Current State Assessment

1.0 -- Executive Summary

Alethea Group was engaged by Twitter to evaluate the state and structure of Twitter's capabilities in countering misinformation and disinformation, with the goal of identifying gaps in its processes, policies, and approach, as well as opportunities to build the organization's ability to safeguard the platforms and its users. This report details the current state of Twitter's misinformation and disinformation capabilities as identified by Alethea Group, based upon internal documents reviewed, stakeholder interviews, and other information gathered as needed. A subsequent report, based on the findings contained in this report, will be delivered in two weeks from final acceptance from the Client in order to make recommendations for how to mature the organization's capabilities to address misinformation and disinformation globally.

Broadly, our assessment found that organizational siloing, a lack of investment in critical resources, and reactive policies and processes have driven Twitter to operate in a constant state of crisis that does not support the company's broader mission of protecting authentic conversation. As a result, Twitter is consistently behind the curve in actioning against disinformation and misinformation threats. Teams identified significant gaps in resource allocation, leading to policies and actions that are often reactive in nature and do not allow the company to think about emerging threats. Twitter does not have a traditional threat intelligence capability that would better position the company to be proactive on misinformation and disinformation and to protect authentic conversation. Ultimately, these gaps mean that although Twitter is a global company with a global mission, it is not currently set up to deliver globally on trust and safety.

Different incentives for different teams working on misinformation and disinformation means Twitter is set up to be reactive, and although it has beneficial partnerships with other social media companies and research institutions, they do not allow Twitter to do proactive analysis that is reflective of the actual threat landscape on the platform or reflective of Twitter's business objectives. These gaps illustrate the extent to which product and growth are prioritized over online user and platform safety. Twitter further lacks sufficient mechanisms to measure progress and impact, therefore it may not be accurately measuring progress or it could be failing to implement lessons learned from the past.

Tools available to Site Integrity to work on these issues are often outdated, "hacked together," or difficult to use, limiting Twitter's ability to effectively enforce policies at scale. A lack of automation and sophisticated tooling means that Twitter relies on human capabilities, which are not adequately staffed or resourced, to address the misinformation and disinformation problem. Further, policies are often written in response to external events, or "fires," rather than being informed by analysis of the current or emerging threats for the platform, without an effective enforcement mechanism and tooling in place. Because policy changes are often implemented

quickly, they often do not incorporate feedback from relevant stakeholders, are not well-executed, and difficult to enforce at scale.

Our assessment found that Site Integrity teams lack diversity, especially gender diversity, across the analytical and managerial level. Additionally, the lack of diverse backgrounds among employees contributed to gaps in foreign-language and on-the-ground contextual capabilities, hindering Twitter's ability to execute its mission and remove harmful content worldwide. Teams in priority growth markets either do not exist, or are not sufficiently staffed or resourced.

Our assessment found that employees in this space are supportive of Twitter's mission and the organization, and have positive perceptions of their teams, teammates, and managers. Despite the challenging subject matter and circumstances, including employees reporting burnout because of a lack of resources, interviewees described managers as receptive to feedback and concerns, and a positive team culture of pulling together to get the work done. The team appears to be dedicated to their mission, believes that Twitter can achieve its goals, and articulated the desire to see the team through this upcoming period of growth.

2.0 Methodology

In order to conduct the current state assessment, Alethea Group interviewed 12 members of Twitter's Trust & Safety, Twitter Services, and Product & Engineering teams , conducted screen sharing exercises to understand Twitter's internal misinformation/disinformation tooling and processes, and reviewed a series of 19 internal documents, retrospectives, and training guides. This assessment does not seek to comprehensively address Twitter's performance, capabilities, or work during the US 2020 election.

3.0 Current State Assessment

3.1 Organization

3.1.1 -- The organizational structure within Twitter that responds to disinformation and misinformation is siloed and not clearly defined. The capabilities were built in an ad hoc manner largely in response to crises. This has contributed to organizational silos, capabilities gaps, and created a culture in which employees must rely on informal relationships across the organization to accomplish work.

Currently, Twitter does not have a clearly defined organization to encompass the functions or offices at Twitter that are dedicated to detecting and mitigating platform harms, and does not

have the ad hoc structures documented in order to support formalization of functions and offices. Efforts to combat misinformation and disinformation on the platform have evolved in an ad hoc manner as a result of external factors, such as the 2016 elections, coronavirus pandemic, and other pressing threats. Because of the ad hoc nature, the informal organization is driven by policy decisions made in a silo, mostly by San Francisco-based staff, and frequently during a time of crisis.

This has consistently meant that relevant entities do not have the opportunity to engage other parts of the organization and key stakeholders responsible for countering disinformation, leading to policies that may be unenforceable at scale or not reflective of the threat landscape on and off the platform. Interviewees said this has also meant that historically, teams across the organization have been inconsistent or slow to respond, especially to information risks or threats that are not easily defined, such as the evolution of QAnon or cases of coronavirus misinformation.

Without a formal organizational structure in the misinformation and disinformation problem set, the holistic solutions required to mature functions that combat platform manipulation are not sufficiently resourced.

3.1.1.1-- Site Integrity, which is responsible for platform policy and enforcement related to platform manipulation matters, works with Health and Twitter Services to collaborate on tools, technical fixes, and policy enforcement, but they lack formal processes and structures that facilitate easy identification of roles and responsibilities and instead rely upon informal cross-functional relationships.

Trust and Safety functions exist in a silo within Security, while actually impacting all parts of the Twitter platform and experience. As a result, Twitter is making critical decisions about new products and product launches without being prepared to mitigate security concerns.

Different parts of the organization are working different pieces of the problem set, but interviewees described a very insular process for their respective teams in which there is a lack of meaningful coordination with other relevant teams and no official mechanism, such as formal working agreements between teams, outlining their authorities and responsibilities to each other. While Site Integrity is responsible for drafting policy, they are unable to adequately respond to threats or enforce new policies at scale because other components of Twitter are not meaningfully engaged. Historically, policies have been created during a crisis or in response to a major platform failure to address misinformation or disinformation, instead of proactively.

Many interviewees credit informal relationships for their ability to make any progress or be able to seek support from other teams, whether engineering or product support. Organizational effectiveness appears to be based on the ability to navigate Twitter versus an intentional organization determined by Twitter's leadership, given the necessary resources and support to achieve its mission.

Example: In one instance, Twitter planned to launch its new product, Fleets, just weeks before the US 2020 election when resources had been pulled from other duties to address the high-profile, high threat election. While SI team members said that they had been involved in a health review of the product throughout, they were not meaningfully involved in the launch of the new product and were not capable or resourced to be able to combat product manipulation. Multiple interviewees reported that they had to "beg" the product team not to launch before the election because they did not have the resources or capabilities to action on disinformation or misinformation on a new product during such a busy, critical time. One interviewee said that SI leadership had to go over the heads of product managers on the Fleets team to help ensure that the product was not launched before the election. According to interviewees, the Fleets example was a serious pain point, underscoring the organizational challenge of new product launches that expose new surfaces which a threat actor can take advantage of. This illustrates the fundamental business challenge of continuing to attract new users while also safeguarding the platform from malign actors, as well as different incentives for different Twitter teams.

3.1.1.2 -- There are components of Twitter that are part of the disinformation and misinformation detection or response that are outside of Site Integrity / Security, and Site Integrity / Security have no access or authority to use these tools absent the good will of other teams.

Through the course of our interviews, we identified multiple teams that were not part of SI/Security yet played a critical role in responding to disinformation and misinformation. SI has no formal authority to require systemic changes or collaborate on key decisions.

For example, as part of a response to disinformation or misinformation, the events teams and curation teams, especially with regards to trending topics, can be partners in mitigating threats by showing Twitter users accurate information. The relationships between the teams with regards to these processes are informal and personality-based versus institutional.

Additionally, with regard to scaled detection of disinformation and misinformation, SI does not have the necessary dedicated engineering support to be able to manage both long-

term priorities and build products that enable threat detection and mitigation at scale, preventing it from being able to focus on proactive activities and instead making them reactive to the crisis of the day.

3.1.1.3 -- Twitter does not have aligned incentives across the organization, and, as a result, priorities with regards to Product Safety.

Product and product managers own all aspects of product development, including risk calculations with regard to product launches. Recently, SI and other Safety components have been included in the design and development process at various check-in points, and provided the opportunity to provide feedback. However, there appear to be no consequences for product managers should their product launches or products increase the workload or costs to Twitter when it falls on SI to develop policies or scale enforcement.

While SI has the authority to make recommendations throughout the product development process, elements of Twitter responsible for identifying threats or security gaps in the products lack the authority to make decisions on product design or roll-out or to hold product teams accountable for failing to mitigate identified risks to the platform, product, and users online.

Interviewees described both the launch of Fleets and Birdwatch as particular pain points for the Trust & Safety team. While product teams do elicit feedback for new product launches, product managers are incentivized to ship products as quickly as possible and thus are willing to accept security risks.

3.1.1.4 -- SI relies on functions that have no accountability to SI in order to piece together solutions.

Interviewees regularly mentioned under-resourced teams, siloing between organizations, and having to borrow resources (such as engineering support), leading to a reliance on the goodwill of other teams leaders or the willingness of Twitter employees to pitch in to support SI in building out its tooling capabilities. This prevents SI from being able to think strategically and develop priorities and goals that are measurable and enable strategy execution.

3.1.2 -- Within, SI, the organizational structure is siloed, with a heavy emphasis and focus on policy enforcement versus threat detection and mitigation.

Within the organizations examined as part of this assessment, there appears to be a lack of alignment and prioritization between teams, and teams appear to be policy focused. Aligning teams to focus on the tactics, techniques, and procedures may cause gaps in Twitter's understanding of adversaries and how they deploy a variety of tactics, techniques, and procedures (TTPs) to carry out an operation or manipulate the platform to achieve a goal.

Silos within the SI may also be contributing to a reactive posture. For example, sophisticated IO actors rely upon misinformation to spread false narratives and often use spamming techniques. Understanding how different threat actors abuse Twitter's platform in a variety of ways (e.g. spam violations to collect data and enhance IO efforts) could help Twitter become predictive, designing holistic tooling, or adding friction to adversarial operations. It is not obvious as to why these teams are split up how they are, other than they are to enforce specific policies. While this may be a good approach in thinking through product features or investigative processes, it silos the threat in such a way that can prevent analysts from piecing together the larger picture.

Importantly, misinformation and disinformation -- which have functionally the same impact on users -- are treated as separate issues and are housed under different teams. Given the fact that misinformation can be leveraged in spam campaigns, state-level information operations, and other types of harms, Twitter's approach has led to siloing, organizational confusion, and slow policy development. Interviewees described several instances in which Twitter was slow to act on misinformation because teams did not see the topic or narrative as falling under their purview or fitting neatly into a particular threat actor they monitored, such as on QAnon or Pizzagate.

One interviewee described the organizational challenges faced by Twitter when dealing with the Pizzagate conspiracy theory and related content. Twitter initially felt as though it was not a disinformation issue because it was not seeded by a foreign actor, was not a child exploitation issue because it included false instances of child trafficking, and was not deemed a spam issue. Twitter could not figure out how to categorize the Pizzagate content, which likely contributed to the narrative's expansion and spread on the platform. In its current posture, the teams are siloed to the degree that it is not always clear who is responsible for what.

3.1.2.1 -- Within SI, there do not appear to be clear priorities from the organization's leadership on how to prioritize threats and thus it is impossible to prioritize resources, goals, and KPIs.

Interviewees said that there is no clear alignment across the teams or prioritization of how to address matters related to platform manipulation. Further, without clear and coherent goals, it is not possible to measure progress against goals in order to mature the organization's capabilities, determine how to allocate resource investment to maximize impact, or sequence the development of tools, resources, and capabilities.

When priorities are developed, it is with a heavy emphasis on English-speaking countries and threats, and whatever goals and metrics are set do not align with the team's observations of the most pressing needs for organizational growth. The team frequently pointed out areas that could be made more efficient through automation, improved processes, and other goals, yet within the organization there does not appear to be a mechanism for meaningful engagement.

3.1.2.2 -- SI sets up strike teams in order to direct resources towards major events, such as elections.

Teams within SI and around Twitter are focused on priority events and providing extra attention to platform matters that are likely to face manipulation. This strike team approach that allows the organization to dedicate additional resources to events appears to be a successful model for addressing threats. However, due to current staffing levels, it requires that teams deprioritize long-term strategic objectives or other responsibilities, and it is not sustainable without increasing resources.

3.1.3 -- Twitter is not poised to deliver on its mission globally, especially in non-English speaking countries.

Twitter lacks the organizational capacity in terms of staffing, functions, language, and cultural nuance to be able to operate in a global context. For example, the misinformation team currently only has two individuals and lacks the sufficient tools to be able to adequately address the threat on a global scale due to a lack of on-the-ground context. This is especially true in priority growth markets, including Africa, Latin America, and Asia. Global teams report a focus on English-language and English-speaking countries. For example, during the 2020 US election, staff were pulled into monitoring, leaving significant vulnerabilities to the regions they support.

The lack of context and understanding has significant implications on the ability to implement policies globally. For example, historically marginalized groups experiencing online threats and harms may not be recognized without an understanding of each country's context, and in some countries it is the government or military that are violating policies, and Twitter is too understaffed to be able to do much other than respond to an immediate crisis. Overseas teams lack the necessary resources to be able to conduct investigations outside of what is already trending or used as a hashtag, making its reactive posture impossible to change without engineering, data science, and investigations support. Twitter expresses a strong preference for fact-checking and labeling content versus removing the content. However, Twitter teams report not having the capacity to fact check in languages other than English.

3.2 Resources

Alethea Group conducted individual interviews, including screen shares, and reviewed internal documentation to determine whether Twitter has the necessary resources, such as tools, datastreams, staff, and skills, to accomplish their tasks.

The lack of sufficient resources, tools, and capabilities has forced SI and TwS to be reactive and largely limit their focus to threats that affect the United States or English-speaking entities. This has ultimately prevented Twitter from proactive threat detection and mitigation to avoid crises. Interviewees described a largely reactive approach to misinformation, disinformation, and spam in which action is taken on content and threats only if it is flagged by reporters or news headlines, partners, or political officials due to the lack of people and sufficient tools to do proactive analysis.

Despite having a global mission, persistent gaps in resources, tools, and capabilities we identified means Twitter does not have the capabilities to operate globally -- including in priority markets -- when it comes to misinformation and disinformation. It also suggests that Twitter is likely spending resources in crisis management and response, rather than investing in capabilities that will allow the company to get ahead of them.

3.2.1 -- Teams in priority growth markets are not sufficiently resourced.

Teams across SI, TwS, and Product prioritize resources to meet primarily US-centric needs. Interviewees across the board said that they do not have the resources, such as staffing and foreign language capabilities, needed to address misinformation and disinformation even in priority markets, such as Asia.

3.2.2 -- Teams have been persistently understaffed.

Twitter has been slow to staff SI teams since 2016. Despite recent team increases, there are currently only two misinformation subject matter experts in SI, both of which are new hires, and four IO investigators to analyze all IO. One interviewee noted that the lack of misinformation expertise was identified as a serious gap in a retrospective from December 2016 about Twitter's lessons learned from Pizzagate. Twitter did not bring on a team member to focus on misinformation until 2019, although existing staff reported that they did focus some of their time to misinformation, however their other responsibilities remained unchanged resulting in staff being asked to do more without additional resources.

Understaffing has meant the teams across Twitter working on the misinformation and disinformation problem set have had to make significant tradeoffs, especially during critical events and surges. For example, Twitter dedicated 100 full-time staff from across SI, TwS, and volunteers from other parts of the company to manage the US 2020 election under the "Election

Squad" framework. As a result, based on interviews and provided documents, SI, Site Policy, Product Trust, and Strategic Response teams had to deprioritize all other work, including work on other critical global events, simply to keep up with the rapid pace of US election-related content.

Site Integrity Headcount for 2021

Team	Total Roles Expected	Roles Filled	Roles Unfilled
Management	7	4	3
API	7	5	2
Misinformation	8	2	6
Identity	5	4	1
IO & Security	10	4	6
Spam	17	12	5

Several interviewees noted personal perceptions about understaffing that may not be accurate, but influence how they view the organization's commitment to filling gaps. For example, one interviewee who had been involved in interviewing candidates for critical roles in SI believed qualified candidates were often rejected by leadership for unimportant reasons. Separately, one interviewee believed understaffing was negatively affecting their team's ability to get resources from Twitter. They noted their belief that funding for internal tool development was decided based on the number of people in the company who would use the tool, which they believed would continue to keep SI teams at a disadvantage; subsequent conversations with Twitter leadership suggested the described process for acquiring funding may not be wholly accurate.

3.2.3 -- SI does not have dedicated engineering support for their tools, so even minor upgrades or changes to existing tools can take months or years to complete.

SI is severely constrained by not having engineers on their teams or engineers dedicated to exclusively supporting their work. Currently, SI must request assistance from engineering teams in other parts of the company to do things like implement even small updates to existing tools or build new ones that could automate more of the process for both policy and investigative analysts. Because these engineering teams do not have an official requirement to support SI and must complete their own work, SI requests are typically put onto a waitlist. That list is then prioritized by SI's immediate engineering support needs for current so-called "fires," such as a critical election. As a result, SI must continue to rely on manual and outdated tools, and individual know-how of its analysts who often must code their own solutions to complete their work. One interviewee called the lack of engineering support dedicated specifically to SI "a real pain point

for internal tooling needs" and said they have had to wait "sometimes years" for minor updates to existing tools they need to do their jobs.

3.2.4 -- SI lacks sufficient dedicated data science support and staff with technical skills.

SI teams rely on the Scaled Enforcement Heuristics (SEH) team to provide data science support, rather than having their own dedicated data scientists. Even though interviewees described receiving excellent work from SEH, they also noted many of the same challenges they have in getting engineering support, namely that SEH has its own work and priorities.

Additionally, in part because many of the tools used by SI require the user to do their own coding and queries, SI lacks sufficient access to technical resources. Having more usable, updated tools with usable UIs would probably reduce the need for some of the technical capabilities.

3.3 -- Tools

SI analysts and managers we interviewed referenced the below range of tools they use to complete their jobs. We were able to personally view the tools that are noted in bold during a screen share or from training materials.

- **Profile Viewer**
- **Batch Action**
- **ClusterDuck**
- **SafetyGraph**
- **Access Search**
- Guano Interface
- **URL Tool**
- Bulk Media Enforcement Tool
- Abuse Triage Tool
- **Botmaker**
- Smyte
- **Semantic Core Editor MisinfoUI**
- **Strato**
- **Thunderbird**
- Hadoop
- Presto
- BigQuery

3.3.1 -- Twitter has not sufficiently invested in developing internal tools to address misinformation and disinformation. As a result, employees must use multiple outdated

and manual tools to do parts and pieces of their investigations, analysis, and enforcement.

In both SI and TwS, interviewees and provided documents described a largely manual process of utilizing multiple outdated, cumbersome, and unreliable tools with poor UIs to do parts of their work, including investigations, analysis, and actioning content. For example, there is currently no comprehensive system for tracking misinformation, from identification to remediation. Existing tools used for surfacing misinformation and disinformation threats are set up so that analysts must go to different tools to manually search for a threat actor or narrative already in mind, rather than the tool using automation and ML to identify potential threats that it then pushes to investigators for analysis.

For IO investigators, one of the most used tools, ClusterDuck, which identifies networks of similar and/or coordinated accounts by country, does not do real-time monitoring and analysis. Data is up to seven days old, and, rather than the tool flagging potentially violative behavior to analysts, users must manually click on a drop-down menu of countries to view results to make a determination on possible coordinated activity. One interviewee described ClusterDuck as "pretty hacked together," and when the assessment team was viewing how the tool operated, it would not load on the first attempt. Another interviewee described ClusterDuck as the only tool really designed specifically for the SI team. A separate tool, AccessSearch, is frequently used by investigators, but its utility is limited by short data storage times (one analyst said it could only store data for two months) that prevent historical research.

Tools used to action on violative content have many of the same problems. For example, according to one interviewee, the process for labeling violative tweets requires using at least five different tools. Tagging tweets in bulk is a manual process that requires the analyst to write a code themselves in a tool called Strato that does not have an easy to use UI. There is also not an easy or automated solution for labelling all tweets that link to a URL that has already been labeled. On Misinformation, SI must manually annotate each new instance of misinformation identified and then moderators manually tag tweets they see with this annotation to apply a warning label.[1] This manual process is especially challenging for large events, such as key elections.

The manual and outdated nature of these tools forces analysts and content moderators to analyze and action against violations tweet by tweet and account by account, a time-consuming process that will keep Twitter reliant on unscalable human power.

3.3.2 -- SI has access to many data sources, but they are spread across several different systems and require largely manual processes to access and analyze.

[1] "Soft Intervention Tool User Manual"

Interviewees in SI suggested they had access to a large number of datastreams with information about on-platform activity. However, they said finding, accessing, and analyzing that data was challenging and time-consuming because it required the use of several different tools and manual processes to search. They also do not have tools aimed at enabling cross-platform analysis. Several analysts also noted having to do their own coding for querying data because many of their tools lack functional UIs. Many of the fixes are small, but would save analysts time and enable more automation. For example, analysts noted having to separately sign-in to external tools, like Domain Tools, to complete a step in the investigative process, when obtaining API access to external tools would allow for integration into internal Twitter capabilities and remove another step in an otherwise manual process.

3.3.3 -- There are existing internal tools in other parts of Twitter that would be useful for the misinformation and disinformation use case, but SI analysts do not have access to them. Analysts also lack access to externally available tools or datastreams that would allow them to do more proactive cross-platform analysis.

Several interviewees noted that other teams at Twitter have internal tools that would be helpful for the misinformation and disinformation use case, but they do not have access to them. For example, one interviewee said Curation uses a tool to create Moments that could potentially help Misinformation and IO analysts proactively identify threats, but they lack access to the tool. SI also does not have access to externally available tools that would allow them to do proactive and more sophisticated analysis and to get insight into emerging threats, such as a social listening tool that provides cross-platform data. One analyst noted that they do not have dedicated staff looking at off-platform activity beyond what external partners provide them, which limits their ability to anticipate possible threats moving on to the Twitter platform.

3.4 -- Capabilities

3.4.1 -- SI does not have a knowledge management system to track and store findings and data. As a result, SI does not have the ability to monitor threat actors or identify changes in their tactics, techniques, and procedures (TTPs) over time, or to measure the impact of SI's work.

Currently, SI does not maintain a knowledge management tool or capability that would enable analysts to save content, data, or their findings. There is no tool or repository where analysts conducting investigations can keep their notes. Most analysts use their own individual Word Documents so that worthwhile investigative notes are individually stored in a way that is not accessible to, or preserved for, their teammates. As a result, analysts are unable to identify and analyze evolving threats or changes in the TTPs of threat actors, or measure the effectiveness of action and enforcement, because information is not being preserved.

IO interviewees noted they have a tasking system housed in Jira to action on leads received from internal teams, such as the Piper Team, or from external partners. However, there is no mechanism by which to save the results of their investigations in a single, automated knowledge management system. Currently, once a tasking is marked as complete in the Jira system, analysts must manually copy their findings into multiple different data source tools or folders to store it, creating an extra step one interviewee said analysts just do not have the time to complete. In some instances, the analyst saves their findings on their own systems, meaning data storage is scattered among different analysts, rather than being preserved in one system accessible to all analysts. This also means Twitter is not feeding its findings into tools or training existing tools to increase automation and ultimately learn from past findings.

3.4.2 -- Twitter does not have traditional threat intelligence capabilities to identify, analyze, and warn about current and future threats, or ingest inputs and intelligence from partnerships.

Twitter does not have a threat intelligence capability internally it can direct based on the company's priorities and to position itself to be proactive in protecting authentic conversation. Misinformation and disinformation teams are currently focused on responding to current threats and so-called "fires" that interviewees said are largely driven by external priorities, such as news headlines, journalist inquiries, or the goodwill of partners.

As a result, Twitter is reactive to events and situations, based on other organizations' goals, interests, or priorities. Relying on civil society cannot scale to meet Twitter's needs, as many priority markets do not have regulatory environments or vibrant civil societies to enable research that, in some cases, may identify government-run influence operations.

3.4.3 -- Twitter does not have the capability to add cost to an adversary attempting to exploit the platform.

In part due to the challenges described above, Twitter has employed a limited set of actions against violative behavior on the platform. Currently, most of Twitter's remediation options have focused on labeling, interstitials, deamplification on a select basis, and removal in response to repeat violations. Twitter leadership has publicly state that account removal could set a bad precedent,[2] and interviewees perceived that removal of accounts or content was considered by Twitter leadership as the option of last resort. However, even removal ultimately does not discourage adversaries from attempting to exploit and leverage the platform, or add costs to their operations because they can quickly adapt. One interviewee did say that Twitter started removing networks piecemeal in order to obfuscate how the network or accounts in question

[2] *https://www.npr.org/2021/01/14/956664893/twitter-ceo-tweets-about-banning-trump-from-site*

were found. Another interviewee estimated that it would realistically take two years before Twitter could build out a strategy and capability to add cost to adversaries by deploying actions like artificial environments.

3.4.4 -- SI and TwS teams lack staff with geographic expertise and foreign language capabilities.

SI and TwS teams do not have sufficient staff with geographic expertise and foreign language capabilities, even in key markets, both of which are needed to understand important cultural and language contexts. Currently, the majority of SI staff are located in the US, with a limited presence in Dublin, and an even smaller footprint in Singapore. The IO team has one staff member with expertise in Russia, one with expertise in Iran, and one with expertise in China, making staffing and coverage, particularly during a crisis, unsustainable. One SI employee noted that the language gap was so significant across Twitter that they regularly receive language support requests from all over the company, not just from the teams responsible for misinformation and disinformation.

The lack of sufficient foreign language skills has hindered work in priority markets. For example, several interviewees and internal policy documents stated that Twitter is limited on fact-checking or debunking to mostly English-language content. One interviewee said that they relied heavily on Google Translate for language capabilities and said that for some countries, such as Thailand, SI is only able to search for trending hashtags for possible exploitation by a threat actor rather than doing investigations because they do not have the language or country expertise on staff.

The lack of language expertise is also affecting Twitter's ability to plan for upcoming priority events. According to internal documentation, Twitter is unable to provide even a scaled-back version of the election support that was deployed for the US 2020 election for the upcoming Japanese election, which has been identified as a priority for the company. According to the "US 2020 Civic Integrity Policy/Ops/Product Reflections" document, that is in large part because there are "no Japanese speakers on the Site Integrity team, only one T&S staff member located in Tokyo, and severely limited Japanese-language coverage among senior TwS Strategic Response staff."

3.5 -- People

3.5.1 -- SI employees are dedicated to the mission and the organization, and feel heard by their immediate SI management.

Interviewees all expressed support for the mission and the organization, as well as positive perceptions about their teams, teammates, and managers. They described pulling together to

meet the demands of each day, and many described a strong commitment to the organization despite challenging circumstances and burnout. Several interviewees also noted that they felt heard by their immediate SI management and felt empowered to raise concerns tothem. At the same time, they were not always confident that action would or could be taken in response to those concerns.

3.5.2 -- SI teams lack diversity, especially gender diversity across both the analyst and management level.

Multiple interviewees expressed a concern for the lack of diversity, particularly gender diversity, on the teams responsible for addressing misinformation and disinformation. According to staffing documents we reviewed, only one-third of SI personnel are women and the majority of management and senior-level positions are held by men. Similarly, several interviewees assessed that the lack of diverse backgrounds among employees contributed to gaps in foreign-language capabilities on the teams and, therefore, the teams focused on primarily Western, English-language content and threats. The lack of diversity almost certainly hinders SI's ability to execute its mission and benefit from the talents and abilities a more diverse workforce provides.

3.5.3 -- SI staff are burned out and do not believe Twitter leadership is aware of it.

Employees in SI reported being burned out. They attributed this in large part due to understaffing, the amount of day-to-day work, frequent policy changes that create confusion, time-consuming manual internal tooling, a lack of strategic planning across all the relevant parts of Twitter, and a consistent crisis state of operating as a result of jumping from one "fire" to the next. These issues have created time-consuming processes and stress on teams where employees are expected to work longer hours when a lack of strategic planning creates a crisis. The majority of interviewees also said they are expected to wear multiple hats, and SI interviewees noted in particular a perceived tendency by leadership to rely on a couple of people for everything. They believed that the fact that those people completed their work was used as justification for not hiring more people.

Most interviewees pointed to the rapid pace of work and the significant workload of the US 2020 election as a recent source of employee burnout. However, many ascribed their burnout to what they saw as a culture of constantly being in a state of "firefighting" or crisis, which they largely saw as driven by external events, such as congressional inquiries or news events. Relatedly, several senior managers across Twitter were expected to be "always on" during the election to address escalations on high-profile accounts because of the company's "low risk tolerance," according to documents we reviewed. A similar-sized effort under the Election Squad construct for another priority election would be unsustainable with current staffing levels.

Interviewees said a lack of strategic planning and coordination between relevant parties in SI, TwS, and Product on product development and deployment had also contributed to staff burnout. For example, SI and TwS interviewees noted that product teams consistently failed to solicit or at least include their feedback on product rollouts, such as Fleets or BirdWatch. They said it had resulted in them having to pull longer hours, often outside of working hours, to address vulnerabilities in products identified sometimes hours before or even after a product rollout.

3.5.4 -- Staffing in SI is top heavy, except for on the Piper Team. Managers are expected to wear multiple hats, including conducting investigations and creating policies, but they spend most of their time with managerial responsibilities, and report spending their days in back-to-back meetings.

SI managers said that they were expected to still conduct IO investigations and lead on developing IO policies, but that they spent the majority of each day in meetings and on personnel management tasks. Some interviewees expressed concern about not having the time to keep up their investigative and technical skills, and one senior manager said they often used what should be their non-work hours to conduct manual investigative work that a more junior employee could do, including finding and suspending large numbers of accounts trying to evade a previous Twitter ban.

3.5.5 -- Content moderators in TwS are not adequately resourced, especially to make determinations on misinformation.

Content moderation is outsourced to vendors, most of whom are located in Manila. One interviewee stated that moderators are "treated like second-class citizens," are "not fully bought-in" to the company, and are underpaid.

Moderators are not properly resourced to take action, especially on misinformation. Several interviewees said that moderators do not have the geographic expertise or language capabilities to understand important cultural or linguistic context, and therefore are not able to make accurate and consistent decisions on what is misinformation. Another interviewee described a long process for training moderators on new policy rollouts and said that managers often did not have sufficient warning about new policies to prepare moderators in time. As a result, full-time TwS employees have had to, at times, do content moderation. Content moderators are also not proactively trained on emerging threats.

3.6 -- Partnerships

SI has prioritized creating official external partnerships with nine companies, largely other social media platforms like Facebook and Google, and more unofficial partnerships with research organizations, such as the Stanford Internet Observatory. These partnerships give SI insight into

misinformation and disinformation trends across social media platforms, provide warning of potential threats on their own platform, allow Twitter to potentially get ahead of news stories, and give the company the opportunity to publicly promote its work on misinformation and disinformation in a way that boosts public perception of its activities. However, Twitter is not fully taking advantage of these existing partnerships and has not established other potential partnerships that would set itself up for more proactive, long-term success in addressing the misinformation and disinformation threat. Additionally, these partnerships contribute to SI staying in reactive mode.

3.6.1 -- There is not a consistent view within SI about the goal of external partnerships.

Judging from the interviews we conducted, teams have different views on what the goals of external partnerships are. Some interviewees suggested partnerships were a way to see what the other platforms were doing or to get ahead of a forthcoming news story. One interviewee characterized partnerships as a "moat to protect the organization" from public criticism. This lack of alignment on the purpose and intent of partnerships may mean that there are other partnership opportunities for SI that can help address some of the gaps in capabilities and resourcing described above.

3.6.2 -- Investigating and actioning on inputs from external partnerships often drives SI's immediate priorities and keeps teams in a constant reactive state. However, findings from other platforms do not necessarily reflect the actual threat landscape on Twitter itself.

Intelligence and leads from its partnerships with other social media platforms gives SI critical insight into cross-platform activity that may also be affecting the Twitter platform. Similarly, working with research organizations like the Stanford Internet Observatory gives SI access to experts and early insight into, and opportunities to collaborate on, forthcoming academic research that may gain media attention upon public release.

However, actioning on the work from these partners means Twitter often prioritizes the findings of other platforms, which are also largely set up to do reactive work and have their own internal priorities and challenges. Similarly, academic organizations face staffing shortfalls, meaning they must prioritize their own work products, primarily resulting in retrospective and targeted research projects rather than Twitter being able to direct research and investigations on its own priorities. As a result, prioritizing investigative inputs from both platform partners and academic partners means SI may not be investing its time in addressing the actual threat landscape on the Twitter platform.

3.6.3 -- SI is currently unable to ingest, action, and store all of the intelligence and leads provided by its existing partnerships. It does not currently have partnerships

that could help fill some of the gaps in being proactive to address Twitter's own threat landscape.

Several SI interviewees said it was a struggle to stay on top of actioning on all of the leads provided by partners or flagged by external parties, such as reporters. They believed that prioritizing those taskings contributed to the teams' inability to do proactive work more reflective of the threat landscape on the platform, including: getting actionable intelligence from outside partners that could be informing long-term planning and decisions, identifying threats, assisting with strategic investigations, and helping to move the company from reactive to proactive on misinformation and disinformation. SI's existing partnerships do not include an ability to task them to conduct targeted analyses or longer-term investigations.

3.7 -- Policies

Alethea Group sought to identify current formal and informal policies and processes in place to help understand Twitter's capabilities to address disinformation/misinformation.

3.7.1 -- Policies are often implemented in response to "fires," rather than being informed by analysis of the current or emerging threats for the platform, without an effective enforcement mechanism in place.

Based on interviews with key stakeholders and a review of internal documentation, policies are often created quickly in response to external events, with no clear strategy for implementation. Team members said that because policies are often reactive in nature, there are significant gaps in the content they cover, and that policies do not address evolving threats.

Interviewees said that major events, including Chrissy Tiegen threatening to leave Twitter because of harassment from users who align with the QAnon movement, or the shooting at Comet Ping Pong (Pizzagate) in 2016, forced Twitter to take a stronger policy and remediation position than they assessed it otherwise would have based on the evolution of the threats alone. But because of the reactive nature of these changes, policies were often rushed, not well-executed, and difficult to enforce.

One interviewee stated: "Twitter only seems to respond to fires, and fires only. We can only handle what is the biggest and loudest fire at that moment." This approach means Twitter is often behind the curve in identifying and responding to misinformation and disinformation.

3.7.2 -- Rapid policy changes often do not incorporate feedback from the relevant stakeholders, making it more difficult to communicate, and ultimately enforce, those policies.

Because policy changes are often implemented quickly, they often do not incorporate feedback from relevant stakeholders, making policies more difficult to communicate and ultimately enforce. For example, in response to a manipulated video of House Speaker Nancy Pelosi in May 2019, Twitter quickly implemented a new policy (Synthetic and Manipulated Media Policy). However, because the policy was rolled out so fast, the organization was unable to effectively enforce it, or train agents on what content was violative. One interviewee stated that feedback was always asked for, and people were "given a seat at the table," but that feedback was not always given or given in a constructive way.

In another instance, interviewees said that policy decisions were not always communicated to the broader global team, making it more difficult for the policy to be widely enforced.

According to the internal document "US 2020 Election – Policy/Ops/Product Reflections," while "communication between policy and enforcement teams was generally solid," during the 2020 election, the "adoption of the decision to stop using interstitials proved to be challenging, as some TwS employees continued to apply the interstitials despite email and Slack notifications about the policy change. A single source of truth on policy enforcement — rather than scattered documents, emails, and announcements — will be vital for future activations." In short, the rapid rollout of policies leads to uneven enforcement from Twitter's moderators.

3.7.3 -- Policies to address misinformation/disinformation often do not address repeat offenders and are applied on a case-by-case basis, leading to a lack of scalability.

Interviewees noted that there is not a sufficient enforcement mechanism for repeat violators of Twitter's policies, and thus, there is little incentive for bad actors to stop posting violative content. One interviewee stated that if 80% of the content that a user posts is misinformation or disinformation, that account should be suspended, adding: "Continuing to address each individual tweet from a user isn't sustainable given staffing shortfalls."

According to the internal document, "US 2020 Election - Policy/Ops/Product Reflections," Twitter's labelling policies "lack any kind of punitive enforcement for repeated misinformation labels. While tweet removals under the Civic Integrity Policy incur a strike (3 strikes resulting in permanent suspension), labels do not accrue strikes, and therefore do not dissuade repeat or malicious behavior."

3.7.4 -- Policies are written for a sophisticated audience, making it difficult for agents on the ground to enforce.

Policies that address misinformation and disinformation at Twitter (e.g. the Civic Integrity Policy, Synthetic and Manipulated Media Policy, and COVID-19 Misleading Information Policy) are often

complicated, highly nuanced, and require significant context for Twitter Services agents to be able to take action. When policy rollouts occur, Twitter trains its agents on those policies, however, many of these agents are located all over the world and may not have sufficient language and/or cultural context to be able to action on specific instances of misinformation. And because of the complicated nature of these policies, remediation and mitigation takes longer and is more difficult to accomplish at scale.

Additionally, when new policies are introduced, content moderators have to manually annotate each new narrative they are seeing, making it impossible to keep track of the content. By creating more digestible policies, moderators would be able to better enforce them.

One interviewee added that policies are often created in a vacuum without the input of subject matter experts and are "therefore not grounded in reality." Another stated that Twitter's issue is "not coming up with new policies, but enforcing the ones that we've already got." Because of the sophistication and nuance in already existing policies, they are not only difficult to enforce at present, but also difficult to enforce at scale.

3.7.5 -- Twitter's US-centric approach to policy decisions makes it difficult to detect and mitigate disinformation and misinformation around the world.

Our assessment found that policy decisions are often made in response to US-based events, such as the 2020 presidential election, QAnon content on the platform, manipulated media of House Speaker Nancy Pelosi, and more.

Because policies are written to address US-based problems, they often do not take into account different ongoing misinformation or disinformation campaigns in other parts of the world. Further, policies that address violative content in a US context are more likely to be enforced because of Twitter's contextual and linguistic capabilities.

According to the internal document, "US 2020 Election - Policy/Ops/Product Reflections," Twitter is "ill-equipped to provide even a scaled-back version of the proactive investigation and remediation efforts we implemented in the US — in no small part because we have no Japanese speakers on the Site Integrity team, only one T&S staff member located in Tokyo, and severely limited Japanese language coverage among senior TwS Strategic Response staff."

Additionally, according to the same document, uneven policy enforcement around the world "creates the potential for accusations of a US-centric bias in Twitter's actions, as well as unequal and ultimately unfair enforcement of our rules."

Because of various factors outlined throughout this assessment, policy teams do not have the ability to plan ahead and write proactive policies in response to known upcoming events. While a certain level of uncertainty will always exist (e.g. COVID-19), there are ample opportunities to

proactively develop policies and capabilities in response to upcoming elections around the world and other major planned events.

3.8 -- Processes

3.8.1 -- While processes exist to elicit feedback from necessary stakeholders, there are no processes to actually incorporate that feedback.

Multiple interviewees explained that while processes exist that elicit feedback from all necessary stakeholders (e.g. product health reviews), feedback often is not incorporated.

Interviewees said that because of existing organizational structures and different incentives across teams (e.g. product teams are incentivized to launch new products), platform and user security are given less consideration than warranted. Further, product teams are not required to incorporate feedback from SI, and because product managers are promoted for launching new products, there is less incentive feedback to be incorporated, and a greater incentive to launch new products quickly.

In launching Twitter's Birdwatch program, members of the SI team said that they were involved in the process throughout, and made suggestions as to how the product could be more secure, including specifically warning that users aligned with QAnon would likely attempt to join. However, feedback was not incorporated in an attempt to keep the product open, leading to a last-minute scramble to secure the product launch. On the evening before Birdwatch launched, Twitter realized that an overt QAnon account had been accepted into the Birdwatch program.

In other instances, interviewees said that the Product Trust team would call out a risk to a product launch, but that the product team would simply "accept the risk" with minimal mitigation efforts. In short, processes don't take into account competing priorities or incentive structures within the company, and when two process owners have competing interests, there isn't a process for deconflicting, at least from a staff perspective.

3.8.2 -- The process for labelling disinformation and misinformation content is largely manual, requires the use of multiple tools, and usually needs to be done on a case-by-case basis.

According to the internal document, "US 2020 Election - Policy/Ops/Product Reflections," even once decisions about enforcement are made, "the process of applying labels is cumbersome," "requires the use of backend interfaces," and the "complex steps involved make scaled application of labels difficult to expand beyond a very small group of highly trained agents."

Alethea Group participated in a screen sharing process with one of the interviewees, and found that no less than five different tools were needed in order to label a single tweet.

3.8.3 -- There is currently no unified system for tracking misinformation and disinformation, from identification to remediation, according to staff interviews and the US 2020 retrospective document.

The organization does not have a system in place to proactively identify or track misinformation or disinformation threats. Leads on violative content often come from user complaints, partner organizations, or independent researchers, but Twitter does not appear to have a systematic approach to identifying these threats on its own. In the case of disinformation content, the IO team is sometimes given leads from the Piper team, but there are no existing formal processes to do so.

It appears that the organization also does not have a formal process in place for what happens after a threat is identified. Investigators stated that while there is a tool (GoIORef) where tickets are submitted and a queue is created, there is an ad hoc system for responding to those claims. And, because of a limited number of subject matter experts working on the IO team, specific team members are often needed to respond to specific disinformation/misinformation threats.

According to the internal document, "US 2020 Election - Policy/Ops/Product Reflections," Twitter's Civic Integrity Policy defines what content the company should enforce on, but "the specifics of particular conspiracies that emerged in the course of the [2020 US] election, whether those conspiracies have been debunked by external sources (and are therefore eligible for remediation), whether we have specific curated resources available for those specific conspiracies, and how to put all the pieces together in practice *is undeveloped and largely ad-hoc.*"

One interview suggested that the misinformation team and the IO team worked together because of personal relationships rather than any formal processes.

3.8.4 -- The process for identifying what civic events (i.e. the Election Assessment Process) are prioritized involves multiple teams who all use different criterion and planning processes. This results in confusion, a lack of coordination, and uneven resource allocation.

According to interviews and internal policy documents, team members from public policy, sales, regulatory, trust and safety, and others are all involved in the process of determining how to prioritize worldwide elections. However, each office has its own criterion to determine what is a priority. Once an election is assigned a priority, or "tier," there appears to be no process in place to determine the resources needed to sufficiently staff that election. Further, while an election

might be considered "tier 1," it does not necessarily receive the same attention or resources as another "tier 1" election.

The result, according to the "US 2020 Election - Policy/Ops/Product Reflections" document, is that "where an election is taking place but doesn't receive the same treatment as the US election (as happened with elections in Brazil in November 2020), in-region teams may become frustrated with limited support and apply considerable pressure to operational and policy teams to enforce rules on an ad-hoc basis, as well as product teams to build ad-hoc experiences, without adequate preparation or resourcing to do so." Because decisions in this space are also made from a US perspective, interviewees felt that elections in other countries were given less priority.

3.8.5 -- Twitter lacks sufficient processes to measure progress and impact, and therefore fails to implement lessons learned from the past.

There are no formal processes to measure the impact of policies on deterring or combatting a threat actor, and Twitter does not have data to determine whether policies are working or need to be modified. While Twitter completes retrospectives on progress to goals (e.g. after Pizzagate), there is no process to measure the effectiveness of the company's remediation attempts. Data is either not retained or not stored in an accessible way team-wide, giving the organization no ability to learn from its past actions.

Re/ Vou

#1 Knowledge mgmt ✓
Investigation mgmt ✓

#2 Staffing strategy