October, 2020

Advertiser Protection Bureau

White Paper: Misinformation/ Disinformation In Focus

CONTENTS

Introduction

Misinformation and disinformation together represent perhaps the most challenging of all brand safety issues, insofar as they are both deceptive and harmful by design. This thought-leadership paper describes special concerns for marketers in this area, defines a clear and actionable definition for misinformation and disinformation and outlines a roadmap for global industry response to these issues designed to help to mitigate negative effects on brands and their stakeholders.

This white paper has been endorsed by:

- The Global Disinformation Initiative (GDI)
- European Association of Communications Agencies (EACA)

Misinformation and disinformation represent major concerns at multiple levels

While the brand safety issues around proximity to misinformation and disinformation are extremely serious, the primary concern in these areas should be the social hazards they are universally acknowledged to represent. Given finite human attention, flooding media with malignant content corrupts the global exchange of ideas, eroding the public knowledge needed in democratic societies to maintain collective societal needs, be they public health, scientific, economic, or social. Economic harm to brands is a major secondary impact, driven largely as a consequence of the negative effects misinformation and disinformation have on consumers. **In view of the desperate need for global consensus and informed choice highlighted by public health concerns around the coronavirus and climate change, we believe disinformation and misinformation represent the most significant source of public harm in the media ecosystem.**

Differentiating between misinformation and disinformation is of limited utility

Misinformation can be differentiated from disinformation through the lens of intent, but the distinction is of limited use in the marketplace. Disinformation is distributed with the intent to deceive. Misinformation is distributed without the intent to deceive. The goal of disinformation is to reproduce and influence or confuse, spreading virally in two ways. The first is via those who recognize the disinformation for what it is and share the disinformers goals, and pass on the disinformation knowingly. The second, as misinformation via those individuals who believe the claims asserted in a piece of disinformation to be true--a far larger audience. Since the harm produced is consistent regardless of the intent, and since intent is known only to the creator or creators of content, we use the two terms together. Instead, intent with regard to disinformation vs. misinformation should be considered in the context of legal measures, outside the scope of this discussion, where intent can be determined in the courts. Note as well, that we are not using the term "fake news" here as it represents only a single facet of the many challenges in this area. It also bears mentioning that the suppression of viable and accurate information is a form of misinformation in its own way.

Misinformation and disinformation challenge definition

Categories of disinformation and misinformation are wide ranging, but include propaganda, conspiracy theories, pseudoscience, memes containing false claims and premises, the re-presentation of old stories as current events, the re-titling of stories, the presentation of pictures with headlines that pretend they were taken at times and in places where they were not, the manufacture of wholly new and fraudulent facts, the distribution of 'deepfakes' which apply AI to make public figures say and do things they have never said or done, and more. Respected academics have considered how best to categorize types of disinformation by their key characteristics, and we have incorporated their thinking into the definition used in this document.

Impacts of proximity to dis/misinformation appear extremely negative for brands

When brand safety is considered, proximity to unsuitable content, including misinformation and disinformation, has been shown to affect brands in a variety of negative ways, including damaging a brands reputation and its bottom line. Consider a pharmaceutical placement running beside discredited claims about hydroxychloroquine,or worse, injected bleach - as COVID-19 cure-alls. Placement beside such unsuitable content can reduce the credibility of the brand in question, and, especially for younger consumers, can significantly affect opinions towards the brand. Research indicates that Millennial and Gen X consumers exposed to brandads beside unsafe placements are three times more likely not to recommend the brand and four times more likely not to consider purchasing from the brand, creating a "negative reach" for paid placements-- worse than not advertising at all. 51% of this group would be less likely to purchase from a brand beside unsafe placements even if the placement wasn't the brand's fault. Misinformation and disinformation about brands themselves clearly represent similar potential sources of damage to brands.

Meanwhile, topics that have suffered heavily from adulteration by misinformation and disinformation, including racial justice and impending political events, are dominating narratives and discussion across platforms, intersecting and sometimes magnifying each other. This situation is creating hazards that brands are not properly equipped to measure, much less avoid. The WHO has referred to the current environment as an "Infodemic" - essentially, a deluge of disinformation and misinformation alongside real information, creating a danger to the public. When sources of misinformation and disinformation are presented beside and as though they are on a par with trusted sources, such presentation also likely damages public trust in those established sources, affecting their brands and the credibility of news and truthful narratives as a whole, and further widening the aperture for future false and misleading information.

While direct effects of negative reach represent real damage to brands, recent events have also brought even more fundamental business-relevant dangers of misinformation and disinformation into focus,in particular, by encouraging the public to disregard the advice of experts in ways that clearly affect the economy and brand profits. For example, the disinformation and misinformation around coronavirus encourages people to disregard necessary safety measures like masks and social distancing, directly affecting the trajectory of the epidemic, the length and severity of lockdown measures and thus the fundamentals of economic activity worldwide. While

stark and immediate in the case of public healtvh, these economic disruptions have the potential to be easily as large or larger when well documented disinformation around topics such as climate change or elections meddling in critical political decisions like Brexit are considered. **Misinformation and disinformation have revealed themselves as fundamental enemies of economic stability and growth, and as such supporting their spread through media investment with partners who serve to distribute misinformation and disinformation at scale must become anathema for advertisers.**

The fight against misinformation and disinformation is not in conflict with Free Speech

When considering how to act against disinformation, legal concerns over the US Communications Decency Act Section 230 and its implications for Web platforms' ability to avoid legal liability for content housed within their services have been cited, as has a commitment to free speech rights. However, these concerns are not in conflict with maintaining freedom of expression. This is supported by the Joint Declaration on Freedom of Expression and Elections in the Digital Age issued in April 2020 by the UN, OAS and OSCE, who state: "Digital media and online intermediaries should make a reasonable effort to address dis-, mis- and mal-information and election related spam, including through independent fact checking and other measures, such as advertisement archives, appropriate content moderation, and public alerts."

Removing untruthful and harmful content on a given platform does not mean that that content cannot be distributed in general, or that the expression of the ideas within that content is criminalized and results in legal consequences for the expressor. Instead, it simply means that a given platform or media owner cannot be used to widely distribute that untrue and harmful content. **Freedom of speech is not freedom of reach**. Some have forgotten that freedom of speech existed in democracies before social media. Previous to its rise, as now, that freedom did not mean that each person was entitled to have their every opinion distributed freely to an international audience, or that any newspaper was obliged to print each opinion piece they received, no matter how incorrect or harmful they judged its content.

In recognition of this principle, platforms are already removing fraudulent and dangerous content flowing from state actors because of the danger they represent to free and fair elections in multiple countries, and there is effort being applied against COVID-19 disinformation. However, the problem of misinformation and disinformation is orders of magnitude larger than the resources currently employed can manage. Fact checking alone is not enough without other systemic changes to manage the scale of the problem and resulting harms. Brands are well within both their rights and their self-interest to avoid advertising in spaces that continue to amplify content that can both damage their brands and ultimately threaten the foundations of stable economic activity on which their business growth depends, and to work together towards a healthier marketplace of ideas.

What is needed to protect advertisers and the public

The Advertiser Protection Bureau believes the five actions below must be taken in order to counter the challenges of misinformation and disinformation.

1. **Define misinformation and disinformation** consistently
2. **Identify, monitor and manage the distribution velocity of potential misinformation and disinformation** within and across platforms and ad buying systems, in real time. Engagement-maximizing algorithms cannot be the chief arbiters of which content achieves wide distribution on major platforms. Reputational history by source and human review must figure much more prominently in distribution and ad algorithms.
3. **Greatly increase the volume, speed and quality of fact checking** to support the identification of misinformation and disinformation, especially in high-velocity content.
4. **Increase the integrity of the fact checking process and strengthen consequences** via measures including deep trusted third party involvement, auditing and rules, greatly increased transparency and regular tracking.
5. **Create a foundation for media tools and approaches to manage adjacency to misinformation and disinformation**, both identified and potential, and report on that adjacency in an auditable manner. This includes **Complete Content-level transparency** prior to media alignment, adjacency, or support, preferably available in near-time and real-time to pace alongside news cycles.

Defining Misinformation

Misinformation and Disinformation are defined as the presentation of verifiably[1] false or misleading claims[2] that are likely to cause harm[3].

[1]**Verifiability:** To be determined by designated reputable fact checking partners (see section 3)

[2]**False or misleading claims** include (source):

- Misleading content: Misleading use of information to frame an issue or individual
- Imposter content: genuine sources that are impersonated
- Fabricated content: New false content
- False connection: headlines, visuals or captions don't support the content
- False Context: genuine content that is shared with false contextual information
- Manipulated content: genuine information or imagery that is manipulated to deceive

This does not include content that is clearly satire or parody, as these are unlikely to cause harm, though they may sometimes fool an audience. It also does not include good faith reporting errors from otherwise trustworthy sources with a consistently applied corrections policy.

[3]**Harm:** The focus of this effort is on societal or public harm, but in some cases, harm to individuals also causes second-order societal harm, such as when journalists or institutional leaders are attacked via the spread of verifiably false information and

claims about them personally. Verifiably false information about brands is also harmful. For these reasons, we have removed "public" from the core definition and believe that this nuance belongs in approaches to focusing fact checking and enforcement resources. Harms include but are not limited to reputational, commercial, electoral, scientific, social, and health damages as well as incitement to violence, individually or collectively.

Priority categories. While all harmful false or misleading claims are a serious concerns that must be addressed, categories for immediate and dramatic improvement are Voting/Elections/Census, Environment, Conspiracies, Racism and Health information.

Key rationale for definition

We have reviewed thinking from a wide variety of academic sources on this topic and focused on the current official definition from the European Commission, differentiated by work from researcher Claire Wardle to describe types of false and misleading claims as enumerated above.

The European Commission definition describes disinformation as: "verifiably false or misleading information that is created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm. Public harm comprises threats to democratic political and policy-making processes, as well as public goods such as the protection of EU citizens' health, the environment or security. Disinformation does not include reporting errors, satire and parody, or clearly identified partisan news and commentary."

We have removed all mentions of intention from the European Commission definition and from the categories of false and misleading claims. Intention is extremely difficult to prove and time consuming to speculate over during fact-checking operations. Instead, we have focused on whether claims are verifiably false or misleading and harmful. We have also omitted reference to partisan news, in order to avoid confusion as to whether verifiably false or misleading harmful claims within partisan news and commentary are excepted; they are not.

Suitability Categories for Misinformation/Disinformation Adjacency in Advertising

Low risk

- Educational, Informative, Scientific treatment of misinformation or disinformation
- News features describing various disinformation campaigns as such

Medium risk:

- Dramatic depiction of misinformation presented in the context of entertainment (e.g.: a sketch comedy show including a person dramatizing injecting bleach as a COVID-19 treatment)

F. Breaking News or Op-Ed coverage of misinformation or disinformation (e.g.: a breaking news story reporting on a celebrity contending that a medically discredited treatment for COVID-19 should be pursued)

High risk: should be removed

- Glamorization/Gratuitous depiction of misinformation or disinformation (e.g.: an influencer video encouraging people to try injecting bleach as a treatment for COVID-19)

Proactively identify, monitor and manage against Misinformation and Disinformation

At first glance, any one of the billions of pieces of content generated daily could be dis/misinformation, the challenge of identification initially appears impossibly large. However, we see a path to dramatic improvement over current levels of client and consumer exposure to misinformation and disinformation via the adoption of the following approaches founded on reputation--not only internally established by the platform, but independently confirmed by trusted third party authorities. Each of these measures will reduce the size of the fundamental challenge, and in combination we believe that challenge can be reduced to a manageable size. The following should apply to all types of content--whether posts, URL's, images, especially memes, videos, including livestreams, or others as they appear. Some of these measures are already in place to varying degrees on the platforms, but urgent progress is needed across the board. Simply focusing on managing distribution and monetization of sources with poor reputation on the basis of historical evidence of misinformation and disinformation has the potential to significantly reduce the total amount of disinformation within the information ecosystem. (For example, GDI has identified over 500 sites publishing high volumes of divisive content.)

A. A reputation a reputation for honesty in the source should fundamentally affect the ease, speed and most importantly reach of content distribution. That reputation should not wholly depend on their history on the platform itself - external sources of reputation, such as agreed third-party authorities in a given field, should figure heavily into what content receives wide distribution as well. For example, sources such as United for News, the Local News Consortium, NewsGuard, GDI, the Trust Project, and the European Digital Media Observatory are qualified foundational sources of externally validated reputation for news organizations. Domain specific trusted arbiters for other key areas like science and health can and should be developed in partnership with reputable fact checking organizations. Content from reputable sources should pass by default through to wide distribution without fact checking. If content from these organizations begins to receive confirmed fact checks via user reporting that constitute a pattern, an established process with input from the source of that third party reputation and fact checking should result. There should be pooled financial support from both platforms and government for the work needed from these organizations to help maintain the foundations for reputation across the information ecosystem, including progressively smaller markets. This applies not only to social platforms, but also to inventory marketplaces and partners both driving traffic to and helping to monetize engagement with disinformation.

B. Reduce the size of the problem by removing violating accounts of both organizations and individuals that have high reach and a history of consistent, confirmed fact checks much more readily. Dedicate resources specifically to fact checking each post from those users that a platform decides to maintain despite violation of this policy. Due process guarantees could include warnings, indication of

the nature of the false content involved and a right to appeal any decision, preferably before it came into effect or at least via a very rapid procedure, to an independent oversight decision-maker.,.

C. Consider UX changes that express the level of trust in the source that a user can expect to have. The false equivalence between a news story from a real source and a fake news story is reinforced by their virtually indistinguishable representations when expressed in social feeds, for example. Once an external source of reputation is agreed on, that reputation can be used to help users visually and intuitively distinguish between reputable and low or unknown reputation sources, via simple means such as the size of embedded post images.

D. Carefully and rapidly monitor content from low-reputation sources achieving high distribution velocity and acceleration. By the time hundreds of thousands of people are rapidly seeing the same piece of content from a low-reputation source (that is, a source with either poor or unknown reputation) it should be prioritized in a queue for review on a topical basis in real time. Turnaround should be based on velocity and acceleration. We believe there is a place for cross-platform coordination around disinformation as well, so that platforms flagging disinformation urls, images, videos, livestreams, and the like, can make others aware of potential issues, including flowing fact checking status of content into programmatic marketplaces. Currently, some platforms depend exclusively on user driven reporting to add items to a fact checking queue, but when low reputation sources are catering directly to their audiences, those items may go unreported for a long time. Proactive monitoring using reputation as a guide could help greatly to manage disinformation in these audiences.

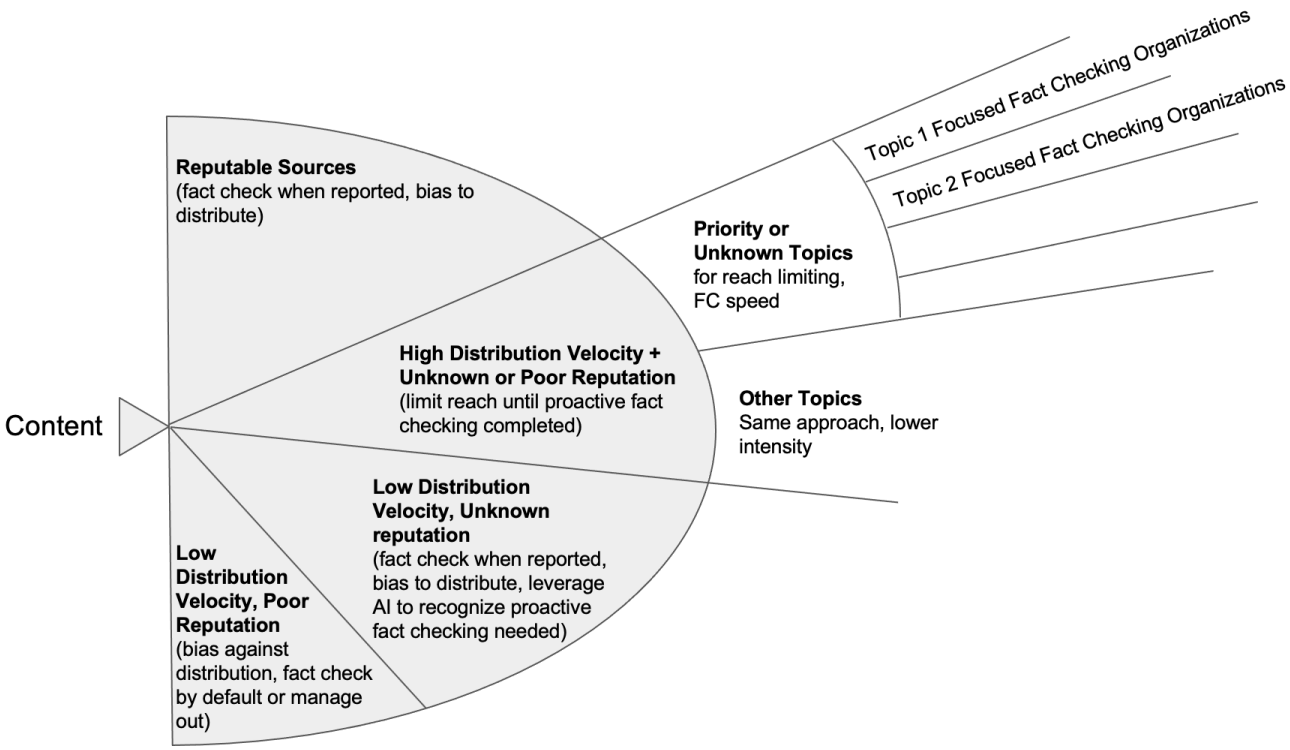
E. Break the task of fact checking down by topic to make it more manageable. Leverage common keywords and machine learning to bucket high-velocity individual items for fact checking to fact checking organizations.

F. Track at the content level in addition to the source level so that the same content posted many times from the same external disinformation sources cannot achieve high viewership through aggregation of many smaller audiences.

G. Bias to hiding or removing confirmed misinformation and disinformation content, not only presenting additional trusted sources alongside it. Recent research has shown that absent these recommendations the reach of disinformation can dwarf that of trusted content.

H. Significantly increase investment in identification and removal of accounts participating in coordinated inauthentic behavior, with an immediate focus on elections, not only in first world countries, and ensure and demonstrate that appropriate oversight for such decisions is in place. Response to such activities must occur in hours, not years, for elections integrity to be maintained. The need is not limited to elections, as documented state level efforts to sow division along racial lines, for example, are of similar concern.

MANAGEABLE PRIORITIZATION OF CONTENT FOR PROACTIVE FACT CHECKING



Greatly increase the volume, speed and quality of fact checking

The level of financial commitment to fact checking efforts to support the identification of misinformation and disinformation on some platforms has been called into question in recent months, with several sources reporting the number of fact checked posts has been only in the hundreds on some platforms as recently as January and later. This appears to be the case on even large topics, when the scope of the problem is clearly orders of magnitude larger than the currently allocated resource. The number of journalists driven out of work by recent economic shocks suggests the problem is not one of potential manpower, and that the situation could be improved by sustained investment in this area.

A. Significantly increase the amount of money directed to fund third party fact checking. Make fact checking a meaningful source of income for those organizations engaged in it in order to help deepen the available number of fact checkers in the ecosystem.

B. Report transparently and regularly on the resources devoted to fact checking and its performance. This includes, dollars allocated by organization, number of items fact checked by each, what those items were, the turnaround on fact checking --how long each item took to get into and out of the queue by those organizations. Work collaboratively to increase throughput.

C. Increase the amount of coordination between fact checking organizations, by enabling organizations to flag to each other which items they are working on non-exclusively.

D. Offer and prioritize items achieving velocity to fact checkers in real time. Some platforms have refreshed fact checking feeds on a weekly basis. This is plainly not nearly fast enough given the speed with which viral content moves. Fact checking feeds must be updated in real time and prioritized based on the velocity and acceleration of content.

E. Provide categories reflective of the types of misinformation and disinformation that commonly appear. For example, conspiracy theories and pseudoscience are not currently categories in all systems. A larger number of categories assist in the designation of various types of disinformation.

F. Create tools to speed and improve fact checking. The use of machine learning to augment the effectiveness and speed of human fact checkers will be critical to winning the arms race against disinformers.

G. Correct the record. This applies not only to those who have shared disinformation, but for those who have been exposed to it.

H. Coordinate across platforms to identify and fact check misinformation and disinformation achieving velocity with a common database of items of concern flagged as misinformation/d isinformation at the content level (text claims, url, image, video, etc). As a report prepared for the Senate Intelligence Committee regarding actions by Russia's Internet Research Agency in 2016 describes, the disinformation effort "operated like a digital marketing agency: develop a brand ... build presences on all channels across the entire social ecosystem, and grow an audience with paid ads as well as partnerships, influencers, and link-sharing. They created media mirages: interlinked information ecosystems designed to immerse and surround targeted audiences." In the face of cross-platform disinformation campaigns, cross-platform coordination appears critical. Legislation suggested by representative Rho Khanna in the US has suggested such a consortium, and researchers from Stanford University have called for a nongovernmental organization to combat disinformation similar to the nonprofit Financial Services Information Sharing and Analysis Center, which helps to avoid shocks in the financial system. Facebook itself has called for a similar approach. Former Facebook Chief Security Officer Alex Stamos told Congress in 2019 that ""[The GIFCT] has been somewhat successful in building capabilities in smaller members while creating a forum for collaboration among the larger members. It is time to follow this initial foray with a much more ambitious coordinating body between tech companies focused on adversarial use of their technologies."

Increase transparency and integrity around critical metrics, fact checking and consequences

Methods include third party involvement, auditing and rules, recognizing challenges from previous coordination efforts, prevalence transparency and regular cadence of tracking.

In the specific context of coordinating to identify and remove or limit the reach of misinformation and disinformation, existing models to counter unsafe content offer important lessons. For example, PhotoDNA, developed by Microsoft, automatically

and preemptively prevents photos of included images of child sexual abuse material (CSAM) from being uploaded to multiple social platforms. The database run by the industry consortium Global Internet Forum to Counter Terrorism (GIFTCT) to prevent the uploading of "terrorist speech" operates similarly. Both, however, have received criticism over being relatively opaque in terms of who has contributed and the impact of their efforts. While there is legitimate concern that governments could label political speech as unacceptable speech and upload it to these database to limit its reach and achieve de-facto censorship, there is a clear roadmap to avoid this, most critically governmental independence, corporate independence, transparency, auditing and credible, accountable third party oversight, inclusive of a robust remediation process and appropriate supporting resources, as scholarship from Evelyn Douek at Knight/Columbia has described in detail. One partial such approach is detailed in "Social Media Councils", a multi-stakeholder accountability model for content moderation on social media. Funding for such an organization could be accomplished via an initial endowment that enables subsequent financial independence. Appeals may proceed to a board composed of representatives from participating reputable NGO's.

A. Report on prevalence of misinformation and disinformation quarterly with third party auditing. The scale of the problem overall must be measured and tracked in order to understand how meaningful the response actually is.

B. Report quarterly on fact checking actions at the content level. ake the raw data and metadata around items removed available to researchers, so that the real impact and accuracy of fact checking efforts can be evaluated and real progress and credibility, not "transparency theatre," achieved. Researcher access, as opposed to general public access, enables transparency without providing a key to those looking to circumvent countermeasures and helps to maintain privacy of users.

C. Collectively create and fund a cross industry collaboration organization to maintain a centralized database of disinformation content. This organization should include a governance partnership with trusted nonprofit organizations in key domains such as health and science education. This database must be transparent, regularly auditable and available to researchers and support a similarly transparent, fast and robust redress process for items flagged incorrectly.

D. In removal decisions, to reclaim credibility, bias to truth, not newsworthiness. Some have suggested that the harmful recommendations of individuals with large existing followings are newsworthy, and thus should remain available or be subject to a less stringent bar of scrutiny. We explicitly and vehemently disagree. For example, health-focused misinformation posted by a famous celebrity is more damaging than that posted by a person with a much smaller following. Harmful, verifiably false claims are always worse coming from those with established followings. Reputation can be a useful filter here as well, as individuals could accumulate fact-checks that should affect the distribution of their content.

Create a foundation for media tools and approaches to manage proximity to misinformation and disinformation, both identified and potential, and report on that proximity in an auditable manner

The approach to adjacency management for media placements should be in line with all other categories in the Brand Safety Floor and Suitability Framework. Reporting should include the type of misinformation or disinformation concerned.

Complete Content-level transparency is needed prior to media alignment adjacency, or support. Given the pace at which social content moves, this should preferably be available in near-time and real-time to pace alongside news cycles. More detail is available via parallel GARM initiatives on adjacency management.

Closing

While the challenge posed by disinformation is massive and multifaceted, the good news is that many of the solutions described in this document, such as improving fact checking and collaboration around a transparent central database of content of concern, are likely to drive significant improvement in other areas of harmful content as well. The 4A's Advertiser Protection Bureau looks forward to working with partners from across the media ecosystem to create a more transparent, collaborative and truthful future.