APPENDIX 2

Twitter: Australian Code of Practice on Disinformation and Misinformation Initial Report

Executive Summary

Twitter's mission is to serve the public conversation. Transparency is fundamental to our work in achieving that mission. This initial report into our commitments and progress under the Australian Code of Practice on Disinformation and Misinformation (the Code) provides an overview of Twitter's efforts to protect the public conversation and uphold the integrity of our service.

We are committed to providing meaningful transparency reporting to the public. We do this both in partnership with government, academics & civil society, under schemes like the Code, and through existing, proactive self reporting initiatives. Under this code, Twitter has made meaningful commitments and progress on all mandatory objectives and applicable opt-ins. Encouragingly, many measures outlined under the Code were already underway through Twitter's proactive policy enforcement and reporting measures, including the Twitter Transparency Report, the Twitter Transparency Center, and our state-backed information operation disclosures.

Key insights from the report

- Overview of ongoing work to protect against platform manipulation and enhance authenticity
- Efforts to disrupt global state-backed information operations to preserve integrity in public conversation
- Trends concerning the enforcement of the Twitter Rules

- Explanation of the range of measures and enforcement options available

As the contours of online behaviours evolve, we continue to iterate on and strengthen our approach to protecting the public conversation on Twitter and our place in the online information ecosystem. We are moving with urgency, purpose, and commitment, as we develop and enforce a range of policy, procedural, and product changes. We trust this report will prove useful in further understanding the challenges presented in the information ecosystem and our work to address these.

Background

Business and Content Context

We share the Australian Government's desire to promote a healthy digital information ecosystem, and Twitter remains focused on protecting and empowering users to participate in the public conversation every day. Twitter is an open service that's home to a world of diverse people, perspectives, ideas and information. We're committed to protecting the health of the public conversation — and we take that commitment seriously.

Twitter is an inherently open, public, real time service. In addition to our suite of advertising products, Twitter's service is primarily composed of user-generated content. In line with our commitments to preserving freedom of expression and the Open Internet, we believe it is vital to strike the right balance between taking proactive steps to protect from harm with human rights and other vital interests, including freedom of expression, privacy, and ensuring we do not act as the sole arbiter of truth.

The Twitter Rules[1] make clear that users are responsible for the content they post and that advertisers must adhere to specific quality guidelines on our service. There are a wide-ranging set of rules that Twitter enforces when content is posted that infringes those rules, which will be explained in more detail throughout this report.

Approach to Disinformation and Misinformation

Any attempts to undermine the integrity of our service is antithetical to our purpose and undermines the core tenets of an open Internet and freedom of expression, the value upon which our company is based. We believe we have a responsibility to protect the integrity of those

[1] Help.twitter.com. Rules and policies. [online] Available at: <https://help.twitter.com/en/rules-and-policiesl> [Accessed April 2021].

conversations from interference and manipulation. We also understand that genuine, open collaboration between industry, government, as well as meaningful involvement of academic experts and civil society organizations is required to address these issues.

As such, we continue to move with urgency, purpose, and commitment, as we develop and enforce a range of policy, procedural, and product changes to combat these complex challenges. We adopt a range of measures to prevent actors with bad intent from creating or maintaining accounts, compromising the accounts of others, or artificially boosting harmful content. This helps us protect the safety, security, and credibility of Twitter accounts and the Twitter service in the context of both organic and promoted content.

Twitter has a comprehensive set of policies addressing a wide range of behaviors that are intended to manipulate the public conversation. These behavioral rules, captured in our Platform Manipulation and Spam Policy, apply across content types, and serve as the basis for our enforcements against all forms of disinformation and coordinated manipulation. Additionally, Twitter's misinformation policies are focused on identifying content that is demonstrably false or misleading and may lead to significant risk of harm. In line with these priorities, our focus is on misinformation related to COVID-19[2], synthetic and manipulated media (SAMM)[3], and civic integrity[4]. Twitter leverages proprietary tools to enforce these policies.

Twitter's approach to enforcement of misinformation and disinformation has significantly evolved over the course of 2020 and 2021. In March 2020, Twitter launched its COVID-19 misleading information policy. Twitter initially enforced this policy through removal of violative content. In May 2020, Twitter expanded its enforcement options to add labels as an option for COVID misinformation[5]. In December 2020, Twitter launched removals of content with, specifically, vaccine misinformation[6], and in March 2021, added labels as an enforcement option for COVID vaccine misinformation[7].

Twitter has also expanded its enforcement options for violations of the civic integrity policy. In September 2020, Twitter added labels for Civic Integrity violations, in addition to content removal. And in January 2021, Twitter adopted a strike policy for Civic Integrity to address repeat offenders.

Globally, Twitter tracks and reports on the reports of violative content and actions taken in the Twitter Transparency Report found in at the Twitter Transparency Center. It should be noted, that while Twitter does not report numbers for enforcement of Twitter's Spam or Platform Manipulation

[2] Help.twitter.com. COVID-19 misleading information policy. [online] Available at: <https://help.twitter.com/en/rules-and-policies/medical-misinformation-policy> [Accessed April 2021].
[3] Help.twitter.com. Synthetic and manipulated media policy. [online] Available at: <https://help.twitter.com/en/rules-and-policies/manipulated-media> [Accessed April 2021].
[4] Help.twitter.com. Civic integrity policy. [online] Available at: <https://help.twitter.com/en/rules-and-policies/election-integrity-policy> [Accessed April 2021].
[5] Blog.twitter.com. Updating our approach to misleading information. [online] Available at: <https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html> [Accessed April 2021].
[6] Blog.twitter.com. COVID-19: Our approach to misleading vaccine information. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2020/covid19-vaccine.html> [Accessed April 2021].
[7] Blog.twitter.com. Updates to our work on COVID-19 vaccine misinformation. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation.html> [Accessed April 2021].

policies, enforcement of these policies are a key cornerstone of Twitter's efforts to combat the spread of misinformation and disinformation. Twitter challenges approximately 5 million accounts per week to prevent these violations.

The next Twitter Transparency Report is scheduled to be published in the second half of 2021 which will cover the period July to December 2020. At the time of this reporting, the following are the approximate global numbers available for the aforementioned time period:

- **All Twitter Rules** (see the Rules Enforcement page for a list of rules)
  - 3.5 million accounts were actioned[8] for violations of the Twitter Rules.
  - 1 million accounts were suspended for violations of the Twitter Rules.
  - 4.5 million pieces of content were removed[9] for violations of the Twitter Rules.
- **COVID-19 misleading information**
  - 3,400 accounts were actioned for violations of the COVID-19 misleading information policy.
  - 600 accounts were suspended for violations of the COVID-19 misleading information policy.
  - 3,900 pieces of content were removed for violations of the COVID-19 misleading information policy.
- **Civic integrity**
  - 6,500 accounts were actioned for violations of the civic integrity policy.
  - More than 50 accounts were suspended for violations of the civic integrity policy.
  - 8,100 pieces of content were removed for violations of the civic integrity policy.

For purposes of this report, we have also compiled the Australia-specific approximate data[10] for July to December 2020.

- **All Twitter Rules** (see the Rules Enforcement page for a list of rules)
  - 37,000 AU accounts were actioned for violations of the Twitter Rules.
  - 7,200 AU accounts were suspended for violations of the Twitter Rules.
  - 47,000 pieces of content authored by AU accounts were removed for violations of the Twitter Rules.
- **COVID-19 misleading information**
  - More than 50 AU accounts were actioned for violations of the COVID-19 misleading information policy.
  - Less than 10 AU accounts were suspended for violations of the COVID-19 misleading information policy.
  - More than 50 pieces of content authored by AU accounts were removed for violations of the COVID-19 misleading information policy.
- **Civic integrity**

[8] "Accounts actioned" reflects the number of unique accounts that were suspended or had some content removed for violating the Twitter Rules.This does not include labels applied.
[9] Content removed includes Tweets and profiles/profile components (avatar, banner, bio, etc)
[10] AU accounts are identified using the account's country designation. The country designation is assigned automatically at sign up, but also can be manually modified by the user.

- More than 40 AU accounts were actioned for violations of the civic integrity policy.
- No (0) AU accounts were suspended for violations of the civic integrity policy.
- About 70 pieces of content authored by AU accounts were removed for violations of the civic integrity policy.

Labelling for misinformation was a new enforcement remedy added in 2020 and 2021. In the last 90 days (28 Jan 2021 - 28 Apr 2021), approximately 25,000 accounts had at least 1 Tweet labeled for misinformation. This metric reflects accounts across all countries.

Approach to monitoring performance

Meaningful transparency between companies, regulators, civil society, and the general public is fundamental to the work we do at Twitter. This transparency is a key tenet of our efforts to preserve and protect the Open Internet. In line with this philosophy, for the past eight years our biannual Twitter Transparency Report has highlighted trends in requests made to Twitter from around the globe.

We believe the open exchange of information can have a positive global impact and through our efforts to provide meaningful transparency, we endeavour to earn public trust, and enable accountability. Recognising that the public as well as policy makers and regulators want to be better informed of our enforcement processes, we launched the new Twitter Transparency Centre in 2020 to make our data easier to understand and analyse for those who access our Transparency Reports.[11] Over time, we have significantly expanded the information we disclose in our reports, adding metrics on platform manipulation, Twitter Rules enforcement, and state-backed information operations.

In line with our long-term commitment to these principles of transparency, and to improving public understanding of inauthentic influence and manipulation campaigns, Twitter also publishes public archives of Tweets and media that we believe resulted from state-backed information operations.[12] We proactively report on these campaigns on an ongoing basis, and have made these public to foster accountability, enable research and monitoring, and empower people through access to information.

Data about trends

In addition to the global and local data shared above, we are continuing to monitor and analyse country and region-specific narratives that have emerged in misleading information, including country-specific information and narratives about COVID-19 outbreaks or vaccine approvals. This supports our work to adapt our policies and enforcement procedures and keep pace with these developments. For example, the differences in approved and experimental vaccines across different regions requires us to develop country-specific enforcement procedures. We will continue to share our progress on this evolving work with government, academic, and civil society partners, as well as the public.

[11]Blog.twitter.com. 2021. Insights from the 17th Twitter Transparency Report. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html> [Accessed April 2021].
[12]Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en/reports/information-operations.html> [Accessed 12 February 2021].

To date, we haven't generally identified specific, large-scale, targeted information operations originating outside of Australia and targeting people and conversations within Australia. While some of our previous disinformation removals and disclosures - such as activity associated with the Russian Internet Research Agency - touched on some themes relevant to Australia,[13] these generally haven't constituted a high-prevalence matter.

Future Initiatives

Our approach to protecting against platform manipulation and inauthentic behaviours continually evolves. As we continue to address behaviours that have the potential to undermine the public conversation, we remain committed to providing meaningful transparency through public reporting. Reporting under the Code, as well as the Twitter Transparency Report and our state-backed information operation disclosures, will continue to highlight trends and new behaviours as we observe them, as well as future measures we introduce to address these and increasing public understanding of misinformation and disinformation.

Objective 1: Safeguards against Disinformation and Misinformation

Outcome 1a:

Signatories. contribute to reducing the risk of Harms that may arise from the propagation of disinformation and misinformation on digital platforms by adopting a range of scalable measures.

Signatories will describe specific actions they have taken to meet this commitment, including any relevant quantitative or qualitative data (including case study examples) about the outcomes of these measures in the Australian context.

As noted, Twitter addresses misinformation and disinformation through a range of policies (discussed in detail in the next section), enforcement actions and product solutions. Under our policies, there are a number of enforcement actions available, including removing or labeling content with misinformation, and locking or suspending accounts that spread misinformation. Proportional to the severity of the policy violation, we may take any, or multiple, of the following actions:

- Remove the content
- Apply a label and/or warning message to the Tweet
- Show a warning to people before they share or like the Tweet;

[13] Sear, T, Jensen, M. 2018, UNSW: Russian trolls targeted Australian voters on Twitter, [online] Available at: <https://newsroom.unsw.edu.au/news/science-tech/russian-trolls-targeted-australian-voters-twitter>

- Reduce the visibility of the Tweet on Twitter and/or prevent it from being recommended;
- Turn off likes, replies, and Retweets; and/or
- Provide a link to additional explanations or clarifications, such as in a curated landing page or relevant Twitter policies.

Twitter is committed to sharing these policies in accessible, plain language. In 2019, we updated the public on our efforts to make our rules easy to understand and recognise the importance of having policies that can be easily interpreted by the general public. This is vital for complex topic areas like defining and identifying platform manipulation, misinformation, and disinformation, which are being continually evaluated and further refined by subject matter experts. Our policies, and corresponding transparency reports, outline a variety of different behaviours that might fall under these categories.

In addition to our policies, which enable us to remove or provide further context on potentially misleading information, we have adopted a number of strategies to signal authenticity, accuracy and authoritative information on the service. These include Events pages, Moments, search prompts, and verification, as well as labeling government and state-sponsored media accounts so that users can evaluate the source of information. We provide this context both through Twitter's Curation team and through trusted partnerships with government, experts, and civil society groups. Our Curators don't act as reporters or creators of original work, they organise and present content that already exists on Twitter in Moments, explanatory content on Trends, in lists and more. They are guided by publicly available principles, including impartiality and accuracy.[14]

For example, in addressing COVID-19 misleading information, our work has focused on removing demonstrably false or potentially misleading content that has the highest risk of causing harm, as well as surfacing credible content from authoritative sources. The latter has been driven through profile verification, our curated COVID-19 tab in Explore, dedicated COVID-19 Event pages, and search prompts for COVID-19 and vaccinations, in partnership like the Department of Health.

Throughout the pandemic, Twitter has continually enhanced its internal and external efforts to build partnerships, protect the public conversation, help people find authoritative health information, and contribute pro bono advertising support to ensure people are getting the right message, from the right source. This work is not limited to organisations focused on fact-checking. We have open lines of communication with critical stakeholders, globally and in Australia, including the US CDC, the World Health Organisation (WHO), Australian State Federal and State health authorities, and public health organisations such as the Australian Medical Association, NACCHO, and the National Health and Medical Research Council, to help amplify their messages, ensure they can troubleshoot account issues, get their experts verified, and seek strategic counsel as they use the power of Twitter to promote healthy outcomes, and mitigate harm.

The launch of our evergreen COVID-19 and immunisation search prompts, built on our existing work and products to guard against the artificial amplification of non-credible content about the safety and effectiveness of vaccines. Thus, when people in Australia search certain keywords

[14] Help.twitter.com, Twitter Moments guidelines and principles, [online], Available at: <https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles>. [Accessed April 2021]

related to COVID-19 or vaccination on Twitter, a prompt appears at the top of the search bar, alerting them to credible sources where they can find the most up to date information from the Australian Federal Department of Health. In April 2021, a timeline prompt also connected people in Twitter in Australia with the COVID-19 vaccine landing page from the Federal Department of Health, and a curated Moment, made by our Curation team, that covered a wide range of topics such as: vaccine safety, effectiveness, vaccines available, distribution plans, how to stay safe before/after vaccinations and more.

As referred above, we also believe Twitter has a responsibility to protect the integrity of the public conversation including via timely disclosure of information about attempts to manipulate Twitter to influence elections and other civic conversations by foreign or domestic state linked entities. In October 2018, we launched our first comprehensive archive of state-backed information operations we have seen on Twitter.[15] We continue to release additional data disclosures as we identify foreign or domestic state-linked activities.

Outcome 1b:

Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

Signatories will include links to published policies and procedures, guidelines and information relating to the prohibition and/or management of user behaviours that may propagate Disinformation and Misinformation via their Services or Products.

Twitter has a range of existing, publicly available, policies that outline our definitions, enforcement options and reporting guidelines for inauthentic behaviour, platform manipulation and misinformation. Our approach to these complex issues is never static; we continually evolve our policies to address new challenges and online behaviours, engaging experts and the public in consultation along the way. The key policies relevant to the operation of the Code are outlined and linked below. Advertising policies will be discussed separately under 'Objective 2'.

Platform manipulation and spam policy

Users may not use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter.

[15] Blog.twitter.com. 2018. Enabling further research of information operations on Twitter. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html> [Accessed April 2021]

In line with Twitter's mission to serve the public conversation, we have engaged in long-term, proactive work to help people find reliable information, and express themselves freely and safely on our service. Under our platform manipulation and spam policy, we do not allow spam or other types of platform manipulation. We define platform manipulation as using Twitter to engage in bulk, aggressive, or deceptive activity that misleads others and/or disrupts their experience.

Platform manipulation can take many forms and our rules are intended to address a wide range of prohibited behavior, including:

- inauthentic engagements, that attempt to make accounts or content appear more popular or active than they are
- coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting
- coordinated harmful activity that encourages or promotes behavior which violates the Twitter Rules

NB: While this policy prohibits fake accounts, it does not apply to those using Twitter pseudonymously or as a parody, commentary, or fan account. The ability to speak anonymously, or to use a pseudonym, has been a core tenet of our service since its inception and we believe that the right to remain anonymous online is essential to preserving free expression.

The consequences for violating this policy depend on the severity of the violation as well as any previous history of violations. Our action is also informed by the type of inauthentic activity that we have identified. The actions we take may include the following:

Anti-spam challenges

When we detect suspicious levels of activity, accounts may be locked and prompted to provide additional information (e.g. a phone number) or to solve a reCAPTCHA.

Denylisting URLs

We denylist or provide warnings about URLs we believe to be unsafe. Read more about unsafe links, including how to appeal if we've falsely identified your URL as unsafe.

Tweet deletion and temporary account locks

- If the platform manipulation or spam offense is an isolated incident or first offense, we may take a number of actions ranging from requiring deletion of one of more Tweets to temporarily locking account(s). Any subsequent platform manipulation offenses will result in permanent suspension.
- In the case of a violation centering around the use of multiple accounts, users may be asked to choose one account to keep. The remaining accounts will be permanently suspended.
- If we believe a user may be in violation of our fake accounts policy, we may require that they provide government-issued identification (such as a driver's license or passport) in order to reinstate their account.

Permanent suspension

For severe violations, accounts will be permanently suspended at first detection. Examples of severe violations include:

- operating accounts where the majority of behavior is in violation of the policies described above;
- using any of the tactics described on this page to undermine the integrity of elections;
- buying/selling accounts;
- creating accounts to replace or mimic a suspended account; and
- operating accounts that Twitter is able to reliably attribute to entities known to violate the Twitter Rules.

People on Twitter who believe their account was locked or suspended in error, can submit an appeal.

COVID-19 misleading information policy

Users may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm.

Even as scientific understanding of the COVID-19 pandemic continues to develop, we've observed the emergence of persistent conspiracy theories, alarmist rhetoric unfounded in research or credible reporting, and a wide range of unsubstantiated rumors, which left uncontextualized can prevent the public from making informed decisions regarding their health, and puts individuals, families and communities at risk.

In this context, content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter. This includes sharing content that may mislead people about the nature of the COVID-19 virus; the efficacy and/or safety of preventative measures, treatments, or other precautions to mitigate or treat the disease; official regulations, restrictions, or exemptions pertaining to health advisories; or the prevalence of the virus or risk of infection or death associated with COVID-19. In addition, we may label Tweets which share misleading information about COVID-19 to reduce their spread and provide additional context.

The consequences for violating our COVID-19 misleading information policy depends on the severity and type of the violation and the account's history of previous violations. In instances where accounts repeatedly violate this policy, we will use a strike system to determine if further enforcement actions should be applied. We believe this system further helps to reduce the spread of potentially harmful and misleading information on Twitter, particularly for high-severity violations of our rules.

Content removal
For high-severity violations of this policy, including (1) misleading information related to the nature or treatment of the COVID-19 virus and (2) pandemic or COVID-19 vaccines that invoke a deliberate conspiracy by malicious and/or powerful forces, we will require you to remove this content. We will also temporarily lock you out of your account before you can Tweet again. Tweet deletions accrue 2 strikes.

Labeling
In circumstances where we do not remove content which violates this policy, we may provide additional context on Tweets sharing the content where they appear on Twitter. This means we may:

- Apply a label and/or warning message to the Tweet
- Show a warning to people before they share or like the Tweet;
- Reduce the visibility of the Tweet on Twitter and/or prevent it from being recommended;
- Turn off likes, replies, and Retweets; and/or
- Provide a link to additional explanations or clarifications, such as in a curated landing page or relevant Twitter policies.

In most cases, we will take all of the above actions on Tweets we label. We prioritise producing Twitter Moments in cases where misleading content on Twitter is gaining significant attention and has caused public confusion on our service. Tweets that are labeled and determined to be harmful will accrue 1 strike.

If we determine that an account is dedicated to Tweeting or promoting a particular misleading narrative (or set of narratives) about COVID-19, this would also be grounds for suspension.

Permanent suspension
For severe or repeated violations of this policy, accounts will be permanently suspended.
Repeated violations of this policy are enforced against on the basis of the number of strikes an account has accrued for violations of this policy:

- 1 strike: No account-level action
- 2 strikes: 12-hour account lock
- 3 strikes: 12-hour account lock
- 4 strikes: 7-day account lock
- 5 or more strikes: Permanent suspension

People on Twitter who believe their account was locked or suspended in error, can submit an appeal.

Civic integrity policy

Users may not use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process. In addition, we may label and reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context.

We also prohibit attempts to use our services to manipulate or disrupt civic processes, including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process. In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion on our service, we may label the Tweets to give additional context.
The consequences for violating our civic integrity policy depends on the severity and type of the violation and the accounts' history of previous violations. In instances where accounts repeatedly violate this policy, we will use a strike system to determine if further enforcement actions should be applied. We believe this system further helps to reduce the spread of potentially harmful and misleading information on Twitter, particularly for high-severity violations of our rules.

The actions we take may include the following:

Content removal
For high-severity violations of this policy, including (1) misleading information about how to participate, and (2) suppression and intimidation, we will require you to remove this content. We will also temporarily lock you out of your account before you can Tweet again. Tweet deletions accrue 2 strikes.

Profile modifications
If you violate this policy within your profile information (e.g., your bio), we will require you to remove this content. We will also temporarily lock you out of your account before you can Tweet again. If you violate this policy again after your first warning, your account will be permanently suspended.

Labeling
In circumstances where we do not remove content which violates this policy, we may provide additional context on Tweets sharing the content where they appear on Twitter. This means we may:

- Apply a label and/or warning message to the content where it appears in the Twitter product;
- Show a warning to people before they share or like the content;
- Turn off people's ability to reply, Retweet, or like the Tweet;
- Reduce the visibility of the content on Twitter and/or prevent it from being recommended;

Synthetic and manipulated media policy
Users may not deceptively promote synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context.

Labeling and removal
In most cases, if we have reason to believe that media shared in a Tweet have been significantly and deceptively altered or fabricated, we will provide additional context on Tweets sharing the media where they appear on Twitter. This means we may:

- Apply a label to the content where it appears in the Twitter product;
- Show a warning to people before they share or like the content;
- Reduce the visibility of the content on Twitter and/or prevent it from being recommended;
- Provide a link to additional explanations or clarifications, such as in a Twitter Moment or landing page; and/or
- Turn off likes, replies, and Retweets.

In most cases, we will take all of the above actions on Tweets we label. Media that meets all three of the following criteria: synthetic or manipulated, shared in a deceptive manner, and is likely to cause harm—may not be shared on Twitter and are subject to removal. Accounts engaging in repeated or severe violations of this policy may be permanently suspended.

Outcome 1c:

Users can report content to Signatories that violates their policies under 5.10 through publicly available and accessible reporting tools.
Signatories will include links to published policies, procedures guidelines that will enable users to report the types of behaviours and content that may propagate Disinformation and Misinformation via their platforms.

Under our **platform manipulation and spam policy**, anyone on Twitter can report accounts or Tweets that violate the criterion defined under the policy or that display inauthentic behaviours, using Twitter's public reporting flow. This is available in-app, on desktop, and via our reporting forms. Respecting that the terms 'disinformation' and 'misinformation' can be unfamiliar to and misunderstood by those without a technical background, our policies clearly outline what inauthentic behaviours look like on Twitter so it's easy to understand the variety of violative content that can be reported and that we can take action on. These reports are then used in aggregate to help refine our enforcement systems and identify new and emerging trends and patterns of behavior.

People using Twitter can also make reports related to Twitter Ads that might potentially violate our policies. These will be assessed against the Twitter Ads Policy, the Twitter Rules and Terms of Service and any enforcement action will be taken in line with these policies.

In addition to public reporting for platform manipulation, we also have partner reporting mechanisms for other priority areas, such as COVID-19 and Civic Integrity. We provide trusted government partners, public health experts and electoral authorities access to a Partner Support Portal, a dedicated reporting flow that allows expert and trusted reporters to escalate potentially violative content and access expedited support and human review by our Support teams. By giving experts, who can readily identify the accuracy of certain information i.e medical misinformation, we are able to address reporting-quality concerns. It is important we can take preventative measures against the potential misuse of public reporting mechanisms as a high-volume of erroneous reports could compromise the efficacy of harm reduction strategies.

Outcome 1d:

Users will be able to access general information about Signatories response to reports made under 5.11

Signatories will give details about how and when they have published information about their responses to reports by users about content that violates their policies.

We believe that transparency is a key principle in our mission to protect the Open Internet, and advancing the Internet as a global force for good. As outlined above, the Twitter Transparency Center provides data about reports received and actions taken on content that violates Twitter policies, including sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations. This is an ongoing reporting effort for Twitter and we will continue to share updates on the trends in violative content on our service and our enforcement.

People who report potentially violative content on Twitter will also receive a response directly from our Support teams about the results of our investigation and any enforcement action taken.

Objective 2: Disrupt advertising and monetisation incentives for disinformation

Outcome 2:

Advertising and/or monetisation incentives for Disinformation are reduced.

Signatories will describe the policies, processes and products that they have developed and/or implemented in order to disrupt advertising and/or monetisation incentives for behaviours that may propagate Disinformation. This section should also contain any relevant qualitative or quantitative data (such as case study examples) about the extent those reduce advertising and/or monetisation incentives for Disinformation in relation to content that is provided to users in Australia.

Promoted content on Twitter must also adhere to our existing Twitter Rules. In addition, we publish specific policies for advertisers that share standards for that are outlined below.

Political content advertising policy
Twitter globally prohibits the promotion of political content. We have made this decision based on our belief that political message reach should be earned, not bought.

Inappropriate content advertising policy
Our policy on inappropriate content prohibits advertising deemed to be dangerous or exploitative, misrepresentative, along with misleading synthetic or manipulated content and content engaged in coordinated harmful activity.

Quality advertising policy
Our quality policy outlines standards for advertisers including that ads should represent the brand or product being promoted and cannot mislead users into opening content by including exaggerated or sensationalised language or misleading calls to action.

Demonetisation of misleading information
Twitter automatically demonetises publisher content monetised through the Amplify Pre-Roll program that receives a misleading information label. To date, there have been no cases of Tweets with ad spend targeting Australian audiences on publisher content labeled as misinformation. Tweets receiving this label also cannot be promoted as ads under our Inappropriate Content policy.

People using Twitter can also make reports related to Twitter Ads that might potentially violate our policies. These will be assessed against the Twitter Ads Policy, the Twitter Rules and Terms of Service and any enforcement action will be taken in line with these policies.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms

Outcome 3:

The risk that Inauthentic User Behaviours undermine the integrity and security of Services and Products is reduced.
Signatories will describe the policies and processes that they have implemented that prohibit or manage the types of user behaviours that may undermine the integrity and security of their services and products. This section should also contain any relevant qualitative or quantitative data (including case study examples) about the extent those measures have reduced the risk that inauthentic user behaviours undermine the integrity and security of their services and products delivered to Australian users.

The relevant policies and strategies for this section are the same policies outlined under 'Objective 1'. An abridged version of these policies is included below.

Platform manipulation and spam policy

Users may not use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter.

NB: While this policy prohibits fake accounts, it does not apply to those using Twitter pseudonymously or as a parody, commentary, or fan account. The ability to speak anonymously, or to use a pseudonym, within the Twitter Rules, has been a core tenet of our service since its inception and we believe that the right to remain anonymous online is essential to preserving free expression.

COVID-19 misleading information policy
Users may not use Twitter's services to share false or misleading information about COVID-19 which may lead to harm. In this context, content that is demonstrably false or misleading and may lead to significant risk of harm (such as increased exposure to the virus, or adverse effects on public health systems) may not be shared on Twitter. This includes sharing content that may mislead people about the nature of the COVID-19 virus; the efficacy and/or safety of preventative measures, treatments, or other precautions to mitigate or treat the disease; official regulations,

restrictions, or exemptions pertaining to health advisories; or the prevalence of the virus or risk of infection or death associated with COVID-19. In addition, we may label Tweets which share misleading information about COVID-19 to reduce their spread and provide additional context.

Civic integrity policy
Users may not use Twitter's services for the purpose of manipulating or interfering in elections or other civic processes. This includes posting or sharing content that may suppress participation or mislead people about when, where, or how to participate in a civic process, such as, a federal or state election. In addition, we may label and reduce the visibility of Tweets containing false or misleading information about civic processes in order to provide additional context.

We also prohibit attempts to use our services to manipulate or disrupt civic processes, including through the distribution of false or misleading information about the procedures or circumstances around participation in a civic process. In instances where misleading information does not seek to directly manipulate or disrupt civic processes, but leads to confusion on our service, we may label the Tweets to give additional context.

Synthetic and manipulated media policy
Users may not deceptively promote synthetic or manipulated media that are likely to cause harm. In addition, we may label Tweets containing synthetic and manipulated media to help people understand their authenticity and to provide additional context.

Objective 4: Empower consumers to make better informed choices of digital content

Outcome 4:

Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.
Signatories detail measures implemented to enable users to make more informed choices about the source of news and factual content accessed via digital platforms. This section should also contain any relevant qualitative or quantitative data about Australian users response to measures such as the extent they have used empowerment tools (including case study examples) in relation to different categories of content (e.g., advertising, news, academic research, search engine results) provided on digital platforms.

Providing context for conversations on Twitter

We recognise the importance of helping users identify trusted information and make informed choices in today's information ecosystem. As outlined above, we have adopted a number of strategies to signal authenticity, accuracy and authoritative information on the Twitter service. These include Events pages, Moments, search prompts, and verification, as well as labeling government and state-sponsored media accounts so that users can evaluate the source of information.

We provide this context both through Twitter's Curation team and through trusted partnerships with government, experts, and civil society groups. Our Curators don't act as reporters or creators of original work, but rather organise and present content that already exists on Twitter in Moments, explanatory content on Trends, in lists and more. They are guided by and adhere to publicly available principles, including impartiality and accuracy.[16]

Government and state-affiliated media accounts

In 2020, we launched government and state media account labels to provide further context on accounts and help people make more informed choices about what they see, and how they engage on Twitter. Labels on state-affiliated media accounts provide additional context about accounts that are controlled by certain official representatives of governments, state-affiliated media entities and individuals closely associated with those entities.

The label appears on the profile page of the relevant Twitter account and on the Tweets sent by and shared from these accounts. Labels contain information about the country the account is affiliated with and whether it is operated by a government representative or state-affiliated media entity. Additionally, these labels include a small icon of a flag to signal the account's status as a government account or an icon of a podium to indicate state-affiliated media.

Currently, Australian accounts are not labelled, however, the initiative will continue to be expanded and additional countries added in each stage of this ongoing effort.

How government accounts are defined

Our focus is on senior officials and entities that are the official voice of the nation state abroad, specifically accounts of key government officials, including foreign ministers, institutional entities, ambassadors, official spokespeople, and key diplomatic leaders. Where accounts are used solely for personal use and do not play a role as a geopolitical or official Government communication channel, we will not label the account.

How state-affiliated media accounts are defined

State-affiliated media is defined as outlets where the state exercises control over editorial content through financial resources, direct or indirect political pressures, and/or control over production and distribution. Accounts belonging to state-affiliated media entities, their editors-in-chief,

[16] Help.twitter.com, Twitter Moments guidelines and principles, [online], Available at: <https://help.twitter.com/en/rules-and-policies/twitter-moments-guidelines-and-principles>. [Accessed April 2021]

and/or their senior staff may be labeled. State-financed media organisations with editorial independence like the ABC in Australia, are not defined as state-affiliated media for the purposes of this policy.

Objective 5: Improve public awareness of the source of political advertising carried on digital platforms

Outcome 5:

Users are better informed about the source of Political Advertising.
Signatories detail policies that provide users with information about the source of Political Advertising carried on digital platforms.

This section is not applicable to Twitter. Twitter globally prohibits the promotion of political content under our Political content advertising policy. We have made this decision based on our belief that political message reach should be earned, not bought.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research

Outcome 6:

Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.
Signatories detail measures to support the efforts of independent researchers to improve the Australians public understanding of Disinformation and Misinformation both online and offline.

This section could for example set out how Signatories are cooperating with the research community including, for example, through funding of research through the provision of tools that facilitate the running of queries by researchers and fact-checkers and enabling independent monitoring and analysis of disinformation trends and assessment of the measures taken by platforms under this code.

Signatories should also discuss issues (such as data protection concerns) that may have impacted on cooperation with the research community. This section should also contain any relevant qualitative or quantitative data (including case study examples) about the extent those measures have assisted research into the experience of Disinformation in Australia.

In line with our commitments to transparency, Twitter is the only major service to make public conversation data proactively available via an application programming interface (API) for the purposes of research.[17] By harnessing the power of the Twitter API, partners are able to tap into the public conversation and study collective issues facing global communities to bring about new insights to universal issues, devise fresh approaches to problems, and foster social good.

Building on this work, this year Twitter launched the Academic Research product track in order to enable academic researchers to access increased data from the public conversation to study topics that are as diverse as the conversations on Twitter. This track provides qualified academics the opportunity to access new endpoints, including the full history of public conversation data, a higher volume of Tweets, and more precise filtering capabilities.[18]Research conducted with the Twitter API must adhere to the Twitter Developer Policy, which is linked in our publicly available information about our approach to providing academic access to data.

We all have a part to play in creating a healthy information ecosystem and we have contributed to education and awareness efforts through sharing our insights on the Twitter blog and through our brand owned and operated Twitter accounts. For example, our blog on understanding the facts about 'bots'[19], an often misinterpreted area related to platform manipulation.

We have also partnered with non-government organisations on global awareness campaigns and initiatives, such as UNESCO for the evergreen, custom emoji activated by the #ThinkBeforeSharing hashtag. #ThinkBeforeSharing aimed to increase comprehension and media literacy and help people learn how to identify, debunk, react to and report on conspiracy theories to prevent their spread.

In line with our principles of transparency and to improve public understanding of inauthentic influence campaigns, Twitter has also published public archives of Tweets and media that we believe resulted from state-backed information operations.[20] It is the only archive of its kind in industry and, since our first comprehensive archive of state-backed information operations on Twitter in 2018[21], we have proactively expanded these datasets with several separate updates to provide more granularity and transparency.[22] We have also collaborated with research and civil

[17]Developer.twitter.com. 2021. Advancing Academic Research with Twitter Data. [online] Available at: <https://developer.twitter.com/en/solutions/academic-research> [Accessed 12 February 2021].
[18]Blog.twitter.com. 2021. Enabling the future of academic research with the Twitter API. [online] Available at: <https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.html> [Accessed 12 February 2021].
[19] Blog.twitter.com. 2020.Bot or not? The facts about platform manipulation on Twitter. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2020/bot-or-not.html> [Accessed May 2021].
[20]Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en/reports/information-operations.html> [Accessed 12 February 2021].
[21] Blog.twitter.com. 2018. Enabling further research of information operations on Twitter. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.html> [Accessed April 2021]
[22] Blog.twitter.com. 2020. Disclosing networks of state-linked information operations we've removed [online] Available at: <https://blog.twitter.com/en_us/topics/company/2020/information-operations-june-2020.html>. [Accessed April 2021]

society partners to increase access, transparency and meaningful interpretation of this information. As an example of our efforts, we partnered with the Australian Strategic Policy Institute (ASPI) and the Stanford Internet Observatory (SIO) to provide them with advance access to the data and enable independent research from subject matter experts to provide analysis and insights to accompany the data disclosure as part of our recent disclosure of state-linked information operations.[23]

Objective 7: Signatories will publicise the measures they take to combat Disinformation

Outcome 7:

The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.
Signatories list and/or provide links to reports made available to government and the public on relevant efforts under this Code.

We have put extensive work into defining platform manipulation behaviors under our Platform Manipulation and Spam policy, in order to improve understanding of our rules and enforcement actions. This ultimately supports and improves public understanding of the wide variety of inauthentic behaviours that can be addressed to protect the integrity of our service.

In addition to committing to annual reporting under the Code, and updating our policies as our rationales, approaches or enforcement options evolve, we will also continue to disclose data under the Twitter Transparency Center and our state-backed operations information disclosures (such as those made in October 2018, January 2019, June 2019, August 2019, September 2019, December 2019, March 2020, April 2020, June 2020, October 2020, February 2021).

As highlighted throughout this initial report, these ongoing, existing self-reporting measures provide insights into trends and actions taken on content that violates Twitter policies, including sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations.

[23] Ibid.

This data is available and open to analysis for government, Twitter users, and the general public. We adopted transparency and open data principles with the establishment of these initiatives and aim to continually improve their accessibility and usefulness to the public by publicising our disclosures on the Twitter blog and through media activity in each reporting period and continually evaluating how easily the language and structure of our reporting can be understood by external audiences.

Concluding Remarks

This section should contain any further observations Signatories wish to make about their response to the Code.

Twitter's mission is to serve the public conversation. We're keenly aware of our responsibility and our part towards protecting and growing an Open Internet, and to fostering a healthy information ecosystem by protecting the integrity of our service. This includes continuing to keep platform manipulation off Twitter, and leading with transparency by sharing regular updates on our progress and learning.

Our approach, as outlined in this initial report, remains closely aligned with our company values as well as the guiding principles of the Code, particularly protecting freedom of expression. We trust this overview of our work to date and future commitments under the Code, provide an understanding of the serious resolve and continuous commitment with which our teams approach protecting the integrity of public conversation.

We look forward to continuing our work with government and academic partners, as well as across industry and civil society to improve public understanding of these complex issues and to take meaningful steps that protect our service, the people who use it, and the Open Internet.

AUSTRALIAN CODE OF PRACTICE ON DISINFORMATION AND MISINFORMATION

1 Preamble

1.1 *Background:* This Code of Practice has been developed by the Digital Industry Group Inc. (DIGI), a non-profit industry association that advocates for the interests of the digital industry in Australia. The Code was developed in response to Government policy as set out in *Regulating in the Digital Age: Government Response and Implementation Roadmap for the Digital Platforms Inquiry*, where Government asked the major digital platforms to develop a voluntary Code of conduct outlining what the platforms will do to address concerns regarding Disinformation and credibility signalling for news content. The Code also takes into account guidance provided by the Australian Communications and Media Authority set out in *Misinformation and News Quality on Digital Platforms in Australia: A Position Paper to Guide Code Development*.

1.2 *Subject matter:* Disinformation and misinformation are aspects of a wider, multifaceted social problem which involves a range of offline and online behaviours which propagate information that threatens to undermine established democratic processes or public goods such as public health. Concepts such as “disinformation”, “misinformation”, and “fake news” mean different things to different people and can become politically charged when they are used by people to attack others who hold different opinions on value-laden political issues on which reasonable people may disagree. The understanding and effects of these concepts varies amongst individuals and is also under-researched.

1.3 *Role of Digital platforms* The digital platforms who have signed this Code recognise their role as important actors within the Australian information ecosystem and have already implemented a range of measures to tackle the propagation of disinformation and misinformation amongst users of their services and products. This Code is designed to express the minimum commitments made by Signatories to address the propagation of Disinformation and Misinformation (as defined in this Code) via digital platforms. Signatories may in their discretion implement policies and processes that contain a more expansive set of obligations than is provided for under this Code.

1.4 *Minimum Standards:* All Signatories commit to meet the minimum commitments outlined in Section 5.2 including the core Objective of providing appropriate safeguards against Harms that may be caused by Disinformation and Misinformation.

1.5 *Opt-in:* The digital industry is highly innovative and diverse, and digital platforms operate vastly different businesses which offer a wide and constantly evolving variety of services and products. As a result, the measures taken by digital platforms to address Disinformation and Misinformation in the context of their respective businesses may vary over time. For example, measures which are taken by a user-generated content platform may differ from those taken by a search engine. To accommodate the need of the Signatories to choose those measures which are most suitable to address instances of Disinformation and Misinformation in relation to different services and products provided by digital platforms, this Code provides Signatories the ability to opt into a range of measures and objectives, additional to the minimum commitments outlined in 1.4.

1.6 *Proportionality:* The types of user behaviours, content and Harms that this Code seeks to address will vary greatly in incidence and impact amongst the diverse range of services and products offered by different digital platforms. Accordingly, the commitments made by Signatories to the Code are intended to enable them to take actions which are proportional responses to their commitments under the Code. Section 6 provides further guidance on the kinds of contextual factors that Signatories may take into account in this regard.

1.7 Need for collaboration and cooperation among all relevant stakeholders: While this Code is intended to apply to digital platforms, the Signatories recognise and emphasise that a range of relevant stakeholders have roles and responsibilities in dealing with Disinformation and Misinformation including public authorities, academia, civil society, influencers, and news organisations. Tackling Disinformation and Misinformation effectively will require concerted effort and collaboration by and among these various stakeholder groups, and not only digital platforms. The Signatories welcome ongoing dialogue with stakeholders about what works well, what does not, and how together we can respond to the evolving challenges of Disinformation and Misinformation.

2 Guiding Principles

2.1 *Protection of freedom of expression*: Digital platforms provide a vital avenue for the open exchange of opinion, speech, information, research and debate and conversation as well as creative and other expression across the Australian community. Signatories should not be compelled by Governments or other parties to remove content solely on the basis of its alleged falsity if the content would not otherwise be unlawful Given its subject matter, the Code gives special attention to international human rights as articulated within the Universal Declaration on Human Rights, including but not limited to freedom of speech. Signatories are encouraged to, in developing proportionate responses to Disinformation and Misinformation to be cognisant of the need to protect these rights.

2.2 *Protection of user privacy:* Digital platforms value their users' privacy. Any actions taken by digital platforms to address the propagation of Disinformation and Misinformation should not contravene commitments they have made to respect the privacy of Australian users, including in terms and conditions, published policies and voluntary codes of conduct as well as by applicable laws. This includes respect for users' expectations of privacy when using digital platforms and in private digital communications. Additionally, any access to data for research purposes must protect user privacy.

2.3 *Policies and processes concerning advertising placements.* Digital platforms recognise the importance of having policies and processes in place with respect to advertisement placements on their services and products to reduce revenues that may reach the propagators of Disinformation.

2.4 *Empowering users:* Digital platforms should empower users to make informed choices about digital media content that purports to be a source of authoritative current news or of factual information.

2.5 *Integrity and security of services and products*. Digital platforms should communicate on the effectiveness of efforts to ensure the integrity and security of their services and products by taking steps to prohibit, detect and take action against inauthentic accounts on their services and products whose purpose is to propagate Disinformation.

2.6 *Supporting independent researchers*: Digital platforms recognise the importance of industry support for research efforts by independent experts including academics that can inform on trends and effective means to counter Disinformation and Misinformation. The Code provides various options for digital platforms to participate in independent research initiatives.

2.7 *Without prejudice commitments* This Code is without prejudice to other initiatives aimed at tackling Disinformation and Misinformation by digital platforms.

3 Glossary

This glossary provides information on some of the key terms used in this Code.

3.1 *Digital Content* is content distributed online on a platform owned and operated by a Signatory to this Code that is targeted at Australian users and includes content that has been artificially produced, manipulated or modified by automated means such as through the use of an artificial intelligence algorithm.

3.2 The aspect of *Disinformation* that this Code focuses on is:

A) Digital Content that is verifiably false or misleading or deceptive by an authoritative or credible source;
B) is propagated amongst users of digital platforms via Inauthentic Behaviours; and
C) the dissemination of which is reasonably likely to cause Harm.

3.3 *Enterprise Services* is software and services including cloud storage and content delivery services which are designed for the use of a specific organisation.

3.4 *Harm* means harms which pose an imminent and serious threat to:

A) democratic political and policymaking processes such as voter fraud, voter interference, voting misinformation; or
B) public goods such as the protection of citizens' health, protection of marginalised or vulnerable groups, public safety and security or the environment.

3.5 *Inauthentic Behaviour* includes spam and other forms of deceptive, manipulative or bulk, aggressive behaviours (which may be perpetrated via automated systems) and includes behaviours which are intended to artificially influence users online conversations and/or to encourage users of digital platforms to propagate Digital Content.

3.6 *Misinformation* means:
A) Digital Content (often legal) that is verifiably false or misleading or deceptive by an authoritative or credible source;
B) is disseminated by users of digital platforms; and
C) is reasonably likely (but may not be clearly intended to) cause Harm.

3.7 *Political Advertising* means paid for advertisements made by, on behalf of a political party or which advocate for the outcome of an election or referendum or about social issues in Australia, or which is regulated as political advertising under Australian law.

3.8 *Search Engines* consist of software systems designed to collect and rank information on the World Wide Web in response to user queries. Search Engines automate their systems in two ways. First, they use software known as “web crawler,” “bots” or “spiders” to discover publicly available webpages and automatically index and collect information from and about these webpages and internet sites. Second, they use ranking systems to return results in a set of links to websites. These ranking systems are made up of a series of algorithms that, ranked based on many factors such as relevance and usability of pages, expertise of sources, and more. The weight applied to each factor may depend based on the nature of the query. “Search Engine” excludes downstream entities that offer search functions on their own platforms, for which the results are powered by third-party search engines, as these downstream entities have no legal or operational control of the search results, the index from which they are generated nor the ranking order in which they are provided.

4 Scope, application and commencement of this Code

4.1 *Scope*: Recognising that the types of user behaviour and content that is subject to the Code will vary greatly in incidence and impact amongst the diverse range of services and products offered by different digital platforms, it is expected that the commitments under this Code will apply to the services and products that deliver to end users in Australia:

A) user-generated (including sponsored and shared) content, and/or
B) content that is selected and ranked by Search Engines in response to user queries.

4.2 *Excluded services and products*: The following are not services and products subject to this Code:

A) private messaging services including those provided via software applications;
B) email services including those provided via software applications;
C) Enterprise Services;

4.3 The list of excluded services and products is not intended to be exhaustive as new services and products are likely to emerge, some of which will not be relevant to the Code.

4.4 *Excluded Content:* Subject to Section 4.5, the following content is excluded from the operation of Sections 5.8 to 5.16 of the Code:

A) content produced in good faith for entertainment (including satire and parody) or for educational purposes;
B) content that is authorised by an Australian State or Federal Government.
C) Political Advertising or content authorised by a political party registered under Australian law.
D) news content that is the subject of a published editorial code which sets out content standards and or/complaints mechanisms.

4.5 Signatories may in their discretion, implement policies and procedures which govern the dissemination by users on their platforms of the types of content excluded from the operation of the provisions of the Code under Section 4.4 where Signatories determine such content is reasonably likely to cause Harm.

4.6 *Application of existing laws*: There are a range of existing laws or regulatory arrangements (such as the *Enhancing Online Safety Act 2015* (*Cth*) as well as prohibitions or restrictions concerning matters as diverse as tobacco, therapeutic goods, online gambling, election advertising, and defamation that may overlap with some of the matters covered by the Code. To the extent of any conflict with this Code, those laws and regulations will have primacy.

4.7 *Application*: The commitments made by each Signatory apply to it, in respect of the commitments it adopts, in respect of the products and services it nominates, from the date that it opts into those commitments.

4.8 *Commencement*: This Code commences on [insert date].

5 Objectives and Measures

5.1 *General:* This section incorporates a range of measures aimed at achieving seven key objectives and ten outcomes which are informed by the purpose and guiding principles of the Code set out in Section 2 above.

5.2 *Signatories Commitments*. All Signatories commit to the core Objective 1 of this Code so as to contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms as set out in Outcome 1b and will provide an annual report as set out in Section 7. Not all objectives and outcomes will be applicable to all Signatories who may adopt one or more of the measures set out in this section 5 in a manner that is relevant and proportionate to their different services and products, in accordance with the guidance in section 6. Signatories recognise that measures implemented under the Code may also evolve to reflect changes in their services and products, technological developments and the information environment.

5.3 *Opt-in*: Section 6 below outlines how Signatories will elect to opt into the commitments.

5.4 *Terminology of measures*: In implementing measures under the Code, Signatories recognise that actions taken aimed at achieving any outcome including the implementation of policies and processes may use terminology other than “Disinformation” and “Misinformation” and may, for example, refer to or a range of prohibited user behaviours or conduct such as making false or misleading representations about the user’s identity, origin or intentions and/or a range of prohibited content such as misleading, deceptive, dangerous or harmful content.

5.5 *Plain language*: Where Signatories commit to publishing their policies, procedures and any relevant community guidelines or additional information on their actions to combat Disinformation and Misinformation, they will use reasonable commercial efforts to do so in plain language and in an accessible, user-friendly format.

5.6 *Restrictions on lawful content or users access:* In seeking to comply with the requirements of this Code, Signatories are not required to (although they may elect to) take measures that require them to delete or prevent access to otherwise lawful content solely on the basis that it is or may be misleading or deceptive or false. Nor will Signatories be required to signal the veracity of content uploaded and shared by their users.

5.7 *Need for transparency to be balanced against disclosure risks:* Signatories recognise that in implementing commitments to promote the public transparency of measures taken under this Code there is a need to balance the need to be open about those measures with the risk that the release of certain information may result in an increase in behaviours that propagate Disinformation and Misinformation, or which increase their virality.

Objective 1: Provide safeguards against Harms that may arise from Disinformation and Misinformation.

Outcome 1a: Signatories contribute to reducing the risk of Harms that may arise from the propagation of Disinformation and Misinformation on digital platforms by adopting a range of scalable measures.

5.8 Signatories will develop and implement measures which aim to reduce the propagation of and potential exposure of users of digital platforms to Disinformation and Misinformation.

5.9 Measures implemented under 5.8, may include, by way of example rather than limitation:

A) policies and processes that require human review of user behaviours or content that is available on digital platforms (including review processes that are conducted in partnership with fact-checking organisations)
B) labelling false content or providing trust indicators of content to users;
C) demoting the ranking of content that may expose users to Disinformation and Misinformation;
D) removal of content which is propagated by Inauthentic Behaviours;
E) providing transparency about actions taken to address Disinformation and Misinformation to the public and/or users as appropriate;
F) suspension or disabling of accounts of users which engage in Inauthentic Behaviours;
G) the provision or use of technologies to identify and reduce Inauthentic Behaviours that can expose users to Disinformation such as algorithmic review of content and/or user accounts,
H) the provision or use of technologies which assist digital platforms or their users to check authenticity or accuracy or to identify the provenance or source of digital content;
I) exposing meta data to users about the source of content;
J) enforcing published editorial policies and content standards; and
K) prioritising credible and trusted news sources that are subject to a published editorial code (noting that some Signatories may remove or reduce the ranking of news content which violates their policies in accordance with section 4.5).

Outcome 1b: Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.

5.10 Signatories will implement and publish policies and procedures and any appropriate guidelines or information relating to the prohibition and/or management of user behaviours and/or content that may propagate Disinformation and/or Misinformation via their services or products.

Outcome 1c: Users can report content or behaviours to Signatories that violates their policies under Section 5.10 through publicly available and accessible reporting tools.

5.11 Signatories will implement and publish policies, procedures and appropriate guidelines that will enable users to report the types of behaviours and content that violates their policies under Section 5.10.

5.12 In implementing the commitment in 5.11 Signatories recognise that the terms Disinformation and Misinformation may be unfamiliar to users and thus policies and procedures aimed at achieving this outcome may specify how users may report a range of impermissible content and behaviours on digital platforms.

Outcome 1d: Users will be able to access general information about Signatories actions in response to reports made under 5.11.

5.13 Signatories will implement and publish policies, procedures and aggregated reports (including summaries of user reports made under 5.11) regarding the detection and removal of content that violates platform policies, including but not necessarily limited to content on their platforms that qualifies as Misinformation and/or Disinformation.

Objective 2: Disrupt advertising and monetisation incentives for Disinformation.

Outcome:**2**: Advertising and/or monetisation incentives for Disinformation are reduced.

5.14 Signatories will implement policies and processes that aim to disrupt advertising and/or monetisation incentives for Disinformation.

5.15 Policies and processes implemented under 5.14 may for example, include:

A) promotion and/or inclusion of the use of brand safety and verification tools;
B) enabling engagement with third party verification companies;
C) assisting and/or allowing advertisers to assess media buying strategies and online reputational risks;
D) providing advertisers with necessary access to client-specific accounts to help enable them to monitor the placement of advertisements and make choices regarding where advertisements are placed; and /or
E) restricting the availability of advertising services and paid placements on accounts and websites that propagate Disinformation.

5.16 Signatories recognise that all parties involved in the buying and selling of online advertising and the provision of advertising-related services need to work together to improve transparency across the online advertising ecosystem and thereby to effectively scrutinise, control and limit the placement of advertising on accounts and websites that propagate Disinformation.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.

Outcome 3: The risk that Inauthentic User Behaviours undermine the integrity and security of Services and Products is reduced.

5.17 Signatories commit to take measures that prohibit or manage the types of user behaviours that are designed to undermine the integrity and security of their services and products, for example, the use of fake accounts or automated bots that are designed to propagate Disinformation.

5.18 To allow for the expectations of some users and digital platforms about the protection of privacy, measures developed and implemented in accordance with this commitment should not preclude the creation of pseudonymous and anonymous accounts.

Objective 4: Empower consumers to make better informed choices of digital content.

Outcome 4: Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify Misinformation.

5.19 Signatories will implement measures to enable users to make informed choices about digital content and to access alternative sources of information on these matters.

5.20 Measures developed and implemented in accordance with the commitment in 5.19 may for example include:

A) the use of technological means to prioritise or rank digital content to enable users to easily find diverse perspectives on matters of public interest;
B) aggregation or promotion of news content subject to an independent editorial code and complaints scheme;

C) the provision or use of technologies which signal the credibility of news sources or which assist digital platforms or their users to check the authenticity or accuracy of online news content or to identify its provenance or source;
D) the promotion of digital literacy; and or
E) the provision of financial support for sustainable partnerships with fact-checking organisations.

Objective 5: Improve public awareness of the source of Political Advertising carried on digital platforms.

Outcome 5: Users are better informed about the source of Political Advertising.

5.21 While Political Advertising is not Misinformation or Disinformation for the purposes of the Code, Signatories will develop and implement policies that provide users with greater transparency about the source of Political Advertising carried on digital platforms.

5.22 Measures developed and implemented in accordance with the commitment in 5.21 may include requirements that advertisers to identify and/or verify the source of Political Advertising carried on digital platforms, the provisions of tools which enable the public to understand whether a political ad has been targeted to them and policies which require that Political Advertisements which appear in a medium containing news or editorial content are presented in such a way as to be readily recognisable as a paid-for communication.

5.23 Signatories may also, as a matter of policy, choose not to target advertisements based on the inferred political affiliations of a user.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.

Outcome 6: Signatories support the efforts of independent researchers to improve public understanding of Disinformation and Misinformation.

5.24 Signatories commit to support and encourage good faith independent efforts to research Disinformation and Misinformation both online and offline. Good faith research includes research that is conducted in accordance with the ethics policies of an accredited Australian University provided such policies require that data collected by the researcher is used solely for research purposes and is stored securely on a university IT system or any research which is conducted in accordance with the prior written agreement of the digital platform.

5.25 Measures taken to implement 5.24 may include, for example, cooperation with relevant initiatives taken by independent fact checking bodies. Other measures may include funding for research and/or sharing datasets, undertaking joint research, or otherwise partnering with academics and civil society organisations.

5.26 Signatories commit not to prohibit or discourage good faith research into Disinformation or Misinformation on their platform.

5.27 Relevant Signatories commit to convene an annual event to foster discussions regarding Disinformation and Misinformation within academia and Civil Society.

Objective 7: Signatories publicise the measures they take to combat Disinformation and Misinformation.

Outcome 7: The public can access information about the measures Signatories have taken to combat Disinformation and Misinformation.

5.28 All Signatories will make and publish the annual report information in Section 7.

5.29 In addition, Signatories will publish additional information detailing their progress in relation to Objective 1 and any additional commitments they have made under this Code.

5.30 Signatories may fulfill their commitment in 5.29 by providing additional reports and/or public updates on areas such as content removals, open data initiatives, research reports, media announcements, user data requests and

business transparency reports. Examples of such information could include, by way of example rather than limitation, blog posts, white papers, in-product notifications, transparency reports, help centres, or other websites.

6 Guidance on platform-specific measures

6.1 **Proportionality of measures under Code:** The measures taken by Signatories pursuant to this Code will be proportionate and relevant to their specific context including the Harm posed by instances of Disinformation and Misinformation. Signatories may take into consideration a variety of factors in assessing the appropriateness of measures including:

A) the actors which are engaged in propagating Disinformation and Misinformation;
B) the nature of the behaviour of users propagating Disinformation and Misinformation, for example, whether the behaviour is automated and intentional and/or maliciously motivated and the extent to which it is coordinated, persistent and at scale;
C) the type of Product or Service via which the content is distributed and whether it has network effects that result in content being widely and rapidly shared amongst users of the platform;
D) Whether the platform may receive a commercial benefit from the propagation of the content (for example, whether the content is sponsored content);
E) the extent to which it is reasonably possible to verify the falsity of relevant content via an authoritative or credible source;
F) the proximity and severity of the Harm that is reasonably likely to result from the propagation of the content;
G) the nature of the online community using the digital platform;
H) the size and nature of the digital platform's business and the resources available to it;
I) the need to protect freedom of expression in balance with other human rights; and
J) the need to protect user privacy.

7 Code administration

7.1 *Opt-in*. In recognition of the variation in business models and product offerings of Digital platforms, this Code is designed to allow a range of businesses to make commitments by way of opt-in arrangements. Upon signing the Code Signatories will nominate the provisions to which they commit using the Opt-in Nominations Form in Appendix 1

7.2 *Withdrawal from Code*. A Signatory may withdraw from the Code or a particular commitment under the Code by notifying DIGI.

7.3 *Annual Report*. In addition to the Opt-in Nomination Form under 7.1, each Signatory will provide an annual report to DIGI setting out its progress towards achieving the outcomes contained in the Code which will be published on the DIGI website. The first report will be in the form of the report Appendix 2 and submitted within three months of the commencement of the Code. Signatories commit to develop and implement, within six months of the commencement of this Code, an agreed format for future annual reports and a guideline that will inform the data and other information to be included in subsequent reports.

7.4 *Complaints.* Signatories agree to establish, within six months of the commencement of this Code, a facility for addressing non-compliance with the Code. The facility will hear appeals of complaints of Code breaches that have not been acted upon by Signatories and a policy describing circumstances in which a non-compliant Signatory may be removed. As part of this process, Signatories will also consider how they can leverage current arrangements with government and relevant regulatory agencies to identify and address instances of Inauthentic Behaviours that propagate Disinformation and are the subject of measures addressed by this Code.

7.5 *Code Administration.* The Administrator of this Code is DIGI who will establish a sub-committee comprising representatives from Signatories and independent members who will meet at six monthly intervals to review the actions of Signatories and monitor how they are meeting their commitments under the Code.

7.6 *Code Review*. The Code will be reviewed after it has been in operation for twelve months, and thereafter at two yearly intervals. The reviews will be based on the input of the Signatories, and on relevant government bodies (including the Australian Communications and Media Authority) and other interested stakeholders including academics and representatives from civil society active in this field.