor, MRC – which is a significant undertaking and achievement.
- GARM is celebrating three years since its launch as a cross-industry initiative to remove harmful content from ad-supported digital and social media.
  - Launch propelled by 'uncommon collaboration' – grown to 122 members, 6 active working groups, monthly Community meetings
- Focus areas for engagement:
  - Bringing advertising technology companies into a Solutions Developers Working Group
  - Broadening membership to National Advertiser Associations
  - Formalising ways of working with NGOs
- Digital media safety:
  - Platform safety – focuses on the product, essential technology design and oversight
  - Brand safety – centred on monetisation safety – how advertising investment is steered away from harmful content and behaviours
- GARM Brand Safety Floor + Suitability Framework
  - Harmful content and its creators threaten the potential for digital media and disrupt the connections everyone seeks. Our first step in safeguarding the positive potential for digital is to provide platforms, agencies, and marketers with the framework with which to define safe and harmful content online.
  - GARM position is that you cannot address the challenge of harmful online content if you are unable to describe it using consistent and understandable language.

- GARM has developed and will adopt common definitions to ensure that the advertising industry is categorizing harmful content in the same way across the board.
- GARM drove an agreement across advertisers, agencies and platforms to set a framework that limits advertising support for harmful content, through the Brand Safety Floor, while providing for a Suitability Framework to manage advertising placement in sensitive content categories, while acknowledging critical nuances across platforms and formats.

- GARM Adjacency Standards Framework
  - Designed to serve as a companion to the GARM Suitability Framework, these standards provide advertising industry participants with a common methodology for evaluating the brand suitability of an ad placement relative to an ad's position to nearby content (i.e. "adjacency") within specific media formats
  - GARM have now defined adjacency standards for News Feed, Stories, In-stream Video, In-stream Audio, and Display overlay:



FORMAT		ADJACENCY STANDARD	MINIMUM ADJACENCY UNIT	NOTES
Feed	This covers content (text, video, image, audio) that is featured in a newsfeed or timeline environment irrespective of the screen being mobile or computer.	Spatial	+/- 1	• Adjacency controls should apply to individual and group feed and timeline (e.g., lists and groups that are public and/or private) • Comments on videos are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Stories	This covers sequenced content from a single creator in a carousel environment, where ads may appear within or between such segments.	Spatial	+/- 1	• Comments on stories are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Video: In-stream	This is prerecorded video content that is uploaded to a website or platform that features ads before, in between or after specified video content	Temporal	+/- 0 Directly Adjacent	• Adjacency solutions and controls should apply to Pre/Mid/Post and Parallel ad units • Comments on videos are not a focus for suitability controls – but platform must be able to uphold Floor on videos
Audio: In-stream	This is pre-recorded audio content that is uploaded to a website or platform that features ads before, in between or after specified video content.	Temporal	+/- n Same as ad unit length (n = ad length)	• Adjacency controls should apply to Pre/Mid/Post and Parallel ad units • Comments on content are not a focus for suitability controls – but platform must be able to uphold Floor on audio content

  - Given the recent incidents involving Livestream formats, GARM are working internally to ensure that there is a robust Safety Floor in place, before advancing those formats into a Suitability Framework.

- In April 2021, GARM developed a common framework to assess industry progress in removing harmful content from advertising-supported media.

- GARM Aggregated Measurement Report – new metrics have been shared that have never been available before ranging from:
  - YouTube's Violative View Rate and Advertising Safety Error Rate
  - Meta's reporting on Prevalence of Hate Speech
  - Snap's Violative View Rate
  - Pinterest's reporting on removal of Misinformation content by views
  - Twitter's sharing of Violative Impressions Rate

- GARM's go-forward – two provocations
  - Provocation 1: Brand Safety isn't a substitute for Platform Safety – continue to be, supportive of progressive, comprehensive regulation that creates a duty of care around design, resourcing, oversight, and moderation
  - Provocation 2: We must get the future right by being proactive – must now help the industry understand safety requirements before commercialization begins in the metaverse, ensure that advertising is aligned with sustainable and responsible growth models

Safety by Design

GARM Adjacency Standards Framework

- We understand that Livestream Audio and Livestream video are currently omitted from the current version of the framework. We would be grateful if the Working Group could share the exploration outcomes.
- Assessment tools – one year on – accessed in 45 countries
- Discussion on how SbD assessment tools + end reports could potentially highlight GARM framework.

Metaverse activities and advertising considerations

- Procured metaverse research findings at a high-level indicated that:
  - 36% of respondents have used any immersive technology – increased to 52% when participants were asked about the specific types of technology they used
  - Confusion about terms used – only 39% previously saying they knew what the metaverse was
  - 22% of those who engage in the metaverse said they had experienced something that made them feel unsafe
- While immersive technologies and the metaverse may be in their early stages of uptake, we can see from these insights the potential for harms to surface and impact users in different ways.
- As part of our work program on immersive technologies and the metaverse, the Safety by Design team are also undertaking a review of our typology of online harms – considerations/developments related to immersive technologies and the metaverse
  - Original typology of online harms informed by cross-office insights and international research/guidance (like the Luxembourg Guidelines)
  - Conducted an internal experts workshop drawing together insights from across eSafety
  - Anticipate conducting further workshop activities to refine the typology of online harms
  - Online risks and harms are multi-layered and multi-dimensional. They can also be understood through different frameworks or lenses.
  - Safety by Design focuses on the rights of users – typology seeks to frame online harms through a human rights lens wherever possible, emphasizing impacts on users.
  - Each type of online harm is not exclusive. For example, an incident may involve multiple types of online harms – therefore whilst these are categorised, there may be overlap across typologies – seek to extend this rationale to our updated typology accounting for the impact of immersive technologies and metaverse environments.

- As part of this work program, eSafety is bolstering considerations of the online harms that are surfacing in immersive environments – including those that may surface through digital advertising.
- Can GARM provide any additional information or insights about advertising in immersive/metaverse spaces, noting updated marketing structures for Web3.0?

BOSE

- Brief update given previous discussion, focussed on next steps and initial notices.
- Interested in any views from GARM on how to effectively communicate outcomes and industry data in a way that can be widely understood/disseminated
- Interested in whether GARM is will to report on areas of “responsibility” even if there isn’t a key issue for advertising (e.g. where a service fails to protect children on parts of its service that there is no advertising or adjacency risk).
- Welcome suggestions for future areas the BOSE could focus on.
- [Welcome thoughts/reflections about the process of auditing metrics through MRC]

AANA Children’s Code

- The Australian Association of National Advertisers (AANA) has launched a review of the Children’s Advertising Code to ensure that it continues to provide a robust framework for the regulation of advertising to children on all media platforms.
- The current code includes standards around the content of advertising but not about broader safety of services on which advertising is placed.
- They have invited eSafety to make a submission. We are considering highlighting the intersections of online safety and advertising, and potentially GARM’s work.
- Any thoughts? Is GARM considering a submission?

Bios

s 47F [redacted] the Global Alliance for Responsible Media	s 47F [redacted]

Potential risks

- None identified

Potential opportunities

- Bolstering of eSafety relationship with GARM and opportunities for future collaboration.

Attachments

20220517 Meeting Brief and Notes - GARM.docx

Meeting Notes – GARM

Company representatives:	s 47F
eSafety representatives:	s 22 [redacted]
Note taker:	s
Meeting date, time and platform:	Thursday 7 July 2022, 11-11:45am via MS Teams [rescheduled]

- Meeting commenced at 11am AEST.

Introductions

- s noted that SWSX is coming to Sydney in 2023
- s provided intro to himself as IAE manager
- s noted thanks for reach out to JIG – and enthusiasm for GARM activities

General Discussion

- s noted work in Aus with DIGI – alignment on mis/disinfo – definition from a monetisation perspective
  - GARM goal – align with Aus govt, EU govt and UK govt goals
  - s noted that platforms are under scrutiny – s noted helpful local work in Aus
- s noted update with Vol 3 of Aggregated Measurement report – important update on YouTube method of monetisation safety
- s suggested potential opportunity for future work with eSafety + AANA + ANZA – how do we get disclosure from the platforms (outside global transparency reports) around sampling methodology (how much content is English language? How much is out of Aus/UK/Canada/other jurisdictions? – drive up disclosure)
  - s noted formal request for info (RFI) to disclose how their global transparency panels are produced – e.g. churn in panels/ representative nature of panels
  - s noted that transparency reports contain globalised/highly averaged numbers – what does safety look like in a market like Australia?

  - Global numbers may look impressive – though these are not close to the ground/local level
- GARM seeking disclosure around transparency through an RFI
- s asked what data we have been requesting at eSafety from platforms at the local level?
  - s suggested the need for a template for incidents at the local level (how much harmful content, how many views, to what extent was it monetised?)
  - s noted desire to consult with eSafety around safety data
- s noted that GARM is seeking feedback on RFI + regulatory request RE sharing types of data/what has been received from the platforms
- s suggested that we would like to see reports done on a local basis too – continue pursuing through local engagement (also work with AANA)
- s suggested that we will also work on gaps in transparency reporting – regulatory processes designed to fill what we see as the 'gaps' in transparency reporting
  - s noted that there are a series of regulatory tools that touch on this – takedown powers are targeted, smaller number of sites involved
  - s noted previous discussion around BOSE – power to ask for reporting/transparency reporting with civil penalties for companies that don't respond
  - s noted initial BOSE focus on CSAM – notices being issued in August – agree process for sequencing in terms of how data and information is published
  - s 22 suggested that eSafety can have an internal discussion about advance sharing of BOSE questions with s (ACTION)
- s noted BOSE focus on gaps in terms of speed of takedown, speed of action for reporting, proactive detection – s noted that with regard to high harms, content should not be available at all – prevalence measures risk downplaying harm done through content sitting online, even if numbers of views/reach is low.
- s suggested that eSafety is keen to work with GARM around metrics which fit objectives but face company pushback (damage to brand, data not being collected etc.)
- s noted Aggregated Measurement Report, which considers:
  - How safe is the platform for consumers? – notion of prevalence
  - How safe is the platform for advertisers?
  - How effective is the platform at enforcing its safety policy? (Pieces of harmful content removed x times viewed)
  - High amounts of removals, low amounts of views – takes some people down a dark hole – possibly lead them to take offline actions
- s noted that if we collaborate on this – could approach issue with following formula: pieces of content removed x views x takedown

- s noted need to look at harms across platforms – consider individuals interacting with content – take harmful content and share it more widely in private groups – s noted current lack of industry incentive to consider the whole picture
- s noted that when Christchurch massacre happened – prompted GIFCT formation – intended to reach across platforms
- s noted cropping up of the metaverse and cross-channel behaviours
    - s noted manual work to detect nudity, spam etc. – noted that detection around hate speech, cyberbullying, planning an attack requires law enforcement work
- s noted work of eSafety on AV – considerations around technological solutions
    - s noted example where children taken from one platform to be groomed – question of how to prevent transfer of contact across platforms?
- s noted eSafety interest in exploring future technologies – e.g. sensory experiences
    - s noted work looking closely at immersive tech/metaverse – through Safety by Design
- s asked s about livestreaming work as part of Adjacency Standards Framework
    - s noted that answers around livestreaming were concerning from GARM's perspective in compiling Adjacency Standards Framework
    - s put forward his hypothesis that livestreaming is the closest thing we have to metaverse right now – as it requires multiple layers of technology
    - s noted that GARM is providing recommendations around content moderation – e.g. 7 second delay, instant human moderation
- s suggested he would like to work more deeply with us on Safety by Design – noted that platforms are enthusiastic for safety by design – work out criteria before commercialisation/ monetisation happens
    - s suggested need to determine criteria for brands before they funnel money into advertising
    - s 47G(1)(b)
    - s noted considerations around reputational, business, revenue loss
- s noted concerns around safety considerations that may or may not be built in + work of gaming platforms – not just social media services that have had previous regulatory scrutiny – interoperability piece is interesting
- s suggested that we would be grateful to be kept in the loop with updates and insights from GARM working groups
- s provided invitation to eSafety to join/participate – alternatively consult with GARM
    - s noted work on current Charter – noted eSafety work with WEF – s noted that we need to determine how we are going to collaborate (what level/which forum)

- s [redacted] noted that eSafety has been a shining line – he has shared eSafety work with European Commission in Brussels – s [redacted] suggested that eSafety work is setting 'new tack' for what a good digital regulator should be doing
- s [redacted] noted discussion around supporting human rights and consumer protection without stymieing industry – s [redacted] noted good balance of eSafety work on this front
- s [redacted] suggested that it would be good to have a 'closed-door meeting' with GARM Board of Directors, with two intentions/considerations:
    - o 1. Get marketers more comfortable with regulators
    - o 2. How do we link up on metaverse
- s 22 [redacted] noted practical advice and guidance in Safety by Design risk assessment tools – industry willingness to adopt guidelines – s [redacted] noted work to update Safety by Design and harms in the metaverse
- s [redacted] noted challenge eSafety has encountered – we see organisational willingness for Safety by Design – however implementation and quantifying the implementation of measures poses a challenge (what does it mean to apply Safety by Design at the practical level?) – BOSE as an opportunity to measure this more accurately.
- s [redacted] asked about whether GARM could use measure put forward by eSafety under BOSE to diagnose 'platform safety'?
    - o s [redacted] noted possible work with eSafety – scorecard/lifting metrics to summarise how GARM would diagnose platform safety
    - o s [redacted] noted that scorecards may be a way off for eSafety – not to say that data we get can't be used to generate scorecards elsewhere – intention is to publish data collected through BOSE as long as it doesn't help bad actors abuse platforms
    - o s [redacted] noted need to dig down into specific parts of the platform – stats in transparency reporting are presented at a high level (stats are aggregated across the company rather than the individual services)
    - o s [redacted] noted need to separate harms in different forums – e.g. newsfeed/groups/ marketplaces
    - o s [redacted] noted need to disaggregate safety metrics – vectors of service vs. surfaces
    - o s [redacted] noted need to consider disaggregation of metrics on the market side – regionally/language-based
    - o s [redacted] noted that lifting the hood helps to break down high level stats
- s [redacted] suggested potential to spend time thinking around three areas:
    - o Sizing up platform safety – dimension of platform safety performance
    - o Commercialisation standards
    - o Disaggregating metrics

- s suggested that these three principles could be taken as guiding principles – 'north star' of where we want industry to go (competitive push to the top)
- s noted that transparency reports tend to give a deliberately good message – but unless companies are truthfully showing where vulnerabilities are – challenge to get objective information
    - s noted conversations around Aggregated Measurement Report with industry
    - s suggested that information is highly contextualised – conversations can be uncomfortable
    - s suggested there is an element of complexity to having these difficult conversations
    - s noted need to frame the conversation positively – help industry to address problem
- s asked about AANA consultation on children's code – noted eSafety is looking at this
    - s noted that he is not directly involved in this consultation – however understands principle/foundation that advertising of certain categories to minors is a 'full stop no'
    - s suggested that the goal is setting a floor – protection of vulnerable audiences/media consumers
    - s noted need to consider whether there are categories that should have hard limits for advertising to children – different ranges of harms – not every media type is the same and not every category of advertising has the same potential for harm
    - s noted that while this is not our core business – working to determine whether it is worthwhile bringing online safety lens to this consultation

Actions/Next steps

eSafety

- Follow up on RFI – feedback on a follow up call
- Check RE sharing BOSE questions in advance of providing first notices

GARM

- Roadmap for scorecards/metrics
- Think tank opportunity – s to provide details – internally work out what role eSafety can play
- Considerations around platform safety footprint
- s suggested hosting standing meetings at the appropriate cadence – connectivity and feedback – how frequently do we want to meet
- s to share some info as thought starters

Questions/AOB

s asked if GARM is contributing to any workstreams for WEF Coalition for Digital Safety? – noted eSafety is on a couple of the WEF working groups

- s noted that WEF encouraging GARM to write a blog post and get think tank up and running – formally join in January 2023
- s noted that as it relates to commercialisation – highly centralised to highly decentralised
- s noted that as it relates to consumer action – highly constrained to highly fluid
- Nine box grid – grade business risk as a result
- s noted that GARM is going to formulate this thinking as an organising principle
- s noted example of Zwift fitness application
- Business interest and advertising business – engagement as a special interest (understand whether consumer protection and safety is being enshrined)
- s noted that GARM will be joining over time

- Meeting concluded at 11:55am AEST.

eSafety FOI 25152
Document 7 of 21

From: Julie Inman Grant s 47E(d) @esafety.gov.au>
Sent: Wednesday, 9 November 2022 10:11 PM
To: Rob Rakowitz
Subject: Re: Twitter [SEC=UNOFFICIAL]
Attachments: Correspondence_to_Twitter_from_eSafety_Commissioner 3.pdf

Hi Rob - I saw the GARM post on LinkedIn and was really pleased to see this proactive approach. I would be very interested in seeing your six points. I believe GARM has significant collective power in helping to hold the platforms to account and I appreciate your informing the importance of both brand and platform safety.

Interesting to hear your perspective on Blue & Birdwatch. It is very hard to piece together what seems like a very erratic and reactive strategy. On one hand, this could be seen as a very basic revenue generating machination and a way to target the bots and scammers by undermining the economic incentive to mass create accounts....but, they clearly haven't thought through the multiple spill on events.

I gather you saw that I wrote to Mr Musk and was questioned about my thoughts at Senate Estimates?

Not sure I will get an immediate response...here is the letter, since it is now in the public record.

As an aside, I'll be in DC next week launching a global online safety regulatory network at FOSI, then in Seattle and San Fran. My team member, s 22, will be up in NYC on the 18th but happy to try and touch base further whilst I'm stateside.

How 'bout this election???????

Thanks for reaching out and for advocating for such important change! Julie

Sent from my iPhone

> On 9 Nov 2022, at 9:47 pm, Rob Rakowitz s 47F @wfanet.org> wrote:
>
> Hi Julie -
>
> I hope you've been well.
> I've issued Twitter a detailed list of questions and I fear we won't hear back on them in a meaningful way.
>
> One of the 6 areas I raised was how collapsing Blue and Birdwatch will increase the quality of content and avoid abuse? As we know bad actors in misinfo are many times stat funded and coordinated $8 accounts to tip a content rating and an algorithm

recommendation engine is a clear issue. The only thing I can thing is perhaps it's intentional.

Happy to connect and share our overall plan and questions.

Let me know if there's a desire to connect.

Best

Rob

Rob Rakowitz
Initiative Lead - Global Alliance for Responsible Media
World Federation of Advertisers

s 47F / s 47F @wfanet.org

Australian Government | eSafetyCommissioner

8 November 2022

Elon Musk
CEO
Twitter HQ
1355 Market Street #900
San Francisco, CA 94103

Dear Mr Musk:

I am Australia's eSafety Commissioner. In this role, I regulate key sectors of the technology industry on behalf of the Australian Government to ensure Australians have safer and more positive experiences online.

Various harms are addressed through Australia's regulatory framework, including child sexual abuse material, terrorist and violent extremist content, child cyberbullying and serious adult cyber abuse, as well as image-based abuse – the non-consensual sharing of intimate images.

I have a range of systemic powers, including enforcement of mandatory industry codes, and transparency powers under legislated Basic Online Safety Expectations provisions.

Among the industry sections subject to regulation under this overall framework are social media services, including Twitter.

Since I commenced my tenure in early 2017, my office has had a constructive relationship with Twitter, including with policy representatives in both Australia and the APAC region, and via direct engagement with Twitter HQ.

While I was disappointed that Twitter failed to adequately address my concerns about child sexual exploitation material on the platform in 2020, I have appreciated Twitter's responsiveness to requests for assistance with combatting other harmful material distributed on the platform. Recently, these requests have related to the distribution on the platform of a video showing the stabbing murder of a Brisbane teenager, links and other material related to the terrorist attack in Buffalo NY, and video depicting a spree shooting in Memphis TN.

I also note that Twitter has been responsive to my notices issued under the Online Safety Act 2021 for removal of tweets intended to cause serious harm to an Australian adult. In the course of my office preparing these, your local policy lead was willing to engage in constructive discussion with my investigations team about the grounds for removal of the material.

Given the events of the weekend, I am deeply concerned about the depth and breadth of recent cuts to Twitter staff across the globe and their potential impact on Twitter's ability to respond to and comply with our regulatory requirements.

The cuts also concern me as I spent more than two years at Twitter on the Public Policy and Philanthropy team championing the practice of trust and safety within the company, including here in Australia and across the Southeast Asia. During that time, I saw firsthand a rapid increase in the volume and complexity of online harms – a trend that has only accelerated during my time as eSafety Commissioner.

According to Yoel Roth, Twitter's Global Head of Trust & Integrity, only about 15% of the Trust and Safety organisation has been affected by the cuts, compared with 50% across the board. He recently posted a graph demonstrating that your team's ability to take moderation actions has been unaffected by the cuts.

But, as someone who understands Twitter's operational ecosystem, I know it is much far more complex than those comments would suggest. Deep cuts to public policy, legal, communications, human rights, ethical AI and transparency teams leave me very concerned that Twitter is removing both the expertise and necessary guardrails to deal with the growing threat of hate, harm, disinformation and other forms of serious online abuse on the platform.

To that end, I wish to raise several questions in this letter to ensure your compliance with Australian regulatory obligations:

1. Will you provide your direct assurance that Twitter will recognise Australia's laws and will continue to be responsive to regulatory actions taken by my office around online harms on your platform?
2. Can you ensure that remaining Twitter personnel will continue to work collaboratively with the eSafety Commissioner to ensure expeditious and effective harms minimisation? Who are those personnel?
3. As I noted above, we have worked with Twitter to build a constructive working relationship over several years. Will the clear and effective reporting channels and escalation paths directly to Twitter's Trust & Safety team remain? If reporting channels change, I would expect your earliest advice to that effect.

I am deeply concerned about the culling the very employees who hold the greatest depth of expertise and experience in trust and safety, and who have specific understanding of Twitter's moderation tools and policies, could do anything other than profoundly undermine safety on the platform.

My team and I welcome the chance to discuss my concerns and Twitter's plans in the trust and safety space in more detail, at a time convenient to you.

Your sincerely,

[redacted]

Julie Inman Grant

eSafety Commissioner

CC:

Yoel Roth, Global Head of Safety & Integrity

Sinead McSweeney, Vice President of Global of Public Policy

From:	Rob Rakowitz s 47F @wfanet.org>
Sent:	Wednesday, 9 November 2022 10:32 PM
To:	Julie Inman Grant
Subject:	Re: Twitter [SEC=UNOFFICIAL]
Attachments:	Twitter GARM Steer Team Observations Action Plan.pdf; Re- GARM- Advice on industry-wide communications.eml

So lovely to hear from you!

I am sharing the attached in confidence with you.
There is an ongoing major intervention.

PDF sets out a plan
Email is a short term RFI

Neither of these are in public domain, yet. And would be good to cross-check them and see how they line up with you.
It would be great to line up together around a shared agenda.

I am also writing a piece now on how brand safety has nothing to do with controlling freedom of speech and everything to do with correcting an imperfect marketplace that has led to things like Molly Russell's suicide.

I'm getting tired of the false debates.

Let's 100% connect – maybe virtual in the short term.
100% would love to see the team in person.

Best as ever – you are an inspiration!
Rob

Rob Rakowitz
Initiative Lead - Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant
Date: Wednesday, November 9, 2022 at 06:11
To: Rob Rakowitz
Subject: Re: Twitter [SEC=UNOFFICIAL]

Hi Rob - I saw the GARM post on LinkedIn and was really pleased to see this proactive approach. I would be very interested in seeing your six points. I believe GARM has significant collective power in helping to hold the platforms to account and I appreciate your informing the importance of both brand and platform safety.

Interesting to hear your perspective on Blue & Birdwatch. It is very hard to piece together what seems like a very erratic and reactive strategy. On one hand, this could be seen as a very basic revenue generating machination and a way to target the bots and scammers by undermining the economic incentive to mass create accounts….but, they clearly haven't thought through the multiple spill on events.

I gather you saw that I wrote to Mr Musk and was questioned about my thoughts at Senate Estimates?

Not sure I will get an immediate response…here is the letter, since it is now in the public record.

As an aside, I'll be in DC next week launching a global online safety regulatory network at FOSI, then in Seattle and San Fran. My team member, s 22 will be up in NYC on the 18th but happy to try and touch base further whilst I'm stateside.

How 'bout this election???????

Thanks for reaching out and for advocating for such important change! Julie

Sent from my iPhone

> On 9 Nov 2022, at 9:47 pm, Rob Rakowitz wrote:
>
> Hi Julie -
>
> I hope you've been well.
> I've issued Twitter a detailed list of questions and I fear we won't hear back on them in a meaningful way.
>
> One of the 6 areas I raised was how collapsing Blue and Birdwatch will increase the quality of content and avoid abuse? As we know bad actors in misinfo are many times stat funded and coordinated $8 accounts to tip a content rating and an algorithm recommendation engine is a clear issue. The only thing I can thing is perhaps it's intentional.
>
> Happy to connect and share our overall plan and questions.
>
> Let me know if there's a desire to connect.
>
> Best
>
> Rob
>
> Rob Rakowitz
> Initiative Lead - Global Alliance for Responsible Media
> World Federation of Advertisers
>
> s 47F / s 47F @wfanet.org

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

eSafety FOI 25152
Document 10 of 21

From: s 47F @wfanet.org>
Sent: Wednesday, 9 November 2022 9:06 PM
To: s 47F
Subject: Re: GARM: Advice on industry-wide communications

Hi guys –

Two things:

1/ The request has come that s 47F note to the Influence Council be sent to the Steer Team members directly by Twitter to allow them to forward it on immediately. Despite all of the calls that you are having there are a lot of experts that you're missing out on in terms of an imperfect cascade (e.g., unit president at company on WPP call unable to fully translate what they hear to a media leader at the same company). So **the request is to send the existing note to the team immediately** [s 47F].

2/ Per the grid of detailed questions below – we really do want to see you onboard this advice. The best practices are what we've seen in YT and then Facebook's responses to safety concerns. Ideally you can **commit to this and start publishing against the list for Monday**. We recognize that there's a lot but starting out and even indicating that there's more to come would be helpful to broad stakeholder groups.

Again – there's a comms gap that is reinforcing perceptions with some real issues underlying them that creates a mushrooming effect.

Thanks,

s
47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: s 47F @wfanet.org>
Date: Tuesday, November 8, 2022 at 17:02
To: s 47F s 47F
Subject: GARM: Advice on industry-wide communications

Hi guys –

It was good seeing you both yesterday.

The Steer Team and I reflected on the discussion and the dynamic, and we've come to the conclusion that there's a communications issue as it relates to brand safety and business continuity. A lot of the conversations at the agency holdco and the Influence Council have been good, but there's still essential details missing that brand safety and media leaders need.

We'd recommend that Twitter take the following steps in terms of corporate communications best practices (I am sorry if this comes off as reinforcing the basics):

1. Transparently post to the web site as the primary point of contact (this is where business leaders will go versus following the conversation on Twitter)
2. Distribute emails to relevant parties to start a cascade to their teams (e.g., by emailing ISBA directly, it allows them to share versus an indirect cascade that may violate sharing)
3. Update steps, changes, or progress on a regular, predictable basis

We recognize that the context for the CSAM incident was different than now, and there is a desire for more documentation and structured conversation.

In terms of the areas for immediate disclosure (note this is based on questions relevant today and this may continue to evolve (note these have been sourced by the GARM Steer Team via their relevant membership base):

TOPIC	**KEY QUESTIONS**
Reduction in Force / Safety Teams	How has the reduction of force impacted key practice areas that advertising industry stakeholders engage with? How has this impacted site integrity teams? How has this impacted brand safety teams? Have there ben organization reporting line shifts relative to site integrity and brand safety? How has any changes in staffing levels impacted content moderation or content monetization? Are there KPIs for service continuity and quality that are being tracked – if so what are they?
Security Access	Has Twitter regained the full security keys and technology stack from the first days of the transfer of ownership? Has this been restored in time for the midterm elections in the US? Has the source of the coordinated attack referenced by Twitter been identified? If safety systems have not been restored, is there a timeline for intended resolution? Are there areas that are vulnerable? Is there a remediation plan for vulnerabilities?
Moderation and Monetization Changes	We understand that Twitter ownership has committed to no changes in moderation and monetization policies until councils are set up and consultative or notice periods are established – when will we receive detail on these councils in terms of composition, governance and remit? Has the rise in antisemitism been attributed to reductions in force or a stasis/moratorium in moderation and monetization policies and enforcement? Has there been any knock-on effects on algorithmic oversight that would have content recommendation engines surface incitement or divisive debate on harmful conspiracy theories? Has monetization been affected by this? How does this affect ads policies?
Brand authenticity + Identity Verification	How will Twitter work with brands to recertify official handles and avoid impersonation? How will the platform prevent brand, corporate, and executive 'Twitter identity theft'?
Monitoring Brand Safety	s 47F presented data that suggests that harmful content is returning to baseline levels. How frequently will this data be refreshed? How can we get assurances on the accuracy of this data share, given that syndicated social listening tools (which we recognize are search query-based v reach-based) show alarming increases in NSFW terms, hate terms directed to blacks and Jews? Will Twitter open API access to a third party to assess the amount of violative content and its reach? If not, how will we gain assurances between two reference points (content availability and content reach)? Will this expedite work with the MRC? Are there other monitoring mechanisms under consideration?

Integrity of Content Quality Plans	Has Twitter pressure tested the combination of Blue and Birdwatch? What is preventing bad actors (many of whom are state funded) from using Blue and content feedback tools mentioned from undermining ownership's plans to surface quality media content? What oversights will exist?
Category conflicts	Have any considerations been made in assuring automakers on the safety of marketing campaign information and acquisition data?

As always, I am here to help prioritize and frame. We need to answer questions above and fill a communications gap.

Finally, can we confirm a standing weekly meeting?

Best,

s
47F

[redacted]
[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

Twitter: GARM Steer Team Acquisition Concerns

1. **Context is sensitive at best, divisive at worst:** The time of the acquisition and transition is marked by several sensitive events; Iranian regime protests, Israeli elections, US midterm elections, Ukraine War, UK government transition, rise in US antisemitism. The polarized views in each of these topics only manifest themselves online – with Twitter being an open feed venue (more open than Instagram, Facebook) and its ability to be exposed to negative behaviors like brigading.

2. **Musk's direction on platform control is inconsistent:** Since acquisition Musk has promised not to reinstate controversial accounts until a content review council is set up (NB content reviews and accounts are usually separated in other platforms). However, Ye has been partially reinstated, and Trump has been approached (although he allegedly refused the idea of returning). Musk's actions and his statements may be at odds.

3. **Layoffs and resignations of key staff leave resourcing questioned:** Musk cleared executives in his first acts (CEO, CFO, Policy/Trust & Safety Lead), makes public a plan to layoff 25% of staff, and has also seen key voluntary departures (CMO, People Officer, Revenue Officer, Customer Officer). Key disciplines like platform trust & safety, brand safety, client management (agency and marketer) are left to middle management. It is unclear of reporting lines and decision-making abilities.

4. **Microaggressions are coming in from fringe networks as a stress test:** Many polarized users view Musk's acquisition as a victory for harmful acts and content. Early reports show a 500% increase in the use of the N-word. An antisemitism watchdog start-up in Israel reports a 600% increase in antisemitic bullying (NB they already work with another GARM platform on moderation). An agency member of GARM a 2x increase in harmful conduct on the platform since the acquisition closed. It is open season and the team previously tasked with moderating the platform has gone from disengaged during the deal period to dismissed post-acquisition

5. **Corporate governance:** Musk has eliminated the Twitter board, and has made himself CEO, despite owning and being CEO in other ventures (Tesla, SpaceX). There are real concerns that Musk can run all three ventures with one requiring a live-time decision making muscle, cognizant of user safety concerns.

Twitter: GARM Steer Team Short Term Action Plan

Context: With the sensitive external environment, platform access concerns, and reduction in force, many advertisers have expressed concern, confusion, or conviction on Twitter's ability to be suitable for advertising investment. Simultaneously a call by NGOs (some affiliated with GARM's NGO Consult Group) for advertisers to boycott Twitter launched on 4 Nov (Stop Toxic Twitter). Worryingly this has triggered Elon Musk to engage in discussions and encourage a counter-boycott of advertisers leaving his platform.

Upholding Commitments + Maintaining Continuity: With the shift in ownership and adjustment to Twitter, we must have Twitter maintain platform and brand safety operations. We must also ensure we monitor the delivery of existing commitments, and communicate any changes to the GARM Community, without filtering and without bias

Recommendation: GARM should take a role of identifying advertiser and agency concerns, identifying challenges in brand safety operations or implementation via fact-based research methods. GARM will drive transparency, and holding Twitter accountable to its prior commitments. GARM should reinforce its position of a standards setting, solutions building and transparency-building forum.

Precedent: GARM did not expel Facebook during its brand safety incidents in 2020. Facebook did agree a reform plan based on GARM's provocations. That reform plan execution is still underway

This stance preserve's GARM's neutrality and avoids a potential adversarial legal backlash that could deleverage advertisers and agencies, and trigger regulatory scope expansion to monetization.

Upcoming milestones: The WFA ExecCo and GARM Steer Team are set to meet with Elon Musk on Dec 1. The Steer Team will also engage with the remaining Twitter Brand Safety Team.

Twitter: GARM Steer Team Short Term Action Plan

While there is bias for action, GARM can help navigate advertiser perceptions and digital safety needs (whether platform or brand safety). GARM will use FACTS and DATA to help . Our timing is to have data ready linked to the WFA ExecCo on 1 Dec.

STEP 1:
Run an advertiser survey via GARM to gauge advertiser perceptions on the transition to New Twitter

Status: In progress
Inputs: Steer Team approval
Partners: 4As, ANA, ISBA
Timing: w/c 7 Nov fielding
w/c 14 Nov release

STEP 2:
Review external ongoing social listening efforts to establish a fact-based narrative on platform and brand safety

Status: In progress
Inputs: N/A
Sources: Agencies, CyberWell
Timing: ongoing
Steer Team to review insights ahead of 1 Dec meeting

STEP 3:
Publish and make public existing GARM Platform Implementation Grids to monitor Twitter's upholding of key obligations

Status: Update complete
Inputs: s 47F GroupM refresh
Platform comment period
Partners: N/A
Timing: Platform comment 7 Nov,
Member distribution 28 Nov

Twitter: GARM Meeting Plan

Date / Time	1 Dec 2022 @ 8a NYC [virtual]
Attendees	WFA ExecCo + GARM Twitter Executive Team [confirmed] + TBD
Format	• GARM to recommend agenda + questions [sent as preread] • Questions to be led by ExecCo members [need to brief] • s 47F to facilitate [impartially]
Preread	Share out of identified issues *with* corroboration sourced from: • Snapshot of Twitter status in GARM • WFA Advertiser Survey • Social Listening
Moderation	Allow for opening remarks s 47F must tightly manage agenda, questions and answers s 47F to prep individual leaders in the ExecCo
Working List of Topics to Probe [to be reformulated with data into the agenda]	1. **Staffing + Resourcing:** Check in on team, resourcing, avoid overuse of tech that has failed on things like CSAM [you can't have tech moderate on nuanced areas like cyberbullying] 2. **Business Model:** Plans to protect negative use of Blue + Birdwatch for moderation and monetization [you can't buy your way into better content moderation] 3. **Relitigating Industry Standards + Commitments:** New narrative on free speech v hate speech needs to be addressed as a slip away from Common Definitions [you can't have new terms that reinvent trust & safety] 4. **Practical Concerns:** Reflect on the positive and negative insights from the survey

From: Julie Inman Grant
Sent: Friday, 11 November 2022 10:20 PM
To: Rob Rakowitz
Cc: s 22
Subject: Re: Twitter [SEC=UNOFFICIAL]

Hi Rob:

Thank you for sharing and well done for leading this important work.

Amazing what can happen in just a few days - talk about a total Twitter meltdown! It makes me so sad for the company I really loved working for - and for my former colleagues.

Your line of questioning is broadly consistent with the concerns I raised in my letter and I imagine you gleaned more following the Spaces call to the advertising community.

You have some very powerful levers at your disposal. We would be grateful if GARM can keep us updated on how Twitter responds and share any information, so we can take into account in our engagement and regulatory decisions.

I'm on my way to DC tomorrow and am looping in s 22 who is on point for these issues on my team. It will certainly be fascinating to watch this all unfold.

Julie

Sent from my iPhone

> On 9 Nov 2022, at 10:59 pm, Julie Inman Grant s 47E(d) @esafety.gov.au> wrote:
>
> That must be soul destroying, Rob!
>
> I worked in the 102nd Congress at a time when members worked across parties on issues of mutual concern. America is not the country of promise I grew up in or left 22 years ago. It makes me sad to see the societal polarisation and deterioration on so many fronts - and frankly, the violence that permeates discourse-in the name of free speech - not to mention the more obvious Second Amendment.
>
> Move to Australia!!!😉
>
> Julie
>
> Sent from my iPhone
>
>> On 9 Nov 2022, at 10:53 pm, Rob Rakowitz s 47F @wfanet.org> wrote:

Yes - we need to connect on that too
I have a playbook and CMO call to action if like your input on

I need to be spending more time with you

Oh and election results
My main thing is I need to see Trump and denials effectively sidelined but I am afraid the contagion is too widespread to protect infection overall

I have little faith in either part establishments - and I say this as a first generation American who is a refugee from anti Jewish persecution in pre-state Israel and Iraq.

Rob Rakowitz
Initiative Lead - Global Alliance for Responsible Media
World Federation of Advertisers

s 47F / s 47F @wfanet.org

From: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Sent: Wednesday, November 9, 2022 6:48:14 AM
To: Rob Rakowitz s 47F @wfanet.org>
Subject: Re: Twitter [SEC=UNOFFICIAL]

Let me know what you hear - I just met with s 47F today, who was here launching a metaverse safety initiative....

Sent from my iPhone

> On 9 Nov 2022, at 10:45 pm, Rob Rakowitz
> s 47F @wfanet.org> wrote:
>
> Also saw
> Asking them now
>
> Rob Rakowitz
> Initiative Lead - Global Alliance for Responsible Media
> World Federation of Advertisers
>
> s 47F / s 47F @wfanet.org
>
> ---
>
> **From:** Julie Inman Grant s 47E(d) @eSafety.gov.au>
> **Sent:** Wednesday, November 9, 2022 6:41:13 AM
> **To:** Rob Rakowitz s 47F @wfanet.org>
> **Subject:** Re: Twitter [SEC=UNOFFICIAL]
>
> Rob - It’s getting late here so I will take a look tomorrow and will keep confidential - will suggest a few times when we

might connect but open to discussing vectors for mutual support.

Just saw that the Meta layoffs hit - hope it doesn’t disproportionately hit trust and safety as well. I suspect this was a bit more about tackle overall “bloat and underperformance” but will watch…

Julie

Sent from my iPhone

On 9 Nov 2022, at 10:32 pm, Rob Rakowitz s 47F @wfanet.org> wrote:

So lovely to hear from you!

I am sharing the attached in confidence with you.
There is an ongoing major intervention.

PDF sets out a plan
Email is a short term RFI

Neither of these are in public domain, yet. And would be good to cross-check them and see how they line up with you.
It would be great to line up together around a shared agenda.

I am also writing a piece now on how brand safety has nothing to do with controlling freedom of speech and everything to do with correcting an imperfect marketplace that has led to things like Molly Russell’s suicide.

I’m getting tired of the false debates.

Let’s 100% connect – maybe virtual in the short term.
100% would love to see the team in person.

Best as ever – you are an inspiration!
Rob

Rob Rakowitz
Initiative Lead - Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with

teams working across multiple time zones.
Although I have sent this at a time that is convenient for me, it is not my expectation that you read,
respond or follow up on this email outside your hours of work.

From: Julie Inman Grant
s 47E(d) @eSafety.gov.au>
Date: Wednesday, November 9, 2022 at 06:11
To: Rob Rakowitz s 47F @wfanet.org>
Subject: Re: Twitter [SEC=UNOFFICIAL]

Hi Rob - I saw the GARM post on LinkedIn and was really pleased to see this proactive approach. I would be very interested in seeing your six points. I believe GARM has significant collective power in helping to hold the platforms to account and I appreciate your informing the importance of both brand and platform safety.

Interesting to hear your perspective on Blue & Birdwatch. It is very hard to piece together what seems like a very erratic and reactive strategy. On one hand, this could be seen as a very basic revenue generating machination and a way to target the bots and scammers by undermining the economic incentive to mass create accounts….but, they clearly haven't thought through the multiple spill on events.

I gather you saw that I wrote to Mr Musk and was questioned about my thoughts at Senate Estimates?

Not sure I will get an immediate response…here is the letter, since it is now in the public record.

As an aside, I'll be in DC next week launching a global online safety regulatory network at FOSI, then in Seattle and San Fran. My team member, s 22, will be up in NYC on the 18th but happy to try and touch base further whilst I'm stateside.

How 'bout this election???????

Thanks for reaching out and for advocating for such important change! Julie

Sent from my iPhone

> On 9 Nov 2022, at 9:47 pm, Rob Rakowitz s 47F @wfanet.org> wrote:

Hi Julie -

I hope you've been well.
I've issued Twitter a detailed list of questions and I fear we won't hear back on them in a meaningful way.

One of the 6 areas I raised was how collapsing Blue and Birdwatch will increase the quality of content and avoid abuse? As we know bad actors in misinfo are many times stat funded and coordinated $8 accounts to tip a content rating and an algorithm recommendation engine is a clear issue. The only thing I can thing is perhaps it's intentional.

Happy to connect and share our overall plan and questions.

Let me know if there's a desire to connect.

Best

Rob

Rob Rakowitz
Initiative Lead - Global Alliance for Responsible Media
World Federation of Advertisers

s 47F /
s 47F @wfanet.org

NOTICE: This email message is for the sole use of the intended recipient(s)
and may contain confidential and privileged information. Any unauthorized
review, use, disclosure or distribution is prohibited. If you are not the
intended recipient, please contact the sender by reply email and destroy all copies of the original message.

<Twitter GARM Steer Team Observations Action Plan.pdf>
<Re- GARM- Advice on industry-wide communications.eml>

> NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

From: Julie Inman Grant
Sent: Thursday, 22 June 2023 5:47 AM
To: s 47F @wfanet.org
Cc: eSafety Parliamentary
Subject: eSafety Commissioner takes regulatory action against Twitter around Online Hate
Attachments: Media release - eSafety demands answers from Twitter about how its tackling online hate.pdf

Dear s 47F

I hope you are keeping well. I wanted to let you know that we have taken this regulatory action today against Twitter on online hate. The brand safety leverage you and GARM have been able to extract is incredible and we hope that this action will help further shine a light on the safety shortcomings currently pervading the platform.

I believe that transparency is vital to ensuring that online services and platforms are safe by design. Without transparency, there can be no meaningful accountability from the global giants shaping our society, enabling our discourse, and facilitating unprecedented communications.

In January last year, stronger modernised online safety protections under the Online Safety Act took effect in Australia. In addition to enhancing eSafety's powers to tackle specific harms such as adult cyber abuse, image-based abuse, child cyberbullying and illegal content, the Act gives me the ability to require information from companies about how they are keeping their users safe.

These Basic Online Safety Expectations ('BOSE') place transparency at the heart of our regulatory model. They are a novel framework of powers. Through their use, eSafety can compel companies to 'show us their working' on specific online safety concerns, rather than being shielded by marketing spin or glossy handouts. By using these powers, eSafety is rapidly developing a strong baseline understanding of where industry is doing well, but where there is more work to do, to harden their services from abuse and malfeasance.

Today, I have issued a BOSE notice to Twitter, challenging the company to explain what they are doing to combat online hate. Twitter has 28 days to respond to the notice and a failure to comply may attract a penalty of up to AUD$687,500 per day.

By taking this step, I aim to shed light on how Twitter is addressing what appears to be a recent surge in hate on the platform, both general and targeted. In particular, I want to understand how Twitter is enforcing its own clear rules prohibiting hateful conduct, and how trust and safety is enabled within the company.

Unfortunately, our experience and that of others suggests that Twitter is falling well short of the mark in both respects.

eSafety has received more complaints about online hate on Twitter than any other service in the last 12 months, with many of these appearing to coincide with the change in ownership last October. The increase overlaps with the platform's reported reinstatement of over 62,000 accounts previously banned for breaching Twitter rules, including 75 with more than 1 million followers. I am concerned that these accounts are playing an outsized role in fuelling the platform's toxicity.

The impact of hate on marginalised communities is not a theoretical concern. New eSafety research has found that 1 in 5 Australians have experienced online hate in the last 12 months, and we know that First Nations people and members of the LGBTQI+ community, face hate at twice the rate of the national average. Overall, one in six adults targeted by online abuse report that their physical health suffered as a result; the figure rises to one in three when emotional and mental wellbeing is considered.

As with previous notices, eSafety will release a report summarising the information we receive. I will keep you updated on the outcome of this process, and our findings.

Thank you again for your important contribution to our collective work of making the internet a safer place for all.

All the best,

Julie

Julie Inman Grant
Commissioner

eSafetyCommissioner

☎ s 47E(d)
[redacted]
esafety.gov.au

eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

Media release **EMBARGOED UNTIL 12.01am 22 JUNE 2023**

eSafety demands answers from Twitter about how it's tackling online hate

Australia's eSafety Commissioner has issued a legal notice to Twitter seeking information about what the social media giant is doing to tackle online hate on the platform.

eSafety received more complaints about online hate on Twitter in the past 12 months than any other platform and has received an increasing number of reports of serious online abuse since Elon Musk's takeover of the company in October, 2022.

The rise in complaints also coincides with a slashing of Twitter's global workforce from 8,000 employees to 1,500 including in its trust and safety teams, coupled with ending its public policy presence in Australia.

This is at the same time a 'general amnesty' was announced by Musk in November, which reportedly saw 62,000 banned or suspended users reinstated to the platform, including 75 accounts with over 1 million followers.

eSafety Commissioner Julie Inman Grant said Twitter's terms of use and policies currently prohibit hateful conduct on the platform, but rising complaints to eSafety and reports of this content remaining publicly visible on the platform, show that Twitter is not likely to be enforcing its own rules.

"We are seeing a worrying surge in hate online," Ms Inman Grant said. "eSafety research shows that nearly 1 in 5 Australians have experienced some form of online hate. This level of online abuse is already inexcusably high, but if you're a First Nations Australian, you are disabled or identify as LGBTIQ+ you experience online hate at double the rate of the rest of the population.

"Twitter appears to have dropped the ball on tackling hate. A third of all complaints about online hate reported to us are now happening on Twitter.

"We are also aware of reports that the reinstatement of some of these previously banned accounts has emboldened extreme polarisers, peddlers of outrage and hate, including neo-Nazis both in Australia and overseas."

eSafety is far from being alone in its concern about increasing levels of toxicity and hate on Twitter, particularly targeting marginalised communities.

Last month, US advocacy group GLAAD designated Twitter as the most hateful platform towards the LGBTQ+ community as part of their third annual social media index.

Research by the UK-based Center for Countering Digital Hate (CCDH) demonstrated that slurs against African Americans showed up on Twitter an average of 1,282 times a day before Musk took over the platform. Afterwards, they more than doubled to an average of 3,876 times a day.

The CCDH also found that those paying for a Twitter Blue Check seemed to enjoy a level of impunity when it came to the enforcement of Twitter's rules governing online hate, compared to non-paying users and even had their Tweets boosted by the platform's algorithms.

The Anti-Defamation League (ADL) also found that antisemitic posts referring to Jews or Judaism soared more than 61 per cent just two weeks after Musk acquired the platform.

“We need accountability from these platforms and action to protect their users and you cannot have accountability without transparency and that’s what legal notices like this one are designed to achieve,” Ms Inman Grant said.

This latest notice on online hate follows a bid in February to get answers from the platform (along with TikTok, Google YouTube, Twitch and Discord) on the steps the company is taking to address child sexual exploitation and abuse, sexual extortion and the promotion of harmful content by its algorithms.

eSafety is currently assessing the responses to those notices and expects to release appropriate information in due course.

If Twitter fails to respond to the most recent notice within 28 days, the company could face maximum financial penalties of nearly $700,000 a day for continuing breaches.

eSafety's regulatory powers under the Online Safety Act cover serious adult online abuse as well as the cyber bullying of children and image-based abuse. In some cases, hate speech may meet the statutory thresholds of adult cyber abuse. eSafety encourages all individuals who feel they have been the target of online abuse to report to the platform and, if the platform fails to act, to report to eSafety at www.esafety.gov.au/report.

eSafety makes its regulatory decisions impartially and in accordance with the legislative test prescribed in the Online Safety Act.

For more information or to arrange an interview, please phone 0439 519 684 or email media@esafety.gov.au

From: s 47F @wfanet.org>
Sent: Thursday, 22 June 2023 2:11 PM
To: Julie Inman Grant
Cc: eSafety Parliamentary
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate
Attachments: GARM Generative AI and Metaverse Playbook[74].pdf

Hi Julie –

Thanks for this.
I am confirming receipt.

Are you and the team available for a call in the coming days?
There are some updates for me to share with you as well.

We are aware of the issues and we have a series of steps already underway, but I am skeptical of the outcomes.
National associations like ISBA, who is a board member of mine, are advising members accordingly.

Transparently, as an industry association with antitrust provisions and developing standards, it is hard for us to manage corrective measures beyond driving transparency on issues and suggested remedies. We are not a watchdog and rely on NGOs like ADL who are on our NGO Consult Group to raise the issue.
You also may have also seen some updates on how the US GOP perceive our work.

Finally, I've attached a playbook we are releasing today on Generative AI and the metaverse. It would be great to get your feedback and discuss a meaningful route forward.
We're eager to understand how we can look at market-facing anticipatory steps in these areas.

Let me know when works for you – I'd really value your personal guidance on some of the issues I am facing into, it'd be good regroup.

Best,

s
47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

*WFA values and encourages flexible working patterns, with teams working across multiple time zones.
Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.*

From: Julie Inman Grant
Date: Wednesday, June 21, 2023 at 21:47
To: s 47F

Cc: eSafety Parliamentary
Subject: eSafety Commissioner takes regulatory action against Twitter around Online Hate

Dear s 47F

I hope you are keeping well. I wanted to let you know that we have taken this regulatory action today against Twitter on online hate. The brand safety leverage you and GARM have been able to extract is incredible and we hope that this action will help further shine a light on the safety shortcomings currently pervading the platform.

I believe that transparency is vital to ensuring that online services and platforms are safe by design. Without transparency, there can be no meaningful accountability from the global giants shaping our society, enabling our discourse, and facilitating unprecedented communications.

In January last year, stronger modernised online safety protections under the Online Safety Act took effect in Australia. In addition to enhancing eSafety's powers to tackle specific harms such as adult cyber abuse, image-based abuse, child cyberbullying and illegal content, the Act gives me the ability to require information from companies about how they are keeping their users safe.

These Basic Online Safety Expectations ('BOSE') place transparency at the heart of our regulatory model. They are a novel framework of powers. Through their use, eSafety can compel companies to 'show us their working' on specific online safety concerns, rather than being shielded by marketing spin or glossy handouts. By using these powers, eSafety is rapidly developing a strong baseline understanding of where industry is doing well, but where there is more work to do, to harden their services from abuse and malfeasance.

Today, I have issued a BOSE notice to Twitter, challenging the company to explain what they are doing to combat online hate. Twitter has 28 days to respond to the notice and a failure to comply may attract a penalty of up to AUD$687,500 per day.

By taking this step, I aim to shed light on how Twitter is addressing what appears to be a recent surge in hate on the platform, both general and targeted. In particular, I want to understand how Twitter is enforcing its own clear rules prohibiting hateful conduct, and how trust and safety is enabled within the company.

Unfortunately, our experience and that of others suggests that Twitter is falling well short of the mark in both respects.

eSafety has received more complaints about online hate on Twitter than any other service in the last 12 months, with many of these appearing to coincide with the change in ownership last October. The increase overlaps with the platform's reported reinstatement of over 62,000 accounts previously banned for breaching Twitter rules, including 75 with more than 1 million followers. I am concerned that these accounts are playing an outsized role in fuelling the platform's toxicity.

The impact of hate on marginalised communities is not a theoretical concern. New eSafety research has found that 1 in 5 Australians have experienced online hate in the last 12 months, and we know that First Nations people and members of the LGBTQI+ community, face hate at twice the rate of the national average. Overall, one in six adults targeted by online abuse report that their physical health suffered as a result; the figure rises to one in three when emotional and mental wellbeing is considered.

As with previous notices, eSafety will release a report summarising the information we receive. I will keep you updated on the outcome of this process, and our findings.

Thank you again for your important contribution to our collective work of making the internet a safer place for all.

All the best,

Julie

Julie Inman Grant
Commissioner

Australian Government | eSafetyCommissioner

s 47E(d)

[redacted]

esafety.gov.au

eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

GARM Global Alliance for Responsible Media

GARM Safe & Suitable Innovation Guic

Assessing brand safety and suitabili concerns for Generative AI + the Me

Introduction

An open, accessible and safe Internet that respects user control is in everyone's interest. Advertising has helped the Internet develop and will likely help develop its next phases, shaped by two new separate technological waves; the metaverse and generative AI.

Over the last four years, the Global Alliance for Responsible Media has been supporting the advertising industry in respect to safety in the digital social media sector. Our focus is on monetization and where ads are placed. We have also provided support with regard to a voluntary roadmap and flexible frameworks which promote improved transparency, controls and consistency.

In the four years that we've operated we've seen new platforms and formats enter the industry landscape. There are two changes emerging where we see the potential to inspire safety as these new digital technologies take shape – namely the metaverse and generative AI.

These are two separate technologies that are different; the metaverse represents an evolution of the digital medium, whereas generative AI represents a new technology. While these are different and distinct and each have their own different levels of concreteness and development, industry stakeholders should consider steps to encourage safe development, safe exploration – as well as safe monetization of these new territories.

This document is meant to help advertising industry stakeholders – whether ad buyers, ad sellers, or enabling partners. This playbook is aimed at providing background and support for industry stakeholders as they consider how to experiment, how to structure and connect teams that straddle advertiser, agencies and platform organizations and learn what to focus on to innovate with brand safety and suitability in mind.

This playbook is the result of collaboration from independent GARM members across a diverse range of industries.

We would like to thank GARM members who helped review research and create this guide, specifically Mars, Unilever, P&G, GroupM, Meta, Google, Publicis, Vodafone, 4As, ISBA, WFA, UM, Roblox, BSI

We recognize that every organization's experience and journey in these areas are unique and will range from the curious to the educated to the exploring to the expert. This playbook is meant to be a flexible resource to organizations and teams, no matter where they are in the journey.

- For the more advanced in these technologies, this playbook can help reaffirm current practice and help align teams and partners.

- For those starting out or curious, this playbook can help educate you on how to navigate advertising opportunities that leverage these technologies.

Finally, we should note that these technologies are energy intensive. We should encourage these technologies to be responsible-by-design and sustainable-by-design. User safety, Brand safety, Privacy and Sustainability are becoming the four table stakes in the media industry and should be extended to the way metaverses and AI are designed and built.

From insight to provocation

In GARM's charter we acknowledge that digital and social media have both a light and a shadow side. Much of our work has been dedicated to supporting that advertising doesn't play into online harms and that the industry has the necessary voluntary guidelines so that content monetization is more transparent and consistent.

As an industry, it's essential that we reflect upon our journey to this point to guide our go-forward in these next areas.

With these learnings and provocations in mind, we can now encourage ourselves to:

1. Test
2. Continue to evolve and build controls
3. Leverage the capabilities and features across the value chain to improve safety

What did we learn?	Where do we see it?	What does it mean as we progress?
Digital media company policies must be clear, consistent and enforced evenly to manage fluid events and innovation	Platform policy exceptions on political officials, celebrities and ensuing oversight boards	Systems, programs and experiences should go through a stress test or dry run before being made generally available or promoted
Comprehensive regulation can embrace industry development of new technology, user empowerment that embraces choice and protection	Digital safety regulation and compliance mechanisms in UK, EU, Australia	Advertisers should do their part in making needs known, while advancing the creation of complimentary frameworks for advertiser controls
Centres of Excellence on digital safety can provide guidance, but governance and compliance measurement are key to effective safety implementation	Imperfect integration of content recommendation engines, user safety and privacy systems	Third-party, audit and accreditation of technology systems and transparency tools (like blockchain) will become increasingly important
Dangers such as deepfakes, misrepresentation, fraud and underage usage will require verification and identity management resources on the technology industry and advertiser side	Fake corporate handles on social media at new product launches, counterfeit NFTs, celebrity deepfakes, Generative AI faked photos in political ads	Platforms, apps, hardware providers will need to work towards verification and appropriately apply trademark and intellectual property standards to protect individuals from fraud. Content and experiences should have appropriate labelling and access, managed by both developers and platforms with safety in mind

Generative AI

There's been a lot of activity in and coverage of Generative AI (GAI) since Chat GPT became public facing on Nov 30, 2022. Within two months of its launch, it reached 100 million active users (what took TikTok 9 months to achieve). There are now 600 new companies joining a rush to develop this new technology, according to an initial scan of CB Insights.

As with most new technology there's been equal calls of euphoria and agony. And most recently G-7 heads of state have discussed the need for a global regulatory framework to address potential harms. There have been calls for a 6-month moratorium in the development of GAI by leading academics and practitioners, but as we have seen in new technologies the locomotive will be hard to stop.

How should the industry begin to think about this new technology?

Advertising support of generative AI is still yet to come, but generative AI will certainly affect the advertising industry – whether realizing the promise of dynamic creative and media optimization, or potentially stripping publishers of traffic that would have otherwise been monetizable.

The impact of Generative AI on advertising and publishing will be real but have yet to be fully understood. The purpose of this playbook is not meant to explore the creative opportunities that GAI will represent to marketing, rather we will focus on helping understand and mitigate some of the risks as it relates to content generation and its impact on brand safety and suitability. Part of the industry's approach can be informed by policies already taken on manipulated media; however, the automation tools and the scale should cause us to reflect on whether the approach is fit for purpose and fit for the future .

Tactically speaking, we have already seen risks in cross-border data transfers for large language models, proliferation of hyper realistic deepfakes in political ads, the frictionless creation of polished made-for-advertising web sites and content sites replacing human writers with automated article generation.

These tactical risks are clear, but what are some of the strategic risks that may arise?

Negative Use Cases: GAI can be weaponized by users with negative intent and the primary threat vectors breakdown into two areas:

1. **Harmful content at scale:** The risks of GAI weaponized on a personal (phishing, cyberbullying), group (hate speech), or system (deepfakes) is real. Advertisers, agencies and platforms may need to safeguard themselves from the presence and monetization of this content, created in an automated way and at scale. The risks for coordinated attack will similarly increase.
2. **Misrepresentation and misinformation at scale:** As discussed in the metaverse section, discerning between parody and trolling will be hard. Similarly, determining fantasy depiction versus propaganda will become equally hard. In essence, intent will be increasingly important – a known area for interpretation and hard for the industry to develop standards around.

Generative AI

Information Sources and Intellectual Property: GAI will also struggle until information sources are addressed in two ways:

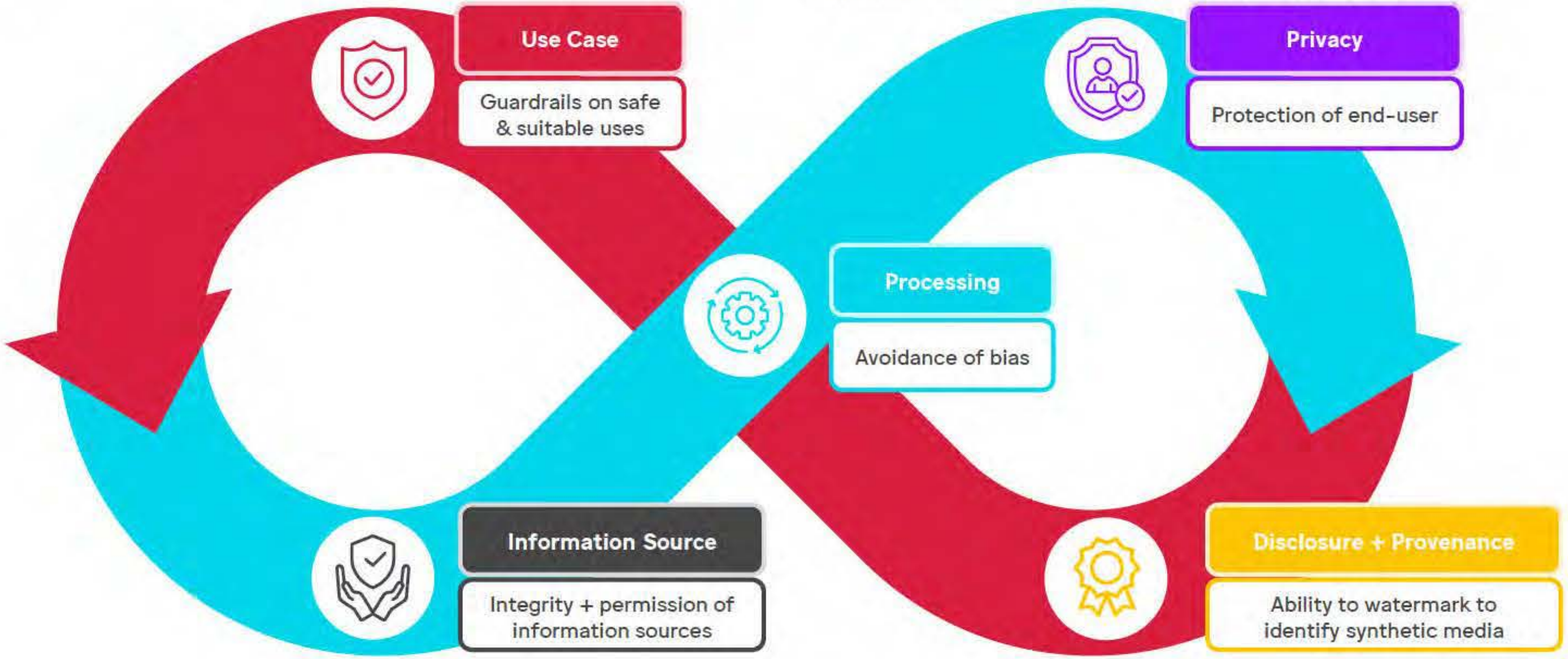
1. **Explainability and governing input sources:** As with all processes – the outputs are usually only as good as the inputs. You can imagine a GAI engine that is reliant on questionable or bad content may produce negative content and conversely quality content into GAI can result in quality content. GAI engines may need to develop selection standards and simultaneously consider user control over content sources.
2. **Fair and mutual business practices with input sources:** GAI may reduce time needed to visit several sources and subscribe to content. Publishing, journalism, music creation and recording are all industries that may be disintermediated in one way or another – will you need to visit said news site, will said singer need to record a new song? Copyrights and intellectual property will become a third rail issue unless a mutual business model is developed.

Bias in processing: Fairness and inclusion data practices in AI to avoid bias in processing: GAI will also need to ensure that processing technologies avoid bias.

When designing and training generative AI systems, it is important to use diverse and representative data sets to reduce the risk of bias and to carefully evaluate the quality and accuracy of the generated output. Developers of GAI will also need to ensure that processing technologies avoid bias. When designing and training generative AI systems, it is important to use diverse and representative data sets to reduce the risk of bias and to carefully evaluate the quality and accuracy of the generated output

Privacy and Confidentiality: GAI app developers will need to provide disclosure on how queries and results are fed back into the engine – are queries and the results indexed and open for review? Users' privacy will need to be clarified as GAI are privy to certain sensitive topics. Enterprise and business users will also need to have confidentiality and security systems.

Disclosure & Transparency in Production: GAI will scale the creation of net new content. To an end user or a publisher, how transparent is the disclosure?



Use Case

Guardrails on safe & suitable uses

Privacy

Protection of end-user

Processing

Avoidance of bias

Information Source

Integrity + permission of information sources

Disclosure + Provenance

Ability to watermark to identify synthetic media

What should the GARM stakeholders expect?

1 Industry-wide efforts to introduce digital watermarking that enhance disclosure & provenance : We have started to see some efforts started by Adobe (Content Authenticity Initiative), Microsoft (Project Origin) and then an umbrella effort to create a standard for content like Coalition for Content Provenance and Authenticity (C2PA) forum that brings together the imaging industry, editing software and news media. These are encouraging signs and is a promising use case of blockchain technology. However, we may need to consider how open and willing platforms will be in accepting these signals into their systems and consider if and how they should be consumer-facing. Further, it will be natural to see potential links in this area with groups focused on security, like the Global Internet Forum for Counter Terrorism (GIFCT) and groups combatting counterfeiting efforts such as the OECD.

2 Co-regulatory frameworks will evolve but the jury is out on if they will be comprehensive and support technology market development and media industry protection: The threat vectors for GAI are clear in Use Case and Information Source and Processing needs. We anticipate regulators and industry to work together to develop a series of standards, based on initial inquiries in the US, UK, EU, Australia individually and more recently at intragovernmental conversations at the G7 Summit in Hiroshima. The advertising industry should assess developments through the framework above to assess whether or not it is comprehensive enough for our industry needs.

3 Tough conversations on mutual business models: GAI engines can upend creative ownership & compensation; for a musician it could be their voice, for publisher it could be their content. GAI's ability to create content could disintermediate creative owners from end-users & commercialization. We can already see some rightsholder concern and the threat of legal challenge in certain key markets, which may spur governments to weigh in on these debates. We've seen similar tensions of content licensing for news publishers in Search and this will likely only be the beginning.

What should GARM do in the midterm?

Outside of GARM	Within GARM
Support the creation of and adoption of voluntary provenance solutions through the content creation, moderation and monetization lifecycle	Seek to understand on an industry-wide and aggregated basis how ad sellers and content targeting companies may spot and assess synthetic media created by GAI tools
Support further independent marketplace marketplace development and safeguards for the advertising industry	Seek to understand on an industry-wide and aggregated basis how GAI content creation engines will be assessed for quality, integrity and risk
Support the industry the industry with its face into tensions on commercialization and disintermediation	Support the exploration of potential common understanding on Use Case, Information Source and Processing that feed into platform safety and brand safety
Continue to monitor how specific applications address privacy, confidentiality and disclosure and consider this as core elements of 'Platform Safety' for Generative AI	
Watch for regulation that addresses risks to users and industry, while allowing for marketplace development	

The Metaverse: defining it, appreciating its variety, understanding responsibilities

The metaverse has been aptly defined by Matthew Ball as

"a massively scaled and interoperable network of real-time rendered 3D virtual worlds which can be experienced synchronously and persistently by an effectively unlimited number of users with an individual sense of presence and with continuity of data, such as identity, history, entitlements, objects, communications and payments."

This definition is robust and rigorous and forward looking as it does not exist in this sense, today.

For the advertising industry stakeholder, we can also describe the metaverse using IAB's description of

"a collection of virtual spaces, or digital worlds, in which users can create content, interact with others as avatars or digital versions of themselves and move freely between worlds."

There are a series of criteria that can be used to determine if an experience should be considered as a metaverse experience:

- Immersive and expansive
- Interoperable and uncapped
- Independent and exchangeable
- Interactive, live, synchronous
- Indefinite and no geography

This guide

The most popular entry points for consumers in the metaverse is gaming, currently. Some of these metaverse platforms are at scale, while several of them are still nascent. Putting scale aside, the most interesting thing to note, is the variation of user controls and behaviours that are possible.

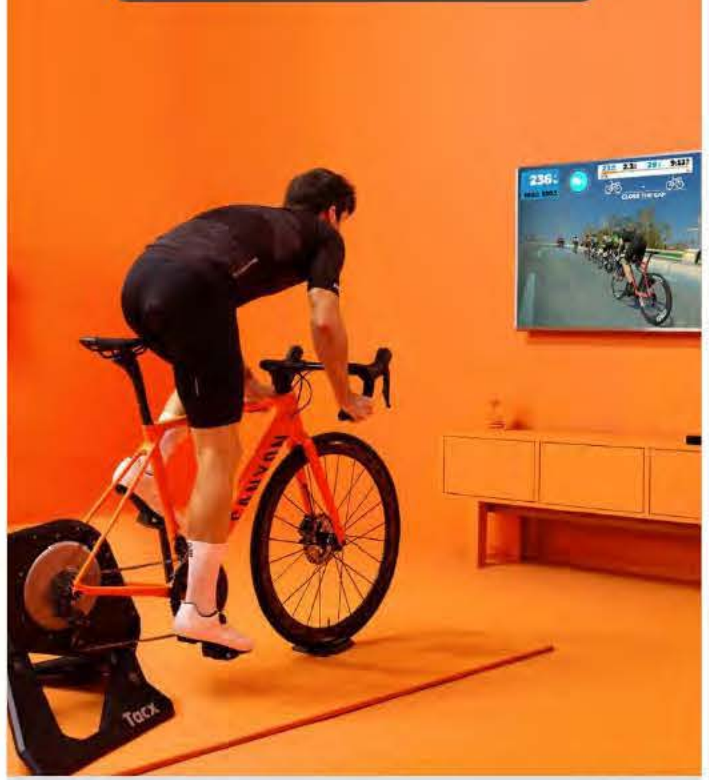
We see a continuum starting to play out in the types of platforms, environments and controls available. There are three gradations of metaverse consumer-controlled experiences, ranging from Fixed to Hybrid to Fluid.

Current points of entry

The most popular entry points for consumers in the metaverse is gaming, currently. Some of these metaverse platforms are at scale, while several of them are still nascent. Putting scale aside, the most interesting thing to note, is the variation of user controls and behaviours that are possible.

We see a continuum starting to play out in the types of platforms, environments and controls available. There are three gradations of metaverse consumer-controlled experiences, ranging from Fixed to Hybrid to Fluid.

Fixed



4mm
Monthly Actives
3.78 miles

Hybrid

100mm
Monthly Actives
18 hours

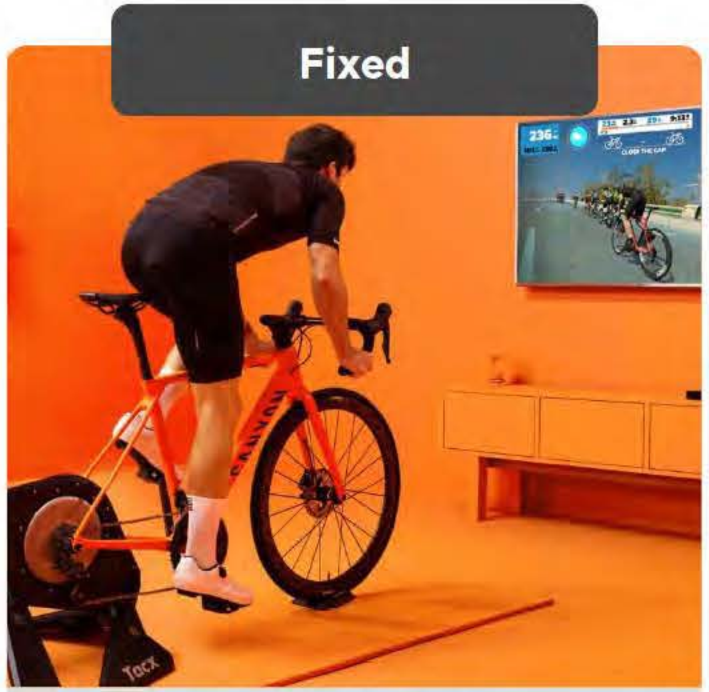
ROBLOX
MINECRAFT
FORTNITE

Fluid

Active
players
300m

Wild variations of consumer autonomy in a synthetic environment and in corresponding community represent different levels of risk. Some harmful incidents have been documented in the early days of each platform and they can be telling on the types of abuses that may be possible in each environment. Therefore, it is essential for us to consider a potential ~~to~~ additional layer to the model to identify general baseline model risk.

Managing risks in the metaverse is achievable but may potentially require a new frame of reference. Given the distributed nature of the experiences, concurrent community use and the multiple layers of technology, an updated layered approach to safety might be considered based on the value chain of how metaverse experiences are delivered.

	Fixed	Hybrid	Fluid
			
Overview	An environment where the user can only do a finite set of actions in the experience that are known and controlled	An environment where users can do a set of actions and the experiences can be user controlled but are limited by software rules	An environment where the user has full autonomy to perform open-ended actions that mimic the real behaviours or enhanced behaviours via things like avatars
Selected Examples	Zwift	Roblox, Minecraft	VR Chat
Risk	LOW	MEDIUM	HIGH

As we look at it, there are two complementary layers of safety relevant to advance:

Platform Safety

Is the platform safe for users?

- Design
- Access
- Recommendation & Moderation

Brand Safety

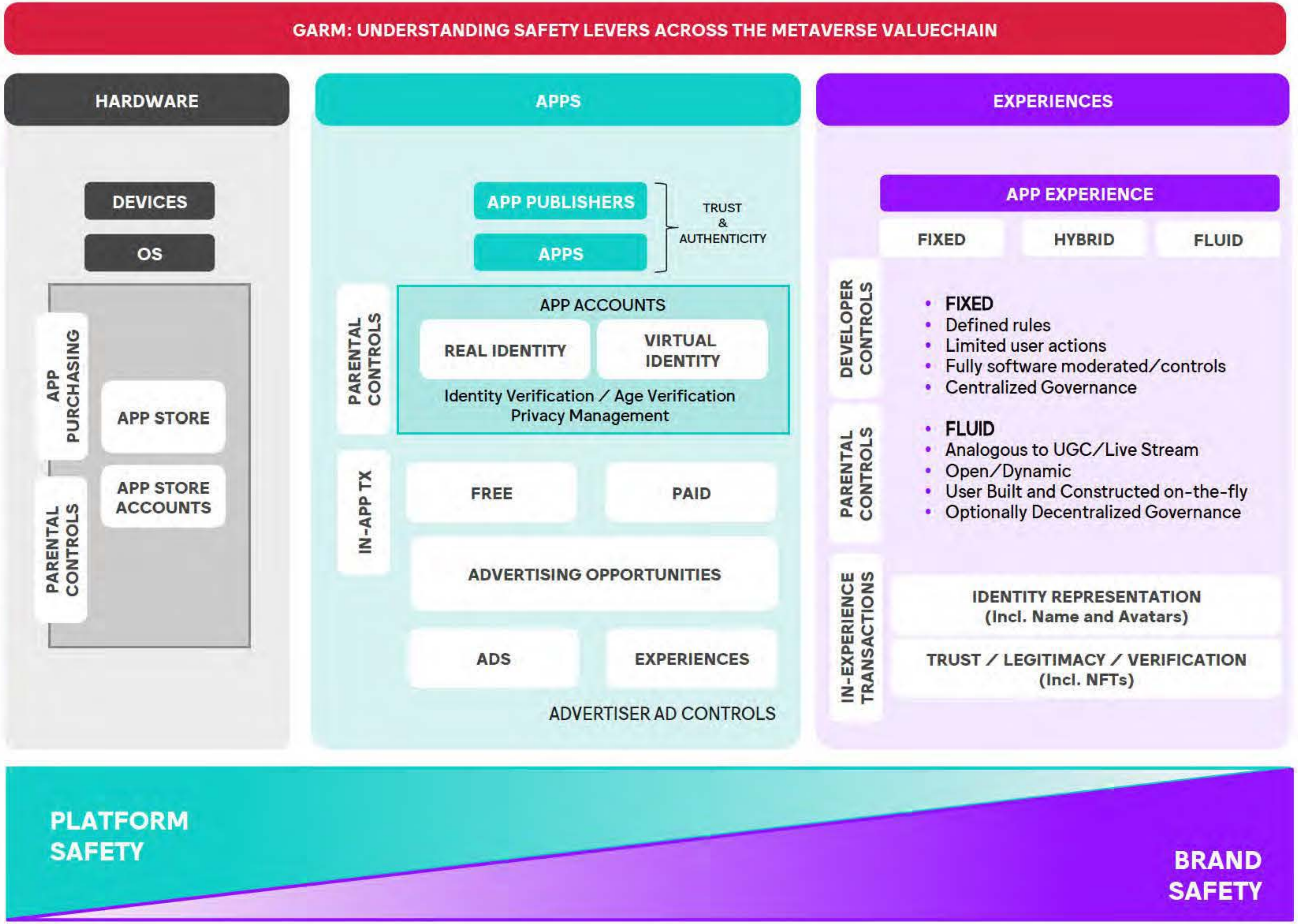
Is the platform safe for advertising support?

- Monetization
- Integration of ad experience (marketing model)

Platform and Brand safety are delivered differently in the metaverse because of the layered approach seen in the graphic.

	Action Required	Responsibilties
Hardware	Devices/ Operating Software	Design safety Access safety
Software	App Store Application	Design safety/ Access safety Behaviour safety
Experience	Marketer, Platform, Agency	Design safety Access safety

A layered and sequenced approach to safety is important to ensure that metaverse experiences are based on user consent & control, keeping communities safe and commercialization via advertising appropriate to the advertiser.



GARM: UNDERSTANDING SAFETY LEVERS ACROSS THE METAVERSE VALUECHAIN

HARDWARE

DEVICES

OS

APP PURCHASING

PARENTAL CONTROLS

APP STORE

APP STORE ACCOUNTS

APPS

APP PUBLISHERS

APPS

TRUST & AUTHENTICITY

PARENTAL CONTROLS

APP ACCOUNTS

REAL IDENTITY

VIRTUAL IDENTITY

Identity Verification / Age Verification
Privacy Management

IN-APP TX

FREE

PAID

ADVERTISING OPPORTUNITIES

ADS

EXPERIENCES

ADVERTISER AD CONTROLS

EXPERIENCES

DEVELOPER CONTROLS

PARENTAL CONTROLS

IN-EXPERIENCE TRANSACTIONS

APP EXPERIENCE

FIXED

HYBRID

FLUID

- **FIXED**
- Defined rules
- Limited user actions
- Fully software moderated/controls
- Centralized Governance

- **FLUID**
- Analogous to UGC/Live Stream
- Open/Dynamic
- User Built and Constructed on-the-fly
- Optionally Decentralized Governance

IDENTITY REPRESENTATION (Incl. Name and Avatars)

TRUST / LEGITIMACY / VERIFICATION (Incl. NFTs)

PLATFORM SAFETY

BRAND SAFETY

CASE STUDY:

Lifestyle apparel creates a virtual sports world to gather its community and offer up unique merchandise [Hybrid Experience + Commerce]

Looking at a real case study, a sports lifestyle apparel brand created an immersive community experience for users. The activation allowed for users to customize their avatars, engage in a series of sports and chat with each other.

The overall activation could be categorized as being a Hybrid Experience.

The advertiser, agency and app identified the following safety and suitability concerns:

1. **Appropriate access:**
   The program needed to ensure that users were 13+ and in the US
2. **Appropriate use:**
   The program needed to ensure that avatar customization and lifestyle apparel was used in appropriate ways
3. **Appropriate interaction:**
   The program needed to ensure community interaction avoids harassment, cyberbullying, teasing, 'trash talk'
4. **Commerce integration:**
   The program's ecommerce capabilities needed to allow for a near frictionless transaction

The advertiser, agency and platform developed a series of responsibilities based on the priorities above and enlisted partners in the value chain to ensure that they were realized through a series of business rules:

Priority	Value chain accountabilities	Business Rule
Access	Devices/ Operating Software	Ensure that device ID's can be age verified Ensure that application can age verify
Use	Developer	Ensure that brand can be used in fixed ways
Interaction	Devices/ Operating Software	Block slurs Moderate comments
Commerce	Developer/ Commerce team (payments, fulfilment)	Ensure that user can transact

Through the shared accountabilities above, they were able to run an activation on the platform that was seen as a successful test and learn program:

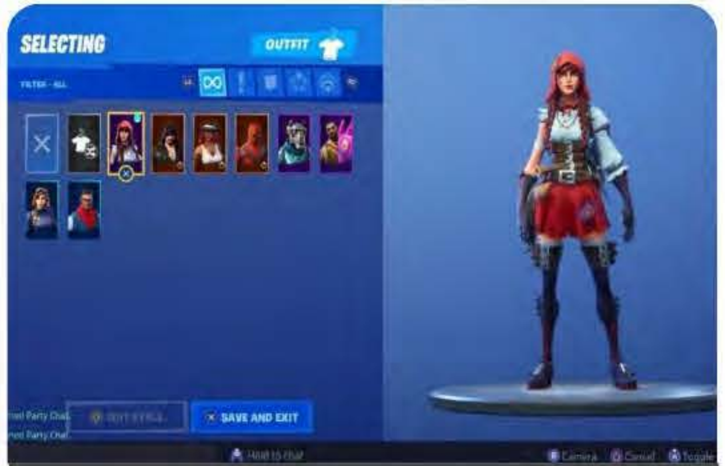
1. The brand and agency were able to learn by doing in the metaverse
2. Consumer and brand safety were delivered via tech-driven and human moderation
3. Limited edition sales of apparel exceeded supply

Marketing models: not every activation is the same, so it is important to manage safety resourcing accordingly

As we saw in the case above, the brand had an involved activation. But not every metaverse marketing campaign will require this level of rigor. There are different models and many of them should be familiar to us from other media campaigns:

Models	Overview
Placement	Placement of a predefined piece of advertising that can be visual and/or audio in nature
Integration	Placement of a product or a branded asset into an application, community environment (inclusive of NFTs)
Participant	Creation of a branded avatar that users and communities can interact with
Experience	Creation of a branded space or event for users and community interaction

Experience



Placement Participant

Integration

Experience

Participant Experience

Experience

Marketing in the metaverse shows diversity in execution and continuity in models

Nikeland

Launched in 2021 on Roblox, Nikeland is one of the metaverse proof of concepts seen to date. This activation has one-part video game, one-part online rewards for offline fitness goals, one-part social community and finally one-part e-commerce. This is truly 'a Swiss Army Knife' execution and set a high bar in engagement and driving a brand benefit (championing consumers reaching of offline fitness via community engagement).

Wendy's

Wendy's metaverse activations span several platforms where they develop a personae and presence in every single major platform – whether Roblox, Horizon World, or Fortnite. Wendy's strategy is to join the community and engage consumers wherever and whenever they are online.

Coca-Cola

Coca-Cola's first activation in the metaverse included NFTs available on OpenSeas in sales to boost donations to Special Olympics International. Since the original charity auction Coca-Cola has built out a distinctive presence spanning more NFTs, in-game presence and developing a limited-edition flavour co-created by Coca-Cola's community in the metaverse, called Sugar Byte.

Forever 21

US fast fashion retailer developed a Roblox-based experience with a metaverse specialist agency. This multiyear activation features an NFT storefront and an ecommerce platform to buy limited-edition real-world clothing co-created in the metaverse.

Dolce & Gabbana

Launched as part of the Italian fashion brand's Fashion Week showcase, D&G unveiled a series of 20 custom wearables featured in Decentraland for a limited time. The fashion items were also then featured in real-world catwalks. This was seen as a PR, critic and consumer success and has propelled the brand to extend work with Mkers and SKNUPs in the eSports and NFT arenas.

Vans

Vans created a virtual skatepark on Roblox in an experience that brings skateboarding, fashion and community together in one experience. The experience spans virtual avatar creation, NFT unlocked via community participation and ecommerce opportunities. Further, the metaverse activations were linked to Vans' physical skateparks in iconic locations such as London.

As such, we can see despite the metaverse being a new medium, the marketing models currently taking shape are easy to identify with from advertising on other channels.

Assessing safety + suitability needs: engagement model x marketing tools

With an understanding of Engagement Models and Marketing Models, we can now start to identify and consider baseline safety and suitability needs.

The assessment here is baseline only and cannot be a substitute for analyzing the content in a metaverse space or community and cannot be a substitute for analyzing how a brand message may appear in a metaverse community.

However, looking at the fluidity of engagement models and marketing models an easy-to-understand model becomes evident.

The matrix below demonstrates that increased fluidity –in engagement and experience increases baseline risk.

MARKETING MODEL	ENGAGEMENT MODEL		
	Fixed Experience	Hybrid Experience	Fluid Experience
	OK for Monetization	Advertiser Controls Needed Human Monitoring Needed	Advertiser Controls Needed Human Moderation Needed
Placement	LOW	MEDIUM	HIGH
Integration	LOW	MEDIUM	HIGH
Participant	LOW	MEDIUM	HIGH
Experience	LOW	HIGH	HIGH

We can make the following assessments accordingly:

1

Fixed Experiences with set user actions have Low Risk and should be viewed as OK for commercialization

2

Hybrid Experiences will predominantly be Medium Risk and should have controls for advertisers with human oversight, which may extend into moderation depending on audience

3

Fluid Experiences will be High Risk and should feature both controls and live moderation

CASE STUDY:

Staging a branded concert series in the metaverse [Fluid Experience + Placement]

Another case study we can observe is a brand sponsored a metaverse-based music concert. In this instance, insurance company signage was featured in a typical digital adaptation of a virtual concert. Therefore the marketing model used was a Placement.

The platform was a Fluid Experience which allowed the crowd to sign, to dance and to interact. Because of this the brand and the platform decided to take a series of steps:

1. Managed audience access via age-gating (registration) and engagement (terms & conditions)
2. Manage audience size into concert sections that could be managed
3. Disclose to users that they were in a live interactive environment and make them aware of reporting tools
4. Monitor crowd interaction by using image and speech recognition software
5. Staff each section with a live moderator to manage the audience accordingly

Because of the steps, the brand was able to sponsor a full series of 9 concerts in the summer of 2022, reaching an average of 28 million attendees per concert.

Testing & learning safely+ suitably: what to do + what to look for

Media experts like to test, learn and experiment. We've also learned that responsibility is high on a media leader's agenda.

In our research, we have identified best practices to consider.

These best practices set out four steps that media leaders should consider as they test and learn in this exciting new space.

There are a series of steps under each of the phases to consider as a team plans a metaverse activation.

Pre-campaign assessment

- Platform and device selection
- Target audience age restrictions
- Target audience geographic restrictions
- Content or behaviour restrictions

In the **Pre-campaign Assessment Step**, teams often consider if the platform and devices involved are appropriate for media investment, considerations covered in GARM's Brand Safety + Suitability Training Bootcamp.

Teams should then consider any audience or behaviour restrictions to their campaign or activation. This assessment will establish clear bounds for the program.

Activation assessment

- Marketing Model Identification
- Engagement Model Identification
- Baseline risk assessment

In the **Activation Assessment Step**, the teams often determine the Baseline Risk Assessment by identifying both the Marketing Model (Placement, Integration, Participant or Experience) and Platform Engagement Model (Fixed, Hybrid, Fluid).

Teams then determine the levels of emphasis needed in verification and moderation resourcing. Teams should also explore and consider.

Safety + suitability guidelines

- Verification requirements + rules identification
- Moderation requirements + guideline development
- Assignment of Verification + Moderation duties

In the **Guidelines Step**, the key output is assignment of verification and moderation roles across the value chain, with clear potential rules. The objective is to ensure that there are clear business rules established to ensure that the metaverse experience is accessed by the right users based on age and location and that the behaviours of the community stay in line with the desired experience. It's important to brief these guidelines to the relevant stakeholders in the value chain necessary for the campaign activation. This is core to ensuring the effectiveness and accountability of these guidelines.

Safety + suitability resourcing

- Align and check user verification layers (hardware, app, experience)
- Align technology moderation (app, experience)
- Resource human moderation (experience team)

In the **Resourcing Step**, the relevant and selected partners are briefed or engaged on the activation to ensure that the right levels of procompetitive collaboration and coordination take place. For instance, the experience team need to ensure that the app is performing using age or location verification to ensure appropriate access. Finally, it is important to ensure that live moderation teams (marketer, agency, platform, or external provider) will be using the appropriate prompts to encourage users or redirect users, as needed.

Managing potential incidents: a framework for assessment + addressing challenges

In GARM's Brand Safety + Suitability Training, we present a series of case studies and frameworks to help industry participants understand, identify and respond to various brand safety and suitability challenges.

The framework on the right encourages the appropriate areas to consider:

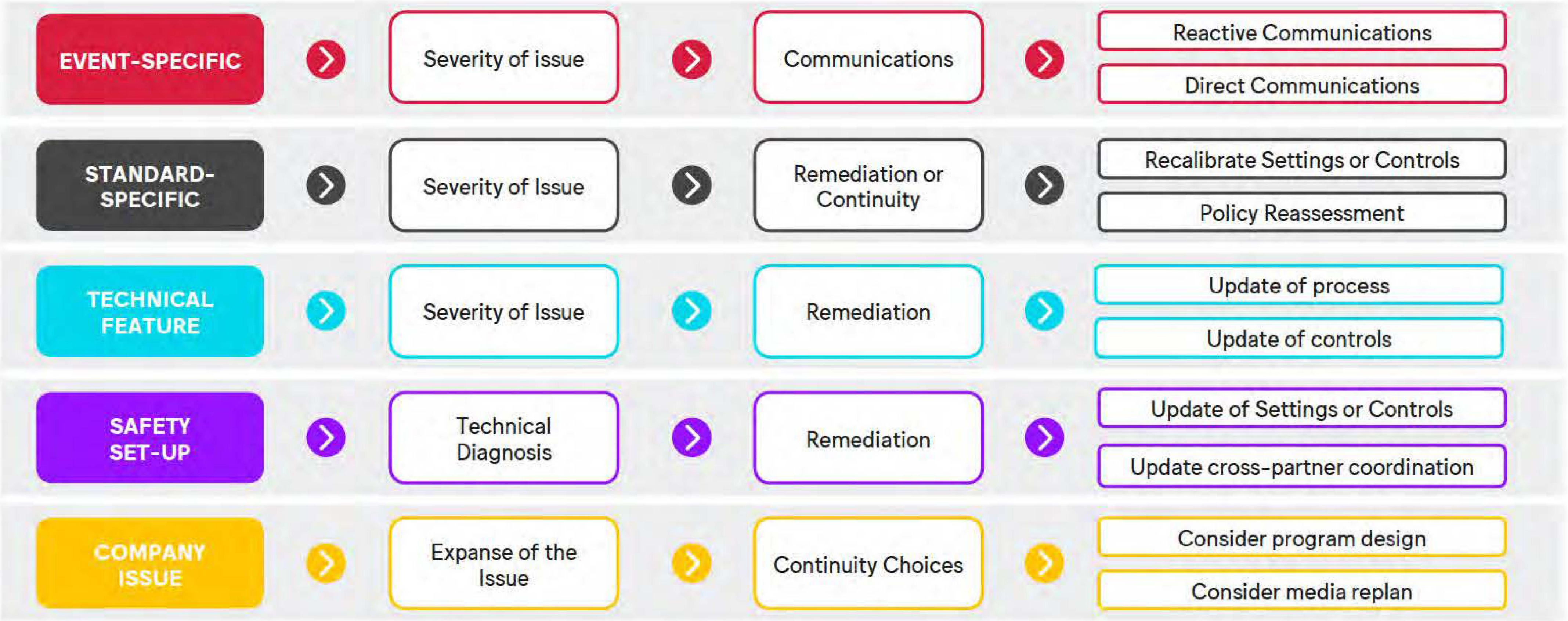
1 Was the incident event-specific? Is the challenge time-specific, is it related to an event inside or outside the activation?

2 Was the incident standard-specific? Is the challenge related to a GARM category or risk level not being upheld?

3 Was the incident specific to a technical feature of the platform or program?

4 Was the incident due to the safety set-up across the program team (advertiser, agency, app/platform) that allowed for the incident to emerge?

5 Was the incident due to a company-specific issue (e.g., incident relative to the marketer)?



EVENT-SPECIFIC	Severity of issue	Communications	Reactive Communications Direct Communications
STANDARD-SPECIFIC	Severity of Issue	Remediation or Continuity	Recalibrate Settings or Controls Policy Reassessment
TECHNICAL FEATURE	Severity of Issue	Remediation	Update of process Update of controls
SAFETY SET-UP	Technical Diagnosis	Remediation	Update of Settings or Controls Update cross-partner coordination
COMPANY ISSUE	Expanse of the Issue	Continuity Choices	Consider program design Consider media replan

Once the incident is diagnosed, the appropriate reactions can be considered:

A Are communications required to address the incident?

B Does the platform or application need to adjust settings, or controls policies?

C Does team resourcing, capabilities, or communications across organizations need to be considered?

D Do the process or controls need to be considered?

E Does cross partner collaboration need to be addressed?

F Does media flighting and continuity need to be considered?

As metaverse activations are live events, best practices have shown that the best plan is "to have a plan."

The following worksheet is meant to help you plan your metaverse activation safety and suitability needs:

Consideration	Self-Assessment	Notes	
Brand-specific Considerations Are there brand or category specific considerations that your metaverse activation needs to take into account?	NO UNSURE YES ↔		Moderation Needs
Content & Behaviour Considerations Are there specific behaviours or content that would be harmful, problematic or embarrassing?	NO UNSURE YES ↔		
User Access: Age Are there age limits or restrictions to the metaverse experience planned? Should this also ask about enforcement of age limits?	NO UNSURE YES ↔		Verification Needs
User Access: Location Are there geographical limits to the metaverse experience planned?	NO UNSURE YES ↔		
Integrated Functions Are there connected experiences (ecommerce, customer service) that need to be integrated into metaverse experience planned?	NO UNSURE YES ↔		

Next Steps:

Share outputs from **Moderation Needs** with the following stakeholders:

1. **Advertiser:** Marketing team, Media team, Agency Teams (creative, media, experiential)
2. **Agency:** Moderation team

Platform: Moderation team

Share outputs from **Verification Needs** with the following stakeholders:

1. **App:** Partnership/Program team
2. **Hardware:** Verification lead
3. **Specialist teams (Ecommerce):** Team leads
4. **Customer service:** Team leads

WFA

WFA
Brussels, London, New York, Singapore

wfanet.org/garm
GARM@wfanet.org

+32 2 502 57 40

linkedin.com/company/wfa
twitter @wfamarketers
youtube.com/wfamarketers

K&S

Competition compliance policy

The purpose of the WFA is to represent the interests of advertisers and to act as a forum for legitimate contacts between members of the advertising industry. It is obviously the policy of the WFA that it will not be used by any company to further any anti-competitive or collusive conduct, or to engage in other activities that could violate any antitrust or competition law, regulation, rule or directives of any country or otherwise impair full and fair competition. The WFA carries out regular checks to make sure that this policy is being strictly adhered to. As a condition of membership, members of the WFA acknowledge that their membership of the WFA is subject to the competition law rules and they agree to comply fully with those laws. Members agree that they will not use the WFA, directly or indirectly, (a) to reach or attempt to reach agreements or understandings with one or more of their competitors, (b) to obtain or attempt to obtain, or exchange or attempt to exchange, confidential or proprietary information regarding any other company other than in the context of a bona fide business or (c) to further any anti-competitive or collusive conduct, or to engage in other activities that could violate any antitrust or competition law, regulation, rule or directives of any country or otherwise impair full and fair competition. Please note that the recommendations included in this document are merely meant as suggestions or proposals. They are not binding in any way whatsoever and members are free to depart from them.

From: Julie Inman Grant
Sent: Thursday, 22 June 2023 5:53 PM
To: s 22; s 47F @wfanet.org
Cc: eSafety Parliamentary
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Thanks s 22 and s 47F I'm finally going on leave - without my laptop! It would be great for you to catch up with my team and perhaps we could try to touch base when I am back? Thanks for all the important work you continue to do! Julie

Get Outlook for iOS

From: s 22 @eSafety.gov.au>
Sent: Thursday, June 22, 2023 5:50:57 PM
To: s 47F @wfanet.org s 47F @wfanet.org>
Cc: Julie Inman Grant s 47E(d) @eSafety.gov.au>; eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 47F

I hope you don't mind me replying on Julie's behalf, as she is just about to go on leave.

We'd be very keen to discuss the notice to Twitter and how we can use it to incentivise higher standards. We have tried to focus the questions to get meaningful answers about how Twitter is enforcing their rules, whether these are applied consistently for all accounts, and the tools and resources in place. We hope the information is helpful to your members, but it'd be great to get your thoughts.

Thank you also for the playbook. We can bring our safety by design/tech trends team to a chat as there is lots going on (including mandatory code & standards that will apply to metaverse & GenAI services), and/or get you any feedback offline.

Would a call at 5pm or 6pm your time (assuming you are on the east coast atm) work one day in the next week?

Best,

s 22

s 22
Manager, Basic Online Safety Expectations
International, Strategy and Futures
s 22

Australian Government | eSafetyCommissioner

esafety.gov.au

From: s 47F [redacted] @wfanet.org>
Sent: Thursday, 22 June 2023 2:11 PM
To: Julie Inman Grant s 47E(d) [redacted] @eSafety.gov.au>
Cc: eSafety Parliamentary s 47E(d) [redacted] @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate

You don't often get email from s 47F [redacted] @wfanet.org. Learn why this is important

Hi Julie –

Thanks for this.
I am confirming receipt.

Are you and the team available for a call in the coming days?
There are some updates for me to share with you as well.

We are aware of the issues and we have a series of steps already underway, but I am skeptical of the outcomes. National associations like ISBA, who is a board member of mine, are advising members accordingly.

Transparently, as an industry association with antitrust provisions and developing standards, it is hard for us to manage corrective measures beyond driving transparency on issues and suggested remedies. We are not a watchdog and rely on NGOs like ADL who are on our NGO Consult Group to raise the issue.
You also may have also seen some updates on how the US GOP perceive our work.

Finally, I've attached a playbook we are releasing today on Generative AI and the metaverse. It would be great to get your feedback and discuss a meaningful route forward.
We're eager to understand how we can look at market-facing anticipatory steps in these areas.

Let me know when works for you – I'd really value your personal guidance on some of the issues I am facing into, it'd be good regroup.

Best,

s 47F [redacted]

[redacted] - Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47E(d) [redacted]

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant s 47E(d) [redacted] @eSafety.gov.au>
Date: Wednesday, June 21, 2023 at 21:47
To: s 47E(d) [redacted] @wfanet.org>

Cc: eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: eSafety Commissioner takes regulatory action against Twitter around Online Hate

Dear s 47F

I hope you are keeping well. I wanted to let you know that we have taken this regulatory action today against Twitter on online hate. The brand safety leverage you and GARM have been able to extract is incredible and we hope that this action will help further shine a light on the safety shortcomings currently pervading the platform.

I believe that transparency is vital to ensuring that online services and platforms are safe by design. Without transparency, there can be no meaningful accountability from the global giants shaping our society, enabling our discourse, and facilitating unprecedented communications.

In January last year, stronger modernised online safety protections under the Online Safety Act took effect in Australia. In addition to enhancing eSafety's powers to tackle specific harms such as adult cyber abuse, image-based abuse, child cyberbullying and illegal content, the Act gives me the ability to require information from companies about how they are keeping their users safe.

These Basic Online Safety Expectations ('BOSE') place transparency at the heart of our regulatory model. They are a novel framework of powers. Through their use, eSafety can compel companies to 'show us their working' on specific online safety concerns, rather than being shielded by marketing spin or glossy handouts. By using these powers, eSafety is rapidly developing a strong baseline understanding of where industry is doing well, but where there is more work to do, to harden their services from abuse and malfeasance.

Today, I have issued a BOSE notice to Twitter, challenging the company to explain what they are doing to combat online hate. Twitter has 28 days to respond to the notice and a failure to comply may attract a penalty of up to AUD$687,500 per day.

By taking this step, I aim to shed light on how Twitter is addressing what appears to be a recent surge in hate on the platform, both general and targeted. In particular, I want to understand how Twitter is enforcing its own clear rules prohibiting hateful conduct, and how trust and safety is enabled within the company.

Unfortunately, our experience and that of others suggests that Twitter is falling well short of the mark in both respects.

eSafety has received more complaints about online hate on Twitter than any other service in the last 12 months, with many of these appearing to coincide with the change in ownership last October. The increase overlaps with the platform's reported reinstatement of over 62,000 accounts previously banned for breaching Twitter rules, including 75 with more than 1 million followers. I am concerned that these accounts are playing an outsized role in fuelling the platform's toxicity.

The impact of hate on marginalised communities is not a theoretical concern. New eSafety research has found that 1 in 5 Australians have experienced online hate in the last 12 months, and we know that First Nations people and members of the LGBTQI+ community, face hate at twice the rate of the national average. Overall, one in six adults targeted by online abuse report that their physical health suffered as a result; the figure rises to one in three when emotional and mental wellbeing is considered.

As with previous notices, eSafety will release a report summarising the information we receive. I will keep you updated on the outcome of this process, and our findings.

Thank you again for your important contribution to our collective work of making the internet a safer place for all.

All the best,

Julie

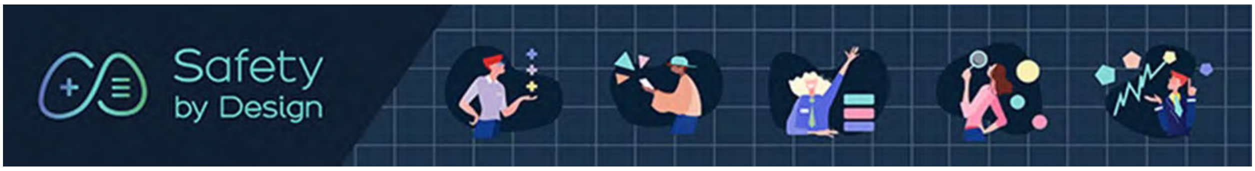
Julie Inman Grant
Commissioner

Australian Government | eSafetyCommissioner

☎ s 47E(d)

[redacted]

esafety.gov.au



eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

From: s 47F @wfanet.org>
Sent: Thursday, 29 June 2023 9:50 PM
To: s 22
Cc: Julie Inman Grant; eSafety Parliamentary
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 22 -

Sure a 5p Eastern call would be great - let me know when works in the coming days.

Best

s 47F

[redacted] - Global Alliance for Responsible Media
World Federation of Advertisers

s 47F / s 47F @wfanet.org

From: s 22 @eSafety.gov.au>
Sent: Thursday, June 22, 2023 3:50:57 AM
To: s 47F @wfanet.org>
Cc: Julie Inman Grant s 47E(d) @eSafety.gov.au>; eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 47F

I hope you don't mind me replying on Julie's behalf, as she is just about to go on leave.

We'd be very keen to discuss the notice to Twitter and how we can use it to incentivise higher standards. We have tried to focus the questions to get meaningful answers about how Twitter is enforcing their rules, whether these are applied consistently for all accounts, and the tools and resources in place. We hope the information is helpful to your members, but it'd be great to get your thoughts.

Thank you also for the playbook. We can bring our safety by design/tech trends team to a chat as there is lots going on (including mandatory code & standards that will apply to metaverse & GenAI services), and/or get you any feedback offline.

Would a call at 5pm or 6pm your time (assuming you are on the east coast atm) work one day in the next week?

Best,

s 22

s 22
Manager, Basic Online Safety Expectations
International, Strategy and Futures
s 22

Australian Government | eSafetyCommissioner

esafety.gov.au

From: s 47F @wfanet.org>
Sent: Thursday, 22 June 2023 2:11 PM
To: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Cc: eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate

You don't often get email from s 47F @wfanet.org. Learn why this is important

Hi Julie –

Thanks for this.
I am confirming receipt.

Are you and the team available for a call in the coming days?
There are some updates for me to share with you as well.

We are aware of the issues and we have a series of steps already underway, but I am skeptical of the outcomes. National associations like ISBA, who is a board member of mine, are advising members accordingly.

Transparently, as an industry association with antitrust provisions and developing standards, it is hard for us to manage corrective measures beyond driving transparency on issues and suggested remedies. We are not a watchdog and rely on NGOs like ADL who are on our NGO Consult Group to raise the issue.
You also may have also seen some updates on how the US GOP perceive our work.

Finally, I've attached a playbook we are releasing today on Generative AI and the metaverse. It would be great to get your feedback and discuss a meaningful route forward.
We're eager to understand how we can look at market-facing anticipatory steps in these areas.

Let me know when works for you – I'd really value your personal guidance on some of the issues I am facing into, it'd be good regroup.

Best,

s
47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Date: Wednesday, June 21, 2023 at 21:47
To: s 47F @wfanet.org>
Cc: eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: eSafety Commissioner takes regulatory action against Twitter around Online Hate

Dear s 47F

I hope you are keeping well. I wanted to let you know that we have taken this regulatory action today against Twitter on online hate. The brand safety leverage you and GARM have been able to extract is incredible and we hope that this action will help further shine a light on the safety shortcomings currently pervading the platform.

I believe that transparency is vital to ensuring that online services and platforms are safe by design. Without transparency, there can be no meaningful accountability from the global giants shaping our society, enabling our discourse, and facilitating unprecedented communications.

In January last year, stronger modernised online safety protections under the Online Safety Act took effect in Australia. In addition to enhancing eSafety's powers to tackle specific harms such as adult cyber abuse, image-based abuse, child cyberbullying and illegal content, the Act gives me the ability to require information from companies about how they are keeping their users safe.

These Basic Online Safety Expectations ('BOSE') place transparency at the heart of our regulatory model. They are a novel framework of powers. Through their use, eSafety can compel companies to 'show us their working' on specific online safety concerns, rather than being shielded by marketing spin or glossy handouts. By using these powers, eSafety is rapidly developing a strong baseline understanding of where industry is doing well, but where there is more work to do, to harden their services from abuse and malfeasance.

Today, I have issued a BOSE notice to Twitter, challenging the company to explain what they are doing to combat online hate. Twitter has 28 days to respond to the notice and a failure to comply may attract a penalty of up to AUD$687,500 per day.

By taking this step, I aim to shed light on how Twitter is addressing what appears to be a recent surge in hate on the platform, both general and targeted. In particular, I want to understand how Twitter is enforcing its own clear rules prohibiting hateful conduct, and how trust and safety is enabled within the company.

Unfortunately, our experience and that of others suggests that Twitter is falling well short of the mark in both respects.

eSafety has received more complaints about online hate on Twitter than any other service in the last 12 months, with many of these appearing to coincide with the change in ownership last October. The increase overlaps with the platform's reported reinstatement of over 62,000 accounts previously banned for breaching Twitter rules, including 75 with more than 1 million followers. I am concerned that these accounts are playing an outsized role in fuelling the platform's toxicity.

The impact of hate on marginalised communities is not a theoretical concern. New eSafety research has found that 1 in 5 Australians have experienced online hate in the last 12 months, and we know that First Nations people and members of the LGBTQI+ community, face hate at twice the rate of the national average. Overall, one in six adults targeted by online abuse report that their physical health suffered as a result; the figure rises to one in three when emotional and mental wellbeing is considered.

As with previous notices, eSafety will release a report summarising the information we receive. I will keep you updated on the outcome of this process, and our findings.

Thank you again for your important contribution to our collective work of making the internet a safer place for all.

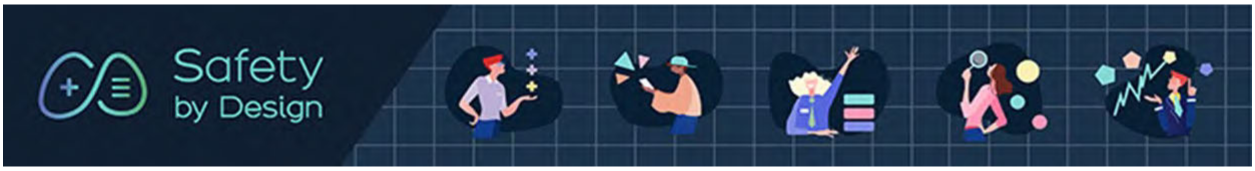
All the best,

Julie

Julie Inman Grant
Commissioner

Australian Government | eSafetyCommissioner

☎ s 47E(d)
📱 [redacted]
🌐 esafety.gov.au



eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

From: s 47F @wfanet.org>
Sent: Friday, 30 June 2023 12:05 AM
To: Julie Inman Grant
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]
Attachments: GARM Brand Safety- Steer Team Call Notes 27 Jun 2023 .eml

Hey –

Thanks for the note – I think you totally earned your time off.
Yes – it's a bit of a cynical move.

I've attached a bit of a reference FYEO on how we are viewing Twitter.
The last line of the table is the area to focus in on.

The changes in policy and lack of transparency has many agencies and brands scratching their heads, rightly so. There needs to be a rewind and a real benchmark of where they stand vis-à-vis voluntary industry standards for transparency sake – advertisers can then decide what to do accordingly.

Hope you're having an amazing time and would really appreciate a 1:1 catch-up for some guidance in general.

Best,

s
47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant
Date: Thursday, June 29, 2023 at 10:01
To: s 47F
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Sorry to miss this conversation with you, s It's been quite a demanding year and I'm taking some time out in Europe with the family - thanks for all that you are doing, or are trying to do. I'm not sure if you know s 47F we'll but presumably she was brought on-board to re-attract advertisers and clearly achieving higher levels of brand safety would be key to that??? Julie

Get Outlook for iOS

From: s 47F
Sent: Thursday, June 29, 2023 12:49:37 PM
To: s 22
Cc: Julie Inman Grant ; eSafety Parliamentary
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 22 -

Sure a 5p Eastern call would be great - let me know when works in the coming days.

Best

s 47F

[redacted] Global Alliance for Responsible Media
World Federation of Advertisers

s 47F / s 47F @wfanet.org

From: s 22
Sent: Thursday, June 22, 2023 3:50:57 AM
To: s 47F
Cc: Julie Inman Grant ; eSafety Parliamentary
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 47F

I hope you don't mind me replying on Julie's behalf, as she is just about to go on leave.

We'd be very keen to discuss the notice to Twitter and how we can use it to incentivise higher standards. We have tried to focus the questions to get meaningful answers about how Twitter is enforcing their rules, whether these are applied consistently for all accounts, and the tools and resources in place. We hope the information is helpful to your members, but it'd be great to get your thoughts.

Thank you also for the playbook. We can bring our safety by design/tech trends team to a chat as there is lots going on (including mandatory code & standards that will apply to metaverse & GenAI services), and/or get you any feedback offline.

Would a call at 5pm or 6pm your time (assuming you are on the east coast atm) work one day in the next week?

Best,

s 22

s 22
Manager, Basic Online Safety Expectations
International, Strategy and Futures
s 22

eSafety

esafety.gov.au

From: s 47F
Sent: Thursday, 22 June 2023 2:11 PM
To: Julie Inman Grant
Cc: eSafety Parliamentary
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate

You don't often get email from s 47F @wfanet.org. Learn why this is important

Hi Julie –

Thanks for this.
I am confirming receipt.

Are you and the team available for a call in the coming days?
There are some updates for me to share with you as well.

We are aware of the issues and we have a series of steps already underway, but I am skeptical of the outcomes.
National associations like ISBA, who is a board member of mine, are advising members accordingly.

Transparently, as an industry association with antitrust provisions and developing standards, it is hard for us to manage corrective measures beyond driving transparency on issues and suggested remedies. We are not a watchdog and rely on NGOs like ADL who are on our NGO Consult Group to raise the issue.
You also may have also seen some updates on how the US GOP perceive our work.

Finally, I've attached a playbook we are releasing today on Generative AI and the metaverse. It would be great to get your feedback and discuss a meaningful route forward.
We're eager to understand how we can look at market-facing anticipatory steps in these areas.

Let me know when works for you – I'd really value your personal guidance on some of the issues I am facing into, it'd be good regroup.

Best,

s
47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Date: Wednesday, June 21, 2023 at 21:47

To: s 47F @wfanet.org>
Cc: eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: eSafety Commissioner takes regulatory action against Twitter around Online Hate

Dear s 47F

I hope you are keeping well. I wanted to let you know that we have taken this regulatory action today against Twitter on online hate. The brand safety leverage you and GARM have been able to extract is incredible and we hope that this action will help further shine a light on the safety shortcomings currently pervading the platform.

I believe that transparency is vital to ensuring that online services and platforms are safe by design. Without transparency, there can be no meaningful accountability from the global giants shaping our society, enabling our discourse, and facilitating unprecedented communications.

In January last year, stronger modernised online safety protections under the Online Safety Act took effect in Australia. In addition to enhancing eSafety's powers to tackle specific harms such as adult cyber abuse, image-based abuse, child cyberbullying and illegal content, the Act gives me the ability to require information from companies about how they are keeping their users safe.

These Basic Online Safety Expectations ('BOSE') place transparency at the heart of our regulatory model. They are a novel framework of powers. Through their use, eSafety can compel companies to 'show us their working' on specific online safety concerns, rather than being shielded by marketing spin or glossy handouts. By using these powers, eSafety is rapidly developing a strong baseline understanding of where industry is doing well, but where there is more work to do, to harden their services from abuse and malfeasance.

Today, I have issued a BOSE notice to Twitter, challenging the company to explain what they are doing to combat online hate. Twitter has 28 days to respond to the notice and a failure to comply may attract a penalty of up to AUD$687,500 per day.

By taking this step, I aim to shed light on how Twitter is addressing what appears to be a recent surge in hate on the platform, both general and targeted. In particular, I want to understand how Twitter is enforcing its own clear rules prohibiting hateful conduct, and how trust and safety is enabled within the company.

Unfortunately, our experience and that of others suggests that Twitter is falling well short of the mark in both respects.

eSafety has received more complaints about online hate on Twitter than any other service in the last 12 months, with many of these appearing to coincide with the change in ownership last October. The increase overlaps with the platform's reported reinstatement of over 62,000 accounts previously banned for breaching Twitter rules, including 75 with more than 1 million followers. I am concerned that these accounts are playing an outsized role in fuelling the platform's toxicity.

The impact of hate on marginalised communities is not a theoretical concern. New eSafety research has found that 1 in 5 Australians have experienced online hate in the last 12 months, and we know that First Nations people and members of the LGBTQI+ community, face hate at twice the rate of the national average. Overall, one in six adults targeted by online abuse report that their physical health suffered as a result; the figure rises to one in three when emotional and mental wellbeing is considered.

As with previous notices, eSafety will release a report summarising the information we receive. I will keep you updated on the outcome of this process, and our findings.

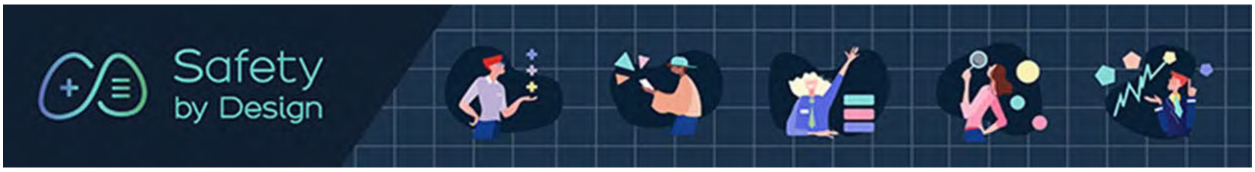
Thank you again for your important contribution to our collective work of making the internet a safer place for all.

All the best,

Julie

Julie Inman Grant
Commissioner

☎ s 47E(d) [redacted]
[redacted]
esafety.gov.au



eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.
NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.

From: s 47F @wfanet.org>
Sent: Thursday, 29 June 2023 8:43 PM
To: s 47F
Cc: s 47F
Subject: GARM Brand Safety: Steer Team Call Notes 27 Jun 2023

Hi all –

Thanks to those who were able to have a flexible schedule and accommodate some of the technical challenges I had on Tuesday.

Here is a recap of the Steer Team call

TOPIC	DISCUSSION / DECISION	NEXT STEP
Cannes Debrief	Team discussed that the Innovation Playbook **didn't get the coverage needed and will do direct outreach with press, partners** (WEF, ADL, AUS eSafety Commissioner, Integrity Institute) and consider others (e.g., Ofcom)	WFA comms team to explore blogpost with DigiDay, AdAge s to reach out to partners directly
Meta CSAM	Team discussed conversations with Meta on the Instagram CSAM incident, and shared that **Meta would be addressing the template and the incident via a direct email, meeting, and then a verbal debrief on the Community Call on Thursday**	Team to review Meta submission Meta to speak on the Community Call
Twitter Acceleration Agenda	Team discussed the meetings Twitter had with GARM in early June. Team discussed Twitter's latest reshuffle of brand safety responsibilities. Team acknowledged the dossier being shared by CAN. Team also reviewed the Acceleration Agenda with the following assessments: 1. **Upholding the Floor:** While Twitter claimed to have demonstrated a 99% upholding of the Brand Safety Floor there are reports from Stanford Internet Observatory, Center for Countering Digital Hate and Conscious Ad Network that **should have Twitter remap their content/consumer and monetization policies to the Brand Safety Floor** 2. **Regular Reporting on Toxic Content**: Twitter appointed Sprinklr to monitor toxic content levels in English only, however there are outstanding questions: **a report has not been produced and we don't have an understanding of content access or methodology.** GroupM has had the latest conversations with Sprinklr and we must ask for clarity around this, and **work with**	s to raise issues to Twitter when they indicate they are ready to meet with the Steer Team – meeting to voluntarily be set by Twitter s to discuss Sprinklr discussions with GroupM GARM Steer Team members to use points as reactive inform to their own consisteuncies

	Sprinklr on these questions with Twitter's understanding. 3. **Prebid Controls:** As discussed previously, **Twitter will ingest GroupM's list with their own, and the Solutions Developers Working Group may decide to augment it with additional terms**. However, confidentially we are aware that adtech firms in GARM have been RFPed for a solution that would represent a way to get advertisers and agencies back on the platform 4. **Transparency Reporting: There is no update on Twitter's plans to outsource self-serve live transparency reporting** 5. **MRC Audit:** Twitter indicated that they hired a head of compliance, and the MRC audit would be part of their remit and that they would engage on any further steps relative to the preassessment delivered by MRC to Twitter in June 2022 (as a reminder MRC confirmed no further progress on that report from our meeting with them in April) **GARM should therefore consider Twitter's Acceleration Agenda as 'on hold and needing reconfirmation'** especially on the Floor given the number of policy changes or exceptions. GARM and WFA will not male a public statement in this regard, but members asking can be debriefed. Additionally, Steer Team members should feel free to debrief their own members with the above points verbally.	

s 47F
[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant
Sent: Friday, 30 June 2023 12:29 AM
To: s 47F
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Thanks for this. Really interesting! We will be releasing our report on Twitter, Twitch, Tik Tok, Google & Discord transparency towards the end of July.

The reason we have these legal compulsion powers is because voluntary transparency has become a bit of a misnomer. Our last BOSE report demonstrated that the companies that signed up to the Five Eyes Voluntary Principles around Combatting Online Child Sexual Abuse were not living up to the VPs they set for themselves.

Revenue and reputation seem to be more important drivers than regulation & fines, which are really drops in the bucket. Advertisers - particularly industry-wide - are likely to be an even more important lever.

I think the CEO gig would be a hard job for anyone- particularly if you care about maintaining your integrity and credibility long-term!

Happy to get on a call when I'm back. Julie

Get Outlook for iOS

From: s 47F @wfanet.org>
Sent: Thursday, June 29, 2023 3:05:09 PM
To: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hey –

Thanks for the note – I think you totally earned your time off.
Yes – it's a bit of a cynical move.

I've attached a bit of a reference FYEO on how we are viewing Twitter.
The last line of the table is the area to focus in on.

The changes in policy and lack of transparency has many agencies and brands scratching their heads, rightly so. There needs to be a rewind and a real benchmark of where they stand vis-à-vis voluntary industry standards for transparency sake – advertisers can then decide what to do accordingly.

Hope you're having an amazing time and would really appreciate a 1:1 catch-up for some guidance in general.

Best,

s 47F

s 47F
Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Date: Thursday, June 29, 2023 at 10:01
To: s 47F @wfanet.org>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Sorry to miss this conversation with you, s 47F It's been quite a demanding year and I'm taking some time out in Europe with the family - thanks for all that you are doing, or are trying to do. I'm not sure if you know s 47F we'll but presumably she was brought on-board to re-attract advertisers and clearly achieving higher levels of brand safety would be key to that??? Julie

Get Outlook for iOS

From: s 47F @wfanet.org>
Sent: Thursday, June 29, 2023 12:49:37 PM
To: s 22 @eSafety.gov.au>
Cc: Julie Inman Grant s 47E(d) @eSafety.gov.au>; eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 22 -

Sure a 5p Eastern call would be great - let me know when works in the coming days.

Best

s 47F

[redacted] Global Alliance for Responsible Media
World Federation of Advertisers

s 47F / s 47F @wfanet.org

From: s 22 @eSafety.gov.au>
Sent: Thursday, June 22, 2023 3:50:57 AM
To: s 47F @wfanet.org>
Cc: Julie Inman Grant s 47F @eSafety.gov.au>; eSafety Parliamentary s 47F @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate [SEC=OFFICIAL]

Hi s 47F

I hope you don't mind me replying on Julie's behalf, as she is just about to go on leave.

We'd be very keen to discuss the notice to Twitter and how we can use it to incentivise higher standards. We have tried to focus the questions to get meaningful answers about how Twitter is enforcing their rules, whether these are applied consistently for all accounts, and the tools and resources in place. We hope the information is helpful to your members, but it'd be great to get your thoughts.

Thank you also for the playbook. We can bring our safety by design/tech trends team to a chat as there is lots going on (including mandatory code & standards that will apply to metaverse & GenAI services), and/or get you any feedback offline.

Would a call at 5pm or 6pm your time (assuming you are on the east coast atm) work one day in the next week?

Best,

s 22

s 22
Manager, Basic Online Safety Expectations
International, Strategy and Futures
s 22

esafety.gov.au

From: s 47F @wfanet.org>
Sent: Thursday, 22 June 2023 2:11 PM
To: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Cc: eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: Re: eSafety Commissioner takes regulatory action against Twitter around Online Hate

You don't often get email from s 47F @wfanet.org. Learn why this is important

Hi Julie –

Thanks for this.
I am confirming receipt.

Are you and the team available for a call in the coming days?
There are some updates for me to share with you as well.

We are aware of the issues and we have a series of steps already underway, but I am skeptical of the outcomes. National associations like ISBA, who is a board member of mine, are advising members accordingly.

Transparently, as an industry association with antitrust provisions and developing standards, it is hard for us to manage corrective measures beyond driving transparency on issues and suggested remedies. We are not a watchdog and rely on NGOs like ADL who are on our NGO Consult Group to raise the issue.
You also may have also seen some updates on how the US GOP perceive our work.

Finally, I've attached a playbook we are releasing today on Generative AI and the metaverse. It would be great to get your feedback and discuss a meaningful route forward.

We're eager to understand how we can look at market-facing anticipatory steps in these areas.

Let me know when works for you – I'd really value your personal guidance on some of the issues I am facing into, it'd be good regroup.

Best,

s 47F

[redacted] Global Alliance for Responsible Media

WFA - World Federation of Advertisers
Brussels • London • New York • Singapore
s 47F

WFA values and encourages flexible working patterns, with teams working across multiple time zones. Although I have sent this at a time that is convenient for me, it is not my expectation that you read, respond or follow up on this email outside your hours of work.

From: Julie Inman Grant s 47E(d) @eSafety.gov.au>
Date: Wednesday, June 21, 2023 at 21:47
To: s 47F @wfanet.org>
Cc: eSafety Parliamentary s 47E(d) @esafety.gov.au>
Subject: eSafety Commissioner takes regulatory action against Twitter around Online Hate

Dear s 47F

I hope you are keeping well. I wanted to let you know that we have taken this regulatory action today against Twitter on online hate. The brand safety leverage you and GARM have been able to extract is incredible and we hope that this action will help further shine a light on the safety shortcomings currently pervading the platform.

I believe that transparency is vital to ensuring that online services and platforms are safe by design. Without transparency, there can be no meaningful accountability from the global giants shaping our society, enabling our discourse, and facilitating unprecedented communications.

In January last year, stronger modernised online safety protections under the Online Safety Act took effect in Australia. In addition to enhancing eSafety's powers to tackle specific harms such as adult cyber abuse, image-based abuse, child cyberbullying and illegal content, the Act gives me the ability to require information from companies about how they are keeping their users safe.

These Basic Online Safety Expectations ('BOSE') place transparency at the heart of our regulatory model. They are a novel framework of powers. Through their use, eSafety can compel companies to 'show us their working' on specific online safety concerns, rather than being shielded by marketing spin or glossy handouts. By using these powers, eSafety is rapidly developing a strong baseline understanding of where industry is doing well, but where there is more work to do, to harden their services from abuse and malfeasance.

Today, I have issued a BOSE notice to Twitter, challenging the company to explain what they are doing to combat online hate. Twitter has 28 days to respond to the notice and a failure to comply may attract a penalty of up to AUD$687,500 per day.

By taking this step, I aim to shed light on how Twitter is addressing what appears to be a recent surge in hate on the platform, both general and targeted. In particular, I want to understand how Twitter is enforcing its own clear rules prohibiting hateful conduct, and how trust and safety is enabled within the company.

Unfortunately, our experience and that of others suggests that Twitter is falling well short of the mark in both respects.

eSafety has received more complaints about online hate on Twitter than any other service in the last 12 months, with many of these appearing to coincide with the change in ownership last October. The increase overlaps with the platform's reported reinstatement of over 62,000 accounts previously banned for breaching Twitter rules, including 75 with more than 1 million followers. I am concerned that these accounts are playing an outsized role in fuelling the platform's toxicity.

The impact of hate on marginalised communities is not a theoretical concern. New eSafety research has found that 1 in 5 Australians have experienced online hate in the last 12 months, and we know that First Nations people and members of the LGBTQI+ community, face hate at twice the rate of the national average. Overall, one in six adults targeted by online abuse report that their physical health suffered as a result; the figure rises to one in three when emotional and mental wellbeing is considered.

As with previous notices, eSafety will release a report summarising the information we receive. I will keep you updated on the outcome of this process, and our findings.

Thank you again for your important contribution to our collective work of making the internet a safer place for all.

All the best,

Julie

Julie Inman Grant
Commissioner

eSafetyCommissioner

☎ s 47F [redacted]
[redacted]
esafety.gov.au

eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the

intended recipient, please contact the sender by reply email and destroy all copies of the original message.
NOTICE: This email message is for the sole use of the intended recipient(s) and may contain confidential and privileged information. Any unauthorized review, use, disclosure or distribution is prohibited. If you are not the intended recipient, please contact the sender by reply email and destroy all copies of the original message.