moonshot

Metrics validation - Final report

The Australian Code of Practice on Disinformation and Misinformation

JUNE **2024**

MOONSHOTTEAM.COM

Overview

In March 2024, Moonshot was contracted by the Australian Communications and Media Authority (ACMA) to assist in their oversight of the **Australian Code of Practice on Disinformation and Misinformation** (the Code). The Code aims to mitigate the risks posed by mis- and disinformation (MDI), and establishes clear Objectives and Outcomes to reduce vulnerability to harm from false or misleading information among the Australian public. Crucial to the success of the Code is the development of standard metrics to measure how platforms are working to achieve its Objectives.

As part of this oversight role, ACMA has developed a consistent set of measures and metrics against which digital platforms can report their efforts to meet the Code's objectives. A thorough evaluation and testing of the metrics is essential to the success of the Code, while ensuring respect for freedom of expression and other important rights.

The aim of this project is to assist ACMA in finalising the metrics developed to assess digital platforms' compliance with the Code. The project was originally divided into two parts:

Part 1
Providing independent expert advice to validate the 64 metrics identified by ACMA in the draft framework.

Part 2
Providing data (where available) against the finalised set of metrics to enable comparisons with future reports provided by signatories.

While implementing Part 1 of the project, Moonshot identified significant barriers to the collection of a comprehensive dataset. As a result of this, the nature of Part 2 was changed, and the data collection exercise was replaced with the provision of independent expert advice to validate an additional four metrics identified by ACMA.

This **final report** provides an overview of the results of Part 1 and 2 of the project, and is submitted alongside an accompanying spreadsheet that details the assessment conducted on each platform. It incorporates feedback provided by ACMA on the **draft report**. As part of the validation exercise, Moonshot assessed the 68 metrics in relation to 15 platforms: Facebook; Instagram; Threads; Google Search; YouTube; Google News; Google Ads; Bing Search; LinkedIn; TikTok; X (formerly Twitter); Reddit; Spotify; Twitch; and Snapchat.

The assessment of each platform was captured in a validation framework, which provides a guide to evaluate the following core questions for each metric:

1. Is the digital platform likely to be able to provide data against each metric?
2. Is the metric likely to provide Australia-specific insights into the success of the measures?
3. Would an alternative metric yield better Australia-specific insights into the success of the measures?
4. Could the information reported by signatories against the metrics be independently verified?
5. Are there any other relevant issues and/or barriers to implementation of the framework?

The approach to the validation exercise mixed analysis of platform infrastructure with reviews of transparency reports, platform policy documents and academic literature. This review was also supplemented by a broader assessment of the metrics to identify any relevant barriers to the implementation of the framework. For a detailed overview of our methodology, including how the platforms in scope for the project were selected, please refer to Appendix 2 at the end of this report.

↘ Table of contents

Platforms

Executive summary

Most platforms were assessed as likely to be able to provide data against each metric.

However, the metrics were found to be significantly less applicable to the search engines evaluated for the project, due to the operational differences between social media platforms and search engines. The likelihood was established by examining the relevance of each metric to a platform's operations and services; whether the platform already provides similar data to other regulators; and whether the provision of data against each metric would violate a platform's policies regarding user privacy or intellectual property, as well as Australian legislation on privacy and freedom of expression.

All platforms were assessed as likely to be able to provide Australia-specific insights into the success of the measures.

A review of each platform's operations and services, as well as reviews of their reporting to other regulators and the public, indicated that platforms are in possession of data related to user location, making them able to provide geographically-disaggregated insights.

No alternative metrics were recommended as yielding better Australia-specific insights.

However, Moonshot identified a number of metrics where changes were recommended, primarily to improve the metrics' clarity, as well as a limited number of metrics that were recommended for removal, either because they are largely duplicates of other metrics or because they were identified as less relevant to their related outcome or not applicable to the service. More information can be found in the recommendations section of this report.

Overall, Moonshot identified no feasible pathway to independent verification of data provided by signatories.

Significant barriers to independent verification were identified. These include a lack of overall data access, through APIs or other means; risks to user privacy rights in collecting a comprehensive dataset; a lack of public information on platform policy violations to enable targeted collection; a lack of standardised definitions and understandings of MDI; and an overarching inability to provide data that would be comparable to that provided by the platforms. These barriers are expanded upon in more detail later in the report.

Aside from challenges related to independent verification, Moonshot did not identify any other major barriers to implementation.

Aside from challenges related to independent verification, Moonshot did not identify any other major barriers to implementation. Despite this, some recommendations have been included in this report for ACMA's consideration. These include: changing the overall approach to request mean instead of median averages in relation to most data points; considering risks to freedom of expression in relation to requests for data connected to content removal; and evaluating the applicability of the framework to platforms where users cannot generate their content, such as search engines.

YouTube

YouTube states that "*certain types of misinformation that can cause real-world harm, certain types of technically manipulated content or content interfering with democratic processes*" are all in violation of the platform's policies.[1] However, it makes no distinction between misinformation and disinformation.

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against all metrics, and has previously provided data similar to 17 of the requested metrics in reporting to the EU and DIGI. A preliminary review of YouTube's policies, platform architecture and recommender systems was conducted, and found all metrics to be applicable to the platform's operations and services, and only two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by the signatory against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that almost all metrics pose challenges for independent verification. Although YouTube has an API, the API does not, in the vast majority of cases, have endpoints that would allow a third party to reliably collect data against the metrics - as even when such endpoints exist, they are not specific to a particular geography. Additionally, a third party would not be able to gather data against a significant portion of the metrics without violating user privacy rights, as some actions undertaken by users (for example, reporting of content in breach of policy) are performed with the understanding that they will not be visible to others. Moonshot identified a limited number of instances where some data could be gathered and provided to ACMA as a case study. These are highlighted in the accompanying spreadsheet. For all metrics, it is important to emphasise that data that could be gathered by a third party would not be fully comparable to what the platform has access to themselves, and could therefore provide in their reporting to ACMA.

[1] YouTube, Misinformation policies, accessed 09 May 2024.

Facebook

Facebook defines misinformation as a concept in its policies, but provides no clear distinction between misinformation and disinformation, which could result in some discrepancies in reporting.[2] Facebook defines coordinated inauthentic behaviour in its policies in line with the definition provided in the Code.[3]

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against all metrics, and has previously provided data similar to 17 of the requested metrics in reporting to the EU and DIGI. A preliminary review of Facebook's policies, platform architecture and recommender systems was conducted and found all metrics to be applicable to the platform's operations and services, and only two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by the signatory against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. The Meta Content Library and API are restricted to academics and NGOs, and access is not provided to private companies, governments, or government bodies. Although there is a significant number of third party tools that allow for data to be collected from Facebook, these tools do not capture most metrics, and are also, for the most part, in breach of Meta's Automated Data Collection policies.[4] Additionally, a third party would not be able to gather data against a significant portion of the metric without violating user privacy rights, as some actions undertaken by users (for example, reporting of content in breach of policy) are performed with the understanding that they will not be visible to others. Moonshot identified a limited number of instances where some data could be gathered and provided to ACMA as a case study. These are highlighted in the accompanying spreadsheet. However, in the case of Facebook, this data would have to be collected manually, and would not be comparable to what the platform could provide in their reporting to ACMA.

[2] Meta, Misinformation, accessed 21 April 2024.

[3] Meta, Inauthentic behaviour, accessed 20 April 2024.

[4] Meta, Automated Data Collection, accessed 22 April 2024.

Instagram

As a Meta product, Instagram defines misinformation in its policies. However, it provides no clear distinction between misinformation and disinformation.[5]

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against all metrics, and has previously provided data similar to 22 of the requested metrics in reporting to the EU and DIGI. A preliminary review of Instagram's policies, platform architecture and recommender systems was conducted, and found all metrics to be applicable to the platform's operations and services, and only two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that almost all metrics pose challenges for independent verification. Although Instagram has an API, the API does not, in the vast majority of cases, have endpoints that would allow a third party to reliably collect data against the metrics - as even when such endpoints exist, they are not specific to a particular geography. Additionally, a third party would not be able to gather data against a significant portion of the metric without violating user privacy rights, as some actions undertaken by users (for example, reporting of content in breach of policy) are performed with the understanding that they will not be visible to others. Moonshot identified a limited number of instances where some data could be gathered and provided to ACMA as a case study. These are highlighted in the accompanying spreadsheet. For all metrics, it is important to emphasise that data that could be gathered by a third party would not be fully comparable to what the platform has access to themselves, and could therefore provide in their reporting to ACMA.

[5] Meta, Misinformation, accessed 21 April 2024.

Threads

Threads' Terms of Use clarify that the platform is "*part of Instagram*" and therefore governed by Instagram's policies. The platform, through Meta's policies, defines misinformation as a concept, but provides no clear distinction between misinformation and disinformation.[6]

Is the digital platform likely to be able to provide data against each metric?
Meta is yet to include Threads in its transparency reporting to the EU and DIGI. However, a preliminary review of Threads' policies, platform architecture and recommender systems found all metrics to be applicable to the platform's operations and services. Similarly to other social media platforms, two metrics were deemed to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
The platform is yet to release its API, making it impossible to establish if there will be any endpoints that would allow a third party to reliably collect data against the metrics. However, given the likelihood that the API will somewhat mirror the APIs of other Meta platforms, there will likely be similar challenges. For this reason, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[6] Meta, Misinformation, accessed 21 April 2024.

TikTok

TikTok's "Integrity and Authenticity" policy covers both disinformation, defined as information that is "*intentionally shared to mislead*", and misinformation, defined as "*harmful misinformation that may not have been shared with the goal of deceiving people*". However, it does not distinguish between the two in its enforcement, stating that the policy is applied to content "*regardless of the poster's intent, as the content's harm is the same either way*".[7]

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against all metrics, and has previously provided data similar to 42 of the requested metrics in reporting to the EU and DIGI. A preliminary review of TikTok's policies, platform architecture and recommender systems was conducted, and found all metrics to be applicable to the platform's operations and services, and only two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that almost all metrics pose challenges for independent verification. TikTok's API does not, in the vast majority of cases, have endpoints that would allow a third party to reliably collect data against the metrics - as even when such endpoints exist, they are not specific to a particular geography. Additionally, TikTok has made amendments to their Research Tools Terms of Service that include further clarity on the process researchers must undertake to gain access to their API.[8] A review of these updates indicates that any research into the proliferation of MDI on TikTok is unlikely to receive approval for access to their API. Third party tools would also not be able to gather data against a significant portion of the metric without violating user privacy rights, as some actions undertaken by users (for example, reporting of content in breach of policy) are performed with the understanding that they will not be visible to others. Moonshot identified a limited number of instances where some data could be gathered and provided to ACMA as a case study. These are highlighted in the accompanying spreadsheet. However, in the case of Facebook, this data would have to be collected manually, and would not be comparable to what the platform could provide in their reporting to ACMA.

[7] TikTok, Community Guidelines: Integrity and Authenticity, accessed 13 May 2024.

[8] TikTok's Research Tools Terms of Service were amended on 14 May 2024. TikTok, Research Tools Terms of Service,, accessed 16 May 2024.

X (formerly Twitter)

X does not provide a definition of disinformation or misinformation per se in its policies. The platform's policies focus instead on "*synthetic and manipulated media*", "*misleading and deceptive identities*" and "*platform manipulation and spam*".[9, 10, 11] This could present challenges in reporting against the metrics should X become a Signatory to the Code.

Is the digital platform likely to be able to provide data against each metric?
Despite the lack of any reference to MDI in X's policies, a preliminary review of the platform architecture and recommender systems found all metrics to be applicable to the platform's operations and services. Notwithstanding, X is unlikely to be able to provide consistent data against the metrics, and has previously provided similar data only against two metrics. The lack of clear definitions, combined with the platform's transparency reporting to the EU also refraining from mentioning MDI, point to the likelihood of X focusing on adjacent policies instead.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that almost all metrics pose challenges for independent verification. Although X has an API, the API does not, in the vast majority of cases, have endpoints that would allow a third party to reliably collect data against the metrics - as even when such endpoints exist, they are not specific to a particular geography. Moreover, X has been limiting third parties' ability to reliably collect data from its platform, even when in possession of Enterprise access, restricting the volumes of data that can be ingested over a certain period of time. Additionally, a third party would not be able to gather data against a significant portion of the metrics without violating user privacy rights or the platform's policies in relation to data collection on their platform. Moonshot identified a limited number of instances where some data could be gathered and provided to ACMA as a case study. These are highlighted in the accompanying spreadsheet. For all metrics, it is important to emphasise that data that could be gathered by a third party would not be fully comparable to what the platform has access to themselves, and could therefore provide in their reporting to ACMA.

[9] X, Synthetic and manipulated media policy,accessed 08 May 2024.

[10] X, Misleading and deceptive identities policy, accessed 08 May 2024.

[11] X, Platform manipulation and spam policy, accessed 08 May 2024.

Twitch

Twitch's policies in relation to misinformation and disinformation are a prime example of how the functionalities of different platforms lead to different moderation efforts and concerns. The platform defines misinformation in its policies and provides an extensive list of the types of misinformation that are not allowed,[12] although it does not distinguish between misinformation and disinformation. However, the platform's focus is primarily on moderating accounts rather than content,[13] arguing that due to its focus on live streaming, "*much of the live violative content is already gone*".[14]

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against most metrics, and has previously provided data similar to 13 of the requested metrics in reporting to the EU. A preliminary review of Twitch's policies, platform architecture and recommender systems was conducted, and found only three metrics to be inapplicable to the platform (metrics 31, 46 and 49) and a further two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics that were deemed applicable to the platform's operations. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
11 metrics were recommended for removal, either because they are considered a duplicate of other metrics, they are less relevant to their related outcome, or do not apply to the platform. We identified an additional 31 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. Although Twitch has an API, the API does not have endpoints that would allow a third party to reliably collect data against any of the metrics. The focus on live streaming of content also makes identifying content or actors in breach of MDI policies a resource intensive and predominantly manual process. Although certain tools allow third parties to export chat rooms that take place alongside livestreams, they require authorisation tokens from Twitch, and are unlikely to receive approval for the independent validation of any of these metrics. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[12] Twitch, Harmful Misinformation Actors, accessed 06 May 2024.
[13] Twitch, Community Guidelines, accessed 06 May 2024.
[14] Twitch, H1 2023 Transparency Report, accessed 06 May 2024.

in Linkedin

Although LinkedIn's policies do not explicitly define MDI, they do include references to "*false or misleading content*"[15] and specify that "*misinformation and inauthentic content is not allowed*".[16] Additionally, Microsoft's reports provided to the EU and DIGI highlight a commitment to the Code, and implicitly appear to interpret MDI in line with the Code's definitions.

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against all metrics, and has previously provided data similar to 20 of the requested metrics in reporting to the EU and DIGI. A preliminary review of LinkedIn's policies, platform architecture and recommender systems was conducted, and found all metrics to be applicable to the platform's operations and services, and only two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. Although LinkedIn has an API, the API does not have endpoints that would allow a third party to reliably collect data against any of the metrics. Any data collection would require the users' consent, and the LinkedIn Partner Program (the platform's version of an Enterprise API) is not applicable to this project's use case. As for external tools, although they do exist, they lack the endpoints specific to most metrics, and LinkedIn has put restrictions in place on the volume of data that can be ingested by third parties. Finally, a third party would not be able to gather data against a significant portion of the metrics without violating user privacy rights, as some actions undertaken by users (for example, reporting of content in breach of policy) are performed with the understanding that they will not be visible to others. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[15] LinkedIn, False and misleading content, accessed 13 May 2024.

[16] LinkedIn, Misinformation and inauthentic behavior, accessed 13 May 2024.

Reddit

Reddit's Content Policy does not explicitly mention MDI. However, the platform provides rules that disallow "*content manipulation*" and impersonation of an individual or entity in a "*misleading or deceptive manner*".[17] This, coupled with the fact that the majority of content moderation on the platform is not centralised but carried out by the moderators of individual subreddits,[18] could create challenges in reporting data under the Code.

Is the digital platform likely to be able to provide data against each metric?

The platform was judged to be capable of providing data against all metrics, and has previously provided data similar to 20 of the requested metrics through internal annual transparency reports (Reddit is not a Signatory of the Code, nor does it reach enough of the EU population to have DSA Transparency Reporting duties).[19] A preliminary review of Reddit's policies, platform architecture and recommender systems was conducted, and found all metrics applicable to the platform's operations and services, and only two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?

A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?

Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?

A preliminary review of the platform's API and architecture highlighted that almost all metrics pose challenges for independent verification. Although Reddit has an API, the API does not, in the vast majority of cases, have endpoints that would allow a third party to reliably collect data against the metrics - as even when such endpoints exist, they are not specific to a particular geography. Additionally, a third party would not be able to gather data against a significant portion of the metric without violating user privacy rights, as some actions undertaken by users (for example, reporting of content in breach of policy) are performed with the understanding that they will not be visible to others. Moonshot identified a limited number of instances where some data could be gathered and provided to ACMA as a case study. These are highlighted in the accompanying spreadsheet. For all metrics, it is important to emphasise that data that could be gathered by a third party would not be fully comparable to what the platform has access to themselves, and could therefore provide in their reporting to ACMA. In May 2024, the platform announced its new "Public Content Policy" and that it is "*building new tools for researchers*" to "*improve their access*" to "*public data*". More extensive data collection could be achieved in the future depending on the nature of the improved access.

[17] Reddit, Content Policy, accessed 14 May 2024.
[18] New America, Everything in Moderation - Case Study: Reddit, accessed 14 May 2024.
[19] Reddit, Transparency Reports, accessed 14 May 2024.

Snapchat

Snapchat's "Harmful False or Deceptive Information" policy prohibits all forms of "*information threats*", including "*misinformation, disinformation, malinformation and manipulated media*". Under the "False information" category, the platform provides an extensive list of examples of content that is not allowed, and it also clarifies that its team takes action against all content that is "*misleading or inaccurate, irrespective of whether the misrepresentations are intentional*".[20]

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against most metrics, and has previously provided data similar to 38% of the requested metrics in reporting to the EU. A preliminary review of Snapchat's policies, platform architecture and recommender systems was conducted, and found only three metrics to be inapplicable to the platform (metrics 31, 46 and 49) and a further two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics that were deemed applicable to the platform's operations. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
11 metrics were recommended for removal, either because they are considered a duplicate of other metrics, they are less relevant to their related outcome, or do not apply to the platform. We identified an additional 31 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification, including due to the platform's privacy settings. Although Snapchat has an API, the API does not have endpoints that would allow a third party to reliably collect data against any of the metrics, as it is primarily for marketing purposes and requires the authorisation of users for any data collection. Third party tools are also not currently capable of collecting Snapchat content and limit data gathering to basic profile information such as bio, location or subscriber count. Additionally, these tools are not authorised by Snapchat, and require manual input. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[20] Snapchat, Harmful False or Deceptive Information, accessed 10 May 2024.

Google Ads

Google Ads' "Misrepresentation" policy refers to the provision of "misleading information" and "attempts to deceive" in the context of advertising for a product, service or business. The policy also addresses "coordinated deceptive practices".[21]

Is the digital platform likely to be able to provide data against each metric?
The platform was judged to be capable of providing data against most metrics, and has previously provided data similar to 34 of the requested metrics in reporting to the EU and DIGI. A preliminary review of Google Ads policies, platform architecture and recommender systems was conducted, and found two metrics to be inapplicable to the platform (metrics 23 and 56) and a further two to be less relevant to their associated measure or outcome (metrics 37 and 38). Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics that were deemed applicable to the platform's operations. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Ten metrics were recommended for removal, either because they are considered a duplicate of other metrics, they are less relevant to their related outcome, or do not apply to the platform. We identified an additional 32 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. Although Google Ads has an API, the API does not have endpoints that would allow a third party to reliably collect data against any of the metrics, as it primarily serves as a means for advertisers to manage their Google Ads accounts and campaigns. Additionally, a third party would not be able to gather data against a significant portion of the metrics without violating user privacy rights or the platform's policies in relation to data collection on their platform. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[21] Google, Misrepresentation, accessed 14 May 2024.

Google Search

Google Search's policies do not explicitly define MDI, but cover "*misrepresentation*",[22] "*manipulated media*" and "*deceptive practices*",[23] including many of the behaviours associated with MDI. Transparency reports provided to the EU and DIGI include recognition of the definitions of misinformation and disinformation.[24]

Is the digital platform likely to be able to provide data against each metric?

A preliminary review identified 19 metrics to be not applicable to the platform's operations and services. In most instances, this was due to the operational differences between social media platforms and search engines. While reporting content[25] is possible through a feedback portal, the platform does not provide a predefined list of violations - prompting users to provide a freeform explanation instead on why they are reporting the content. Additionally, no evidence of a user appeals process was identified. The lack of applicability of many of the metrics is reflected in the historic provision of data in transparency reporting, with only 11 instances of data similar to the requested metrics featuring in reporting to the EU and DIGI. Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?

A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics that were deemed applicable to the platform's operations. Given the architecture of search engines, it may be technically feasible for the platform to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?

23 metrics were recommended for removal, either because they were deemed to be inapplicable to the platform's operations and services or because they largely duplicated other metrics. We identified an additional 28 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?

A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. Although Google Search has an API, the API does not have endpoints that would allow a third party to reliably collect data against the metrics. A preliminary review of available external tools also did not identify any technical solutions to gather data against the metrics. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[22] Google, Policies for Content Posted by Users on Search, accessed 09 May 2024.

[23] Google, Content policies for Google Search, accessed 09 May 2024.

[24] Google, Google Annual Transparency Report, May 2023, accessed 09 May 2024.

[25] For the purpose of assessing search engines, content has been interpreted as search results.

Bing Search

Although Bing's policies do not explicitly define MDI, they do cover "*misleading information*"[26] and "*fraudulent, false or misleading*" activity.[27] Additionally, Microsoft's reports to the EU and DIGI highlight a commitment to the Code, and implicitly appear to interpret MDI in line with the Code's definitions.

Is the digital platform likely to be able to provide data against each metric?
Similarly to Google Search, a preliminary review identified 19 metrics as not applicable to Bing's operations and services. In most instances, this was due to the operational differences between social media platforms and search engines. While reporting content is possible through the "report a concern" portal and the "feedback" tool, the platform does not provide a predefined list of violations that includes MDI. Additionally, no evidence of a user appeals process was identified. The lack of applicability of many of the metrics is reflected in the historic provision of data in transparency reporting, with only five instances of data similar to the requested metrics featuring in reporting to the EU and DIGI. Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics that were deemed applicable to the platform's operations. Given the architecture of search engines, it may be technically feasible for the platform to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
23 metrics were recommended for removal, either because they were deemed to be inapplicable to the platform's operations and services or because they largely duplicated other metrics. We identified an additional 28 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. Although Bing Search has an API, the API does not have endpoints that would allow a third party to reliably collect data against the metrics. A preliminary review of available external tools also did not identify any technical solutions to gather data against the metrics. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[26] Microsoft, Microsoft Services Agreement, accessed 09 May 2024.

[27] Bing, How Bing delivers search results, accessed 09 May 2024.

Spotify

Spotify's policies do not explicitly define MDI. The platform's rules specify that "*deceptive content*" that "*promotes manipulated and synthetic media as authentic*" is in breach,[28] and so is content that "*promotes dangerous false or dangerous deceptive medical information that may cause offline harm*".[29] This could present challenges in reporting against the metrics should Spotify become a Signatory to the Code.

Is the digital platform likely to be able to provide data against each metric?
No evidence of Spotify providing similar data in any transparency reporting was identified. A preliminary review of Spotify's policies, platform architecture and recommender systems was conducted, and found all metrics to be applicable to the platform's operations and services. However, Spotify is unlikely to be able to provide consistent data against the metrics. The lack of clear definitions, combined with the platform not releasing any transparency reporting that mentions MDI, points to the likelihood of Spotify focusing on adjacent policies instead. Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal.

Is the metric likely to provide Australia-specific insights into the success of the measures?
A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?
Eight metrics were recommended for removal, primarily because they are considered a duplicate of other metrics or because they are less relevant to their related outcome. We identified an additional 34 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?
A preliminary review of the platform's API and architecture highlighted that all metrics pose challenges for independent verification. Although Spotify has an API, the API does not have endpoints that would allow a third party to reliably collect data against the metrics, and requires the user's consent for data to be collected against their profile. A preliminary review of available external tools also did not identify any technical solutions to gather data against the metrics. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[28] Spotify, Spotify Platform Rules, accessed 10 May 2024.
[29] Ibid.

Google News

Content on Google News' must follow all Google Search Content Policies. Therefore, the platform does not explicitly define MDI, but it does cover "*misrepresentation*",[30] "*manipulated media*" and "*deceptive practices*",[31] including many of the behaviours associated with MDI. The platform also has Google News-specific policies that include similar restrictions to those mentioned above.[32] Transparency reports provided to the EU and DIGI include recognition of the definitions of misinformation and disinformation.[33]

Is the digital platform likely to be able to provide data against each metric?

A preliminary review identified two metrics as not applicable to the platform's operations and services due to the operational differences between a news aggregator and social media platforms. A further five metrics were deemed only applicable to the platform to an extent. These metrics all relate to "actors", which in the case of this platform would be news outlets. Therefore, particular care and consideration should be taken in terms of the potential impact that the provision of data against any of the metrics may have in endangering freedom of speech. The platform was found to have only provided similar data in the past only against three metrics. Potential risks to freedom of expression were identified in relation to 11 metrics, either connected to content removal or to "actors" for the reasons explained above.

Is the metric likely to provide Australia-specific insights into the success of the measures?

A preliminary review of the platform's architecture, transparency reporting and policies in relation to user privacy found that Australia-specific data could be provided against all metrics that were deemed applicable to the platform's operations. However, it is unlikely that the platform would be able to provide more granular data for specific Australian states and territories.

Would an alternative metric yield better Australia-specific insights?

Ten metrics were recommended for removal, either because they are considered a duplicate of other metrics, they are less relevant to their related outcome, or do not apply to the platform. We identified an additional 32 metrics that could be rephrased, either for clarity or to provide a better measurement for their outcome.

Could the information reported by signatories against the metrics be independently verified?

A preliminary review of the platform's architecture highlighted that all metrics pose challenges for independent verification. Google News does not currently provide an API, and it is unlikely that it will in the future. Although several third-party tools were identified, none of them provide endpoints relevant to the collection of data against the metrics. For these reasons, although independent verification may be re-evaluated in the future, it does not appear possible at the time of assessment.

[30] Google, Policies for Content Posted by Users on Search, accessed 09 May 2024.
[31] Google, Content policies for Google Search, accessed 09 May 2024.
[32] Google, Google News-specific Policies, accessed 09 May 2024.
[33] Google, Google Annual Transparency Report, May 2023, accessed 09 May 2024.

Recommendations and barriers to implementation

Independent verification

Overall, the main barrier identified is related to independent verification of data provided by the services. Upon review of the metrics, we identified the following challenges:

Lack of standardised definitions of MDI and opaque enforcement of MDI policies

Overall, one of the main challenges to independent verification relates to a lack of a standardised definition of MDI across services. Definitions of MDI, where they do exist, tend to be vague and, in a lot of instances, subject to individual interpretation. This is primarily due to a lack of governmental policies that clearly define what MDI is as well as the fluid and ever-changing nature of the issue. Because of this reason, any data collection of MDI by a third party will need to rely on an external interpretation of what content the platform may deem in breach of its policies - but without a “black and white” framework that defines which content meets the MDI threshold. This is further complicated by the fact that services do not always specify which policy was breached to trigger moderation - referencing breaches of “Terms of Service” more broadly instead. This means that any third party would lack the necessary information to collect data on moderation efforts, as only the platform would be privy to this. Any data collection would be based on educated assumptions of which content is eligible for moderation based on what the services include in their policies, as there is no record - public or through the API - to reliably collect cross-platform data on this.

Lack of overall data access, through APIs or other means

Almost all metrics request access to data that only the service would be privy to. This includes, for example, data related to: content that is removed before appearing on the platform; individuals employed by the services; levels of engagement with the service’s policies; the specific policy that was violated and led to content being moderated, and how the violation was identified by the service; or appeals and reports filed by users in response to content moderation. Additionally, most platforms do not provide a level of disaggregation for the action of simply viewing content that would allow a distinction between “reach” and “impressions” to be made. Although in broad terms some data could be gathered against a limited number of the metrics, the geographical specificity required presents further challenges, as API endpoints, where they exist, rarely allow for any geographical disaggregation whatsoever. This is true for third party tools as well, which present additional challenges as in most instances their data access is not authorised by the services and in breach of their Terms of Service.

User privacy

In some cases, data gathering by a third party also presents risks to user privacy rights. This particularly applies to actions undertaken by users with the assumption that they would not be made public, such as reporting content in breach of policies or appealing moderation by services. More broadly, the metrics’ emphasis on “active” users presents additional challenges, as privacy settings would prevent a third party from accurately assessing whether a user that saw or engaged with a piece of content is active on the site or not without gaining access to their private profile first. Additionally, to collect a comprehensive dataset, any data collection effort would require gathering posts from private spaces - which presents further concerns related to user privacy.

Inability to provide data that would be comparable to that provided by the services

Even in the limited number of instances where some data could be gathered, the data provided by any third party would not be fully comparable to that provided by the services themselves. This is because, in the absence of specific API endpoints that allow for cross-platform data collection, any data gathering exercise would be limited to a case study - a limited collection exercise that focuses on collecting data for some of the metrics in relation to specific and predefined MDI narratives during a short time frame. These collections would not provide all-encompassing data that could be used for any baselining exercise, or to carry out comparisons with the service's reporting. Additionally, users' privacy settings would also prevent gathering any content that is not made public, presenting further challenges to the provision of a comprehensive dataset.

Freedom of expression

All metrics were evaluated against potential risks to freedom of expression. Potential risks to freedom of expression were identified in relation to seven metrics, all connected to content removal. Asking platforms to report on content takedowns can risk incentivising them to remove more content in an effort to appear like they are addressing the problem. This can end up restricting broader content that may neither be demonstrably harmful nor false.

There have been reports of content moderation enforcement leading to a number of instances where free speech has been restricted - ranging from YouTube removing content documenting the civil war in Syria to Facebook removing articles accompanied by the photo of Phan Thị Kim Phúc running naked after being burned on her back by a South Vietnamese napalm attack.[34] This is further complicated by the increasing use of artificial intelligence to moderate content, which creates both practical and principled limitations on how MDI is identified and dealt with. ACMA's inclusion of metrics aimed at measuring appeals of content removed by services may help gather data to better assess the effects of content removal as a method to address the harm of MDI.

Engagement and impressions

Several metrics related to understanding the impact of MDI request for platforms to provide data concerning the volume of engagement and impressions with certain content. Although we agree that engagement is the most pertinent metric to understand virality, we believe that impressions also provide a useful measurement of the overall harm caused by MDI, especially if understood in conjunction with engagement. We would recommend retaining both.

[34] Rasmus Kleis Nielsen, Reuters Institute, How to respond to disinformation while protecting free speech, accessed 14 May 2024.

Requesting mean instead of median averages

A significant number of metrics require services to report certain data points by calculating the median, either in terms of reach, impressions, engagement or follower count. We recommend a broader change in approach to request mean averages or totals instead. Although we understand ACMA's desire to minimise the impact of outliers, we believe that outliers are an important indicator of the overall harm of MDI on online audiences - and that they should, as a consequence, be included in how the service compiles the metrics. This is because a single piece of content that achieves virality, or an account with a particularly high volume of followers, can have a disproportionate impact on online users, especially within the context of MDI. Excluding these values could result in a partial picture of the overall harm caused within the reporting period. Alternatively, ACMA might also opt to request totals instead of mean or median averages for certain metrics, and then calculate averages internally using data reported by the services. Should ACMA decide to proceed with this option, we would still recommend to then internally calculate mean instead of median averages. Requesting totals instead of averages would not impact a platform's ability to report against any of the metrics.

Suggestion of alternative metrics and changes to existing ones

Overall, we suggested changes to several metrics, primarily to improve the metrics' clarity and allow for consistent understanding across services. In a limited number of cases, metrics were combined to reduce the overall number of data points that services would have to provide. Finally, some metrics were recommended for removal, either because they are largely duplicates of other metrics or because they were identified as less relevant to their related outcome or not applicable to the service. A breakdown of the recommended changes can be found in the accompanying spreadsheet, as there were differences across platforms depending on the nature of the service and the platform's architecture. Additionally, Moonshot did not identify any additional metrics that could be incorporated in the framework which may be more suited to independent verification, as any metric related to MDI will inevitably face the same barriers explored above.

Differences between search engines and social media platforms

Overall, we found the metrics to be highly applicable to social media platforms, or platforms where users can generate their content more broadly. However, the metrics were significantly less applicable to the search engines evaluated for the project. This was due to a high volume of metrics focusing on "actors" or "accounts", which do not strictly apply to search engines. Although we recognise the importance of a consistent set of metrics for all services regardless of their structure or operations, we recommend considering how to adapt some of the metrics to be more applicable to a variety of platforms.

Case study

Moonshot conducted a case study examining an instance where MDI has occurred to test the metrics against a real-world scenario. For the case study, we examined the Australian Aboriginal Voice Campaign (AAVC) on Facebook, selected due to the significant volume of MDI narratives that were shared during the campaign. We focused on Facebook as the platform emerged as one of the most promising when assessing the applicability of the metrics, and also due to its partnership with the Australian-based fact-checking RMIT University FactLab, which has demonstrated that they have recognised the existence of MDI in relation to the campaign and have taken efforts to attempt to mitigate its spread.[35]

This case study is theoretical in nature and outlines potential methodologies for data collection and challenges that would be encountered. Moonshot was unable to collect data for any of the metrics and, consequently, the description that follows describes what could be done if Moonshot was provided with a different level of data access.

Baselines

The importance of establishing baselines when assessing how a platform deals with MDI is clear. However, collecting data for any of these baseline metrics within the context of this case study is complex. For example, Meta does not currently release the number of active Australian end-users (including children) in a given month. Establishing this figure independently would require the collection of all public posts on Facebook in a given month, which would then need to be filtered to only include posts made by Australian users and used to calculate the number of active Australian end-users. Regardless of the feasibility, the process would still result in a number that does not capture users who posted privately on their profile, making the figure not comparable to what the platform could provide. Online media monitoring companies such as Meltwater or DataReportal estimate that there are roughly 19.7 million monthly active Australian end-users of Facebook.[36] However, these are figures based on ad reach data published by Facebook itself.[37] Establishing the number of active Australian end-users that violated Facebook's policies in relation to MDI, the reach of this content, and the actions taken against it all face similar complications. Similarly, the number of reviewers employed by Facebook would only be available should the platform decide to divulge it.

The scale, nature, reach and influence of MDI

Measuring the scale, nature, reach and influence of MDI is vital to understanding the risks it poses on a given platform. A range of researchers have explored the MDI narratives that emerged in response to the campaign.[38] However, none of the investigations examined by Moonshot featured an indication of the total volume of MDI content in this context. Consequently, no data relating to the reach or influence of this content could be gleaned.

[35] RMITFactLab, Debunks, accessed 03 May 2024.

[36] Meltwater, 2024 Social Media Statistics for Australia, accessed 01 May 2024.

[37] DataReportal, Digital 2024: Australia, accessed 01 May 2024

[38] Insikt Group, Malign Narratives Oppose "the Voice" Ahead of Australia's Referendum, accessed 07 May 2024; Fivecast, HOW DISINFORMATION DICTATED THE VOICE REFERENDUM, accessed 07 May 2024; Amnesty International, SORTING FACT FROM FICTION IN THE VOICE TO PARLIAMENT REFERENDUM, accessed 07 May 2024.

To measure the total number of posts containing MDI narratives in relation to the AAVC, the following methodology would have to be implemented:

1. The creation of a database of terms and phrases associated with MDI in relation to the AAVC.
2. The creation of a database of manipulated media associated with MDI in relation to the AAVC.
3. The creation of a database of public spaces where data would be collected from.
4. The development of a methodology that collects MDI data from the public spaces identified using the two databases.
5. The collection of evidence of MDI and the subsequent enrichment and analysis of that data to understand its reach and impact.

On Facebook, due to the challenges outlined in the relevant section of this report, data collection would be mostly manual, although the process could be automated on platforms such as YouTube or X. Regardless of the platform where evidence of MDI is being collected from, the lack of access to private data would limit any third party's ability to gather a comprehensive dataset. Additionally, discrepancies in how individual platforms conceptualise and define MDI would lead to further complications in establishing a comparable dataset.

The identification of coordinated inauthentic behaviour (CIB) in relation to the AAVC could be accomplished through a similarly sampled approach. To establish the presence of CIB, Moonshot would additionally incorporate: temporal analysis, such as posting time, posting frequency and account creation dates; analysis of similarities in text and images posted by other accounts on the same or other platforms; and identification of links between the source and known propagators of disinformation. Once again, it is important to bear in mind that the lack of access to private data would limit any third party's ability to gather a comprehensive or comparable dataset.

Platform mitigation strategies

As previously discussed in this report, measuring the number of user reports or appeals in relation to MDI on a given platform is reliant on the platform providing that data itself. However, the methodology detailed above could be adapted to conduct limited "spot-checks" in relation to content removal. For example, implementing the methodology above would provide a dataset of posts containing MDI that are possibly in breach of Facebook's policies. The dataset could then be reviewed at different time intervals to establish if the posts have been removed or are still available on the platform. However, it would not be possible to verify why the content has been removed, how it was identified or assessed, or identify any content that was removed before it appeared on the platform itself.

Platform functions and policies

Gathering data for metrics connected to platforms' recommender systems or policies remains entirely reliant on platforms opting to divulge the information. Collecting data against these metrics is predominantly technically impossible. In the rare instances where it could be technically feasible, it could not be achieved without violating user privacy rights.

↘ Appendix 1: Literature review

Conceptualising misinformation and disinformation

There has been increasing interest in understanding the causes, prevalence and consequences of MDI. Despite this leading to a sometimes scattered research field, with terms often used interchangeably,[39] the key difference between misinformation and disinformation is commonly understood as whether the false and misleading information was intended to be misleading.

This conceptual focus on the motivation for producing or distributing MDI is captured in the Digital Industry Group Inc's (DIGI) definition. It defines disinformation as "*digital content that is verifiably false or misleading or deceptive*", is "*propagated amongst users of digital platforms via Inauthentic Behaviours*" and "*the dissemination of which is reasonably likely to cause Harm*". Misinformation is also defined as "*digital content (often legal) that is verifiably false or misleading or deceptive*", is "*propagated by users of digital platforms*", and "*the dissemination of which is reasonably likely (but may not be clearly intended to) cause Harm*".[40]

This is in line with definitions provided by the UK Government, which conceptualises disinformation as "*the deliberate creation and spreading of false and/or manipulated information that is intended to deceive and mislead people, either for the purposes of causing harm, or for political, personal or financial gain*"[41] and misinformation as the "*inadvertent spread of false information*." Similarly, the EU Commission defines disinformation as "*false or misleading content that is spread with an intention to deceive or secure economic or political gain, and which may cause public harm*",[42] whereas misinformation is defined as "*false or misleading content shared without harmful intent though the effects can be still harmful.*"

Current governmental approaches to addressing MDI

Approaches employed by Ofcom

Following the Royal Assent of the UK Online Safety Act 2023 on 26 October 2023, the Office of Communications (Ofcom) officially commenced its role as the UK's online safety regulator. The Act "*aims to increase user safety and to improve users' ability to keep themselves safe online. All regulated services must protect users from illegal content that reaches the criminal threshold*".[43] While the Act does not explicitly identify disinformation or misinformation as specific harms that need to be addressed by online services, Ofcom names several offences and provisions that are relevant to disinformation, such as:[44]

- **Terms of Service duty:** Ofcom is able to hold Category 1 services (the largest platforms) to account for any Terms of Service they may have in this area (i.e., removing misinformation or disinformation content that meets the thresholds set out in their own policies).

[39] Elena Broda & Jesper Strömbäck, Annals of the International Communication Association, Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review, accessed 19 April 2024.
[40] DIGI, Australian Code of Practice on Disinformation and Misinformation, accessed 13 May 2024.
[41] UK Government, Fact Sheet on the CDU and RRU, accessed 18 April 2024.
[42] European Commission, Tackling online disinformation, accessed 18 April 2024.
[43] House of Commons Library, Preventing misinformation and disinformation in online filter bubbles, accessed 14 May 2024.
[44] Ofcom, Letter from Dame Melanie Dawes to Parliamentarians, accessed 18 April 2024.

- **False communications offence:** the Act sets out a new false communication offence, whereby a person commits an offence if they send a message that conveys information they know to be false and which is intended to "*cause non-trivial psychological or physical harm to a likely audience*".
- **Transparency reporting:** Ofcom has powers to tailor transparency notices to individual services, and they can include a range of information as set out in the Act, including measures taken to comply with the Terms of Service duties. To date, no transparency reports have been publicly released.
- **Advisory committee:** Ofcom is required to establish and maintain an advisory committee on MDI, which will provide advice to Ofcom about issues such as how providers of regulated services should deal with MDI.

There has been debate among stakeholders as to the adequacy or necessity of the policies. Several organisations have criticised the Act's reliance upon a platform's Terms of Service, arguing that it could mean that "tech companies" in effect "*decide what is and what isn't legal*".[45]

Approaches employed by the European Union

The Digital Services Act establishes transparency duties, risk analysis requirements and independent auditing processes for Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) reaching at least 45 million users in the EU (representing 10% of the population). The DSA introduces a number of obligations to tackle the spread of misinformation and disinformation:[46]

- **Risk assessments:** VLOPs and VLOSEs will be required to perform risk assessments on various elements of their services, including risks stemming from their design, functioning or use. This includes disinformation campaigns. The assessment should consider how the services are used to disseminate or amplify misleading information. Based on the risk assessments, online platforms are obliged to implement risk mitigation measures.
- **Crisis response mechanisms:** VLOPs and VLOSEs need to have a crisis response mechanism, which includes measures to take when their platform is used for the spread of disinformation.
- **A code of practice:**[47] The DSA encourages platforms to sign up to the voluntary code of practice on disinformation.
- Public advertisement repository: VLOPs and VLOSEs are required to maintain a public advertisement repository to help researchers study emerging risks, including disinformation.
- **Transparency reporting:** The DSA obliges VLOPs and search engines to inform users about content moderation decisions they make. These decisions and their justifications have to be submitted to the EU Commission and are hosted in the DSA Transparency Database.

[45] Full Fact, The Online Safety Act and Misinformation: What you need to know, accessed 22 April 2024, and Open Right Group, ORG Warns of Threat to Privacy and Free Speech as Online Safety Bill is Passed, accessed 22 April 2024.

[46] European Commission, Questions and answers on the Digital Services Act, accessed 19 April 2024.

[47] European Commission, The 2022 Code of Practice on Disinformation, accessed 22 April 2024.

Current approaches to mitigating and safeguarding against MDI

Removals and restrictions

Perhaps the most obvious counter measure for MDI is the removal of content or accounts consistently spreading fake or misleading information. Companies like Meta have used machine learning to detect and remove Facebook accounts consistently amplifying MDI content by recognising account behavioural patterns.[48] Although some level of content removal is likely to remain a necessity, academics warn that placing the impetus of removal on platforms essentially makes these companies "*arbiters of truth*".[49]

This is further complicated by the fact that there is no shared definition of what constitutes MDI across platforms, and that platforms are likely to prefer a "light touch" approach to content moderation in order to preserve an appearance of political neutrality.[50] These factors mean that MDI removal policies are likely to be inconsistently enforced and defined.

The overall efficacy of removals as a strategy to counter MDI has also been questioned for failing to address the "root causes" of why this content spreads and is attractive to users.[51] Removals may also cause users to migrate to smaller, less regulated social media platforms, where MDI content is more easily accessible. This effect was observed following large-scale removals of QAnon content on mainstream platforms like Facebook and X, which enabled Telegram and Gab to attract new users and emerge as "central hubs" of QAnon conspiracy content.[52] Removals also have the potential to restrict legitimate speech and infringe on freedom of expression. For example, Kate Jones, CEO of the Digital Regulation Cooperation Forum, has cautioned that incentivising removals "*may give undue weight to vested interests in restricting comment*", and highlighted the lack of transparency in relation to the decision-making behind removals as posing risks to freedom of speech.[53]

Alternatives to removal include the deprioritisation or "shadow banning" of MDI content, where content is shown only to a reduced audience.[54] Platforms like Meta and TikTok may enforce variations of this policy, limiting the audience of unverifiable content or ensuring that content is not promoted on users' feeds.[55, 56]

However, this can, again, place the impetus of deciding what constitutes MDI on platforms - opening up avenues for abuse.[57] Other alternative measures include the prohibition of political campaigns that target small groups based on consumer data and demographics, which has historically enabled MDI content to be targeted towards very niche audiences.[58] However, definitions of what constitutes political content vary significantly between platforms, leading to these policies being inconsistently enforced.[59]

[48] Vasu et al, Nanyang Technological University Singapore, Fake News: National Security in the Post-Truth Era, accessed 18 April 2024.

[49] David Moschella, Information Technology & Innovation Foundation, We Shouldn't Ask Technologists To Be Arbiters of "Truth", accessed 18 April 2024.

[50] Dallas Flick, SMU Science and Technology Law Review, Combatting Fake News: Alternatives to Limiting Social Media Misinformation and Rehabilitating Quality Journalism, accessed 18 April 2024.

[51] Roger McNamee, Time, Social Media Platforms Claim Moderation Will Reduce Harassment, Disinformation and Conspiracies. It Won't, accessed 18 April 2024.

[52] Jordan Wildon and Marc-André Argentino, Global Network on Extremism & Technology, QAnon is not Dead: New Research into Telegram Shows the Movement is Alive and Well, accessed 18 April 2024.

[53] Kate Jones, Chatham House, Online Disinformation and Political Discourse: Applying a Human Rights Framework, accessed 18 April 2024.

[54] Emily Saltz and Claire Leibowicz, NiemanLab, Shadow bans, fact-checks, info hubs: The big guide to how platforms are handling misinformation in 2021, accessed 18 April 2024.

[55] TikTok, Combating harmful misinformation, accessed 19 April 2024.

[56] Meta, How fact-checking works, accessed 19 April 2024.

[57] Yen-Shao Chen and Tauhid Zaman, PLoS ONE, Shaping opinions in social networks with shadow banning, accessed 18 April 2024.

[58] Karen Kornbluh et al, The German Marshall Fund of the United States, Safeguarding Digital Democracy, accessed 18 April 2024.

[59] Paddy Leerssen et al, Internet Policy Review, Platform ad archives: promises and pitfalls, accessed 18 April 2024.

Fact-checking as an alternative to removals

Alongside content removals and restrictions, a number of platforms employ fact-checking as a measure to counter MDI. Fact-checking is a mitigation strategy with three main aims: first, the identification of false or misleading information; second, the verification of these claims; and third, the issuing of timely and effective corrections.[60]

Companies like X use crowdsourcing to place the impetus of fact-checking directly on the users, asking them to flag posts they feel are misleading and to provide context in "Community Notes". If enough users from a diversity of ideological perspectives rate notes as helpful, they are publicly shown on posts.[61] In practice, this policy enables MDI content to be publicly available, and potentially even viral, before it has been fact-checked. It also relies on users having high levels of digital literacy, and leaves the door open for users to 'fact-check' content using out-of-date or inaccurate information.[62] Meta attempts to mitigate bias and inaccuracy by outsourcing fact-checking to non-partisan third parties certified through the International Fact-Checking Network, and using AI to enable fact-checking to take place at scale.[63]

The use of AI in fact-checking remains controversial. Distinguishing between MDI content and "controversial facts" remains a sensitive matter, which is likely to continue to require human input.[64] Moreover, fact-checking is reactionary by nature, and applied to MDI content inconsistently.[65] The timing of fact-checking is also crucial, with posts often being evaluated only after they have been allowed to spread on social media.[66]

The effectiveness of fact-checking as a method to counter the impact of MDI is also contested. Fact-checking may be ineffective against narratively interesting MDI content, which viewers simply find more engaging.[67] In some instances, fact-checking also runs the risk of creating a 'boomerang' effect, by causing some users who feel "targeted" by fact-checking labels to react by sharing content at higher rates and thus amplifying MDI.[68] As with content removals, this may push communities to move to less regulated platforms. However, as advances in AI continue, opportunities for de-escalation and reimagining fact-checking may emerge. A recent study found that individuals' levels of belief in conspiracies can be significantly decreased by short facts-based conversations with an AI chatbot instructed to lessen belief in a particular conspiracy.[69] While these early findings are promising, it should be noted that AI is also a useful tool for creating and disseminating engaging MDI content at scale.[70]

[60] Lucy Graves, Reuters Institute and University of Oxford, Understanding the Promise and Limits of Automated Fact-Checking, accessed 17 April 2024.
[61] X, Community Notes: a collaborative way to add helpful context to posts and keep people better informed, accessed 19 April 2024.
[62] Madison Czopek, Poynter, Why Twitter's Community Notes feature mostly fails to combat misinformation, accessed 19 April 2024.
[63] Meta, How Meta's third-party fact-checking program works, accessed 19 April 2024.
[64] Ludovica Vecchio, Misinformation: A Threat to Digital Governance, accessed 17 April 2024.
[65] Karen Kornbluh et al, The German Marshall Fund of the United States, Safeguarding Digital Democracy, accessed 18 April 2024.
[66] DISARM Foundation, Companion Guide to the 2019 'Blue' workshop output, accessed 18 April 2024.
[67] Aleksandra Lazić and Iris Žeželj, Public Understanding of Science, A systematic review of narrative interventions: Lessons for countering anti-vaccination conspiracy theories and misinformation, accessed 18 April 2024.
[68] Sam Levin, The Guardian, Facebook promised to tackle fake news. But the evidence shows it's not working, accessed 18 April 2024.
[69] Costello et al, PsyArXiv Preprints, Durably reducing conspiracy beliefs through dialogues with AI, accessed 18 April 2024.
[70] Menz et al, British Medical Journal, Current safeguards, risk mitigation, and transparency measures of large language models against the generation of health disinformation: repeated cross sectional analysis, accessed 18 April 2024.

Digital literacy: from debunking to prebunking

In addition to tackling MDI at a content level, a number of countries have adopted a "whole-of-society" approach, focusing on fostering enhanced digital literacy and building resilience to MDI at a population level.[71] Some countries have incorporated media and digital literacy classes into their school curriculums, with the aim of equipping young people with the ability to identify and combat MDI.[72]

In line with a digital literacy perspective, "pre-bunking" campaigns have shown promise in lessening individuals' receptivity to believing MDI. Pre-bunking generally utilises inoculation theory, which posits that exposing individuals to weakened or fake versions of MDI can strengthen their resistance to genuine MDI in the future.[73, 74] Games like Roozenbeek et al's Bad News[75] or Moonshot's Gali Fakta[76] have been particularly successful in utilising inoculation theory to improve players' digital literacy skills. These games simulate a social media context and expose users to controlled versions of MDI in order to test and train their ability to recognise fake or misleading information. Notably, in Moonshot's testing, users were found likely to engage with gamified educational inoculation content for far longer than similar content hosted on static websites.[77]

[71] Department of the Prime Minister and Cabinet of New Zealand, Strengthening resilience to disinformation online, accessed 18 April 2024.

[72] Ludovica Vecchio, Invictus Corporation Ltd, Misinformation: A Threat to Digital Governance, accessed 17 April 2024.

[73] Josh Compton et al, Social and Personality Psychology Compass, Inoculation theory in the post-truth era: Extant findings and new frontiers for contested science, misinformation, and conspiracy theories, accessed 18 April 2024.

[74] Lewandowsky and van der Linden, European Review of Social Psychology, Countering Misinformation and Fake News Through Inoculation and Prebunking, accessed 18 April 2024.

[75] Roozenbeek et al, The Harvard Kennedy School MIsinformation Review, Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures, accessed 18 April 2024.

[76] Matthew Facciani, KISIP, Playing Gali Fakta "Inoculates" Indonesia Participants Against False Information, accessed 18 April 2024.

[77] Moonshot, Advancing Media Literacy in Indonesia, accessed 18 April 2024.

↘ Appendix 2: Methodology

Overview

In March 2024, Moonshot was contracted by the Australian Communications and Media Authority (ACMA) to assist in their oversight of the Australian Code of Practice on Disinformation and Misinformation (the Code). The Code aims to mitigate the risks posed by MDI, and establishes clear Objectives and Outcomes to reduce vulnerability to harm from false or misleading information among the Australian public. Crucial to the success of the Code is the development of standard metrics to measure how platforms are working to achieve its Objectives.

As part of this oversight role, ACMA has developed a consistent set of measures and metrics against which digital platforms can report their efforts to meet the Code's objectives. A thorough evaluation and testing of the metrics is essential to the success of the Code, while ensuring respect for freedom of expression and other important rights.

The aim of this project is to assist ACMA in finalising the metrics developed to assess digital platforms' compliance with the Code. The project was originally divided into two parts:

- **Part 1** - Providing independent expert advice to validate the 64 metrics identified by ACMA in the draft framework.
- **Part 2** - Providing data (where available) against the finalised set of metrics to enable comparisons with future reports provided by signatories.

While implementing Part 1 of the project, Moonshot identified significant barriers to the collection of a comprehensive dataset. As a result of this, the nature of Part 2 was changed, and the data collection exercise was replaced with the provision of independent expert advice to validate an additional 4 metrics identified by ACMA.

This **final report** provides an overview of the results of Part 1 and 2 of the project, and is submitted alongside an accompanying spreadsheet that details the assessment conducted on each platform. It incorporates feedback provided by ACMA on the **draft report**. As part of the validation exercise, Moonshot assessed the 68 metrics in relation to 15 platforms, covering both Signatories and non-Signatories to the Code. The assessment of each platform was captured in a validation framework, which provides a guide to evaluate the following core questions for each metric:

- Is the digital platform likely to be able to provide data against each metric?
- Is the metric likely to provide Australia-specific insights into the success of the measures?
- Would an alternative metric yield better Australia-specific insights into the success of the measures?
- Could the information reported by signatories against the metrics be independently verified?
- Are there any other relevant issues and/or barriers to implementation of the framework?

The approach to the validation exercise mixed analysis of platform infrastructure with reviews of transparency reports, platform policy documents and academic literature. This review was also supplemented by a broader assessment of the metrics to identify any relevant barriers to the implementation of the framework.

Platform selection

Moonshot was provided with a list of services and associated platforms of interest to ACMA for this project. These included:

Current Signatories to the Code

- Adobe: All Adobe Products
- Apple: Apple News
- Google: Google Search, YouTube, Google News, Google Ads and Google AdSense
- Meta: Facebook, Instagram and Threads
- Microsoft: Bing Search, LinkedIn, Microsoft Advertising and Microsoft Start
- Redbubble: Redbubble
- TikTok: TikTok
- Twitch: Twitch

Additional platforms

- X Corp: X (formerly Twitter)
- Snap: Snapchat
- Advance Publications: Reddit
- Spotify: Spotify

These services and associated platforms were assessed according to: relevance of the services offered (both in terms of global and Australian user-base); known presence and scale of MDI; presence of an API that could enable data collection; and ACMA's priorities. As a result of this process, 15 platforms were selected for the project: Facebook; Instagram; Threads; Google Search; YouTube; Google News; Google Ads; Bing Search; LinkedIn; TikTok; X (formerly Twitter); Reddit; Spotify; Twitch; and Snapchat. The remaining six platforms were deemed out of scope. Google AdSense, Apple News, all Adobe products, Microsoft Advertising, Microsoft Start and RedBubble were all found to either feature little evidence of MDI or have no identifiable API to enable data collection.

Development of the framework

The framework used for the validation exercise was developed to establish the following in relation to the five key research questions:

Is the digital platform likely to be able to provide data against each metric?

- Applicability: Determining if the metric is applicable to the platform's operations and services.
- Relevance: Assessing if the metric is relevant to the associated measure or outcome.
- Availability: Examining if the platform already collects and provides the required data.
- Feasibility: Analysing if it is feasible for the platform to gather the data.
- Compliance: Ensuring that platforms can provide the data requested while also remaining compliant with their own policies regarding user privacy and intellectual property, as well as wider regulations on freedom of expression and data protection.

Is the metric likely to provide Australia-specific insights into the success of the measures?

- Availability of information on user location: Determining if the platform provides data on the location of users - whether disclosed by users or identified by the platform.
- Disaggregation at geographic level: Assessing if metrics can be broken down to the geographic level, allowing for the analysis of MDI trends specific to Australian regions.

Would an alternative metric yield better Australia-specific insights?

- Assessing each metric against the related outcomes of Objectives 1 and 3.
- Outlining if the metric can yield Australia-specific insights and proposing changes if not.

Could the information reported by signatories against the metrics be independently verified?

- Availability of API: Assessing whether platforms have an Application Programming Interface (API) with endpoints capable of collecting the required data by third parties.
- Privacy compliance: Ensuring that data collection methods for publicly-available information by third parties could be achieved without violating user privacy rights.
- Sources of independent verification: Investigating whether proprietary technology or off-the-shelf software can be developed/deployed to gather data on the metrics.
- Methodological robustness: Considering the robustness and consistency of the methodology required for independent verification, including the feasibility of collecting MDI at scale in a way that would make the independently gathered data comparable to data provided by the platforms.
- Means of independent verification: Identifying the specific means and methodologies for independently verifying reported information.

Are there any other relevant issues and/or barriers to implementation of the framework?

- Establishing if the assessments of any of the other questions highlights any significant barriers or issues with the implementation of the framework.

Validation of the metrics

The approach to the validation exercise incorporated an assessment and review of:

- Transparency reports (or equivalents thereof) published by the platforms.
- Platforms' policies, architecture and operating systems.
- Data shared with other regulators, particularly geographically disaggregated data.
- API documentation and available third party tools for independent data gathering.
- Relevant Australian freedom of expression and data privacy legislations to ensure that providing data against the metrics would not risk endangering the protection of crucial rights.

The assessment conducted for each platform was summarised in the accompanying spreadsheet. All coding conducted by the analysts was quality assured by the project manager and project oversight. The metric-specific review was also supplemented by a broader assessment to identify any relevant barriers to the implementation of the framework. The results of the assessment have been included in this report.

Analysis and reporting

Once the validation of the metrics was complete, Moonshot assessed the results for each platform, as well as the results of the validation as a whole, to provide ACMA with platform-specific profiles and recommendations. In order to guarantee reliability and precision in our analysis and reporting, we implemented a comprehensive Quality Assurance (QA) process. This process encompasses all facets of our research. By adopting this QA process, we apply the principles of accuracy, objectivity and a strictly non-political approach. Through evaluation and scrutiny, we are able to maintain the validity of our research, ensuring accurate and consistent analysis, and providing a solid foundation for reliable insights and informed decision-making.

Case study

Part one

Moonshot conducted a case study examining an instance where mis or disinformation has occurred to test the metrics against real-world scenarios. For the case study, we examined the Australian Aboriginal Voice Campaign (AAVC) on Facebook, selected due to the significant volume of MDI narratives that emerged during the campaign. We decided to focus on Facebook as the platform was one of the most promising when assessing the applicability of the metrics, and also due to its partnership with the Australian-based fact-checking RMIT University FactLab, which has demonstrated that they have recognised the existence of MDI in relation to the campaign and have taken efforts to attempt to mitigate its spread.[78]

As part of the case study, the 64 metrics were divided into four categories: metrics related to establishing a baseline; metrics related to measuring the scale, nature, reach and influence of MDI; metrics related to platform functions and policies; and metrics related to platform mitigation strategies. The categorisation of each metric can be found in the spreadsheet submitted alongside this report. Dividing the metrics into these categories allowed for an overarching analysis of their applicability, as well as how data against metrics could be collected and the barriers to this collection.

The research carried out as part of this case study incorporated open source investigation, review of relevant academic and journalistic outputs, and the development of theoretical data collection methodologies. The case study was theoretical in nature and outlines potential methodologies for data collection and challenges that would be encountered. Moonshot was unable to collect data for any of the metrics and, consequently, the case study describes what could be done if Moonshot was provided with a different level of data access.

[78] RMITFactLab, Debunks, available online, accessed 03 May 2024.

Part two

During the validation exercise, Moonshot identified a limited number of instances where some data could be gathered. The metrics that were originally selected for Part 2 were:

- Metric 54 - If not proactively removed, what was the number of Australian inauthentic users/accounts that appeared on the service.
- Metric 57 - Median number of posts from Australian inauthentic users if they appeared on the service.
- Metric 61 - Number of CIB operations identified by the signatory emerging out of Australia.
- Metric 62 - Number of CIB operations identified by the signatory targeting Australia or Australians.

Moonshot selected YouTube, Tiktok, Facebook, Instagram, Reddit and X as suitable platforms for the data collection exercise, primarily due to levels of data access. To provide data against these four metrics, the following approach was devised:

Selection of MDI narratives for targeted data collection

To collect data against the metrics, Moonshot identified a number of MDI narratives that have been designated as such by independent fact-checkers. This step was necessary to establish some parameters for data collection, as each narrative centres around a coherent subject that can be captured through the use of keywords. The MDI narratives selected for data collection were:

- The Australian Aboriginal Voice Campaign (AAVC).
- The targeting of Australian politicians through state-sponsored disinformation.
- Australian-specific COVID-19 and broader anti-vaccine conspiracies.
- State-sponsored disinformation in relation to Russia's invasion of Ukraine.

Database development and identification of spaces for data collection

To collect data across the six platforms, Moonshot started to enhance existing databases of keywords relating to each of the four MDI narratives, with a focus on tailoring them to Australia. Moonshot also began to identify Australian public spaces across the six platforms where posts containing these keywords could be collected from. This phase was underway when a decision was made to discontinue Part 2 of the project in its original form.

Data collection

Concurrently, Moonshot also began the data collection set up process. Proprietary technology that relies on API access was used for X, Reddit, YouTube, and Instagram. For Facebook, the set up included manual data collection due to a lack of available API. A review of TikTok's API documentation identified that the platform had altered their Terms of Service on 14 May 2024, meaning that coverage of that platform would also need to rely on manual collection. Given the timeframe of the project, data was due to be collected for March and April 2024.

Identification of CIB

If the data collection had been implemented, Moonshot would have then assessed the dataset of posts containing MDI narratives for evidence of CIB. Establishing the presence of CIB requires a mixed-methods approach that involves reviewing posts and user accounts for a wide range of indicative behaviours, including:

- Duplicates and patterns in the post text, including hashtag sequencing.
- The times at which users post and the presence of similar behaviours across multiple accounts at the same time.
- The generation of high volumes of content with no user engagement.
- The repurposing of accounts to generate MDI.
- High volumes of posts linking to established CIB actors and spaces.
- High levels of engagement with posts from accounts with low follower counts.

Collection of data in relation to each metric

Users assessed as demonstrating evidence of CIB would have subsequently been reviewed to establish: the number of accounts that were demonstrably based in Australia; the overall number of CIB operations, originating in or targeting Australia; and the number of posts they published.