Scoping the Institute for Research on the Information Environment

Nils B. Weidmann Margaret E. Roberts Zachary

Steinert-Threlkeld Sebastian Hellmeier

June 17, 2022

Executive Summary

Making scientific progress in the study of the information environment is more important than ever. This report has four aims. First, **we identify the most pressing research questions** that scholars need to address in the area of misinformation and disinformation on social media. We believe that research must be better able to (i) identify digital mis- and disinformation and provide estimates of its quantity, (ii) gauge the impact that this content has on citizens' attitudes and behavior, and (iii) develop countermeasures and test their effectiveness. Second, **we review existing data and tools** for the study of misinformation, and argue that these tools are insufficient to fully understand and address the challenges of misinformation. **We discuss two major kinds of impediments** that limit scientific progress in the third part of our report: (i) technical challenges, where researchers face difficulties to acquire, store and process the data needed for their research, and (ii) ethical and organizational challenges preventing them from using existing data in the most effective way possible.

In the fourth part of our report, **we propose a new Institute for Research on the Information Environment** that addresses these challenges. We envision an infrastructure that **offers a number of novel data products** that would allow researchers access to social media data and metadata across platforms, including tools to identify mis-information, understand its spread, flag disingenuous accounts, and conduct representative surveys of social media users. We also propose products that would enable researchers to understand what types of countermeasures are implemented across platforms and assess their effectiveness. The institute would offer **a stage-wise access model**, where teams first develop their procedures from an "outer circle," which provides access to sample data and interfaces from the outside, without the need to be physically present at the Institute. For analysis of real data that come with strong legal or privacy limitations, researchers temporarily move to the Institute. This work is done on the Institute's secluded infrastructure, without the possibility to share data with the outside world. Aggregated and anonymized results can be shared and exported, subject to prior legal and ethical approval. The Institute should have **pre-established ethical, legal and scientific protocols**, defined and regularly adjusted by legal experts and a scientific advisory board. The latter is also responsible for **selecting projects based on a competitive application process**. Finally, we recommend that the Institute be set up in a modular fashion, where **satellite institutes can be established in other geographic regions**. This will ensure that scientific work can be done in different legal environments (for example, the EU), but will also contribute to a more adequate global representation of research and researchers.

1 Introduction: Studying the Information Environment

Access to information has never been easier than today. With the diffusion of the Internet, people around the globe have turned from information receivers in a one-to-many environment (e.g., TV, newspapers) to senders and receivers in many-to-many communication systems (e.g., social media). The share of the global population using the Internet regularly has risen from around 10% a decade ago to more than 60% in 2020 (International Telecommunication Union, 2021). According to recent data, 97% of U.S. citizens own a cellphone, and 93% use the Internet (Pew Research Center, 2021), allowing them to access and distribute information anywhere and

anytime. Even though geographical and demographic divides persist, access to information and communication technology has increased dramatically over the last decades.

The diffusion of information and communication technology (ICT) and the emergence of social media platforms such as Facebook, Twitter, and TikTok have created a complex in formation environment defined as "[t]he aggregate of individuals, organizations, and systems that collect, process, disseminate, or act on information" (Committee on National Security Systems, 2015). While these developments in the information environment positively affected economic growth (Stanley, Doucouliagos and Steel, 2018), knowledge sharing, interpersonal relations, and public goods provision, they created new political challenges for democracies and autocracies alike (Tucker et al., 2017).

Early hopes that ICTs would serve as a "liberation technology" (Diamond, 2010) that allow people to coordinate and mobilize against state abuses of power were premature. Despite the importance of social media for uprisings around the globe, recent work centers around the use of ICTs as "repression technology" (see e.g., Rød and Weidmann (2015)) and the rise of "digital authoritarianism" or "digital repression" (Feldstein, 2021). Authoritarian regimes but also some democracies assert control over the information environment by surveilling dissidents, censoring content, and spreading propaganda (Deibert et al., 2008; Roberts, 2018).

One of the biggest challenges of today's information environment is the spread of misinformation, understood as "incorrect factual beliefs" (Jerit and Zhao, 2020). While misinformation has always been a feature of human communication, some scholars claim that we live in an "misinformation age" (O'Connor and Weatherall, 2019) where the accuracy of information is of little importance to many people. Online communication channels facilitate the increasing prevalence of misinformation (Guess, Nyhan and Reifler, 2020). Twitter data from the U.S. showed that almost 0.1% of users shared 80% of fake news around the 2016 presidential election (Grinberg et al., 2019). While misinformation may be unintentional, in some cases, actors use misinformation strategically in "disinformation" campaigns defined as "deliberate propagation of false information," (Tucker et al., 2018, 3). One example is fake news, "news articles that are intentionally and verifiably false, and could mislead readers." (Allcott and Gentzkow, 2017, 213).

Misinformation and disinformation campaigns have been documented in such varied settings as during elections in the U.S. (Shao et al., 2018) and France, during the Covid-19 pan demic since 2020 (Bursztyn et al., 2020), in discussions about climate change (Van der Linden et al., 2017) and Russian activities in Ukraine (Mejias and Vokuev, 2017), to name a few. As online information becomes the primary news source around the world, these campaigns threaten the foundations of democratic institutions. In societies with high levels of polarization, misinformation can spread quickly and spur further polarization (Tucker et al., 2018). Misinformation can incite hate and increase harassment and discrimination online and offline (Williams et al., 2020). As the examples of Myanmar and Ethiopia show, social media can help coordinate violent action with detrimental consequences, particularly for marginalized groups. Finally, disinformation has the potential to shake the core of liberal democracy. Authoritarian regimes try to manipulate the information environment abroad, e.g., by interfering in elections (Bradshaw and Howard, 2018). Aspiring autocrats can erode democracy by spreading false information and manipulating the media landscape. De liberation, a vital element of democratic decision-making, could be severely hampered by misinformation and disinformation campaigns.

We believe that not all negative consequences need to manifest and science can contribute to stopping the flow of disinformation. Fundamental to preventing the harms caused by

misinformation is gaining a better understanding of its prevalence, understanding when and where it has most impact, and gauging the effectiveness of countermeasures against it. This report provides a social science perspective on the study of the information environment focusing on the consequences from a global perspective. We begin by identifying pressing issues and open questions in the study of the information environment, such as how to identify and measure misinformation and its impact on society. Then we provide information on the infrastructure and data that would be required to answer these questions. Finally, we outline technical and ethical challenges in the study of the information environment and provide recommendations for a potential new research center for the study of misinformation.

2 Open Questions: What the Academic Community Needs To Study

Despite the increasing attention being devoted to the scientific study of misinformation, many important issues remain unaddressed. In this section, we discuss a number of open research questions related to misinformation.

2.1 Identifying and Quantifying Disinformation

All empirical research must start with a concise operational definition of what "misinformation" is, which can then be applied to actual content. This is challenging to apply in practice. Misinformation is commonly defined as statements that are factually wrong, and are referred to as "disinformation" if they are shared in an attempt to deceive others.[1] Thus, for a piece of digital information to be classified as "disinformation," we must establish (i) that it is factually false, and (ii) that the sender/author was led by malign intent. Both pose considerable challenges:

Checking factual veracity is possible for information that contains a factual statement, but not for others. For example, conspiracy theories oftentimes rely on factual events and occurrences, but differ in their interpretation (Brotherton and Son, 2021). This interpretation may oftentimes seem extremely improbable (such as, for example, the existence of a secret world government), but it is oftentimes not possible to refute it with evidence to the contrary. Also, misinformation oftentimes relies on a strong partisan position, making it a political matter to decide whether it is right or wrong. This "motivated reasoning" has been shown to occur

[1]See Wu et al. (2019) for a more extensive categorization.

frequently (Taber and Lodge, 2016). Hence, scholarly work has to work with independent definitions, and potentially very narrow, definitions of misinformation, to counter criticisms of defining misinformation "in the eye of the beholder."

Existing research relies on different shortcuts to identify misinformation and its consequences. Some rely on clearly identifiable but very narrow instances of false information, as for example the "birther" conspiracy belief (Richey, 2017). While this has the advantage of being a clearly defined and provably false claim, its spread is confined mostly to the US, where its salience may also change over time. This approach is therefore not suitable for more general analyses of the spread of misinformation across different topics and countries. Other research codes entire outlets (websites, portals) as those spreading false information, rather than examining each and every message (Allcott, Gentzkow and Yu, 2019). This has the advantage of being easily applicable to larger quantities of content, but lacks a fine-grained distinction at the

level of individual reports or posts. Yet other research codes misinformation at a fine-grained level (Jiang and Wilson, 2018) with the help of fact-checking websites like Snopes. This approach is cumbersome, since the respective posts have to be matched manually with the data produced by Snopes. This makes timely and large-scale coding impossible.

Coding attribution and intent is equally difficult. In several cases, the origin of malicious content can be clearly determined, as for example in the case of the Russian-sponsored TV/online channel *Russia Today* (Yablokov, 2015), or even for larger sets of channels (Allcott, Gentzkow and Yu, 2019). Here, malign intent and the goal to deceive readers are assumed, since the false content has been created explicitly for public consumption. In many other cases, the creators of these messages and their origin remain unclear. In many social networks, malicious content becomes salient because users like or share it, and without complete data about the history of a post, it is difficult to determine who created it in the first place (let alone identify the person or organization around it). Hence, attribution – which is a widespread problem in cyber-research (Keremoglu and Weidmann ˇ , 2020) – remains a fundamental problem in research on misinformation.

While misinformation is narrowly defined as a feature of individual pieces of content, re search needs to study it also at a more aggregate level, for example websites or users that regularly post it. This is closely related to the identification of disinformation *campaigns*, which are repeated, orchestrated attempts to systematically spread misinformation for a particular purpose by a given actor or organization. Rather than studying individual messages and their diffusion, this requires the identification of *patterns* in posting, sharing or upvoting of particular content across messages and users. Existing research along these lines typically ignores the actors and users, and defines campaigns as the set of messages with a common aim, regardless of the provenance (Wu et al., 2019). This way, we lack knowledge about the actors that systematically produce misinformation, as well as their target audiences.

Overall, the challenges mentioned above have made it difficult to produce research that conclusively answers the following questions:

1. How much misinformation is circulating? So far, we have no good estimates of the extent of misinformation that is spreading via online channels. Existing research is confined to extremely narrow topics, making generalization difficult. This also impedes comparisons across countries, since existing codings of misinformation rely on particular topics and issues that are oftentimes country-specific. The general difficulty of attribution in cyber-research makes it possible to identify the creators of misinformation, and to find out whether misinformation is part of larger, orchestrated campaigns.

2. How can we identify newly emerging misinformation? Existing work proceeds by retrospectively coding content or campaigns, typically long after they have started. It remains very difficult to identify misinformation in real time, which makes timely scientific analysis and quick counter-measures impossible. With a quick and general way to identify misinformation, it would also be possible to know which topics are likely to become potential targets of misinformation, in order to prepare countermeasures.

3. How can we detect “disinformation by omission”? So far, research proceeds by identifying content that satisfies a particular definition of misinformation. This neglects that readers can also be deceived by the omission of particular facts in a report. These instances of misinformation are very difficult to identify, since we would need the complete set of facts so that we can identify which ones are missing.

2.2 Assessing the Impact of Misinformation

Despite increasing public and scholarly attention to misinformation on digital media, we do not know enough about its societal and political impact. In general, attempts to estimate the impact of digital information and its manipulation mostly rely on experimental approaches. Some research has studied the determinants for why people believe in false information. This research has shown, for example, that endorsements by trusted peers make people more likely to believe in misinformation, irrespective of its content (Mena, Barbe and Chan-Olmsted, 2020). Also, motivated reasoning plays a major role, but politically sophisticated people are better able to identify false information (Vegetti and Mancosu, 2020).

Much research on the effects of misinformation focuses on particular issues and topics. One that has received considerable attention is the Covid-19 pandemic, and in particular the measures employed by governments to counter it. For example, research has shown that people that are receptive to misinformation and conspiracy beliefs are also that are particularly critical of pandemic-related governmental measures (van Mulukom et al., 2020; Gemenis, 2021), which echoes results that indicate a similar relationship between support for conspiracy theories and authoritarianism (Richey, 2017). Does this matter for actual behavior? Some research suggests so. For example, results show that people receptive to misinformation show less willingness to reduce their carbon emissions, or to get vaccinated (Jolley and Douglas, 2014*a*,*b*). The latter finding was confirmed in an observational study, which shows that the consumption of misinformation in the US strongly correlates with lower vaccination rates (Pierri et al., 2021).

Overall, while we see some research that examines the impact of misinformation on attitudes and behavior, there are two main challenges in this work. First, much of this work takes an individualist focus and examines individuals, their media consumption, as well as their attitudes and (intended) behavior. Oftentimes relying on small experimental or survey samples, this research is necessarily confined to narrow issues and populations, and does not generalize well. Here, there is a clear need to conduct more comparative research across different populations and countries. Second, causal inference remains a great challenge in particular in studies beyond narrow lab or survey experiments. Oftentimes, studies reveal a correlation between the consumption of misinformation and some behavioral outcome, as for example vaccinations (Pierri et al., 2021). This relationship is not necessarily causal, since media consumption patterns in the online sphere are oftentimes defined by choice, where people actively seek out those channels and outlets they want to follow. To summarize, there are several questions regarding the impact of misinformation that we believe should be studied more:

1. Are people persuaded by misinformation, rather than only consuming it? This question is important, since we should expect changes in their behavior only in the former case. More generally, research needs to become more comparative and study the effects of misinformation across many different societal groups and countries. This would allow us to understand better which groups are particularly vulnerable to misinformation *within* countries, but also how this generalizes to *other national contexts*.

2. What is the causal effect of misinformation on individual political perceptions and behavior? Research needs to improve when it comes to establishing causal effects of misinformation on outcomes we are most interested in. A particular challenge is news consumption is a choice, which is why correlations with attitudes or behavior may be confounded. Also, many analyses on the behavioral effects of media consumption

estimate intention-to-treat (ITT) effects, not average treatment effects that researchers are typically interested in. Therefore, the effects found in these studies may actually underestimate individual-level effects.

3. How to bridge the micro-macro gap? Even if we know more about the effects of misinformation on individuals from research in political psychology, research needs to improve when it comes to the implications at the macro level. For example, does the quality of democratic debate and competition really decline when some narrow segment of the population becomes exposed to misinformation (Kuklinski et al., 2000)? Are autocratic regimes effective in the aggregate when they target opposition movements with misinformation (Weidmann and Geelmuyden Rød, 2019)? Overall, answering these questions requires us to study not only the impact of misinformation on individuals, but also how this affects aggregate societal outcomes.

2.3 Countermeasures and Their Effectiveness

In light of increasing amounts of disinformation online, different countermeasures are being employed to fight it. This can be done by flagging it, changing its presentation and/or appearance, or by removing it.

One way to reduce malicious effects of misinformation is to flag particular outlets or users on social media. Twitter, for example, introduced such a policy, which requires accounts related to official state institutions or government-affiliated media outlets to be labeled accordingly. This policy seems to be able to make users more cautious (Liang, Zhu and Li, 2022), although we have evidence only for a limited context (China). Another way to counter misinformation is by altering the information that is presented to users online, for example by algorithmically de-prioritizing information that is deemed problematic such that users are less likely to consume it.

A much more prominent way to reduce misinformation and contain its effects is by means of content moderation. Content moderation refers to the removal of information that is considered to be in violation of legal restrictions or platforms' terms of use. It is done for many differ ent reasons, only one of which is to prevent the spread of disinformation on social media. Content moderation involves large amounts of resources and difficult decisions, oftentimes with the need to balance the prevention of online hatred and propaganda, and freedom of speech (Gillespie, 2018). Content moderation can be highly controversial – while many "easy" tasks can be automated, it is more difficult to detect misinformation (Gorwa, Binns and Katzenbach, 2020), not least because of the challenges of defining it in the first place (see above). The precise mechanisms of content moderation that social media platforms employ are mostly hidden from users (and researchers). Some platforms use "shadow banning," where content or entire accounts are blocked without users receiving a notification about it. In most cases, however, users are notified that some of their content is in violation of guidelines, and therefore remains blocked. While there is some work about this impact this can have on individual users (Myers West, 2018), we know less about how measures of this kind can prevent the amount and the spread of misinformation. At the user level, providing explanations for why content was removed seems to improve future user behavior and reduce the sharing of problematic content (Jhaver, Bruckman and Gilbert, 2019). Also, removing automated "bots" on Twitter seems to reduce the reach of radical right and Eurosceptic actors, although the effect remains small (Silva and Proksch, 2021).

Another way to counter online disinformation is by means of actively correcting it with factual information. Research has generally documented beneficial effects of fact-checking in general (Walter et al., 2020), but also in particular domains such as climate change (Bene gal and Scruggs, 2018) or health misinformation (Walter et al., 2021). Still, this research has also shown that effects are highly conditional. For example, motivated reasoning plays a major role, with people accepting corrected information primarily when it fits their partisan view (Walter et al., 2020). Even more, corrections can lead to backfire effects and increase, rather than reduce misinformation (Nyhan and Reifler, 2010; Nyhan, Reifler and Ubel, 2013), although a recent meta-analysis suggests that evidence for backfire effects is weak (Swire-Thompson, DeGutis and Lazer, 2020). Overall, it must be kept in mind that the consultation of fact-checking sources and websites usually requires users to take action themselves; this is why its overall impact will remain smaller as in the controlled scientific studies, where the presentation of corrected information was experimentally controlled. Overall, research on potential counter measures against misinformation remains limited and often inconclusive. We therefore believe that future work should focus on the following questions:

1. What content is being moderated? Given the almost complete opacity of social media platforms as regards their content moderation mechanisms, it is difficult to find out whether misinformation campaigns are currently successful because little is done to stop them, or because the existing measures are simply ineffective. At the moment, researchers are left with only few insights into how content moderation works, often having to reverse-engineer moderation practices from the outside.

2. Are existing measures against disinformation – such as content moderation or fact checking – effective? Existing research suffers from a number of limitations. For example, most work is tied to particular contexts, most often the US (Walter et al., 2020). There is evidence that political content moderation in authoritarian regimes reduces public knowledge of censored topics (Roberts, 2018), but this is a very different context than disinformation in democratic environments. It is therefore difficult to generalize these results to other world regions and/or topics. Similar questions arise as regards other measures against misinformation, as for example the labeling or slowing down of misinformation. This has been studied in narrow contexts, such as Twitter's labeling of government-affiliated accounts in China, generalizations to other countries and platforms are difficult.

3. What other new counter-measures are there? So far, research has centered around existing mechanisms such as content moderation that platforms employ for a number of reasons (prevention of misinformation is only one of them). Future research should focus on the development of new techniques to improve online communication, some of which are described in recent studies by the *Partnership for Countering Influence Operations* (PCIO) (see for example Yadav, 2021; Bateman et al., 2021). For example, it would be possible to integrate fact-checking directly into social media platforms at the level of individual posts.

3 Data Availability for Research

3.1 Social Media

Meta (previously Facebook) is the most widely used social media platform worldwide. However,

accessing data from Meta has become increasingly challenging over time. Up until the late 2010s, there used to be multiple APIs that provided an entry point to Facebook data. However, the Cambridge Analytica scandal in 2018/2019, when large amounts of personal information were obtained from Facebook and sold for political purposes, led to the closure of most Facebook APIs (Freelon, 2018). Consequently, Facebook put up technical and legal barriers to prevent third parties from accessing personalized information on the platform. Scraping data from Facebook constitutes a violation of the company's Automated Data Collection Terms, and Facebook relies on software to recognize and shut down web scraping applications.

Nevertheless, through a series of initiatives, Meta provides access to selected samples of their data without giving out personal information about users. First, researchers can run ad campaigns on Facebook and track aggregate information on the impact of their campaigns via Meta's marketing API. For example, researchers could examine if users react differently to specific posts. Second, in an attempt to increase transparency about ad campaigns on Facebook, Meta allows access to their "Ad library", allowing researchers to gather data on third-party campaigns (amount of spending, geographical origin) that are currently running on the platform. Third, Meta offers the tool "CrowdTable" that provides engagement metrics and analytics for public pages, public groups, verified profiles, and public Instagram accounts, allowing researchers to gain insights into the spread of information and the popularity of specific posts. Even though the tool was made for professionals like influencers and media outlets, academics can use the tool for research on the information environment. Fourth, Meta has published several curated data sets in its "Data for Good at Meta" initiative. These data sets use aggregate data from Meta's platforms and include, for instance, fine-grained population density estimates, mobility measures, or results from several surveys related to Covid 19. Finally, Meta has partnered with academic researchers, exemplified by the project "Social Science One" with Harvard's Institute for Quantitative Social Science. In a recent project, the company granted access to a large sample of shared URLs on Facebook to a selected group of researchers.

Several of Meta's data-sharing initiations and published data sets can be fruitful for studying the information environment. However, the company's current data protection policies prevent access to individual-level data that academics require for carefully designed studies. At the same time, the company is willing to engage in partnerships with selected academic institutions. From a strategic perspective, the new infrastructure should, therefore, try to build sustainable ties to Meta and other companies in order to shape the company's agenda and influence company-academic partnerships. Fellowships at the PhD or faculty level could be used to establish trust between companies and the new infrastructure.

Furthermore, Meta's transparency and data protection policies are in flux, and significant changes are to be expected soon. According to recent reports, Meta has intensified its collaborations with academics, and the company is testing innovative ways to allow analysis with individual-level data while keeping the data safe. One example is the so-called digital cleaning room, i.e., a space where researchers can access Facebook APIs via a VPN connection and run their analysis in a safe environment. In the end, researchers will only be able to export their results, not the raw data.

In addition to that, the legal environment to access data on social media is undergoing substantive changes. There are ongoing legal battles over the legality of web scraping. For example, in 2021 Meta shut down accounts of NYU researchers affiliated with the Ad Observatory Project who examined Meta's approach to identify misinformation in political ads for alleged violations of its terms of service. A recent ruling from the U.S. Ninth Circuit of Appeals in a dispute between LinkedIn and the data company hiQ Labs stipulated that certain

forms of web scraping are legal. Further changes in the regulatory environment could open new avenues for researchers studying the information environment. As a result, the shared infrastructure should monitor ongoing developments in the regulatory framework and provide legal assistance to help researchers comply with current regulations.

Twitter though this project should explore paying for enhanced access to Twitter data, it is important to understand the data that are available for free via the Academic Research product, Twitter's name for the enhanced access given to approved academic researchers. The Academic Research product was introduced as part of Twitter's transition from its v1.1 to v2.0 endpoints.

The best way to acquire free Twitter data is via the Academic Research product, which is Twitter's name for enhanced access to its API it provides to approved academic researchers. A random sample of tweets are available in real time via the GET /2/tweets/sample/stream endpoint, which provides approximately 1% of all public tweets, or via *post hoc* queries to the GET /2/tweets/search/all endpoint. The search endpoint is especially powerful because a researcher can ask for any still public tweet since the first one on March 26, 2006, a capability not available to researchers for free until the Academic Research product's launch on Jan uary 26, 2021. In addition to tweet text, other useful data made available include user profile information (screen name, user identification number, self-reported location, number of followers and following, profile media, and others) and accounts' followers and following sorted in reverse chronological order.

Free is never truly free, of course, and the Academic Research product is no exception: Twitter does not allow any account to download more than 10 million tweets per month. Since about 500 million tweets *per day* are created, this cap is in fact quite low. The only exception is the random sample of the stream, meaning a research team can collect approximately 5 million tweets per day for free. Since streamed tweets are delivered based on the millisecond of their creation, maintaining multiple connections across a research team will not increase the number of unique tweets delivered (Kergl, Roedler and Seeber, 2014). Since most tweets are anodyne and therefore not relevant for studying the information environment, the random sample is likely to be of limited utility, though it should still be collected. The monthly quota is especially limiting because the relevance of streamed tweets can be improved by connecting instead to the GET /2/tweets/search/stream endpoint. This endpoint allows one to pass filters Twitter uses before returning tweets; tweets can be filtered based on their location, language, keywords, mentioned users, whether they contain media, among others. Importantly, Twitter returns all tweets matching filters until the volume of tweets matches the 1% quantity, meaning one can theoretically capture every tweet containing a keyword or about or from a set of users.

Separate from the Academic Research product, a key source of Twitter data that should be ingested as part of this project are the releases of "all the accounts and related content associated with potential information operations that we have found on our service since 2016" (Gadde and Roth, 2018). Since October 2018, Twitter periodically releases datasets of tweets and media in the tweets. As of December 2021, these releases amount to 200 million tweets from 17 countries and includes 9 terabytes of separate media data (Twitter Safety, 2021). While Twitter defines an information campaign in a way that does not mean every campaign contains misinformation, this content, both the tweets and media, are much more likely to contain misinformation than a random subset of tweets. The data are also useful because they include campaigns associated with countries that span a range of state capabilities, regions, electoral institutions, and socioeconomic variables: Armenia, Bangladesh, Cuba, Ecuador, Egypt, Ghana,

Honduras, Indonesia, Iran, Mexico, Nigeria, the People's Republic of China, Russia, Saudi Arabia, Serbia, Spain, Tanzania, Thailand, Turkey, Uganda, the United Arab Emirates, and Venezuela.

This data should be used to acquire training data to recognize misinformation campaigns as well as learn how to profile accounts affiliated with state backed information operations. This data will help understand behaviors that can be analyzed in non-Twitter datasets in order to link misinformation campaigns across platforms. Since online platforms structure behavior such that not all behavioral features from Twitter will necessarily exist on other platforms (Kreiss, Lawrence and McGregor, 2018), the media data are especially useful since they may be more similar within misinformation campaigns across platforms than account behavior across platforms. The data are also useful because academics can access the original account names (the public releases use anonymized ones). These names can then be used to measure the extent to which past and ongoing research using different datasets is affected by information operation campaigns. The ability to purify researchers' data would be a major contribution of the research infrastructure in its own right.

Reddit is a website organized into thematic subreddits whose content users provide and moderate. Reliance on users means it has the scale of social network platforms, but it is not a social media platform since users do not follow each other; it is best thought of as Wikipedia with more active users providing less persistent content. Its reliance on user provided content means it is also one of the most visited websites in the world, 15^{th} or 20^{th} depending on the source (Alexa and Wikipedia, respectively). It is also a source and amplifier of much misinformation, making its data a valuable, probably necessary, resource for this project. Sub reddits can be directly scraped, though the project would most likely be better off using the pushshift.io API or the data dumps push shift provides.[2]

Instagram since Instagram belongs to Meta, similarly restrictive policies apply. However, Meta offers a series of APIs that allow users to extract metadata, most tailored to clients and professional users. For example, the Instagram Graph API allows professional users to track reactions to their posts, followers, etc. Instagram Basic Display API users would be able to access basic profile information, photos, and videos on their Instagram accounts. Again, the available tools allow for access to the profile of public accounts and access metadata, but currently, no individual-level information can legally be obtained from Instagram. Libraries like Mineur allow you to crawl the Instagram website tags and users as if it was a simple API client.

TikTok compared to platforms operated by Meta, accessing public data from TikTok is easier, and there are several tools to scrape public information. One example is the Unofficial TikTok API in Python by David Teather, which enables users to identify trending posts and fetch user-specific information. In addition to that, there are multiple providers of fee-based providers that offer to scrape TikTok profiles, comments, hashtags, posts, URLs, numbers of shares, followers, and music-related data, such as Apify or Bright Data. Again, scraping specific user-specific information may violate TikTok's user terms, and the platform tries to block web scraping attempts.

YouTube Accessing individual pages on YouTube is relatively straightforward using existing web scraping tools like the Python library Selenium. With a few lines of code, information such as content, comments, likes, shares, and views can be obtained. The same toolset can be used to

assess all content in a specific channel or from a particular user. In addition to that, YouTube offers a series of APIs that facilitate interaction with the platform for channel owners. While the YouTube Analytics API and the YouTube Reporting API are tailored toward channel and content owners, the Data API allows users to interact with YouTube in many ways. For example, it is possible to obtain lists with search results, channel IDs, or comments. Users need to take quota limits into account when running search queries. Further, even though the YouTube API is relatively accessible, important information like the trajectory of viewership cannot be reconstructed retrospectively and therefore must be collected in real time.

Parler is a social media platform modeled after Twitter and promotes ”free speech”, i.e., users are allowed to post any content they like without content moderation. The platform gained prominence after the 2020 elections in the US (Aliapoulios et al., 2021). There are claims that the platform was used to prepare and broadcast the Capitol attack on Jan. 6, 2021. In

January 2021, activists were able to scrape 70TB of data directly from the Parler API due to a lack of security measures. The data set includes followers' personal information, (deleted) posts, and geolocations. While the platform has made changes to the API in response to the data breach in 2021, there are APIs available. For example, there is the unofficial API in Python by Konrad Iturbe that allows access to individual user profiles, a user's Parleys and echoes, trending posts and hashtags. And there is also the official Python API Parler with similar features.

3.2 Chat Applications

Chat applications such as Whatsapp and Telegram are more difficult to access compared to social media since most traffic consists of user-to-user communication, which is typically encrypted. Whatsapp, for example, is part of Meta and the company's restrictive approach has been described above. APIs are only available to business clients and are built for professional purposes and are thus of limited use for social science research. They help advertisers to monitor interactions with potential customers but do not have individual-level information that could be used for social science. However, there are ways to access public channels, usually used to broadcast messages without interactions with the users and open chat groups. Previous work has explored communication dynamics in public Whatsapp groups. For example, Machado et al. (2019) examine videos and images from about 130 Whatsapp groups around the 2018 Presidential elections in Brazil. Saha et al. (2021) identify so-called fear speech, i.e. messages that have the potential to trigger offline violence, in Whatsapp groups in India.

One tool to interact with Telegram is Telethon, a Python library that interacts with the Telegram API. Based on this approach, Baumgartner et al. (2020) created the Pushift data set, which contains hundreds of millions of posts from more than 30,000 public Telegram channels related to right-wing extremist politics and cryptocurrencies. Thus, once a subset of relevant accounts is identified, Telegram can be useful for studying information disseminated by those accounts.

[2]For more information on the push shift data, see (Baumgartner et al., 2020).

3.3 Existing Tools to Study Misinformation

Finally, there are a series of tools that facilitate the dissemination and assessment of data in the information environment. One of such tools is Hoaxy by the Indiana University Network Science Institute (IUNI) and the Center for Complex Networks and Systems Research (CNetS). Hoaxy traces the dissemination of articles on Twitter and allows for the identification of networks via which a specific article is sent. The platform identifies low-quality senders that regularly disseminate misinformation and estimates the probability that a certain piece of information is trustworthy. Another helpful tool is the so-called Botometer (formerly BotOrNot) that checks the activity of a Twitter account and gives it a score that represents the probability that a given account is not a human user (Davis et al., 2016).

Similar to the Botometer, Bot Sentinel is a tool that uses machine learning to identify Twitter accounts that violate the platform's terms of use by spreading disinformation. Based on the analysis of several hundred tweets for each account Bot sentinel produces a score between 0 and 100 that represent the likelihood of an account engaging in harmful behavior. The underlying model was trained on several thousand accounts and, according to the creators, can detect bots with 95% accuracy. To identify toxic behavior, the creators focused on accounts that repeatedly violated Twitter's rulers without any further assessment of whether the posted content was, in fact, problematic. As the creators emphasize, "Ideology, political affiliation, religious beliefs, geographic location, or frequency of tweets are not factors when determining the classification of a Twitter account."

In addition to automated approaches, there are efforts to rely on qualitative assessments of websites to study disinformation. For example, the The Global Disinformation Index is a US-based not-for-profit organization that aims at raising awareness of disinformation in ad campaigns. The goal is to help businesses identify disinformation campaigns and collect a list of offending websites and apps in different countries. The so-called "Dynamic Exclusion List" contains major offenders and offers help to companies to assess the risk of collaborating with specific websites and apps based on qualitative research. Unfortunately, the list was not publicly available when writing this report. Another non-profit, Ranking Digital Rights scores tech companies on the clarity of their rules around content moderation, commitment to user privacy, and algorithmic transparency.

There are also several tools that allow for the study of visual content. Forensically, for example, is a free image verification tool where users can upload images to check whether they were manipulated in some way. Among other things, the tool allows users to zoom into the image, find similar areas and examine compressed versions of the image. The tool also helps to identify suspicious changes in brightness and shows hidden metadata such as geotags. Ghiro is an open source tool with similar functionality.

4 Technical Challenges

This section considers the technical challenges that researchers face when studying misinformation. We focus on the acquisition of data, storing data, and preprocessing data, each of which take significant resources.

4.1 Acquiring Data

In order to acquire data from social media companies, researchers must either scrape the data, connect to the Application Programming Interface (API), or purchase the data. These all pose

their own unique challenges. While most direct, scraping data directly from the social media websites may violate terms of service and run into anti-scraping software. As discussed above, programs to scrape data are in a legal gray zone and can cause conflict between researchers and social media companies.

As noted in the previous section, not all social media companies have APIs meant for researcher access. When they do, acquiring data from social media companies' APIs often requires multiple IP addresses. For example, Twitter only allows one connection to its API per IP address. If the research infrastructure collects Twitter data in real time and will also need to access other endpoints, then it will need as many IP addresses as there are unique processes. Multiple IP addresses are relatively easy to acquire. For example, computer devices like Blackberry Pis or cloud solutions like Amazon EC2 instances cost less than $100, the former as a one time cost and the latter as a recurring expense. One can also purchase proxy IP addresses and route requests through them. Alternatively, processes can be carefully coordinated to minimize the number of IP addresses simultaneously required. The main drawback of using APIs is that the quantity of data delivered is usually a small percentage of a company's actual data, and there are often types of data only available via purchase. These limitations allow companies to charge for access to more and different data. Prices for purchasing data are unpublished for most companies and require talking to sales representatives; they should be assumed to run into at least the low four figures per month. These costs can be prohibitive from researchers, particularly those without large research budgets. Though purchasing is more expensive than relying on APIs, it saves engineering time. Since the engineering required to route requests through dozens of IP addresses represents time not spent on analysis, the cost of purchasing data may be offset with the labor cost saved. Given that data is acquired, an additional difficulty is that some universities impose restrictions on data ingestion. University shared computing facilities are generally designed to ingest, process, and export large amounts of data *within the university domain itself*. They are not designed to connect to sources of data external to the university and download those data. Security concerns drive this limitation.

Any resulting solution will therefore need to be constructed in close collaboration with the appropriate campus information technology organizations or reside outside of a university. Campus IT security will probably want to monitor all inbound traffic and restrict queries so they only come from predefined IP addresses. If the acquiring infrastructure resides outside of the university, it can still send acquired data to storing and processing infrastructure.

4.2 Storage

After data is acquired, they need to be stored in an accessible manner. This storage occurs separately from the original data to protect them from accidental or malicious manipulations. Traditionally, relational databases have been the implementations that make the underlying data available, but the rise of digital data, much of which is unstructured or semi-structured, has made relational databases suboptimal for many types of data analysis. This obsolescence is especially true for transactional data, data that does not need constant validation and syncing. Since the data this infrastructure uses are transactionless, as are most if not all academic datasets, databases will not be the primary method of storing data.

It is most likely the case that data will be stored using a distributed file system from which a few specialized databases are created. "Distributed file system" means the files containing the original data, or files that are copies of the original data, are accessed directly, and items in each

file are extracted for subsequent analysis. Files that contain the same type of data, such as tweets or Reddit posts, can then be accessed in parallel. Hadoop was the original distributed file system, but Apache Spark is the new standard in this space.

The infrastructure will still need some databases, especially to store metadata, and they will increase the amount of storage space that needs to be provisioned. This point is discussed in further detail in the processing subsection.

It is difficult to estimate how much space the data will require, but it is probably on the order of hundreds of terabytes or petabytes. For example, the 1% of tweets that Twitter delivers for free via its streaming endpoints - approximately 5 million per day - is about 1.6 gigabytes per day or 584 gigabytes per year, compressed. Uncompressed, those numbers are approximately 10.4 gigabytes and 3.8 terabytes. Downloading additional data to construct panel, network, or image datasets can easily match and exceed the size of the streaming tweets. Images from tweets that Twitter delivers are about 300 kilobytes uncompressed, and approximately 10% of tweets contain images; a heuristic is that every 10,000 images from Twitter require 1 gigabyte of storage. As of April 2022, all of Wikipedia is approximately 177.44 gigabytes compressed.[3] Reddit posts require about 10 compressed gigabytes per month and comments on those post 20.

Once an estimate of the hard drive space needed for the first years' of the project is determined, a backup plan is essential. Options here include cloud solutions such as AWS S3 or Glacier and its competitors from Microsoft and Google. Those require sending the project's data to a third party, which may not be ideal from security and ethical perspectives. If the project's implementation has several physical locations for accessing data, those locations are necessarily equivalent and therefore act as backups. If the data needs to remain in one location, various Redundant Array of Inexpensive Disks (RAID) configurations generate backups. Any solution chosen will double, at a minimum, the amount of storage required. Determining the exact backup plan should be done in consultation with information technology experts.

In actuality, all of the details discussed up to this paragraph should be finalized with academic and information technology experts in the relevant domains. For example, campus IT personnel will best know the specific requirements governing connecting to external services, strengths and weaknesses of specific data backup solutions, and which databases to use for each type of data. Especially if the infrastructure spans multiple physical locations, the technical implementation details should be separated from the overall architecture of the infrastructure, and these latter should be the primary focus of the academic team at this time (Gilardi et al., 2022).

Finally, the infrastructure may need to be able to update data in response to notifications from data providers. For example, Twitter now provides a compliance endpoint that it expects developers to use so that only public tweets and accounts are used in a product. One submits tweet or user identifiers, and Twitter indicates whether either is deleted, deactivated, private, suspended, or geographic information should be removed. Using this endpoint, and equivalents at other platforms, can also provide useful research data.

4.3 Data Processing

Once acquired and stored, the data needs to be processed to facilitate researchers' analysis. The processing step has two primary goals: providing researchers metadata with which to structure search queries and using these metadata to recognize probable misinformation. Metadata can be generated from the text, image, and location information in content.

Natural language processing (NLP) is important to understand the intended meaning of a piece of text. One challenge of creating a shared infrastructure is that exactly what measures a researcher needs will depend on the dataset and question of study (Grimmer, Roberts and Stewart, 2021). Text classifiers that may work well for one particular definition of misinformation may work poorly for another, or classifiers that work for one platform or modality may work poorly on another platform. Thus all analysis of text must be tuned to the particular application and dataset a researcher is using. Still, there are general tools that a research infrastructure

[3]English Wikipedia represents 11.66% of Wikipedia and is 20.69 gigabytes compressed. $20.69 * \frac{1}{.1166} = 177.44$.

could provide such as pre-trained models or hand-labeled datasets and codebooks that could be repurposed and fine-tuned by researchers for their own application. Another challenge of a shared infrastructure in text processing is the numerous languages that social media operates in. Ideally, the researchers could analyze text in any language. In practice, the availability of off the shelf tools positively correlates with the amount of content in each language. Deciding on which language to create models necessitates grappling with thorny issues of representation.

Image and video analysis is also likely to be necessary because memes - images with overlaid text designed to make an argument - and short videos are becoming the primary vector of misinformation. How to proceed with image and video analysis in turn depends on exactly how images and videos will be used in research. Downloading all images and videos is often not feasible because of the bandwidth and storage space that would be required; one image requires as much storage space as 50,000-1.5 million words, depending on its size. For example, images from 3.6 million tweets from two years in Venezuela require 220 gigabytes of storage.

Research projects therefore need to determine if every image and video possible should be downloaded. If the answer is no, the question becomes how to choose what to download. To know which images and videos to download, it is useful to have an idea of what information they contain. Knowing what an image is about requires seeing the image, which negates the gains from not downloading irrelevant images. A shared infrastructure on social media could enable this decision making by providing metadata on images and video to enable selection based on image and video attributes.

5 Ethical and Organizational Challenges

5.1 Privacy

While social media data is a treasure trove for social scientists because it allows for unprecedented access to postings of people all over the world, it also poses significant ethical challenges for those using it in research. While some social media platforms like Twitter are primarily public, much interaction on social media takes place on platforms where messages are shared privately with a group of friends or acquaintances. Some social media exchanges might be considered semi-public, where users can opt into a group message, and then are privy to posts by users in the group. In each case – private, semi-private, and public – researchers must grapple with the ethical issues of the consent and expectations of those whose information appears in the data.

Much of interaction on social media occurs in non-public spaces. Users on Facebook, for example, share social media posts to their timelines for a large group of friends or within groups where people have to be admitted to join. Platforms such as WeChat, WhatsApp, and Instagram

have significant content that is shared privately.

Any shared research infrastructure to study misinformation will have to grapple with what, if any, data could be used that is not public. While social media companies have access to this data for research because of consent that users provide in the terms of service agreement, they may be less willing to share this data with researchers outside of their system, particularly if they have no control over the research questions it is used for. Existing models, such as Social Science One, allow for researchers to use differentially private or aggregated data that obscures the identity of the users. Yet adding noise to the data to obscure identity and aggregating the data to higher levels can also erode its usefulness for research questions.

Even if researchers were to find a legal avenue for gaining access to this private data, researchers and their institutions' Institutional Review Boards will have to grapple with the ethical questions of using private data without explicit consent of each user. Even in the best case that a provision were added to the terms of service agreement for cooperating social media companies to allow for user data to be used by selected researchers, questions remain about whether users read and understand terms of service (Obar and Oeldorf-Hirsch, 2020) and whether users are able to opt out of specific provisions and still use the service (Hans, 2012). Users may check the box that their data could be used for research, but not be aware of or even object to the types of studies it ends up being used for (e.g. for facial recognition). Further, access to identifiable private data in a shared research infrastructure significantly increases the risk that it could be misused by researchers in ways that are not intended by the shared infrastructure, as the 2016 case of Cambridge Analytica makes clear.

While the use of private data by researchers poses significant challenges, even the use of public data has ethical challenges. Research has shown that many people are not aware that their public social media postings are or could be used for research, and would not approve of their data being used in a study (Fiesler and Proferes, 2018). A portion of users also think their accounts are private when they are public (McClain et al., 2021). In many cases, public postings can include information about people who themselves have not consented to them being posted, for example, users may post pictures of their friends without their consent online.

Many Institutional Review Boards consider studies using public social media data either exempt from human subjects review or not human subjects research, as long as the researchers are not interacting with users on the platform or studying private data. But clearly, a higher standard needs to be set than simply accepting all research that uses public social media data from ethical review, based both on user expectations and potential harms that could result in collecting potentially identifiable public data. Institutional review boards and other ethics reviews should offer more guidance on when the benefits of the research outweigh the risks of using the data. In addition to overall guidance on the ethics on particular research designs, there are several areas where more guidance could be given:

First, researchers could use more ethical guidance in how to think about ethically posting replication data for studies that use publicly available social media data. While replication data is useful for ensuring research transparency, it makes identifiable data more public, and can be in tension with the terms of service of social media companies. For example, Twitter requires that researchers not post the raw data of each social media post, but instead post message IDs, allowing users to delete content and not have it archived in replication files. Further, replication data when paired with research that algorithmically or manually assigns user posts or accounts to labels runs the risk of publicly labeling individuals in a way that they may not agree with, or could be incorrect (Benton, Coppersmith and Dredze, 2017; Chancellor et al., 2019).

Second, IRBs could offer guidance as to when it would be ethical for researchers to link

public social media data with other public data (McKee, 2013). Users may also not intend to have their data linked, for example to have social media accounts linked across platforms, or their social media accounts linked to other public information. Data linkage can expose information about users in ways that can de-identify them (Conway et al., 2014).

5.2 Researcher Access

Any shared infrastructure for social media research will need to confront the issue of deciding which researchers will be able to access the platform and which researchers will have the power to influence the direction of the infrastructure. Existing work on social media and politics, both in academic and within social media platforms, is dominated by U.S. institutions. These researchers do not reflect the demographic diversity of the U.S. public, much less the world as a whole. This lack of diversity can have negative impacts on the questions that we ask and on the ability to understand how social media affects communities around the world.

A new shared infrastructure could easily become an "insider's club" where only researchers from particular institutions or with particular academic pedigrees can get access to the shared infrastructure. Instead, diversity of thought both in terms of the demographics of the researchers involved, the location and country of the institution they work at, and the types of questions they are interested in asking should be prioritized. To make this successful, a shared infrastructure should actively seek out diverse scholars from a wide variety of institutions. Success in this area would require making funding available for researchers from less well-resourced institutions.

5.3 Conflict of Interests and Company Cooperation

Any shared infrastructure developed to study social media will require cooperation between researchers and social media companies. Some of the previous work on social media required internal access to companies, which often required pre-publication review. This can drastically reduce the usefulness and credibility of the findings. Any shared infrastructure for researchers would need to limit pre-publication review by social media companies, or at least these companies' power in influencing the integrity of the research.

Even without pre-publication review, because the findings of the research could influence the public image of the platform, any shared infrastructure will need clear conflict of interest and disclosure requirements, both for proposals and for publications. Since many academics consult for social media companies, clear rules need to be established to manage perceived conflicts and communicate any conflicts to readers.

More broadly, if voluntary cooperation from social media companies is required for the maintenance of the shared infrastructure, then any shared infrastructure will need to grapple with how this need for cooperation will affect the research pursued. Steps will need to be taken to mitigate influence of collaborating companies over the research process and output.

5.4 Open Science and Replication

Any shared infrastructure will need to find a way that research results can be replicated and verified by other researchers in a way that both protects the infrastructure and the privacy of the individuals within the data. Privacy concerns as well as proprietary algorithms used in data analysis could make replication more difficult for outside researchers. Finding a way to give

verified outside users who want to replicate results will be important at ensuring that the results can be built upon by others. Limiting the use of proprietary infrastructure that can not be shared publicly would also help ensure that others can repeat work completed in the shared infrastructure.

5.5 Studying Misinformation in the Academy

The study of misinformation spans multiple fields, industry, academia, and policy. While this provides enormous opportunities, it also poses challenges under the traditional academic publishing model. Interdisciplinary research can be undervalued by universities, especially when disciplines have different publishing norms. For example, computer scientists prioritize publishing in conferences, which are less recognized in social science. Computer scientists often have a more inclusive definition of authorship, where author order signals contribution, where social scientists are known to publish alphabetically with a higher bar for authorship. These norms affect the evaluation and promotion of individuals working at the intersection of these fields, and therefore their incentives to participate.

Further, traditional fields in both social and technical sciences may not value the pressing policy questions that we outlined above. For example, political science and economics often values carefully crafted causal questions linked to larger social science theories. In contrast, many of the questions we need to answer are purely descriptive (Munger, Guess and Hargittai, 2021).

Last, social media and social media policy is constantly changing. Research might be relevant to decisions today, but outdated in a few months. By contrast, peer review in the social sciences can take up to a year. How do we make research actionable and relevant to the fast pace of social media, while also ensuring it stands up to the scrutiny of the peer review process? Coordinating the timeline of policy decisions and research could be one contribution of a collaborative, shared infrastructure.

6 The Road Ahead: An International Institute for Research on the Information Environment

In this section, we lay out product proposals and a general organizational structure that we think should guide the creation of an institute for the study of social media and misinformation. To advance social-scientific research on misinformation, we propose five specific research products that we think would significantly enhance the researcher community's ability to solve the most pressing problems in misinformation. Creating these products would require solving some of the challenges we addressed above, including data access, computation and storage, and legal and ethical challenges. Since solving these challenges as individual researchers is prohibitive, we believe that by doing this collectively, as a community of researchers, we could make significant progress in understanding misinformation – its prevalence, impact, and policies that should govern it.

We first lay out the five specific product proposals and discuss their challenges. Then we discuss considerations around access to these products with some recommendations of how that could be structured. Next, we discuss how legal and ethical challenges could be handled organizationally by such an institute. And last, we propose a long-term vision for increasing the international scope of the center, so as to address important problems of misinformation

worldwide.

6.1 New Data Products

In this section, we propose a set of products that we think the institute should prioritize to make the most progress on the questions we discussed above. The motivation for proposing these specific products draws from our own assessment of which data products would be most useful in studying misinformation on social media. It also draws on insightful commentary from Pasquetto et al. (2020).

6.1.1 Misinformation and Featurization Products

The first category of product that we think is necessary is a product that helps researchers label and identify misinformation. The first challenge is definitional. For research on misinformation to be comparable and as cumulative as possible, we consider it necessary that scholars at the institute openly discuss, and possibly even agree on, a common definition of misinformation. Current research is hampered by the fact that studies differ in their definition and identification of online misinformation, which is likely to have considerable effects on empirical results. A new joint effort to study misinformation should therefore strive to arrive at a common definition and a coding standard for misinformation, which can then be used throughout the research at the institute. This definition and the corresponding coding protocol must be jointly developed among a group of experts, and must be openly communicated. Also, we fully acknowledge that these conventions are unlikely to remain static, so it is necessary to allow for revisions and adaptations over time as new types and topics of misinformation emerge.

To provide a fast way to identify misinformation, a misinformation product would include a tool that labels misinformation based on the above definition through machine coding. Here it could be useful to build on recent attempts to automate fact-checking. For example, the CheckThat! Workshop brings together computer scientists working on automatic tools for this. Providing such an automated service via an API, either within the institute only, or possibly even to the larger research community, would foster the adoption of a common standard, while at the same time making revisions to that standard easy to implement (since only the central classification service would have to be adjusted). The API should allow for the submission of content to be checked, with results returned in a standardized, machine-readable way. Both adoption and revision of coding standards would be much more difficult for manually annotated and verified content, since the path dependence of researchers is significantly higher. A common standard would also facilitate competition of researchers or research teams, which could stimulate progress in misinformation research.

For researchers to be able to tune their own models for identifying misinformation or related topics tuned to their own research project, the misinformation and featurization product could also include a tool that returns features social media text, images, and video that are typically useful for identifying the topic, category, or important metadata from social media that identifies misinformation. For example, such a tool could provide pre-trained topic models or pre-trained classifiers that could be fine-tuned to particular applications using transfer
learning. A tool could also create features for video and image data, for example text tags that identify objects in an image or video, location information, or geo-tagging that could enable researchers to conduct their own analyses. For example, places are often mentioned in text, so analyzing content from the same account for location names provides a close approximation of

an account's location (Ryoo and Moon, 2014). The same capability exists for images that contain landmarks (Zhou et al., 2018).

6.1.2 Social Media Data and Engagement Products

A social media data and engagement product would give researchers access to social media data from a variety of different platforms and information about the spread and popularity of particular posts on the platform. We imagine that this product would have different features. First, it would have a simple query feature which would allow researchers to query public social media data based on keyword, time, or other post and account metadata. This is similar to the existing Twitter API, however, researchers require data from more platforms. As recent years have shown, online exchange and debate can quickly switch from one platform to another. For example, many users switched to Parler after former President Trump was banned from Twitter. Also, different platforms attract different user bases, and the concentration of research efforts onto a few selected ones can bias results and make generalizations difficult.

Second, this product would provide information about how much a given post spread on social media and who was most likely to have seen the post. This product could come in two forms. First, researchers could query the social media stream for posts that were most seen and engaged with by certain types of users. This data would not be limited to engagement metrics like likes and shares, but would also include whether the post was in a user's news feed or if the user was likely to have read it. Second, researchers could give the product a set of social media posts and receive information about each posts' spread among which types of users and during which time periods.

These features would be crucial to the study of misinformation. The social media query feature would allow researchers to characterize the types of posts that are written on platforms, including the degree to which they contained misinformation. The spread feature would help researchers understand what types of misinformation becomes popular, and in doing so understand the effects of this information on the public.

6.1.3 Account Flag Product

As discussed above, one pressing challenge of identifying disinformation is knowing the intent of the actors behind spreading it. Existing tools like Botometer allow researchers to get the probability that an account is an automated bot. However, these tools could be scaled to provide a number of different useful account flags. Like Botometer, it would be useful to know which accounts are likely automated and which accounts have content that is posted by humans. Such a tool could be generalized for platforms beyond Twitter.

In addition, this tool could tag certain types of institutional accounts, including media accounts and government accounts. In collaboration with platforms, it could also provide information about accounts that are likely to be systematically coordinated with one another, such as accounts that reprint the same content in a coordinated way or large numbers of accounts that are initiated from the same device or IP address.

6.1.4 Content Moderation Products

To better understand how countermeasures affect misinformation, the institute should prioritize

collaborating with platforms to build a content moderation product. In its ideal form, this product would give researchers access to a representative stream of posts that were deleted or downgraded by the social media company because its content violated the platform's policies. In an alternative version of the product, researchers could provide posts to the product and the tool could provide a likelihood that that post would have been removed or down weighted at a given point in time. This product would allow researchers to understand what types of content platforms are responding to in which time periods and would allow them to begin to assess the effectiveness of these counter-measures.

6.1.5 Survey Products

Last, the institute should seek to recruit a representative set of social media users on each platform who would be willing to answer surveys that are linked with the social media data. This would require two features. One, the institute should try to keep and update descriptive statistics about the metadata that characterizes a representative set of users on each platform. This information in and of itself would be useful to researchers who might wonder if representative of their own datasets are of a platform and who could use this information to weight their data to increase the generalizability of their research findings.

Second, the institute could recruit a set of users who are representative of each platform to participate in surveys or donate other metadata that could be linked with their social media data. Attempts have been made to do this. The survey company YouGov, for example, recently launched their 'Safe" product, which allows researchers to track the online behavior of users through access to their browsing history and cookies. YouGov also has panels of survey respondents who have agreed to share their social media data. Similar data is available from the Mozilla Rally project, where users voluntarily donate data on their online behavior for research. This is of course available only for a limited panel of users which have given their consent, but it may be worthwhile to rely on a similar approach for the institute. In particular, this would allow researchers to observe online activity and interaction "in the field," which is generally extremely difficult. Such surveys would allow researchers to answer broader questions about the ways in which social media might influence political beliefs and political behavior.

6.2 Different Levels of Access

The new infrastructure needs to be widely accessible in order to facilitate research advances, but most of the data gathered will have terms of service that restrict its sharing. The project's ambition and innovation is therefore also a source of weakness. This tension is irreducible, though there are at least four designs that trade off different amounts of access and security. We discuss these designs below, in order of increasing access granted and decreasing security, before making a recommendation of how they could be used for the new infrastructure.

One secure location The most secure option is to require researchers to be physically present at the infrastructure in order to use it. This approach is the most secure because it does not require distributing copies of the project's data, it is easier to monitor individuals for security risks when they are physically present, and the number of people accessing the data at any moment in time can be tightly controlled. This approach is similar to how access is governed for large natural science tools like the Large Hadron Collider. While the active work with the raw and sensitive data happens at the central facility only, it must be possible for aggregated,

anonymized and otherwise reduced results to be exported, for the purpose of (partial) replication (as required by many journals) and further analysis.

Several secure locations The next most secure option is to provide secure access at certain locations via secure data enclaves. A data enclave is a restricted access room with its own hardware that provides access to sensitive data. An important decision when designing a data enclave is whether or not it maintains an internet connection to the original data or is air gapped (has no internet connection). Air gapping is most secure but requires periodic physical delivery of data so the enclave's data matches the project's. By making copies of sensitive data available in multiple locations, researchers' travel costs are reduced and more researchers can benefit from the project's data. The 31 Federal Statistical Research Data Centers the United States Census Bureau maintains is a prominent example of the secure data enclave approach.

Project-provided laptops for access The third option is to allow researchers to access the project's data without traveling to an enclave. This access is granted via a laptop the project team provides, and the laptop can be configured with however much security is desired. For example, internet access could be restricted to only visit certain domains, such as the project's and Stack Overflow, software installation can be forbidden, and the laptop can be inoperable between certain hours. Provisioning laptops is the most common method for private companies working with researchers. Similarly, the project could use a virtual data enclave (VDE), i.e. providing access to the project's data via remote desktop software. Using remote desktop software allows for strict controls over data access, export, and analysis just like if a researcher accesses the data in person or at a secure data enclave. VDEs are slightly less secure than secure data enclaves or provisioning laptops because users can take screenshots or use screen recording software, capabilities which can be removed on project specific hardware. Prominent providers of VDEs are the Inter-university Consortium for Political and Social Research, Bureau of Labor Statistics, and the Women's Health Initiative from the National Heart, Lung, and Blood Institute.

Researchers use their own equipment The fourth option would allow researchers to access the data and the APIs described above using their own machines. Given the sensitive nature of the data this project will collect and what are likely to be restrictive terms of service, this option cannot be used to access much of the data researchers need to study the information environment (see above). This approach will need to be careful to stay within each data provider's terms of service, and the data it could provide will not be as useful as granting researchers access to the underlying raw data.

6.2.1 Recommendation: The Inner and the Outer Circle

We recommend that the new Institute use a combination of the different access levels described above. Most work with data from commercial platforms will only be possible under the maximum level of data protection and security provided by the single location access model. We call this the "inner circle" of the Institute's environment. For the Institute to be successful in establishing collaborations with providers, this model is likely the only one that providers will agree to, if they agree to data sharing and processing outside their own infrastructure at all. At

the same time, however, this access model poses the highest practical obstacles to researchers. Not only will people have to relocate temporarily to this facility, they also need to adjust to a new working environment without access to their own equipment, development tools etc. For this reason, we recommend a two-stage process, where researchers work within an outer circle before entering the inner circle. The outer circle provides researchers and their teams with access to the different products described above, but without these products delivering full access to the underlying data they are based on. This way, researchers can develop methods and code to work with these products, but without access to sensitive data. This allows them to quickly gain speed once they have moved to the inner circle, which reduces the time frame of their research stays at the Institute. Through this combination of different access levels, we believe that the Institute can effectively maximize data access, while keeping practical obstacles for researchers as low as possible.

6.3 Pre-defined Ethical, Legal and Scientific Protocols

An infrastructure of the scale proposed that can work across platforms, countries, and questions, needs its own internal legal, ethical, and scientific protocols specifically designed to address the challenges of studying social media and misinformation. Having a centralized ethics board, legal team, and agreement on scientific standards would have many benefits.

Collaboration with commercial platforms Having a lab-specific legal team could allow for the researchers at the lab to collectively negotiate agreements between the lab and social media companies that would facilitate researcher access to data. Such a team would also be necessary to give advice on the constantly changing set of regulations on data and data sharing worldwide, helping researchers conduct cross-national studies. Finally, researchers could use this resource on seeking advice on auditing social media company's platforms using automated scraping.

Access by researchers Last, to establish high scientific standards, the lab should have clear scientific protocols for the research conducted in the lab and a transparent peer review process for those seeking access to the lab. Such protocols should cover whether to require pre-analysis plans and clear standards of transparency, documentation, and reproducibility of the research. They would also cover proper procedures to mitigate conflicts of interest, such as those described earlier in this document. A scientific advisory board could be established to select projects to be conducted at the lab. Individuals and teams need to be verified before accessing the lab (at least for full access from the inner circle). This could be done through an open call for projects, where researchers submit project proposals detailing the research question(s) they intend to pursue, as well as descriptions of the data they require. The board needs to ensure that selected projects adequately represent institutional, geographic, demographic, and career stage diversity. Given that misinformation is especially impactful outside of OECD countries, it is normatively important to ensure access for researchers from those parts of the world. To make this possible, there should be no financial charge for performing research at the Institute.

Ethics review A centralized ethics board would develop a set of ethical standards specific to social media research, above and beyond Institutional Review Boards, which were developed mainly with biomedical research in mind. Such a board could offer advice on the ethical

conundrums offered above, such as on the ethics of posting replication data, linking social media data with external datasets, and the distinction between public and private data. Beyond research at the lab, these ethical standards could also serve as a blueprint for social media research conducted worldwide.

Scientific collaboration and publication To properly incentivize those participating in creating the infrastructure, the lab should create clear norms around publication and authorship. Scientists employed by the lab to collect and process data should be included on publications when their work makes significant contributions to these publications, which should be clearly laid out in advance to establish publication expectations. To ensure the timely distribution of research, the lab should help publicize research findings and incentivize publication and distribution of research.

6.4 Toward a Global Network

While the new infrastructure will be set up in the US and its main activities will be carried out in the United States, we recommend creating a global network with smaller regional branches worldwide. Building on existing ties between researchers at the main center and top scholars abroad, we envision establishing satellite centers that facilitate collaboration and the dissemination of knowledge and promote equal access to products developed at the Institute. This global network could create synergies to benefit the Institute and the scientific community by offering access to the new data products proposed in this report. One example of the kind of global network we have in mind is the system of regional centers established in the Varieties of Democracy (V-Dem) project that coordinate data collection efforts and organize outreach activities. While the project headquarters is located in Sweden, there are eight V-Dem regional centers in the Balkans, Central Asia, East Asia, Eastern Europe and Russia, North America, South Africa, and Southern Europe. A similar setup could strengthen the infrastructure's position in the international academic community.

Facilitating collaboration The proposed satellite centers would strengthen the Institute in two ways. First, they would aggregate pressing research questions at a regional level and thereby increase the Institute's impact in a world where societal challenges require a global perspective. The satellite centers connect researchers to local knowledge and pave the way for innovative projects, especially with researchers from non-Western countries. Consequently, the Institute's research agenda would take a global perspective and cater to the needs of a worldwide scientific community. Second, new ties created by the satellite centers would help recruit academics from those locations and partner with social media companies located in these regions. Since legal requirements for scientific research on the information environment vary a lot from country to country, the proposed satellite centers could support joint research projects and serve as a first point of contact for stakeholders such as social media companies, civil society or state agencies. On top of that, satellite centers might open up new funding opportunities from local agencies.

Knowledge transfer In addition, the proposed satellite centers would support the core institution in its outreach activities and the dissemination of knowledge. Top scholars from differ ent regions affiliated with the proposed satellite centers would act as multipliers of research output.

The aforementioned regional centers in the V-Dem project regularly host outreach events such as symposia or workshops with a regional focus.

Promoting equal access As described in section 6.2, we recommend different levels of access to products produced by the Institute due to safety concerns. While necessary, safety measures such as the requirement to work on-site make access to the Institute's infrastructure unequal and create gatekeepers. Researchers without sufficient funds to travel to the secure location(s) cannot access the Institute's data products. This applies particularly to researchers from the Global South, whose mobility is often additionally constrained by visa issues. The satellite centers could reduce this challenge in multiple ways. First, they could serve as data enclaves (see section 6.2) and host products that do not require full security clearance while directly copying other scientific and legal procedures developed at the main center. Second, the satellite centers would help manage access to the central infrastructure by pre-screening potential collaborators or vetting research proposals to be carried out at the Institute. As a side effect, the establishment of satellite centers would strengthen academic institutions abroad.

References

Aliapoulios, Max, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini and Savvas Zannettou. 2021. A Large Open Dataset from the Parler Social Network. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15 pp. 943–951.

Allcott, Hunt and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *Journal of Economic Perspectives* 31(2):211–36.

Allcott, Hunt, Matthew Gentzkow and Chuan Yu. 2019. "Trends in the diffusion of misinformation on social media." *Research & Politics* 6(2):1–8. Publisher: SAGE Publications Sage UK: London, England.

Bateman, Jon, Elonnai Hickok, Jacob N. Shapiro, Laura Courchesne and Julia Ilhardt. 2021. "Measuring the Efficacy of Influence Operations Countermeasures: Key Findings and Gaps From Empirical Research.". Partnership for Countering Influence Operations, Carnegie Endowment for International Peace.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire and Jeremy Blackburn. 2020. "The Pushshift Reddit Dataset." *Proceedings of the International AAAI Conference on Web and Social Media* 14(1):830–839.

Benegal, Salil D and Lyle A Scruggs. 2018. "Correcting misinformation about climate change: The impact of partisanship in an experimental setting." *Climatic Change* 148(1):61–80.

Benton, Adrian, Glen Coppersmith and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the first ACL workshop on ethics in natural language processing*. pp. 94–102.

Bradshaw, Samantha and Philip N Howard. 2018. "The global organization of social media disinformation campaigns." *Journal of International Affairs* 71(1.5):23–32.

Brotherton, Robert and Lisa K. Son. 2021. “Metacognitive Labeling of Contentious Claims: Facts, Opinions, and Conspiracy Theories.” *Frontiers in Psychology* 12.
URL: *https://www.frontiersin.org/article/10.3389/fpsyg.2021.644657*

Bursztyn, Leonardo, Aakaash Rao, Christopher P Roth and David H Yanagizawa-Drott. 2020. “Misinformation during a pandemic.”. National Bureau of Economic Research. NBER Working Paper No. 27417.

Chancellor, Stevie, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. pp. 79–88.

Committee on National Security Systems. 2015. “Glossary.” CNSSI No. 4009.
URL: *https://www.niap-ccevs.org/Ref/CNSSI 4009.pdf*

Conway, Mike et al. 2014. “Ethical issues in using Twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature.” *Journal of Medical Internet Research* 16(12):e3617.

Davis, Clayton Allen, Onur Varol, Emilio Ferrara, Alessandro Flammini and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on the world wide web*. pp. 273–274.

Deibert, Ronald, John Palfrey, Rafal Rohozinski and Jonathan Zittrain. 2008. *Access denied: The practice and policy of global internet filtering*. The MIT Press.

Diamond, Larry. 2010. “Liberation technology.” *Journal of Democracy* 21(3):69–83.

Feldstein, Steven. 2021. *The Rise of Digital Repression: How Technology is Reshaping Power, Politics, and Resistance*. Oxford: Oxford University Press.

Fiesler, Casey and Nicholas Proferes. 2018. “‘Participant’ perceptions of Twitter research ethics.” *Social Media+ Society* 4(1):1–14.

Freelon, Deen. 2018. “Computational Research in the Post-API Age.” *Political Communication* 35(4):665–668.

Gadde, Vijayya and Yoel Roth. 2018. “Enabling further research of Information Operations on Twitter.”. Available at https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter.

Gemenis, Kostas. 2021. “Explaining Conspiracy Beliefs and Scepticism around the COVID-19 Pandemic.” *Swiss Political Science Review* 27(2):229–242.

Gilardi, Fabrizio, Lucien Baumgartner, Clau Dermont, Karsten Donnay, Theresa Gessler, Mael ¨ Kubli, Lucas Leemann and Stefan Muller. 2022. “Building research infrastructures to study ¨ digital technology and politics: Lessons from Switzerland.” *PS: Political Science & Politics* 55(2):354–359.

Gillespie, Tarleton. 2018. *Custodians of the Internet*. Yale University Press.

Gorwa, Robert, Reuben Binns and Christian Katzenbach. 2020. “Algorithmic content moderation:

Technical and political challenges in the automation of platform governance." *Big Data & Society* 7(1):2053951719897945.

Grimmer, Justin, Margaret E Roberts and Brandon M Stewart. 2021. "Machine learning for social science: An agnostic approach." *Annual Review of Political Science* 24:395–419.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson and David Lazer. 2019. "Fake news on Twitter during the 2016 U.S. presidential election." *Science* 363:374–378.

Guess, Andrew M, Brendan Nyhan and Jason Reifler. 2020. "Exposure to untrustworthy websites in the 2016 US election." *Nature Human Behaviour* 4(5):472–480.

Hans, Gautam S. 2012. "Privacy policies, terms of service, and FTC enforcement: Broadening unfairness regulation for a new era." *Mich. Telecomm. & Tech. L. Rev.* 19:163.

International Telecommunication Union. 2021. "Measuring digital development. Facts and Figures 2021.". Available at https://www.itu.int/en/ITU-D/Statistics/Documents/facts/FactsFigures2021.pdf.

Jerit, Jennifer and Yangzi Zhao. 2020. "Political misinformation." *Annual Review of Political Science* 23:77–94. Publisher: Annual Reviews.

Jhaver, Shagun, Amy Bruckman and Eric Gilbert. 2019. "Does transparency in moderation really matter? User behavior after content removal explanations on reddit." *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW):1–27.

Jiang, Shan and Christo Wilson. 2018. "Linguistic signals under misinformation and fact checking: Evidence from user comments on social media." *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):1–23.

Jolley, Daniel and Karen M Douglas. 2014*a*. "The Effects of Anti-vaccine Conspiracy Theories on Vaccination Intentions." *PloS one* 9(2):e89177.

Jolley, Daniel and Karen M Douglas. 2014*b*. "The Social Consequences of Conspiracism: Exposure to Conspiracy Theories Decreases Intentions to Engage in Politics and to Reduce One's Carbon Footprint." *British Journal of Psychology* 105(1):35–56.

Keremoglu, Eda and Nils B. Weidmann. 2020. "How Dictators Control the Internet: A Review ˇ Essay." *Comparative Political Studies* 53(10-11):1690–1703.

Kergl, Dennis, Robert Roedler and Sebastian Seeber. 2014. On the endogenesis of Twitter's Spritzer and Gardenhose sample streams. In *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE pp. 357–364.

Kreiss, Daniel, Regina G. Lawrence and Shannon C. McGregor. 2018. "In Their Own Words: Political Practitioner Accounts of Candidates, Audiences, Affordances, Genres, and Timing in Strategic Social Media Use." *Political Communication* 35(1):8–31.

Kuklinski, James H., Paul J. Quirk, Jennifer Jerit, David Schwieder and Robert F. Rich. 2000. "Misinformation and the currency of democratic citizenship." *The Journal of Politics* 62(3):790–816. Publisher: University of Texas Press.

Liang, Fan, Qinfeng Zhu and Gabriel Miao Li. 2022. “The Effects of Flagging Propaganda Sources on News Sharing: Quasi-Experimental Evidence from Twitter.” *The International Journal of Press/Politics* . Online first.

Machado, Caio, Beatriz Kira, Vidya Narayanan, Bence Kollanyi and Philip Howard. 2019. A Study of Misinformation in WhatsApp groups with a focus on the Brazilian Presidential Elections. In *Companion proceedings of the 2019 World Wide Web conference*. pp. 1013–1019.

McClain, Colleen, Regina Widjaya, Gonzalo Rivero and Aaron Smith. 2021. “The behaviors and attitudes of US adults on Twitter.” *Pew Research Center* .

McKee, Rebecca. 2013. “Ethical issues in using social media for health and health care re search.” *Health policy* 110(2-3):298–301.

Mejias, Ulises A and Nikolai E Vokuev. 2017. “Disinformation and the media: the case of Russia and Ukraine.” *Media, culture & society* 39(7):1027–1042.

Mena, Paul, Danielle Barbe and Sylvia Chan-Olmsted. 2020. “Misinformation on Instagram: The impact of trusted endorsements on message credibility.” *Social Media + Society* 6(2):1– 9.

Munger, Kevin, Andrew M Guess and Eszter Hargittai. 2021. “Quantitative description of digital media: A modest proposal to disrupt academic publishing.” *Journal of Quantitative Description* (1):1–13.

Myers West, Sarah. 2018. “Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms.” *New Media & Society* 20(11):4366–4383.

Nyhan, Brendan and Jason Reifler. 2010. “When Corrections Fail: The Persistence of Political Misperceptions.” *Political Behavior* 32(2):303–330.

Nyhan, Brendan, Jason Reifler and Peter A Ubel. 2013. “The hazards of correcting myths about health care reform.” *Medical care* 51(2):127–132.

Obar, Jonathan A and Anne Oeldorf-Hirsch. 2020. “The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services.” *Information, Communication & Society* 23(1):128–147.

O’Connor, Cailin and James Owen Weatherall. 2019. *The misinformation age*. Yale University Press.

Pasquetto, Irene V, Briony Swire-Thompson, Michelle A Amazeen, Fabr´ıcio Benevenuto, Na dia M Brashier, Robert M Bond, Lia C Bozarth, Ceren Budak, Ullrich KH Ecker, Lisa K Fazio et al. 2020. “Tackling misinformation: What researchers could do with social media data.” *The Harvard Kennedy School Misinformation Review* .

Pew Research Center. 2021. “Mobile Fact Sheet.”. Available at https://www.pewresearch.org/internet/fact-sheet/mobile/.

Pierri, Francesco, Brea Perry, Matthew R. DeVerna, Kai-Cheng Yang, Alessandro Flammini, Fil ippo Menczer and John Bryden. 2021. “The impact of online misinformation on U.S. COVID 19 vaccinations.” *CoRR* abs/2104.10635.

URL: *https://arxiv.org/abs/2104.10635*

Richey, Sean. 2017. "A Birther and a Truther: The Influence of the Authoritarian Personality on Conspiracy Beliefs." *Politics & Policy* 45(3):465–485.

Roberts, Margaret E. 2018. *Censored*. Princeton University Press.

Rød, Espen Geelmuyden and Nils B Weidmann. 2015. "Empowering activists or autocrats? The Internet in authoritarian regimes." *Journal of Peace Research* 52(3):338–351.

Ryoo, KyoungMin and Sue Moon. 2014. "Inferring Twitter user locations with 10 km accuracy." *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion* pp. 643–648.

Saha, Punyajoy, Binny Mathew, Kiran Garimella and Animesh Mukherjee. 2021. "Short is the Road that Leads from Fear to Hate": Fear Speech in Indian WhatsApp Groups. In *Proceedings of the Web Conference 2021*. pp. 1110–1121.

Shao, Chengcheng, Pik-Mai Hui, Lei Wang, Xinwen Jiang, Alessandro Flammini, Filippo Menczer and Giovanni Luca Ciampaglia. 2018. "Anatomy of an online misinformation network." *PLOS ONE* 13(4):e0196087.

Silva, Bruno Castanho and Sven-Oliver Proksch. 2021. "Fake it 'til you make it: A natural experiment to identify European politicians' benefit from Twitter bots." *American Political Science Review* 115(1):316–322.

Stanley, Tom D, Hristos Doucouliagos and Piers Steel. 2018. "Does ICT generate economic growth? A meta-regression analysis." *Journal of Economic Surveys* 32(3):705–726. Publisher: Wiley Online Library.

Swire-Thompson, Briony, Joseph DeGutis and David Lazer. 2020. "Searching for the backfire effect: Measurement and design considerations." *Journal of Applied Research in Memory and Cognition* 9(3):286–299.

Taber, Charles S and Milton Lodge. 2016. "The Illusion of Choice in Democratic Politics: The Unconscious Impact of Motivated Political Reasoning." *Political Psychology* 37:61–85.

Tucker, Joshua A, Andrew Guess, Pablo Barbera, Cristian Vaccari, Alexandra Siegel, Sergey ´ Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social media, political polarization, and political disinformation: A review of the scientific literature.". Available at https://ssrn.com/abstract=3144139orhttp://dx.doi.org/10.2139/ssrn.3144139.

Tucker, Joshua A, Yannis Theocharis, Margaret E Roberts and Pablo Barbera. 2017. "From ´ liberation to turmoil: Social media and democracy." *Journal of Democracy* 28(4):46–59.

Twitter Safety. 2021. "Disclosing state-linked information operations we've re moved.". Available at https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed.

Van der Linden, Sander, Anthony Leiserowitz, Seth Rosenthal and Edward Maibach. 2017. "Inoculating the public against misinformation about climate change." *Global Challenges*

1(2):1600008.

van Mulukom, Valerie, Lotte Pummerer, Sinan Alper, Hui Bai, Vladimira Cavojova, Jessica E M ´ Farias, Cameron S Kay, Ljiljana B Lazarevic, Emilio J C Lobato, Gaelle Marinthe and et al. ¨ 2020. "Antecedents and Consequences of COVID-19 Conspiracy Beliefs: A Systematic Review." PsyArXiv Preprint.

Vegetti, Federico and Moreno Mancosu. 2020. "The impact of political sophistication and motivated reasoning on misinformation." *Political Communication* 37(5):678–695.

Walter, Nathan, John J Brooks, Camille J Saucier and Sapna Suresh. 2021. "Evaluating the im pact of attempts to correct health misinformation on social media: A meta-analysis." *Health Communication* 36(13):1776–1784.

Walter, Nathan, Jonathan Cohen, R Lance Holbert and Yasmin Morag. 2020. "Fact-checking: A meta-analysis of what works and for whom." *Political Communication* 37(3):350–375.

Weidmann, Nils B. and Espen Geelmuyden Rød. 2019. *The Internet and Political Protest in Autocracies*. Oxford Studies in Digital Politics. New York: Oxford University Press.

Williams, Matthew L, Pete Burnap, Amir Javed, Han Liu and Sefa Ozalp. 2020. "Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime." *The British Journal of Criminology* 60(1):93–117.

Wu, Liang, Fred Morstatter, Kathleen M. Carley and Huan Liu. 2019. "Misinformation in social media: definition, manipulation, and detection." *ACM SIGKDD Explorations Newsletter* 21(2):80–90. Publisher: ACM New York, NY, USA.

Yablokov, Ilya. 2015. "Conspiracy theories as a Russian public diplomacy tool: The case of Russia Today (RT)." *Politics* 35(3-4):301–315.

Yadav, Kamya. 2021. "Platform Interventions: How Social Media Counters Influence Operations.". Partnership for Countering Influence Operations, Carnegie Endowment for International Peace.

Zhou, Bolei, Agata Lapedriza, Aditya Khosla, Aude Oliva and Antonio Torralba. 2018. "Places: A 10 million Image Database for Scene Recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(6):1452–1464.