Code of Practice on Disinformation – Report of Twitter for the period H2 2022

Executive summary

Twitter's Evolving Approach to Countering Disinformation

The focus of Twitter's evolving approach is *Community Notes*. This product represents an important evolution in how we mitigate both misinformation and disinformation. Community Notes exemplifies the adoption of a new content moderation model – one that relies on user participation rather than centralised enforcement. This transition is ongoing and the impact of this work will continue to grow, in parallel with enforcing the Twitter rules, particularly around manipulation and spam.

Community Notes aims to create a better-informed world by empowering people on Twitter to collaboratively add helpful notes to Tweets that might be misleading. Contributors can leave notes on any Tweet and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a Tweet. Community Notes is an inherently scalable and localised response to the challenge of disinformation. By making this feature an integral and highly visible part of the Twitter product, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that can be truly global in its application. It also reduces reliance on forms of content moderation that are more centralised, manual and bespoke; or which require intensive and time-consuming interactions with third parties.

Effectiveness & Research
Where launched to date, Community Notes is effective. According to the results of four surveys run in the United States at different times between August, 2021 and August, 2022, a person who sees a Community Note is, on average, 20-40% less likely to agree with the substance of a potentially misleading Tweet than someone who sees the Tweet alone. Survey participation ranged from 3,000 to more than 19,000 participants, and the results were consistent throughout the course of the year, even as news and Tweet topics changed.

We also see that Community Notes informs sharing behaviour. Analysing our internal data, we've found that a person on Twitter who sees a note is, on average, 15-35% less likely to choose to *Like* or *Retweet* a Tweet than someone who sees the Tweet alone.

In our most recent survey, notes were found to be informative regardless of a person's self-identified political party affiliation — there was no statistically significant difference in average informativeness across party identification.

We've published a research paper on Community Notes that provides more detail on how we've been measuring efficacy. All Community Notes contributions are publicly available on the Community Notes site Download Data page so that anyone has free access to analyse the data, identify problems, and identify product enhancement opportunities.

Expansion & Localisation
Community Notes are now publicly visible to everyone. Users in the US, the UK, Ireland, Canada, Australia and New Zealand can now contribute to the program. Over the coming months, users in more markets will be able to contribute notes and the product will be localised further. We currently have around 20,000 contributors and we aim to expand this number by 10% each week. Over time, users in any EU member state, writing in any language, should be able to contribute to Community Notes and the most helpful contributions will be surfaced to inform readers.

The technology-first strategy evidenced by Community Notes is reflective of how we intend to approach content moderation going forward. This approach has advantages over more centralised

methods of content moderation, which have always faced the same two challenges: speed and scale.

The Code of Practice on Disinformation

The Code of Practice on Disinformation asks participants to make progress in many areas. Twitter is making real advancements across the board. For example:

- We have clear policies that prohibit manipulative or spammy advertising.
- Our approach to issue-based and political advertising is changing. We will provide the transparency that people expect with these forms of advertising. The Twitter Ads Transparency Centre will be reinstated.
- Our Transparency Centre and, in particular, our page on the fight against state-backed information operations demonstrates how we effectively tackle coordinated platform manipulation. Our policies on platform manipulation and spam are robust and our Threat Disruption team continues its work in parallel with our development of Community Notes.
- Under Twitter's new ownership, the goal is to maximise "unregretted user minutes". In other words, Twitter users should feel like their time on the platform is informative and worthwhile. We're advancing this goal by improving product design.
    - We recently added the capacity to intuitively toggle between the algorithm that suggests content on Twitter and a reverse chronological feed. In the coming months, we will be open-sourcing the algorithm that recommends content in the timeline.
    - We also launched the subscription service, Twitter Blue, which is designed, in part, to authenticate user identities and thereby reduce the prevalence of spam and viral disinformation.
    - Community Notes is an agile and dynamic response to disinformation and misinformation. It directly empowers users by enabling them to be part of the solution. Third-party experts can avail of the program both through direct participation and by analysing the data that's made freely available on a daily basis.
- Twitter has been among the most open actors in the platform sector with regards to sharing data for academic research. Large data sets detailing extensive state-backed information operations have been made available to the global academic community. Twitter's API program is also used widely among academic researchers.
- Twitter published its first transparency report in 2012, over a decade ago. Since then, the Twitter Transparency Centre has become more detailed with almost every subsequent publication, now offering country-level data on both legal requests and Terms of Service violations.
- Twitter has long engaged with civil society organisations and we will continue to do so.
- Finally, Twitter has shared data that demonstrates that hateful speech accounts for less than 0.1% of all English-language Tweet impressions.

In some areas, Twitter is unable to provide granular data due to resource constraints and data limitations. In other areas, there are issues that are not applicable to Twitter's service. These issues were highlighted during the drafting of the Code.

Looking ahead, Twitter is engaging with the relevant stakeholders as to the best way to provide details on Twitter's compliance with the Digital Services Act. We expect that this reporting template and structure will be reviewed as part of those conversations and that services are afforded the opportunity to provide responses that are reflective of their broader approach to the DSA, their respective product and policy models, and proportionate to the risks faced and resources available.

II. Scrutiny of Ad Placements	
Commitment 1 Relevant signatories participating in ad placements commit to defund the dissemination of disinformation, and improve the policies and systems which determine the eligibility of content to be monetised, the controls for monetisation and ad placement, and the data to report on the accuracy and effectiveness of controls and services around ad placements	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter's comprehensive advertising policies can be found here: https://business.twitter.com/en/help/ads-policies.html There are 21 areas covered by the company's advertising policies. They are as follows: ● Adult Sexual Content ● Alcohol Content ● Copyright ● Counterfeit Goods ● Drugs and Drug Paraphernalia ● Endangered Species ● Financial Products and Services ● Gambling Content ● Hateful Content ● Healthcare ● Inappropriate Content ● Malware and Software Downloads ● Political Content ● Prohibited Content for Minors ● Quality ● State Media ● Tobacco and Tobacco Accessories ● Trademark ● Unauthorised Ticket Sales ● Unacceptable Business Practices

	• Weapons and Weapon Accessories Please note that, as with all of Twitter's policies, these are subject to continuous development and iteration. You can read about Twitter's approach to offboarding advertisers that violate our policies here: https://business.twitter.com/en/help/ads-policies/about-twitter-ads-offboarding.html To promote transparency around what can and cannot be advertised on Twitter, users can monitor this live log of changes made to our policies: https://business.twitter.com/en/help/ads-policies/ads-policy-update-log.html Twitter's informational page around brand safety can be found here: https://business.twitter.com/en/help/ads-policies/brand-safety.html#policies-that-lead Twitter recently launched a Brand Safety initiative in the US (which may expand to Europe). Twitter's brand safety measurement solutions with industry leaders DoubleVerify and Integral Ad Science are now generally available to advertising customers. These services monitor and quantify the prevalence of ad placement adjacent to English-language content deemed either unsafe or unsuitable for monetization by the Global Alliance for Responsible Media (GARM) in Twitter's Home Timeline. These feed-based solutions are the first of their kind to be made broadly available, and underscore our commitment to independent validation of Twitter's efforts to uphold industry brand safety standards. You can read more on the initiative here: https://business.twitter.com/en/blog/third-party-brand-safety-measurement.html The initiative described above is an extension of the work that Twitter has been doing to empower advertisers with more Adjacency Controls and third-party measurement. See full article here: https://business.twitter.com/en/blog/adjacency-controls-third-party-measurement.html

	Accounts that advertise on Twitter must meet certain criteria. That criteria is set out here: https://business.twitter.com/en/help/ads-policies/campaign-considerations/about-eligibility-for-twitter-ads.html Finally, Twitter may, during the course of a sensitive event, pause advertisements serving to and/or from a particular location. Advertisers in such locations, or targeting such locations, may be temporarily ineligible for Twitter ads. Twitter took this action when the Russia-Ukraine conflict started.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

Measure 1.1								
QRE 1.1.1	Outline relevant actions [suggested character limit: 2000 characters]							
SLI 1.1.1 – Numbers by actions enforcing policies above (specify if at page and/or domain level)	Methodology of data measurement [suggested character limit: 500 characters]							
	Type of Action 1 [linked to the policy mentioned in QRE]		Type of Action 2 [linked to the policy mentioned in QRE]		Type of Action 3 [linked to the policy mentioned in QRE]		Type of Action 4 [linked to the policy mentioned in QRE]	
Level	Page	Domain	Page	Domain	Page	Domain	Page	Domain
Member States								
Austria								
Belgium								
Bulgaria								
Croatia								
Cyprus								
Czech Republic								
Denmark								

Estonia				
Finland				
France				
Germany				
Greece				
Hungary				
Ireland				
Italy				
Latvia				
Lithuania				
Luxembourg				
Malta				
Netherlands				
Poland				
Portugal				
Romania				
Slovakia				
Slovenia				
Spain				
Sweden				
Iceland				
Liechtenstein				
Norway				
Total EU				
Total EEA				
Bulgarian				
Croatian				
Czech				
Danish				
Dutch				
English				
Estonian				
Finnish				
French				
German				
Greek				
Hungarian				

Irish				
Italian				
Latvian				
Lithuanian				
Maltese				
Polish				
Portuguese				
Romanian				
Slovak				
Slovenian				
Spanish				
Swedish				
Icelandic				
Norwegian				

Measure 1.2				
QRE 1.2.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 1.2.1	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of policy reviews	Nr of update to policies	Nr of accounts barred	Nr of domains barred
Member States				
List actions per member states and languages (see example table above)				
Measure 1.3				
QRE 1.3.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 1.4				
QRE 1.4.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 1.5				
QRE 1.5.1	Outline relevant actions [suggested character limit: 2000 characters]			
QRE 1.5.2	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 1.5	Outline relevant actions [suggested character limit: 2000 characters]			
QRE 1.5.1	Outline relevant actions [suggested character limit: 2000 characters]			
QRE 1.5.2	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 1.6				

QRE 1.6.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 1.6.2	Outline relevant actions [suggested character limit: 2000 characters]
QRE 1.6.3	Outline relevant actions [suggested character limit: 2000 characters]
QRE 1.6.4	Outline relevant actions [suggested character limit: 2000 characters]
SLI 1.6.1	Methodology of data measurement [suggested character limit: 500 characters]
	In view of steps taken to integrate brand safety tools: % of advertising/ media investment protected by such tools
Member States	
List actions per member states and languages (see example table above)	

II. Scrutiny of Ad Placements				
Commitment 2 Relevant Signatories participating in advertising commit to prevent the misuse of advertising systems to disseminate Disinformation in the form of advertising messages.				
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**See Twitter's Ad Policies and other initiatives outlined in Commitment 1 that are also applicable here.**			
If yes, list these implementation measures here [short bullet points].				
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]				
If yes, which further implementation measures do you plan to put in place in the next 6 months?				
Measure 2.1				
QRE 2.1.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 2.1.1 – Numbers by actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]			
	Type of Action 1 [linked to the policy mentioned in QRE]	Type of Action 2 [linked to the policy mentioned in QRE]	Type of Action 3 [linked to the policy mentioned in QRE]	Type of Action 4 [linked to the policy mentioned in QRE]
Member States				

List actions per member states and languages (see example table above)				
Measure 2.2				
QRE 2.2.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 2.3				
QRE 2.3.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 2.3.1	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of ads removed (as well as reach of ads before they were successfully removed)		Number of ads prohibited	
Member States				
List actions per member states and languages (see example table above)				
Measure 2.4				
QRE 2.4.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 2.4.1	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of appeals		Proportion of appeals that led to a change of the initial decision	
Member States				
List actions per member states and languages (see example table above)				

II. Scrutiny of Ad Placements	
Commitment 3 Relevant Signatories involved in buying, selling and placing digital advertising commit to exchange best practices and strengthen cooperation with relevant players, expanding to organisations active in the online monetisation value chain, such as online e-payment services, e-commerce platforms and relevant crowd-funding/donation systems, with the aim to increase the effectiveness of scrutiny of ad placements on their own services.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**See Twitter's Ad Policies and other initiatives outlined in Commitment 1 that are also applicable here. The brand safety work, in particular, involves partnerships with organisations along the revenue chain, including media agencies, and brand safety measurement bodies such as DoubleVerify and Integral Ad Science. Twitter is also a member of the Global Alliance for Responsible Media (GARM).**

If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

Measure 3.1	
QRE 3.1.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 3.2	
QRE 3.2.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 3.3	
QRE 3.3.1	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 4 Relevant Signatories commit to adopt a common definition of "political and issue advertising".	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	

If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 4.1	
Measure 4.2	
QRE 4.1.1 (for measures 4.1 and 4.2)	Outline relevant actions [suggested character limit: 2000 characters]
QRE 4.1.2 (for measures 4.1 and 4.2)	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 5 Relevant Signatories commit to apply a consistent approach across political and issue advertising on their services and to clearly indicate in their advertising policies the extent to which such advertising is permitted or prohibited on their services.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 5.1	
QRE 5.1.1	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising				
Commitment 6 Relevant Signatories commit to make political or issue ads clearly labelled and distinguishable as paid-for content in a way that allows users to understand that the content displayed contains political or issue advertising				
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**			
If yes, list these implementation measures here [short bullet points].				
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]				
If yes, which further implementation measures do you plan to put in place in the next 6 months?				
Measure 6.1				
QRE 6.1.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 6.2				
QRE 6.2.1	Outline relevant actions [suggested character limit: 2000 characters]			
QRE 6.2.2	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 6.2.1 – numbers for actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]			
	Number of ads accepted & labelled according to 6.2	Amounts spent by labelled advertisers	Other relevant metrics	Other relevant metrics
Member States				
List actions per member states and languages (see example table above)				
Measure 6.3				

QRE 6.3.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 6.4	
QRE 6.4.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 6.5	
QRE 6.5.1	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 7 Relevant Signatories commit to put proportionate and appropriate identity verification systems in place for sponsors and providers of advertising services acting on behalf of sponsors placing political or issue ads. Relevant signatories will make sure that labelling and user-facing transparency requirements are met before allowing placement of such ads.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 7.1	
QRE 7.1.1	Outline relevant actions [suggested character limit: 2000 characters]

SLI 7.1.1 – numbers for actions enforcing policies above (comparable metrics as for SLI 6.2.1)	Methodology of data measurement [suggested character limit: 500 characters]	
	Nr of ads rejected	Other relevant metrics
Member States		
List actions per member states and languages (see example table above)		

Measure 7.2	
QRE 7.2.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 7.2.2	Outline relevant actions [suggested character limit: 2000 characters]
Measure 7.3	
QRE 7.3.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 7.3.2	Outline relevant actions [suggested character limit: 2000 characters]
Measure 7.4	
QRE 7.4.1	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 8 Relevant Signatories commit to provide transparency information to users about the political or issue ads they see on their service.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	

Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 8.1	
Measure 8.2	
QRE 8.2.1 (for measures 8.1 & 8.2)	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 9 Relevant Signatories commit to provide users with clear, comprehensible, comprehensive information about why they are seeing a political or issue ad.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 9.1	

Measure 9.2	
QRE 9.2.1 (for measures 9.1 & 9.2)	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 10 Relevant Signatories commit to maintain repositories of political or issue advertising and ensure their currentness, completeness, usability and quality, such that they contain all political and issue advertising served, along with the necessary information to comply with their legal obligations and with transparency commitments under this Code.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 10.1	
Measure 10.2	
QRE 10.2.1 (for measures 10.1 & 10.2)	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising

Commitment 11	
Relevant Signatories commit to provide application programming interfaces (APIs) or other interfaces enabling users and researchers to perform customised searches within their ad repositories of political or issue advertising and to include a set of minimum functionalities as well as a set of minimum search criteria for the application of APIs or other interfaces."	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 11.1	
Measure 11.2	
Measure 11.3	
Measure 11.4	
QRE 11.1.1 (for measures 11.1-11.4)	Outline relevant actions [suggested character limit: 2000 characters]
QRE 11.4.1	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising
Commitment 12
Relevant Signatories commit to increase oversight of political and issue advertising and constructively assist, as appropriate, in the creation, implementation and improvement of political or issue advertising policies and practices.

In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 12.1	
Measure 12.2	
Measure 12.3	
QRE 12.1.1 (for measures 12.1-12.3)	Outline relevant actions [suggested character limit: 2000 characters]

III. Political Advertising	
Commitment 13 Relevant Signatories agree to engage in ongoing monitoring and research to understand and respond to risks related to Disinformation in political or issue advertising.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Broadly, Commitments 4-13 are not relevant to Twitter's current approach to political and issue advertising in Europe at the time of writing. This may change going forward.** **In addition, under the DSA, Twitter will relaunch its Advertising Transparency Centre, which had been operational until late 2019 and gave users an accessible interface to search for accounts that advertise on the platform.**

If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 13.1	
Measure 13.2	
Measure 13.3	
QRE 13.1.1 (for measures 13.1-13.3)	Outline relevant actions [suggested character limit: 2000 characters]

IV. Integrity of Services

Commitment 14

In order to limit impermissible manipulative behaviours and practices across their services, Relevant Signatories commit to put in place or further bolster policies to address both misinformation and disinformation across their services, and to agree on a cross-service understanding of manipulative behaviours, actors and practices not permitted on their services. Such behaviours and practices, which should periodically be reviewed in light with the latest evidence on the conducts and TTPs employed by malicious actors, such as the AMITT Disinformation Tactics, Techniques and Procedures Framework, include:

The following TTPs pertain to the creation of assets for the purpose of a disinformation campaign, and to ways to make these assets seem credible:

- 1. Creation of inauthentic accounts or botnets (which may include automated, partially automated, or non-automated accounts)
- 2. Use of fake / inauthentic reactions (e.g. likes, up votes, comments)
- 3. Use of fake followers or subscribers
- 4. Creation of inauthentic pages, groups, chat groups, fora, or domains
- 5. Account hijacking or impersonation

The following TTPs pertain to the dissemination of content created in the context of a disinformation campaign, which may or may not include some forms of targeting or attempting to silence opposing views. Relevant TTPs include:

- 6. Deliberately targeting vulnerable recipients (e.g. via personalized advertising, location spoofing or obfuscation)
- 7. Deploy deceptive manipulated media (e.g. “deep fakes”, “cheap fakes”...)
- 8. Use “hack and leak” operation (which may or may not include doctored content)

• 9. Inauthentic coordination of content creation or amplification, including attempts to deceive/manipulate platforms algorithms (e.g. keyword stuffing or inauthentic posting/reposting designed to mislead people about popularity of content, including by influencers) • 10. Use of deceptive practices to deceive/manipulate platform algorithms, such as to create, amplify or hijack hashtags, data voids, filter bubbles, or echo chambers • 11. Non-transparent compensated messages or promotions by influencers • 12. Coordinated mass reporting of non-violative opposing content or accounts	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter's approach to platform manipulation and spam is set out here: https://help.twitter.com/en/rules-and-policies/platform-manipulation** As mentioned in the Executive Summary, Twitter's Trust and Safety and Threat Disruption teams continue to enforce our policies in this area, often investigating and challenging networks of accounts that can number in the thousands. Twitter does 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 behaviour, including: • commercially-motivated spam, that typically aims to drive traffic or attention from a conversation on Twitter to accounts, websites, products, services, or initiatives; • 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; and • coordinated harmful activity that encourages or promotes behaviour which violates the Twitter Rules. **What is in violation of this policy?** Under this policy we prohibit a range of behaviours in the following areas:

	Multiple accounts and coordination You can't mass-register Twitter accounts or use automation to create Twitter accounts. You can't artificially amplify or disrupt conversations through the use of multiple accounts or by coordinating with others to violate the Twitter Rules. This includes: ● Overlapping accounts – operating multiple accounts with overlapping use cases, such as identical or similar personas or substantially similar content; ● Mutually interacting accounts – operating multiple accounts that interact with one another in order to inflate or manipulate the prominence of specific Tweets or accounts; and ● Coordination – creating multiple accounts to post duplicative content or create fake engagement, including: ○ posting identical or substantially similar Tweets or hashtags from multiple accounts you operate; ○ engaging (Retweets, Likes, mentions, Twitter Poll votes) repeatedly with the same Tweets or accounts from multiple accounts that you operate; ○ coordinating with or compensating others to engage in artificial engagement or amplification, even if the people involved use only one account; and ○ coordinating with others to engage in or promote violations of the Twitter Rules, including violations of our abusive behavior policy. *Engagement and metrics* You can't artificially inflate your own or others' followers or engagement. This includes: ● selling/purchasing Tweet or account metric inflation – selling or purchasing followers or engagements (Retweets, Likes, mentions, Twitter Poll votes); ● apps – using or promoting third-party services or apps that claim to add followers or add engagements to Tweets;

	• reciprocal inflation – trading or coordinating to exchange follows or Tweet engagements (including but not limited to participation in "follow trains," "decks," and "Retweet for Retweet" behavior); and • account transfers or sales – selling, purchasing, trading, or offering the sale, purchase, or trade of Twitter accounts, usernames, or temporary access to Twitter accounts. *Misuse of Twitter product features* You can't misuse Twitter product features to disrupt others' experience. This includes: Tweets and Direct Messages • sending bulk, aggressive, high-volume unsolicited replies, mentions, or Direct Messages; • posting and deleting the same content repeatedly; • repeatedly posting identical or nearly identical Tweets, or repeatedly sending identical Direct Messages; • repeatedly posting Tweets or sending Direct Messages consisting of links shared without commentary, so that this comprises the bulk of your Tweet/Direct Message activity; and • Tweeting an existing phrase or content in a duplicative manner, whether individually or in concert with other accounts. Learn more in our copypasta and duplicate content policy. Following • "follow churn" – following and then unfollowing large numbers of accounts in an effort to inflate one's own follower count; • indiscriminate following – following and/or unfollowing a large number of unrelated accounts in a short time period, particularly by automated means; and • duplicating another account's followers, particularly using automation

	Engagement ● aggressively or automatically engaging with Tweets to drive traffic or attention to accounts, websites, products, services, or initiatives. ● aggressively adding users to Lists or Moments. Hashtags ● using a trending or popular hashtag with an intent to subvert or manipulate a conversation or to drive traffic or attention to accounts, websites, products, services, or initiatives; and ● Tweeting with excessive, unrelated hashtags in a single Tweet or across multiple Tweets. URLs ● publishing or linking to malicious content intended to damage or disrupt another person's browser (malware) or computer or to compromise a person's privacy (phishing); and ● posting misleading or deceptive links; e.g., affiliate links and clickjacking links.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 14.1	
QRE 14.1.1	**See above re Commitment 14**
QRE 14.1.2	**In the cases and investigations summarised below, Twitter offers an insight into how its**

	Threat Disruption team has worked collaboratively with stakeholders in the second half of 2022 to disrupt coordinated efforts to manipulate the platform, efforts that were often aimed at distributing disinformation. Please note that the cadence of this work is variable depending on the prevalence of certain types of coordinated inauthentic activity on the platform. *Investigation: 2022-09-06* *Disruption: 2022-09-23* *Actioned Assets: 149 accounts* On 6 September 2022, Meta shared 1133 Facebook accounts that they identified to be involved in coordinated inauthentic behaviour and publicly disclosed them in their quarterly report. Meta noted that a portion of this network had been described in German-language press already. Meta asserted the content that the accounts primarily shared was part of a Russian influence operation (IO). Twitter investigated the shared impersonation media entities allegedly set up by the Russian government-affiliated actors. Our analysis revealed that these accounts had a tendency to use Russian (RU) infrastructure, but no links to previously RU-origin activity. Behaviorally they also targeted RU geopolitical interests. *Investigation: 2022-09-28* *Disruption: 2022-09-30* *Actioned Assets: 15 accounts* An investigation into a Russian disinfo campaign using fake/deceptive media outlets, linked to Sputnik, and targeting European audiences. Twitter's investigation revealed that all accounts were technically linked to various Sputnik assets, showing signs of identity deception via location mismatches. The accounts occasionally attempted to obfuscate languages and locations, but content was consistently in alignment with RU geopolitical interests. *Investigation: 2022-06-30*

	Disruption: 2022-07-01 *Actioned Assets: 7 accounts* An investigation into politically motivated inauthentic behaviour that Google attributed to the IRA, that resulted in Twitter investigating and suspending some fake Russian accounts. Google's team shared additional seed information for the further investigation into potential Internet Research Agency activity which was based on pro-Putin and anti-American content. A small number of accounts were identified and suspended by Twitter.			
Measure 14.2				
QRE 14.2.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 14.2.1 – Numbers of instances and actions related to each TTP listed, enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of instances of identified TTP	Type of action taken addressing identified TTP	Other relevant metrics/information on type of content	
Member States				
List actions per member states and languages (see example table above)				
SLI 14.2.2 – for each TTP identified and action taken	Methodology of data measurement [suggested character limit: 500 characters]			
	Views/ impressions before action	Interaction/ engagement before action	Views/ impressions after action	Interaction/ engagement after action
Member States				
List actions per member states and languages (see example table above)				
SLI 14.2.3 – once available	Methodology of data measurement [suggested character limit: 500 characters]			
	Penetration and impact on genuine users	Trends on targeted audiences	Trends on narratives used	Other relevant metrics
Member States				
List actions per member states and languages (see example table above)				
SLI 14.2.4 – estimation for each TTP identified	Methodology of data measurement [suggested character limit: 500 characters]			

	TTP related content in relation to overall content on the service	Views/ impressions of TTP related content (in relation to overall views/impressions on the service)	Interaction/ engagement with TTP related content (in relation to overall interaction/engagement on the service)	Other relevant metrics
Member States				
List actions per member states and languages (see example table above)				

IV. Integrity of Services	
Commitment 15 Relevant Signatories that develop or operate AI systems and that disseminate AI-generated and manipulated content through their services (e.g. deep fakes) commit to take into consideration the transparency obligations and the list of manipulative practices prohibited under the proposal for Artificial Intelligence Act.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter has a dedicated policy on synthetic and manipulated media. It is as follows: You may not share synthetic, manipulated, or out-of-context media that may deceive or confuse people and lead to harm ("misleading media"). In addition, we may label Tweets containing misleading media to help people understand their authenticity and to provide additional context. In order for content with misleading media (including images, videos, audios, gifs, and URLs hosting relevant content) to be labelled or removed under this policy, it must: ● Include media that is significantly and deceptively altered, manipulated, or fabricated, or ● Include media that is shared in a deceptive manner or with false context, and ● Include media likely to result in widespread confusion on public issues, impact public safety, or cause serious harm We use the following criteria as we consider Tweets and media for labelling or removal under this policy as part of our ongoing work to enforce our rules and ensure healthy and safe conversations on Twitter: **1. Is the content significantly and deceptively altered, manipulated, or fabricated?**

	In order for content to be labelled or removed under this policy, we must have reason to believe that media are significantly and deceptively altered, manipulated, or fabricated. Synthetic and manipulated media take many different forms and people can employ a wide range of technologies to produce these media. Some of the factors we consider include: ● whether media have been substantially edited or post-processed in a manner that fundamentally alters their composition, sequence, timing, or framing and distorts their meaning; ● whether there are any visual or auditory information (such as new video frames, overdubbed audio, or modified subtitles) that has been added, edited, or removed that fundamentally changes the understanding, meaning, or context of the media; ● whether media have been created, edited, or post-processed with enhancements or use of filters that fundamentally changes the understanding, meaning, or context of the content; and ● whether media depicting a real person have been fabricated or simulated, especially through use of artificial intelligence algorithms We will not take action to label or remove media that have been edited in ways that do not fundamentally alter their meaning, such as retouched photos or colour-corrected videos. In order to determine if media have been significantly and deceptively altered or fabricated, we may use our own technology or receive reports through partnerships with third parties. In situations where we are unable to reliably determine if media have been altered or fabricated, we may not take action to label or remove them. **2. Is the content shared in a deceptive manner or with false context?** We also consider whether the context in which media are shared could result in confusion or suggests a deliberate intent to deceive people about the nature or origin of the content, for example, by falsely claiming that it depicts reality. We assess the context provided alongside media to see whether it provides true and factual information. Some of the types of context we assess in order to make this determination include:

	● whether inauthentic, fictional, or produced media are presented or being endorsed as fact or reality, including produced or staged works, reenactments, or exhibitions portrayed as actual events; ● whether media are presented with false or misleading context surrounding the source, location, time, or authenticity of the media; ● whether media are presented with false or misleading context surrounding the identity of the individuals or entities visually depicted in the media; ● whether media are presented with misstatements or misquotations of what is being said or presented with fabricated claims of fact of what is being depicted We will not take action to label or remove media that have been shared with commentary or opinions that do not advance or present a misleading claim on the context of the media such as those listed above. In order to determine if media have been shared in a deceptive manner or with false context, we may use our own technology or receive reports through partnerships with third parties. In situations where we are unable to reliably determine if media have been shared with false context, we will not label or remove the content. **3. Is the content likely to result in widespread confusion on public issues, impact public safety, or cause serious harm?** Tweets that share misleading media are subject to removal under this policy if they are likely to cause serious harm. Some specific harms we consider include: ● Threats to physical safety of a person or group ● Incitement of abusive behaviour to a person or group ● Risk of mass violence or widespread civil unrest ● Risk of impeding or complicating provision of public services, protection efforts, or emergency response ● Threats to the privacy or to the ability of a person or group to freely express themselves or participate in civic events, such as: ○ Stalking or unwanted and obsessive attention ○ Targeted content that aims to harass, intimidate, or silence someone else's voice ○ Voter suppression or intimidation

	We also consider the time frame within which the content may be likely to impact public safety or cause serious harm, and are more likely to remove content under this policy if immediate harm is likely to result. Tweets with misleading media that are not likely to result in immediate harm but still have a potential to impact public safety, result in harm, or cause widespread confusion towards a public issue (health, environment, safety, human rights and equality, immigration, and social and political stability) may be labelled to reduce their spread and to provide additional context. While we have other rules also intended to address these forms of harm, including our policies on violent threats, civic integrity, and hateful conduct, we will err toward removal in borderline cases that might otherwise not violate existing rules for Tweets that include misleading media.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 15.1	
QRE 15.1.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 15.2	
QRE 15.2.1	Outline relevant actions [suggested character limit: 2000 characters]

IV. Integrity of Services
Commitment 16

Relevant Signatories commit to operate channels of exchange between their relevant teams in order to proactively share information about cross-platform influence operations, foreign interference in information space and relevant incidents that emerge on their respective services, with the aim of preventing dissemination and resurgence on other services, in full compliance with privacy legislation and with due consideration for security and human rights risks.			
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter has worked in partnership with other companies and platforms for several years now, with regular engagement between Threat Disruption and Site Integrity teams in peer organisations. Some of the information operation case studies highlighted elsewhere in this report demonstrate our commitment to such ongoing partnership and collaboration.**		
If yes, list these implementation measures here [short bullet points].			
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]			
If yes, which further implementation measures do you plan to put in place in the next 6 months?			
Measure 16.1			
QRE 16.1.1	Outline relevant actions [suggested character limit: 2000 characters]		
SLI 16.1.1 – Numbers of actions as a result of information sharing	Methodology of data measurement [suggested character limit: 500 characters]		
	Nr of actions taken (total)	Type of detected content	Other relevant metrics
Member States			
List actions per member states and languages (see example table above)			
Measure 16.2			
QRE 16.2.1	Outline relevant actions [suggested character limit: 2000 characters]		

V. Empowering Users

Commitment 17

In light of the European Commission's initiatives in the area of media literacy, including the new Digital Education Action Plan, Relevant Signatories commit to continue and strengthen their efforts in the area of media literacy and critical thinking, also with the aim to include vulnerable groups.

In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter has had a global partnership with UNESCO on the promotion of media and information literacy since 2018. See blog here from Twitter: https://blog.twitter.com/en_us/topics/company/2019/twitter-launches-new-media-literacy-handbook-for-schools See blog here from UNESCO: https://www.unesco.org/en/articles/unesco-and-twitter-team-media-and-information-literacy The flagship piece of work is a resource called 'Teaching & Learning with Twitter', a resource aimed at educators seeking to offer lessons on digital and media literacy in the classroom. See link here: https://about.twitter.com/content/dam/about-twitter/en/tfg/download/teaching-learning-with-twitter-unesco.pdf This handbook was localised into over 10 languages and distributed internationally. We also launched custom emojis for the hashtags #ThinkBeforeSharing and #ThinkBeforeClicking to encourage users to pause and assess content before they distribute it on the platform, or boost the virality of mis/disinformation. Twitter looks forward to continuing this partnership with UNESCO.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 17.1	
QRE 17.1.1	Outline relevant actions [suggested character limit: 2000 characters]

SLI 17.1.1 – actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]				
	Total count of the tool's impressions	Interactions/ engagement with the tool	Other relevant metrics	Other relevant metrics	
Member States					
List actions per member states and languages (see example table above)					
Measure 17.2					
QRE 17.2.1	Outline relevant actions [suggested character limit: 2000 characters]				
SLI 17.2.1 – actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]				
	Nr of media literacy/ awareness raising activities organised/ participated in	Reach of campaigns	Nr of participants	Nr of interactions with online assets	Nr of participants (etc)
Member States					
List actions per member states and languages (see example table above)					
Measure 17.3					
QRE 17.3.1	Outline relevant actions [suggested character limit: 2000 characters]				

V. Empowering Users	
Commitment 18 Relevant Signatories commit to minimise the risks of viral propagation of Disinformation by adopting safe design practices as they develop their systems, policies, and features.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	***Product Innovation: Community Notes*** The centrepiece of Twitter's new approach to offering context and surfacing credible information is Community Notes. We believe that this product represents a fundamental shift in how we mitigate disinformation.

	Community Notes aims to create a better-informed world by empowering people on Twitter to collaboratively add helpful notes to Tweets that might be misleading. Contributors can leave notes on any Tweet and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a Tweet. We believe that Community Notes is an inherently scalable and localised response to the challenge of disinformation. By making this feature an integral and highly visible part of the Twitter product, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that can be truly global in its application. It also reduces our reliance on forms of content moderation that are more centralised, manual and bespoke; or which require intensive and time-consuming interactions with third parties. Here's how it works: ● **Contributors write and rate notes:** Contributors are people on Twitter who sign up to write and rate notes. The more people that participate, the better the program becomes. ● **Only notes rated helpful by people from diverse perspectives appear on Tweets:** Community Notes doesn't work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings. Learn more about how Community Notes handles diverse perspectives. ● **Twitter doesn't choose what shows up, the people do:** Twitter doesn't write, rate or moderate notes (unless they break the Twitter rules.) We believe giving people a voice to make these choices together is a fair and effective way to add information that helps people stay better informed. ● **Open-source and transparent:** It's important for people to understand how Community Notes works, and to be able to help shape it. The program is built on transparency: all contributions are published daily, and our ranking algorithm can

	be inspected by anyone. Learn more about how it works through our dedicated Community Notes Guide. We're keenly aware that a product like this can be vulnerable to abuse and manipulation. You can read more here on how we're thinking about quality control, guardrails, circuit breakers, and the various remediations we have in place to challenge bad actors. *Effectiveness & Research* We already know that Community Notes is effective. According to the results of four surveys run at different times between August, 2021 and August, 2022, a person who sees a Community Note is, on average, 20-40% less likely to agree with the substance of a potentially misleading Tweet than someone who sees the Tweet alone. Survey participation ranged from 3,000 to more than 19,000 participants, and the results were consistent throughout the course of the year, even as news and Tweet topics changed. We also see that Community Notes informs sharing behaviour. Analysing our internal data, we've found that a person on Twitter who sees a note is, on average, 15-35% less likely to choose to Like or Retweet a Tweet than someone who sees the Tweet alone. In our most recent survey, notes were found to be informative regardless of a person's self-identified political party affiliation — there was no statistically significant difference in average informativeness across party identification. We've published a research paper on Community Notes that you can read here. It goes into more detail on how we've been measuring efficacy. In addition, all Community Notes contributions are publicly available on the Download Data page of the Community Notes site so that anyone has free access to analyse the data, identify problems, and spot opportunities to make the product better. *Expansion & Localization*

	Community Notes are now publicly visible to everyone. Users in the US, the UK, Ireland, Canada, Australia and New Zealand can now contribute to the program. Over the coming months, users in more markets will be able to contribute notes and the product will be localised further. We currently have around 20,000 contributors and we aim to expand this number by 10% each week. Over time, users in any EU member state, writing in any language, should be able to contribute to Community Notes and the most helpful contributions will be surfaced to inform readers. Eventually, we can see a future where attempts to spread disinformation are consistently flagged by conscientious users seeking to share important context and facts with citations. The technology-first strategy evidenced by Community Notes is reflective of how we intend to approach content moderation going forward. We believe that this approach has obvious advantages over more centralised methods of content moderation, which have always faced the same two challenges: speed and scale. This is an open and transparent process. That's why we've made the Community Notes algorithm open source and publicly available on GitHub, along with the data that powers it so anyone can audit, analyse or suggest improvements.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 18.1	
QRE 18.1.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 18.1.2	

QRE 18.1.3				
SLI 18.1.1 - actions proving effectiveness of measures and policies	Methodology of data measurement [suggested character limit: 500 characters]			
	Reduction of prevalence of disinformation	Reduction of views/ impressions of disinformation	Increase in visibility of authoritative information	Other relevant metrics
Member States				
List actions per member states and languages (see example table above)				
Measure 18.2				
QRE 18.2.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 18.2.1 - actions taken in response to policy violations	Methodology of data measurement [suggested character limit: 500 characters]			
	Total no of violations	Metric 1: indicating the impact of the action taken	Metric 2: indicating the impact of the action taken	Metric 3: indicating the impact of the action taken
Member States				
List actions per member states and languages (see example table above)				
Measure 18.3				
QRE 18.3.1	Outline relevant actions [suggested character limit: 2000 characters]			

V. Empowering Users	
Commitment 19 Relevant Signatories using recommender systems commit to make them transparent to the recipients regarding the main criteria and parameters used for prioritising or deprioritising information, and provide options to users about recommender systems, and make available information on those options.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Since 2018, Twitter has allowed people to switch between a ranked timeline or a reverse-chronological feed, through what was known as the 'sparkle' icon. Twitter recently altered the Timeline to make this easier to navigate, allowing people to choose between 'For you' and 'Following'. The 'For you' timeline contains content that's recommended based on your interests and engagement on the platform. The 'Following' timeline is the classic reverse chronological experience, where you only see content from the accounts you've chosen to follow.

	We also announced that the app will soon revert to your chosen timeline each time you open it: https://twitter.com/elonmusk/status/1616594332907372544 In addition, the company has committed to open-sourcing the recommendation algorithm over the coming months: https://twitter.com/elonmusk/status/1613995936585519104 By open-sourcing the algorithm, the company will significantly increase transparency around how content is surfaced on Twitter. The company is also open to feedback on how the algorithm can be improved.			
If yes, list these implementation measures here [short bullet points].				
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]				
If yes, which further implementation measures do you plan to put in place in the next 6 months?				
Measure 19.1				
QRE 19.1.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 19.2				
SLI 19.2.1 – user settings	Methodology of data measurement [suggested character limit: 500 characters]			
	No of times users actively engaged with these settings			
Member States				
List actions per member states and languages (see example table above)				

V. Empowering Users
Commitment 20 Relevant Signatories commit to empower users with tools to assess the provenance and edit history or authenticity or accuracy of digital content.

In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter recently introduced the capacity to edit Tweets for users subscribed to Twitter Blue. When a Tweet has been edited on the platform, an annotation appears on the content to show you when it was last edited. This annotation can be clicked through to see the original version of the Tweet. See blog here: https://blog.twitter.com/en_us/topics/product/2022/twitter-new-edit-tweet-feature-only-test
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 20.1	
QRE 20.1.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 20.2	
QRE 20.2.1	Outline relevant actions [suggested character limit: 2000 characters]

V. Empowering Users	
Commitment 21 Relevant Signatories commit to strengthen their efforts to better equip users to identify Disinformation. In particular, in order to enable users to navigate services in an informed way, Relevant Signatories commit to facilitate, across all Member States languages in which their services are provided, user access to tools for assessing the factual accuracy of sources through fact-checks from fact-checking organisations that have flagged potential Disinformation, as well as warning labels from other authoritative sources.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	***Product Innovation: Community Notes*** The centrepiece of Twitter's new approach to offering context and surfacing credible information is Community Notes. We believe that this product represents a fundamental shift in how we mitigate disinformation.

	Community Notes aims to create a better-informed world by empowering people on Twitter to collaboratively add helpful notes to Tweets that might be misleading. Contributors can leave notes on any Tweet and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a Tweet. We believe that Community Notes is an inherently scalable and localised response to the challenge of disinformation. By making this feature an integral and highly visible part of the Twitter product, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that can be truly global in its application. It also reduces our reliance on forms of content moderation that are more centralised, manual and bespoke; or which require intensive and time-consuming interactions with third parties. Here's how it works: ● **Contributors write and rate notes:** Contributors are people on Twitter who sign up to write and rate notes. The more people that participate, the better the program becomes. ● **Only notes rated helpful by people from diverse perspectives appear on Tweets:** Community Notes doesn't work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings. Learn more about how Community Notes handles diverse perspectives. ● **Twitter doesn't choose what shows up, the people do:** Twitter doesn't write, rate or moderate notes (unless they break the Twitter rules.) We believe giving people a voice to make these choices together is a fair and effective way to add information that helps people stay better informed. ● **Open-source and transparent:** It's important for people to understand how Community Notes works, and to be able to help shape it. The program is built on transparency: all contributions are published daily, and our ranking algorithm can

	be inspected by anyone. Learn more about how it works through our dedicated Community Notes Guide. We're keenly aware that a product like this can be vulnerable to abuse and manipulation. You can read more here on how we're thinking about quality control, guardrails, circuit breakers, and the various remediations we have in place to challenge bad actors. *Effectiveness & Research* We already know that Community Notes is effective. According to the results of four surveys run at different times between August, 2021 and August, 2022, a person who sees a Community Note is, on average, 20-40% less likely to agree with the substance of a potentially misleading Tweet than someone who sees the Tweet alone. Survey participation ranged from 3,000 to more than 19,000 participants, and the results were consistent throughout the course of the year, even as news and Tweet topics changed. We also see that Community Notes informs sharing behaviour. Analysing our internal data, we've found that a person on Twitter who sees a note is, on average, 15-35% less likely to choose to Like or Retweet a Tweet than someone who sees the Tweet alone. In our most recent survey, notes were found to be informative regardless of a person's self-identified political party affiliation — there was no statistically significant difference in average informativeness across party identification. We've published a research paper on Community Notes that you can read here. It goes into more detail on how we've been measuring efficacy. In addition, all Community Notes contributions are publicly available on the Download Data page of the Community Notes site so that anyone has free access to analyse the data, identify problems, and spot opportunities to make the product better. *Expansion & Localization* Community Notes are now publicly visible to everyone. Users in the US, the UK, Ireland, Canada, Australia and New Zealand can now contribute to the program. Over the coming

	months, users in more markets will be able to contribute notes and the product will be localised further. We currently have around 20,000 contributors and we aim to expand this number by 10% each week. Over time, users in any EU member state, writing in any language, should be able to contribute to Community Notes and the most helpful contributions will be surfaced to inform readers. Eventually, we can see a future where attempts to spread disinformation are consistently flagged by conscientious users seeking to share important context and facts with citations. The technology-first strategy evidenced by Community Notes is reflective of how we intend to approach content moderation going forward. We believe that this approach has obvious advantages over more centralised methods of content moderation, which have always faced the same two challenges: speed and scale. This is an open and transparent process. That's why we've made the Community Notes algorithm open source and publicly available on GitHub, along with the data that powers it so anyone can audit, analyse or suggest improvements.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 21.1	
QRE 21.1.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 21.1.1 - actions taken under measure 21.1	Methodology of data measurement [suggested character limit: 500 characters]

Total impressions of fact-checks	Ratio of impressions of fact-checks to original	Reach of labels/ fact-checkers and	Other pertinent metric

		impressions of fact-checked content	other authoritative sources	
Member States				
List actions per member states and languages (see example table above)				
SLI 21.1.2 – actions taken under measure 21.1	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of articles published by independent fact-checkers	Nr of labels applied to content, such as on the basis of such articles	Meaningful metrics such as the impact of 21.1. measures on user interactions with, or user re-shares of, content fact-checked as false or misleading	
Member States				
List actions per member states and languages (see example table above)				
QRE 21.2.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 21.3				
QRE 21.3.1	Outline relevant actions [suggested character limit: 2000 characters]			

V. Empowering Users	
Commitment 22 Relevant Signatories commit to provide users with tools to help them make more informed decisions when they encounter online information that may be false or misleading, and to facilitate user access to tools and information to assess the trustworthiness of information sources, such as indicators of trustworthiness for informed online navigation, particularly relating to societal issues or debates of general interest.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	***Product Innovation: Community Notes*** The centrepiece of Twitter's new approach to offering context and surfacing credible information is Community Notes. We believe that this product represents a fundamental shift in how we mitigate disinformation. Community Notes aims to create a better-informed world by empowering people on Twitter to collaboratively add helpful notes to Tweets that might be misleading.

	Contributors can leave notes on any Tweet and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a Tweet. We believe that Community Notes is an inherently scalable and localised response to the challenge of disinformation. By making this feature an integral and highly visible part of the Twitter product, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that can be truly global in its application. It also reduces our reliance on forms of content moderation that are more centralised, manual and bespoke; or which require intensive and time-consuming interactions with third parties. Here's how it works: ● **Contributors write and rate notes:** Contributors are people on Twitter who sign up to write and rate notes. The more people that participate, the better the program becomes. ● **Only notes rated helpful by people from diverse perspectives appear on Tweets:** Community Notes doesn't work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings. Learn more about how Community Notes handles diverse perspectives. ● **Twitter doesn't choose what shows up, the people do:** Twitter doesn't write, rate or moderate notes (unless they break the Twitter rules.) We believe giving people a voice to make these choices together is a fair and effective way to add information that helps people stay better informed. ● **Open-source and transparent:** It's important for people to understand how Community Notes works, and to be able to help shape it. The program is built on transparency: all contributions are published daily, and our ranking algorithm can be inspected by anyone. Learn more about how it works through our dedicated Community Notes Guide.

	We're keenly aware that a product like this can be vulnerable to abuse and manipulation. You can read more here on how we're thinking about quality control, guardrails, circuit breakers, and the various remediations we have in place to challenge bad actors. *Effectiveness & Research* We already know that Community Notes is effective. According to the results of four surveys run at different times between August, 2021 and August, 2022, a person who sees a Community Note is, on average, 20-40% less likely to agree with the substance of a potentially misleading Tweet than someone who sees the Tweet alone. Survey participation ranged from 3,000 to more than 19,000 participants, and the results were consistent throughout the course of the year, even as news and Tweet topics changed. We also see that Community Notes informs sharing behaviour. Analysing our internal data, we've found that a person on Twitter who sees a note is, on average, 15-35% less likely to choose to Like or Retweet a Tweet than someone who sees the Tweet alone. In our most recent survey, notes were found to be informative regardless of a person's self-identified political party affiliation — there was no statistically significant difference in average informativeness across party identification. We've published a research paper on Community Notes that you can read here. It goes into more detail on how we've been measuring efficacy. In addition, all Community Notes contributions are publicly available on the Download Data page of the Community Notes site so that anyone has free access to analyse the data, identify problems, and spot opportunities to make the product better. *Expansion & Localization* Community Notes are now publicly visible to everyone. Users in the US, the UK, Ireland, Canada, Australia and New Zealand can now contribute to the program. Over the coming months, users in more markets will be able to contribute notes and the product will be localised further. We currently have around 20,000 contributors and we aim to expand this number by 10% each week.

	Over time, users in any EU member state, writing in any language, should be able to contribute to Community Notes and the most helpful contributions will be surfaced to inform readers. Eventually, we can see a future where attempts to spread disinformation are consistently flagged by conscientious users seeking to share important context and facts with citations. The technology-first strategy evidenced by Community Notes is reflective of how we intend to approach content moderation going forward. We believe that this approach has obvious advantages over more centralised methods of content moderation, which have always faced the same two challenges: speed and scale. This is an open and transparent process. That's why we've made the Community Notes algorithm open source and publicly available on GitHub, along with the data that powers it so anyone can audit, analyse or suggest improvements. **Account Labels** Labels on government accounts provide additional context for accounts heavily engaged in geopolitics and diplomacy. Labels on state-affiliated accounts provide additional context about accounts that are controlled by certain official representatives of governments, state-affiliated media entities and individuals 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 and of a podium for state-affiliated media.

If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 22.1	
QRE 22.1.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 22.1.1 – actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]
	Percentage of users that have enabled the trustworthiness indicator
Member States	
List actions per member states and languages (see example table above)	
Measure 22.2	
QRE 22.2.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 22.3	
QRE 22.3.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 22.4	
QRE 22.4.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 22.4.1 – actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]
	Volume of traffic to trustworthy sources generated thanks to the outlined trustworthiness indicators
Member States	
List actions per member states and languages (see example table above)	
Measure 22.5	
QRE 22.5.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 22.5.1 – actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]
	Total nr of instances when a publisher's rating changed from untrustworthy to trustworthy following a hearing before a rating/updated rating is issued
Member States	
List actions per member states and languages (see example table above)	

SLI 22.5.2 - actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]			
	Total nr of publishers who improved their score under the trustworthiness indicator			
Member States				
List actions per member states and languages (see example table above)				
Measure 22.6				
QRE 22.6.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 22.6.1 - actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]			
	Relevant statistics and analysis on engagement and conformity assessment			
Member States				
List actions per member states and languages (see example table above)				
Measure 22.7				
QRE 22.7.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 22.7.1 - actions enforcing policies above	Methodology of data measurement [suggested character limit: 500 characters]			
	Impressions	Clicks	CTR	Shares
Member States				
List actions per member states and languages (see example table above)				

V. Empowering Users	
Commitment 23 Relevant Signatories commit to provide users with the functionality to flag harmful false and/or misleading information that violates Signatories policies or terms of service.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Anyone can report, either in-app or through the Twitter help centre, accounts that are seeking to manipulate Twitter, including violations of our impersonation or spam policies. ***Product Innovation: Community Notes*** The centrepiece of Twitter's new approach to offering context and surfacing credible information is Community Notes. We believe that this product represents a fundamental shift in how we mitigate disinformation.

	Community Notes aims to create a better-informed world by empowering people on Twitter to collaboratively add helpful notes to Tweets that might be misleading. Contributors can leave notes on any Tweet and if enough contributors from different points of view rate that note as helpful, the note will be publicly shown on a Tweet. We believe that Community Notes is an inherently scalable and localised response to the challenge of disinformation. By making this feature an integral and highly visible part of the Twitter product, and by ensuring that the user interface is simple and intuitive, we are investing in a tool that can be truly global in its application. It also reduces our reliance on forms of content moderation that are more centralised, manual and bespoke; or which require intensive and time-consuming interactions with third parties. Here's how it works: ● **Contributors write and rate notes:** Contributors are people on Twitter who sign up to write and rate notes. The more people that participate, the better the program becomes. ● **Only notes rated helpful by people from diverse perspectives appear on Tweets:** Community Notes doesn't work by majority rules. To identify notes that are helpful to a wide range of people, notes require agreement between contributors who have sometimes disagreed in their past ratings. This helps prevent one-sided ratings. Learn more about how Community Notes handles diverse perspectives. ● **Twitter doesn't choose what shows up, the people do:** Twitter doesn't write, rate or moderate notes (unless they break the Twitter rules.) We believe giving people a voice to make these choices together is a fair and effective way to add information that helps people stay better informed. ● **Open-source and transparent:** It's important for people to understand how Community Notes works, and to be able to help shape it. The program is built on transparency: all contributions are published daily, and our ranking algorithm can be inspected by anyone. Learn more about how it works through our dedicated Community Notes Guide.

	We're keenly aware that a product like this can be vulnerable to abuse and manipulation. You can read more here on how we're thinking about quality control, guardrails, circuit breakers, and the various remediations we have in place to challenge bad actors. *Effectiveness & Research* We already know that Community Notes is effective. According to the results of four surveys run at different times between August, 2021 and August, 2022, a person who sees a Community Note is, on average, 20-40% less likely to agree with the substance of a potentially misleading Tweet than someone who sees the Tweet alone. Survey participation ranged from 3,000 to more than 19,000 participants, and the results were consistent throughout the course of the year, even as news and Tweet topics changed. We also see that Community Notes informs sharing behaviour. Analysing our internal data, we've found that a person on Twitter who sees a note is, on average, 15-35% less likely to choose to Like or Retweet a Tweet than someone who sees the Tweet alone. In our most recent survey, notes were found to be informative regardless of a person's self-identified political party affiliation — there was no statistically significant difference in average informativeness across party identification. We've published a research paper on Community Notes that you can read here. It goes into more detail on how we've been measuring efficacy. In addition, all Community Notes contributions are publicly available on the Download Data page of the Community Notes site so that anyone has free access to analyse the data, identify problems, and spot opportunities to make the product better. *Expansion & Localization* Community Notes are now publicly visible to everyone. Users in the US, the UK, Ireland, Canada, Australia and New Zealand can now contribute to the program. Over the coming months, users in more markets will be able to contribute notes and the product will be

	localised further. We currently have around 20,000 contributors and we aim to expand this number by 10% each week. Over time, users in any EU member state, writing in any language, should be able to contribute to Community Notes and the most helpful contributions will be surfaced to inform readers. Eventually, we can see a future where attempts to spread disinformation are consistently flagged by conscientious users seeking to share important context and facts with citations. The technology-first strategy evidenced by Community Notes is reflective of how we intend to approach content moderation going forward. We believe that this approach has obvious advantages over more centralised methods of content moderation, which have always faced the same two challenges: speed and scale. This is an open and transparent process. That's why we've made the Community Notes algorithm open source and publicly available on GitHub, along with the data that powers it so anyone can audit, analyse or suggest improvements.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 23.1	
QRE 23.1.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 23.2	
QRE 23.2.1	Outline relevant actions [suggested character limit: 2000 characters]

V. Empowering Users				
Commitment 24 Relevant Signatories commit to inform users whose content or accounts has been subject to enforcement actions (content/accounts labelled, demoted or otherwise enforced on) taken on the basis of violation of policies relevant to this section (as outlined in Measure 18.2), and provide them with the possibility to appeal against the enforcement action at issue and to handle complaints in a timely, diligent, transparent, and objective manner and to reverse the action without undue delay where the complaint is deemed to be founded.				
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter has had an enforcement appeals process for several years. See form here: https://help.twitter.com/en/forms/account-access/appeals/redirect For more on how the company approaches enforcement actions, including suspensions, see this article: https://help.twitter.com/en/rules-and-policies/enforcement-options#:~:text=Violators%20can%20appeal%20permanent%20suspensions,that%20the%20account%20has%20violated. Twitter will soon launch the ability for users to appeal labels and enforcement actions that affect content visibility.			
If yes, list these implementation measures here [short bullet points].				
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]				
If yes, which further implementation measures do you plan to put in place in the next 6 months?				
Measure 24.1				
QRE 24.1.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 24.1.1 - enforcement actions	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of enforcement actions	Nr of actions appealed	Metrics on results of appeals	Metrics on the duration and effectiveness of the appeal process
Member States				

List actions per member states and languages (see example table above)				

V. Empowering Users	
Commitment 25 In order to help users of private messaging services to identify possible disinformation disseminated through such services, Relevant Signatories that provide messaging applications commit to continue to build and implement features or initiatives that empower users to think critically about information they receive and help them to determine whether it is accurate, without any weakening of encryption and with due regard to the protection of privacy.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**NA**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 25.1	
QRE 25.1.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 25.1.1	Methodology of data measurement [suggested character limit: 500 characters]
	Our company would like to provide following data:
Member States	
List actions per member states and languages (see example table above)	
Measure 25.2	
QRE 25.2.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 25.2.1 - use of select tools	Methodology of data measurement [suggested character limit: 500 characters]
	Metrics on the use and impact of tools, features and campaigns deployed to meet Measures 25.2 and 25.2
Member States	

List actions per member states and languages (see example table above)	

VI. Empowering the research community	
Commitment 26 Relevant Signatories commit to provide access, wherever safe and practicable, to continuous, real-time or near real-time, searchable stable access to non-personal data and anonymised, aggregated, or manifestly-made public data for research purposes on Disinformation through automated means such as APIs or other open and accessible technical solutions allowing the analysis of said data.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter has long had an industry-leading API program. Researchers can apply for access for various levels of API access. In addition, Twitter has made several disclosures on state-backed information operations. Some of the networks included in these disclosures feature thousands of accounts. Researchers assess and analyse this data to identify the strategies and tactics of state actors in the platform space. API program link: https://developer.twitter.com/en/products/twitter-api Information Operations: https://transparency.twitter.com/en/reports/moderation-research.html Twitter has also conducted its own research into issues such as political bias in algorithmic content recommendations. See here: https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

Measure 26.1						
QRE 26.1.1	Outline relevant actions [suggested character limit: 2000 characters]					
QRE 26.1.2	Outline relevant actions [suggested character limit: 2000 characters]					
SLI 26.1.1 - e uptake of the tools and processes described in Measure 26.1	Methodology of data measurement [suggested character limit: 500 characters]					
	Nr of users of public access		Other quantitative information on public access		Other quantitative information on public access	
Member States						
List actions per member states and languages (see example table above)						
Measure 26.2						
QRE 26.2.1	Outline relevant actions [suggested character limit: 2000 characters]					
QRE 26.2.2	Outline relevant actions [suggested character limit: 2000 characters]					
QRE 26.2.3	Outline relevant actions [suggested character limit: 2000 characters]					
SLI 26.2.1 - meaningful metrics on the uptake, swiftness, and acceptance level of the tools and processes in Measure 26.2	Methodology of data measurement [suggested character limit: 500 characters]					
	No of monthly users	No of applications received	No of applications rejected	No of applications accepted	Average response time	Other metrics
Member States						
List actions per member states and languages (see example table above)						
Measure 26.3						
QRE 26.3.1	Outline relevant actions [suggested character limit: 2000 characters]					

VI. Empowering the research community	
Commitment 27 Relevant Signatories commit to provide vetted researchers with access to data necessary to undertake research on Disinformation by developing, funding, and cooperating with an independent, third-party body that can vet researchers and research proposals.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter has long had an industry-leading API program. Researchers can apply for access for various levels of API access. In addition, Twitter has made several disclosures on state-backed information operations. Some of the networks included in these disclosures feature thousands of accounts. Researchers assess and analyse this data to identify the strategies and tactics of state actors in the platform space.

	API program link: https://developer.twitter.com/en/products/twitter-api Information Operations: https://transparency.twitter.com/en/reports/moderation-research.html Twitter has also conducted its own research into issues such as political bias in algorithmic content recommendations. See here: https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 27.1	
QRE 27.1.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 27.2	
QRE 27.2.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 27.3	
QRE 27.3.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 27.3.1 - research projects vetted by the independent third-party body	Methodology of data measurement [suggested character limit: 500 characters] Nr of research projects for which they provided access to data
Member States	
List actions per member states and languages (see example table above)	
Measure 27.4	
QRE 27.4.1	Outline relevant actions [suggested character limit: 2000 characters]

VI. Empowering the research community

Commitment 28 Relevant Signatories commit to support good faith research into Disinformation that involves their services.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter has long had an industry-leading API program. Researchers can apply for access for various levels of API access. In addition, Twitter has made several disclosures on state-backed information operations. Some of the networks included in these disclosures feature thousands of accounts. Researchers assess and analyse this data to identify the strategies and tactics of state actors in the platform space. API program link: https://developer.twitter.com/en/products/twitter-api Information Operations: https://transparency.twitter.com/en/reports/moderation-research.html Twitter has also conducted its own research into issues such as political bias in algorithmic content recommendations. See here: https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 28.1	
QRE 28.1.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 28.2	
QRE 28.2.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 28.3	
QRE 28.3.1	Outline relevant actions [suggested character limit: 2000 characters]

Measure 28.4	
QRE 28.4.1	Outline relevant actions [suggested character limit: 2000 characters]

VI. Empowering the research community	
Commitment 29 Relevant Signatories commit to conduct research based on transparent methodology and ethical standards, as well as to share datasets, research findings and methodologies with relevant audiences.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**See response to Commitment 26 – also see the transparency offered in the datasets and research related to Community Notes under Commitment 18.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 29.1	
QRE 29.1.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 29.1.2	Outline relevant actions [suggested character limit: 2000 characters]
QRE 29.1.3	Outline relevant actions [suggested character limit: 2000 characters]
SLI 29.1.1 - reach of stakeholders or citizens informed about the outcome of research projects	Methodology of data measurement [suggested character limit: 500 characters] Reach of stakeholders or citizens informed about the project
Member States	
List actions per member states and languages (see example table above)	
Measure 29.2	
QRE 29.2.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 29.2.2	Outline relevant actions [suggested character limit: 2000 characters]

QRE 29.2.3	Outline relevant actions [suggested character limit: 2000 characters]
SLI 29.2.1	Methodology of data measurement [suggested character limit: 500 characters] Reach of stakeholders or citizens informed about the project
Member States	
List actions per member states and languages (see example table above)	
Measure 29.3	
QRE 29.3.1	Outline relevant actions [suggested character limit: 2000 characters]
SLI 29.3.1 - reach of stakeholders or citizens informed about the outcome of research projects	Methodology of data measurement [suggested character limit: 500 characters] Reach of stakeholders or citizens informed about the project
Member States	
List actions per member states and languages (see example table above)	

VII. Empowering the fact-checking community	
Commitment 30 Relevant Signatories commit to establish a framework for transparent, structured, open, financially sustainable, and non-discriminatory cooperation between them and the EU fact-checking community regarding resources and support made available to fact-checkers	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Not applicable.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 30.1	

QRE 30.1.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 30.1.2	Outline relevant actions [suggested character limit: 2000 characters]
QRE 30.1.3	Outline relevant actions [suggested character limit: 2000 characters]
SLI 30.1.1 - Member States and languages covered by agreements with the fact-checking organisations	Methodology of data measurement [suggested character limit: 500 characters]
	Nr of agreements with fact-checking organisations
Member States	
List actions per member states and languages (see example table above)	
Measure 30.2	
QRE 30.2.1	Outline relevant actions [suggested character limit: 2000 characters]
QRE 30.2.2	Outline relevant actions [suggested character limit: 2000 characters]
QRE 30.2.3	Outline relevant actions [suggested character limit: 2000 characters]
Measure 30.3	
QRE 30.3.1	Outline relevant actions [suggested character limit: 2000 characters]
Measure 30.4	
QRE 30.4.1	Outline relevant actions [suggested character limit: 2000 characters]

VII. Empowering the fact-checking community	
Commitment 31 Relevant Signatories commit to integrate, showcase, or otherwise consistently use fact-checkers' work in their platforms' services, processes, and contents; with full coverage of all Member States and languages.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Not applicable.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

Measure 31.1				
Measure 31.2				
QRE 31.1.1	Outline relevant actions [suggested character limit: 2000 characters]			
SLI 31.1.1 - use of fact-checks	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of fact-checked articles published	Reach of fact-checked	Nr of content pieces reviewed by fact-checkers	Other
Member States				
List actions per member states and languages (see example table above)				
SLI 31.1.2 - impact of actions taken	Methodology of data measurement [suggested character limit: 500 characters]			
	Nr of pieces of content labelled	Impact of said measures on user interactions with information labelled as false or misleading	Other	
Member States				
List actions per member states and languages (see example table above)				
SLI 31.1.3 – Quantitative information used for contextualisation for the SLIs 31.1.1 / 31.1.2	Methodology of data measurement [suggested character limit: 500 characters]			
	Denominator to be decided within the TF ahead of the baseline report			
Member States				
List actions per member states and languages (see example table above)				
Measure 31.3				
QRE 31.3.1	Outline relevant actions [suggested character limit: 2000 characters]			
Measure 31.4				
QRE 31.4.1	Outline relevant actions [suggested character limit: 2000 characters]			

VII. Empowering the fact-checking community

Commitment 32

Relevant Signatories commit to provide fact-checkers with prompt, and whenever possible automated, access to information that is pertinent to help them to maximise the quality and impact of fact-checking, as defined in a framework to be designed in coordination with EDMO and an elected body representative of the independent European fact-checking organisations.

In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**See responses above regarding access to Twitter's API program.**		
If yes, list these implementation measures here [short bullet points].			
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]			
If yes, which further implementation measures do you plan to put in place in the next 6 months?			
Measure 32.1			
Measure 32.2			
QRE 32.1.1	Outline relevant actions [suggested character limit: 2000 characters]		
SLI 32.1.1 - use of the interfaces and other tools	Methodology of data measurement [suggested character limit: 500 characters]		
	Monthly users	Other	Other
Member States			
List actions per member states and languages (see example table above)			
Measure 32.3			
QRE 32.3.1	Outline relevant actions [suggested character limit: 2000 characters]		

VII. Empowering the fact-checking community	
Commitment 33 Relevant Signatories (i.e. fact-checking organisations) commit to operate on the basis of strict ethical and transparency rules, and to protect their independence.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Not applicable.**
If yes, list these implementation measures here [short bullet points].	

Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]		
If yes, which further implementation measures do you plan to put in place in the next 6 months?		
Measure 33.1		
QRE 33.1.1	Outline relevant actions [suggested character limit: 2000 characters]	
SLI 33.1.1 - number of European fact-checkers that are IFCN-certified	Methodology of data measurement [suggested character limit: 500 characters]	
	Nr of fact-checkers IFCN-certified	Nr of members of CPI
Member States		
List actions per member states and languages (see example table above)		

VIII. Transparency Centre	
Commitment 34 To ensure transparency and accountability around the implementation of this Code, Relevant Signatories commit to set up and maintain a publicly available common Transparency Centre website	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter confirms its commitment.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 34.1	
Measure 34.2	

Measure 34.3	
Measure 34.4	
Measure 34.5	

VIII. Transparency Centre	
Commitment 35 Signatories commit to ensure that the Transparency Centre contains all the relevant information related to the implementation of the Code's Commitments and Measures and that this information is presented in an easy-to-understand manner, per service, and is easily searchable.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter's commitment was fulfilled.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 35.1	
Measure 35.2	
Measure 35.3	
Measure 35.4	
Measure 35.5	
Measure 35.6	

VIII. Transparency Centre

Commitment 36 Signatories commit to updating the relevant information contained in the Transparency Centre in a timely and complete manner.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter's commitment was fulfilled.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 36.1	
Measure 36.2	
Measure 36.3	
QRE 36.1.1 (for the Commitments 34-36)	Outline relevant actions [suggested character limit: 2000 characters]
QRE 36.1.2 (for the Commitments 34-36)	Outline relevant actions [suggested character limit: 2000 characters]
SLI 36.1.1 - (for Measures 34 and 36) meaningful quantitative information on the usage of the Transparency Centre, such as the average monthly visits of the webpage.	Methodology of data measurement [suggested character limit: 500 characters] Our company would like to provide following data:
Member States	
List actions per member states and languages (see example table above)	

IX. Permanent Task-Force
Commitment 37

Signatories commit to participate in the permanent Task-force. The Task-force includes the Signatories of the Code and representatives from EDMO and ERGA. It is chaired by the European Commission, and includes representatives of the European External Action Service (EEAS). The Task-force can also invite relevant experts as observers to support its work. Decisions of the Task-force are made by consensus.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter has participated in the Task-Force and relevant subgroups.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 37.1	
Measure 37.2	
Measure 37.3	
Measure 37.4	
Measure 37.5	
Measure 37.6	
QRE 37.6.1	Outline relevant actions [suggested character limit: 2000 characters]

X. Monitoring of Code	
Commitment 38 The Signatories commit to dedicate adequate financial and human resources and put in place appropriate internal processes to ensure the implementation of their commitments under the Code.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter is in the process of recalibrating its resources with a focus on our development of Community Notes as a central user and product-focused disinformation risk mitigation commitment, and compliance with the Digital Services Act.

If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 38.1	
QRE 38.1.1	Outline relevant actions [suggested character limit: 2000 characters]

X. Monitoring of Code	
Commitment 39 Signatories commit to provide to the European Commission, within 1 month after the end of the implementation period (6 months after this Code's signature) the baseline reports as set out in the Preamble.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**Twitter's commitment was fulfilled.**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

X. Monitoring of Code
Commitment 40

Signatories commit to provide regular reporting on Service Level Indicators (SLIs) and Qualitative Reporting Elements (QREs). The reports and data provided should allow for a thorough assessment of the extent of the implementation of the Code's Commitments and Measures by each Signatory, service and at Member State level.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter is engaging with the relevant stakeholders as to the best way to provide details on Twitter's compliance with the DSA.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 40.1	
Measure 40.2	
Measure 40.3	
Measure 40.4	
Measure 40.5	
Measure 40.6	

X. Monitoring of Code	
Commitment 41 Signatories commit to work within the Task-force towards developing Structural Indicators, and publish a first set of them within 9 months from the signature of this Code; and to publish an initial measurement alongside their first full report. To achieve this goal, Signatories commit to support their implementation, including the testing and adapting of the initial set of Structural Indicators agreed in this Code. This, in order to assess the effectiveness of the Code in reducing the spread of online disinformation for each of the relevant Signatories, and for the entire online ecosystem in the EU and at Member State level. Signatories will collaborate with relevant actors in that regard, including ERGA and EDMO.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter is engaging with the relevant stakeholders as to the best way to provide details on Twitter's compliance with the DSA.

If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	
Measure 41.1	
Measure 41.2	
Measure 41.3	

X. Monitoring of Code	
Commitment 42 Relevant Signatories commit to provide, in special situations like elections or crisis, upon request of the European Commission, proportionate and appropriate information and data, including ad-hoc specific reports and specific chapters within the regular monitoring, in accordance with the rapid response system established by the Taskforce.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	**NA this period**
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

X. Monitoring of Code

Commitment 43	
Signatories commit to produce reports and provide data following the harmonised reporting templates and refined methodology for reporting and data disclosure, as agreed in the Task-force.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter is engaging with the relevant stakeholders as to the best way to provide details on Twitter's compliance with the DSA.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	
If yes, which further implementation measures do you plan to put in place in the next 6 months?	

X. Monitoring of Code	
Commitment 44	
Relevant Signatories that are providers of Very Large Online Platforms commit, seeking alignment with the DSA, to be audited at their own expense, for their compliance with the commitments undertaken pursuant to this Code. Audits should be performed by organisations, independent from, and without conflict of interest with, the provider of the Very Large Online Platform concerned. Such organisations shall have proven expertise in the area of disinformation, appropriate technical competence and capabilities and have proven objectivity and professional ethics, based in particular on adherence to auditing standards and guidelines.	
In line with this commitment, did you deploy new implementation measures (e.g. changes to your terms of service, new tools, new policies, etc)? [Yes/No]	Twitter is engaging with the relevant stakeholders as to the best way to provide details on Twitter's compliance with the DSA.
If yes, list these implementation measures here [short bullet points].	
Do you plan to put further implementation measures in place in the next 6 months to substantially improve the maturity of the implementation of this commitment? [Yes/No]	

If yes, which further implementation measures do you plan to put in place in the next 6 months?	

Reporting on the service's response during a period of crisis			
Covid-19 pandemic			
Overview of the main threats observed, such as crisis related disinformation campaigns, spread of misinformation, coordinated manipulative behaviours, malicious use of advertising products, involvement of foreign state actors, etc.: [suggested character limit: 2000 characters].			
Executive summary of the company's main strategies and actions taken to mitigate the identified threats and react to the crisis: [suggested character limit: 2000 characters].			
Best practices identified for future crisis situations: [suggested character limit: 2000 characters].			
Future measures planned within the next six months: [suggested character limit: 2000 characters].			
[Note: Signatories are requested to provide information relevant to their particular response to the threats and challenges they observed on their service(s). They ensure that the information below provides an accurate and complete report of their relevant actions. As operational responses to crisis situations can vary from service to service, an absence of information should not be considered a priori a shortfall in the way a particular service has responded. Impact metrics are accurate to the best of signatories' abilities to measure them].			
Changes in Policy Framework			
	Policies	**Rationale**	
Policies newly introduced for addressing the crisis			
Policies adapted for addressing the crisis	**Twitter's policy regarding misleading information for COVID-19 was deprecated on November 23, 2022.**		
Actions to mitigate the crisis impact on the service			
Type of mitigation	**Intervention or action (short summary)** [suggested character limit: 500 characters]	**Intervention or action (explanation and implementation)** [suggested character limit: 2000 characters]	**Impact metrics**

Actions taken against dis- and misinformation content (for example deamplification, labelling, removal etc.)	*Intervention applied*	*Implementation and enforcement action(s) corresponding to Intervention*	*Performance Metrics (either as a subset of SLI 18.1.1. or 18.1.2. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Promotion of authoritative information, including via recommender systems and products and features such as banners and panels	*Source promoted, Product deployed or Initiative taken*	*Implementation measures*	*Performance Metrics (either as a subset of SLI 18.2.1. or 22.7.1. for the crisis context, or other meaningful performance metrics available for the referenced intervention)*
Cooperation with independent fact-checkers in the crisis context, including coverage in the EU	*Implementation measure (Agreement with fact-checker in Member state)*	*Approach of cooperation with Fact-checkers in specific country*	*Performance Metrics (either as a subset of SLI 21.1.1., 21.1.2., 30.1.1., 31.1.1., 31.1.2., or 32.1.1. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Measures taken to demonetise disinformation related to the crisis	*Intervention applied*	*Implementation measures*	*Performance Metrics (either as a subset of SLI 1.1.1., 1.2.1. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Measures taken to prevent malicious advertising	*Intervention applied*	*Implementation measure*	*Performance metrics (either as a subset of SLI 2.1.1., 2.3.1., 2.4.1. for the crisis context or*

			other meaningful performance metrics available for the referenced intervention)
Measures taken in the context of the crisis to counter manipulative behaviours/TTCs	*Intervention applied*	*Implementation measures*	*Performance metrics (either as a subset of SLI 14.1.1., 14.2.1., 14.2.2., 14.2.3., 14.2.4. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Measures taken to support research into crisis related misinformation and disinformation	*Program supported*	*Implementation measures*	*Performance Metrics (either as a subset of SLI 26.1.1., 26.2.1. or 27.3.1. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Relevant changes to working practices to respond to the demands of the crisis situation and/or additional human resources procured for the mitigation of the crisis	*Changes to working practices*	*Actions carried out*	*Any available meaningful metrics for the referenced changes*

Reporting on the service's response during a period of crisis

War of aggression by Russia on Ukraine

Below is an overview of Twitter's response to the Russia-Ukraine conflict and its manifestation on the platform. A version of the content below was shared in a letter to the Commission on December 21, 2022. With the development and continued rollout of Community Notes, the company's strategy with respect to disinformation in conflict environments will be subject to change in the future.

Finally, see our input for QRE 14.1.2 where we provide detail on several information operations that are also relevant to this section.

Our Approach to Coordinated Inauthentic Activity

We use both manual and automated reviews to identify and remove inauthentic coordinated behaviour related to the current war in Ukraine. We also use a combination of technology and human review to proactively identify misleading information. More than 65% of violative content is surfaced by our automated systems, and the majority of remaining content we enforce on is surfaced through regular monitoring by our internal

teams and our work with trusted partners. Early detection, manual review, and automation of detection/enforcement have allowed us to effectively stop inauthentic coordinated behaviour from reaching a broader audience. For example, in the months after the outbreak of the conflict, we removed more than 75,000 accounts through proactive screening to curb platform manipulation. Note that these accounts represent a wide range of attempts at platform manipulation, including opportunistic, financially-motivated spam and attempts to fraudulently solicit donations (particularly in cryptocurrency) — and don't represent a specific, coordinated campaign associated with a government actor. Our teams continue to take action on such content under our Terms of Service.

Our Approach to Synthetic & Manipulated Media

All parties are keen to promote their narratives, resulting in a high volume and velocity of information. We noted a significant increase in escalations on our Synthetic and Manipulated Media policy (video game footage, videos/images of other global conflicts, or military manoeuvres). These are typically intended to mislead the public about the conflict, or unintentionally shared without verification. While this content often originates on other platforms, some of it has circulated in a misleading context on Twitter. We will remove content from the platform based on an assessment of harms associated with the media. Reviews of violations of this policy are extremely time-consuming and we had more than 44,000 cases labelled or removed for violating our policy in the first month or so of the conflict.

Our Approach to State Propaganda

We endeavour to ensure people on Twitter have as much context as possible about the conversations they're seeing. This includes labels on government and state-affiliated media accounts to give people context about the Tweet they're seeing. 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. Twitter will not recommend or amplify state-affiliated media entities accounts or their Tweets with these labels to people. Since August 2020, we've labelled and de-amplified state-affiliated accounts belonging to the Russian Federation, in addition to 20 other countries to provide important context about who they represent. We expanded the list of outlets and countries again in 2021. We continue to review and update the lists of accounts labelled as Russian affiliated state media to reflect outlets' use of Twitter and the creation of new accounts. Twitter currently has 100 state-affiliated media labels in place for Russian media.

Earlier this year, we saw more than 45,000 Tweets a day from individuals on Twitter sharing these links to such state-affiliated media — meaning that the overwhelming majority of content from state-affiliated media is coming from individuals sharing this content, rather than accounts we've been labelling for years. We made the decision to expand our policy and apply a label to Tweets sharing links to designated state-affiliated

media outlets. These Tweets sharing state-affiliated media content won't be amplified — they won't appear in Top Search and won't be recommended by Twitter.

The European Union (EU) sanctions legally require us to withhold certain content in EU member states, and we are complying accordingly.

Our Approach to Information Operations Detection & Data Access

Since 2018, Twitter has provided industry-leading access to data about government-backed platform manipulation campaigns, sharing 37 datasets of attributed platform manipulation campaigns originating from 17 countries, spanning more than 200 million Tweets and nine terabytes of media. Should we find evidence to suggest that any inauthentic coordinated behaviour is the result of a state actor's efforts, we have historically published our findings to Twitter's information operations archive. Since 2006, academic researchers have used data from the public conversation to study topics as diverse as the conversation on Twitter itself - from state-backed efforts to disrupt the public conversation to floods and climate change, from attitudes and perceptions about COVID-19 to efforts to promote healthy conversation online. Today, academic researchers are one of the largest groups of people using the Twitter API.

Our Approach to Monetization

Content that discusses or focuses on the Russia-Ukraine conflict is not eligible for monetization under Twitter's Brand Safety Policy. Content that is considered false or misleading under the Twitter Rules is also not eligible for monetisation. Additionally, we are demonetizing Search terms related to the Russia-Ukraine conflict, preventing ads from appearing on the Search results pages for these words. Beyond not recommending or amplifying accounts or Tweets of State-affiliated media, advertisements and the promotion of content from state affiliated news media is also prohibited on Twitter. On Russia Today (RT) and Sputnik specifically, in 2017 Twitter made the policy decision to ban advertising from all accounts owned by RT and Sputnik. This decision was based on the retrospective work we did around the 2016 US election and the US intelligence community's conclusion that both RT and Sputnik attempted to interfere with the election on behalf of the Russian government. Twitter re-invested the $1.9 million in revenue from RT advertising into external research on civic integrity in the online space.

Further Proactive Steps

In addition to the measures noted above, we have taken a number of additional proactive measures to protect the health of the service, including:

- We review Tweets to detect platform manipulation (or other inauthentic behaviour) and take enforcement action against synthetic and manipulated media that presents a false or misleading depiction of what's happening.

- We're actively monitoring vulnerable high-profile accounts, including journalists, activists, and government officials and agencies to mitigate any attempts at a targeted takeover or manipulation. We have proactively scanned our service for accounts in the range of relevant Russian/Ukrainian individuals, organisations, and messages that may violate our impersonation policy. As a result of this manual review, we have suspended over 1,000 accounts under this policy.

- For people using Twitter in Ukraine and Russia, we also paused some Tweet recommendations from people they don't follow on Home Timeline to reduce the spread of abusive content.

- We're working across features like Topics, Lists, and Spaces to ensure the policies and measures in place can ensure the safety of these products, so they can continue to be resources people trust.

[Note: Signatories are requested to provide information relevant to their particular response to the threats and challenges they observed on their service(s). They ensure that the information below provides an accurate and complete report of their relevant actions. As operational responses to crisis situations can vary from service to service, an absence of information should not be considered a priori a shortfall in the way a particular service has responded. Impact metrics are accurate to the best of signatories' abilities to measure them].

Changes in Policy Framework		
	Policies	**Rationale**
Policies newly introduced for addressing the crisis		
Policies adapted for addressing the crisis		

Actions to mitigate the crisis impact on the service			
Type of mitigation	**Intervention or action (short summary)** [suggested character limit: 500 characters]	**Intervention or action (explanation and implementation)** [suggested character limit: 2000 characters]	**Impact metrics**

Actions taken against dis- and misinformation content (for example deamplification, labelling, removal etc.)	*Intervention applied*	*Implementation and enforcement action(s) corresponding to Intervention*	*Performance Metrics (either as a subset of SLI 18.1.1. or 18.1.2. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Promotion of authoritative information, including via recommender systems and products and features such as banners and panels	*Source promoted, Product deployed or Initiative taken*	*Implementation measures*	*Performance Metrics (either as a subset of SLI 18.2.1. or 22.7.1. for the crisis context, or other meaningful performance metrics available for the referenced intervention)*
Cooperation with independent fact-checkers in the crisis context, including coverage in the EU	*Implementation measure (Agreement with fact-checker in Member state)*	*Approach of cooperation with Fact-checkers in specific country*	*Performance Metrics (either as a subset of SLI 21.1.1., 21.1.2., 30.1.1., 31.1.1., 31.1.2., or 32.1.1. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Measures taken to demonetise disinformation related to the crisis	*Intervention applied*	*Implementation measures*	*Performance Metrics (either as a subset of SLI 1.1.1., 1.2.1. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Measures taken to prevent malicious advertising	*Intervention applied*	*Implementation measure*	*Performance metrics (either as a subset of SLI 2.1.1., 2.3.1., 2.4.1. for the crisis context or*

			other meaningful performance metrics available for the referenced intervention)
Measures taken in the context of the crisis to counter manipulative behaviours/TTCs	*Intervention applied*	*Implementation measures*	*Performance metrics (either as a subset of SLI 14.1.1., 14.2.1., 14.2.2., 14.2.3., 14.2.4. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Measures taken to support research into crisis related misinformation and disinformation	*Program supported*	*Implementation measures*	*Performance Metrics (either as a subset of SLI 26.1.1., 26.2.1. or 27.3.1. for the crisis context or other meaningful performance metrics available for the referenced intervention)*
Relevant changes to working practices to respond to the demands of the crisis situation and/or additional human resources procured for the mitigation of the crisis	*Changes to working practices*	*Actions carried out*	*Any available meaningful metrics for the referenced changes*