Twitter January update: Code of Practice on Disinformation

In light of the upcoming European Parliament elections, we welcome the opportunity to provide updates on the progress made on our commitments to the European Commission Code of Practice on Disinformation and we will continue to strengthen Twitter against attempted manipulation, spam and disinformation, and abuse. The people who use our service must have confidence in the integrity of the information found on the platform, especially with respect to information relevant to elections and the democratic process. We continue our efforts to address the threats posed by hostile domestic and foreign actors, and work to foster an environment conducive to healthy, meaningful conversations on our service.

We intend to provide greater transparency leading up to the European elections on our own efforts around election integrity, including how we are working with our partners, including governments, industry peers, civil society, and research partners.

In each month preceding the European Parliament election, we will provide a report outlining our efforts in that month, not every report will focus on the same measures, but collectively they will provide a comprehensive view of our approach to the commitments we have made to ensuring the integrity of our services and enhancing access to healthy democratic discourse around the EU elections. This report will elucidate upon our work throughout January 2019 and provide insights into upcoming progress in the coming months.

Developments in January

- **Transparency on lessons learnt:** We continue to build on past efforts around election integrity leading up to the EU elections. In January, we shared a comprehensive review of our efforts to protect the integrity of the public conversation on Twitter regarding the 2018 US midterm elections including new threats, domestic and foreign information operations, partnerships,and insights into political advertising.
- **Investigating potential foreign information operations on Twitter:** In January, we added five new account sets to our archive of potential foreign information operations on Twitter, which we found based on continued contextual and semantic analysis from our investigations teams, one of the core components in our effort to protect the integrity of our service.
- **Empowering further research into potential information operations:** The archive of accounts and content we have made publically available has been accessed by thousands of researchers, governments, and people interested in learning more about foreign information operations. We have already seen valuable reporting and research produced on the basis of these datasets, including by major universities and think tanks in the EU and around the world, and will provide vital learning opportunities leading up to and following the EU elections.

- **New research partners to improve Machine Learning:** In January we partnered with researchers to establish a new initiative focused on studying and improving the performance of machine learning in social systems, such as the impact of algorithmic decisions and machine learning models on behaviours and social dynamics online.

Below you will find more details on our updates.

Scrutiny of ad placements

Twitter is committed to ensuring that promoted accounts and paid advertisements are free from bad faith actors, including foreign state actors seeking to manipulate our service around the EU elections. The key principles of our advertising policy are rooted in keeping users safe and promoting honest content to ensure high editorial standards for the Twitter Ads content created across the EU.

We continue to engage with advertisers to keep them informed of about the Twitter Ads processes and brand safety measures, and through these monthly reports, provide updates when available in these areas.

Transparency of advertising

Expansion of the Ads Transparency Centre

Our team continues to work on the expansion of the Ads Transparency Centre (ATC) across Europe. This will provide transparency for political advertising ahead of the European Elections. In February, we will announce the next steps to this process to ensure that all stakeholders within the EU have a clear understanding of the implementation process for the EU leading up to the May elections. Below we have provided some details and an overview of how our ATC has been implemented in the U.S. to provide a better understanding of the product leading up to the EU launch.

Following the original launch of the ATC, and from our work on the U.S Midterm Elections, we have been able to provide comprehensive data on paid electioneering communications on the service, this includes purchases made by a specific account, all past and current ads served on the service for a specific account, targeting criteria and results for each advertisement, the number of views each advertisement received, and certain billing information associated with the account, thereby increasing transparency and promoting accountability in the ads served to Twitter customers.

To illustrate the type and level of details we will be able to provide regarding political advertising in the European Elections, as of December 3, 2018, Twitter received 407 applications to register as political advertisers and fully approved 236 in the United States. We also received 28 applications deemed to be a news outlet thereby exempting the organisations from the Ads Transparency Center requirements.

In 2018, 96 political advertisers spent nearly 2.3 million USD to purchase 2,267 ads that resulted in nearly 170 million impressions.

Our monthly reports will provide updates and transparency on our process going forward over the coming months and once the Ads Transparency Center has been rolled out across Europe, the information and data will be accessible to both Twitter users and non-Twitter users.

Tackling malicious actors on Twitter (integrity of services)

What we learned: Lessons from 2018 U.S. Midterm Election in build on for the EU 2019 elections

To better inform our ongoing efforts to protect the integrity of the public conversation on Twitter in light of the upcoming European Elections, we published the results of an in depth review of our work in the 2018 U.S. midterm elections in January. This has been a highly valuable exercise as it provides vital insight in to potential threats in the context of elections, and has helped equip us with the right tools and perspectives to prepare for any potential challenges, such as we may face in the European Elections. By evaluating challenges such as foreign interference in the U.S Midterm Election and scruitising the efficacy of our responses, we have reinforced our preparedness for the European Elections.

The public conversation occuring on Twitter is never more important than during elections. Our service shows the world what is happening, democratises access to information and — at its best — provides people around the globe with insights into a diversity of perspectives on critical election issues. Any attempts to undermine the integrity of our service erode the core tenets of freedom of expression online, the value upon which our company is based. This issue affects all of us and is one that we care deeply about as individuals, both inside and outside the company.

What we have derived from this exercise that can be applied to our approach to the integrity of the public conversation surrounding the upcoming European Elections is that our efforts need to be collaborative, indeed as the internet evolves, so too do the challenges and opportunities society faces.

- Collaborative partnerships with governmental bodies, law enforcement agencies, civil society, and our peer companies make us better.
- Our greatest partner in making the service healthier continues to be the public who challenge us, hold us accountable, and bring potentially problematic content to our attention. We also need to be agile in our response as the tactics of bad actors are ever evolving.
- In the case of the U.S. midterm elections we identified that the dominant threat was voter suppressive content of primarily domestic origin, we need to remain one step ahead of potential new challenges that will arise from within the EU.
- In contrast to 2016 US elections, we identified much less platform manipulation from bad-faith actors located abroad. That said, as part of our ongoing review we found limited operations that have the potential to be connected to sources within Iran, Venezuela, and Russia. The majority of these accounts were proactively suspended in advance of Election Day due to the increasingly robust nature of our technology and internal tooling for identifying platform manipulation.
- Furthermore, we employ a cross-functional approach to our work on elections integrity, and draw expertise from a wide variety of teams, the diversity of perspective and backgrounds is absolutely critical. We all care deeply about elections and work passionately to protect the service. From engineering to data science to legal, these enforcements and disclosures touch upon many core areas of our company.

As we move forward, we will continue with this model of bringing in additional expertise and personnel who can augment our approach, growing the level of experience from one critical election to the next.We hope this will provide useful insight and best practises that can be applied in the context of the upcoming European Elections.

Investigating potential foreign information operations on Twitter

Through the transparency of our actions and access to information, we aim to help improve the public understanding of alleged foreign influence campaigns, this will enable users to better discern the origin and intent of various types of content, and will be crucial to our work around the EU elections in May.

In the latter half of January we added five new account sets to our archive of potential foreign information operations on Twitter. This is a publicly available archives of Tweets and media that we believe resulted from potentially state-backed information operations on our service. These accounts were found based on continued contextual and semantic analysis from our investigations teams, one of the core components in our effort to protect the integrity of our service. The datasets are comprised of, **2617** accounts in Iran, **15** accounts in Bangladesh, **418** accounts in Russia and two account sets from Venezuela comprised of **1196** and **764** accounts respectively.

These cases are illustrative of potential foreign information operations such as attempted influence campaigns and coordinated platform manipulation that could be deployed in any geography, and will provide vital learning opportunities for Europe.

Manipulation of information for national or geopolitical ends is part of human history and transcends ideological viewpoints. The medium of communication is what has changed. The behavior is against our values as a company. For our part, we are learning, evolving, and building a technological and personnel-driven approach to combating it. We hope that holistic, transparent disclosures such as this can help us all learn and build the necessary societal defenses and capacities to protect public conversation.

The process of investigating suspected foreign influence and information campaigns is an ongoing one. Although the volume of malicious election-related activity that we could link to Russia was relatively small, we strongly believe that any such activity on Twitter is unacceptable. We remain vigilant about identifying and eliminating abuse on the service perpetrated by hostile foreign actors, and we will continue to invest in resources and leverage our technological capabilities to do so.

Empowering the research community

Empowering further research into potential information operations

In line with our principles of transparency and to improve public understanding of alleged foreign influence campaigns, we made the aforementioned archive of potential foreign information operations on Twitter public and searchable. This is so that members of the public, governments, and researchers can investigate, learn, and build media literacy capacities for the future. So far, this archive of accounts and content has been accessed by thousands of researchers, governments, and people interested in learning more about foreign information operations. We have already seen valuable reporting and research produced on the basis of these datasets.

If we identify additional attempted information operations on Twitter in the future, we will release similar datasets in a timely fashion after we complete our investigations. We may also release incremental additions to existing datasets if we believe the additional information could materially impact research findings.

New research partners to improve Machine Learning

In January we partnered with researchers to establish a new research initiative focused on studying and improving the performance of Machine Learning in social systems (such as Twitter). The initiative will be lead by Professor Moritz Hardt and Professor Ben Recht. The team at UC Berkeley will closely collaborate with a corresponding team inside Twitter. As a company, Twitter is able to bring data and real-world insights to the table, but by partnering with UC Berkeley we can create a research program that has the right mix of fundamental and applied research components to make a real practical impact across industry.

Machine Learning plays a key role in powering Twitter. From onboarding users on the platform to preparing their

timeline and everything in between, a multitude of machine learning models help power the experience. Thus, making Twitter more healthy requires making the way we practice machine learning more fair, accountable and transparent.

The consequences of exposing algorithmic decisions and machine learning models to hundreds of millions of people are poorly understood. Even less is known about how these algorithms might interact with social dynamics: people might change their behaviour in response to what the algorithms recommend to them, and as a result of this shift in behaviour the algorithm itself might change, creating a potentially self-reinforcing feedback loop. We also know that individuals or groups will seek to game or exploit our algorithms and safeguarding against this is essential.

By bringing together academic expertise with our industry perspective, we are looking to do fundamental work in this nascent space and apply it to improve Twitter. This will have a positive influence platform wide.