Digital Disinformation

A PRIMER

By Tim Hwang

Atlantic Council
EURASIA CENTER

Konrad Adenauer Stiftung

Revelations around Russian efforts to shape the 2016 US presidential election through the use of disinformation, bots, and hacking have thrust the problems of "fake news" and social media manipulation into the public spotlight.

This primer is an introduction to this phenomenon, laying out the key terms, major actors, and potential legislative actions that might be taken.

September 2017

Key Terms

Artificial Intelligence (AI)

Colloquially, the field of computer science research focused on enabling machines to mimic or reproduce cognitive function.

Bot

In the context of social media, a bot is a user account that is controlled autonomously by software. Often, these accounts purport to be genuine users on platforms like Twitter and Facebook.

CDA 230

Section 230 of the Communications Decency Act—an important legal provision that shields online platforms from liability for user-generated, user-uploaded content. CDA 230 is perceived as a disincentive for platforms to mediate or filter content, as well as a factor discouraging intervention in instances of harassment or hate speech.

Computational Propaganda

The use of algorithms, automation, and human curation to purposefully distribute misleading information over social media networks.

Disinformation

Intentional actions by individuals and groups that—either knowingly or unknowingly—result in the spread of false or misleading information.

DMCA

The Digital Millennium Copyright Act. Among many other provisions, the act provides a "safe harbor" in which online platforms will not be liable for user-generated or user-uploaded content. The DMCA is seen as a disincentive for platforms to mediate or filter content, as well as a factor discouraging intervention in instances of harassment or hate speech.

Echo Chamber

Social spaces in which ideas, assumptions, and beliefs are continually repeated and reinforced among a group of similarly minded members.

Machine Learning

The subfield of artificial intelligence that specifically studies systems that improve themselves with data. Recent breakthroughs in this subfield have driven much of the excitement around artificial intelligence, and the two are often synonymous in everyday usage.

MADCOM

Machine driven communications. The use of AI and machine learning techniques to fabricate text, audio, and video content for distribution online, particularly in the context of an effort to spread disinformation.

Metadata

Data that describes other data. Analysis of metadata generated by users of online platforms has become an important focus in trying to gain an understanding of consumer behavior for advertisers, researchers, and the platforms themselves.

Platform

In this context, companies that own and maintain online services that typically include hosting and sharing user-generated content (e.g., Facebook, Twitter, Instagram), as well as curating collections of third-party content responsive to a query by a user (e.g., Google).

Wisdom of the Crowd

The notion that the aggregated observations of many users will help to weed out inaccuracies and falsehoods. Implies that with a sufficient number of users, the user-generated content on a platform will essentially be self-filtered for truthful information. This idea informed the design of several platforms—notably Twitter and Reddit—during the 2000s.

Why Does Disinformation Matter?

Nearly 60 million Americans have used the Internet to help them "make big decisions or negotiate their way through major episodes in their lives in the previous two years."[1] It is also increasingly the channel through which news is disseminated and understood. As the Internet has become a widely relied upon channel for learning about the world, effective manipulation by malicious actors can shape perceptions, threatening democratic processes, markets, and national stability. Over the past few years, researchers and journalists have observed efforts by a range of different actors to do precisely this.

It is important to note that high-profile campaigns of disinformation may be corrosive regardless of their effectiveness. By eroding trust in information generally, these campaigns may decrease the credibility of reliable sources of information online as well. To that end, disinformation can undermine the ability of journalists and others to ensure accountability and transparency throughout society.

1 John B. Horrigan and Lee Rainie, The Internet's Growing Role in Life's Major Moments, Pew Research Center, April 19, 2006, http://www.pewinternet.org/2006/04/19/the-internets-growing-role-in-lifes-major-moments/.

Key Players and Vulnerabilities

Although the threat posed by disinformation is shaped by public receptiveness, mass media, and many other factors, Internet platforms have emerged as key channels through which these campaigns spread their messages. There are several major entities that are hosts to the largest numbers of users in the United States and have attracted some of the highest-profile efforts to spread disinformation:

Twitter

Founded in 2006, Twitter is an online news and social networking service where users post and interact with messages, "tweets," restricted to 140 characters. Only registered users can post tweets, but most content can be read by anyone online.

Facebook

Founded in 2004, users create profiles indicating their name, occupation, education, and other information. Users then connect with other users, exchange messages, post status updates and photos, and share videos and news. The News Feed, a core part of the site, curates this activity automatically as a user's network of "friends" engages with the platform.

Google

Founded in 1998 and most known for its search engine that surfaces relevant content across the web in response to a query, Google maintains a range of web products including user-generated video hosting (YouTube) and a service that algorithmically gathers news stories from across the web (Google News).

Reddit

Founded in 2005, Reddit is a site for news aggregation and discussion. Reddit's registered users submit content such as text posts and images, and can submit direct links to outside sources including video. This content is then voted on by the community to generate a list of "hot" content on the front page of the site.

Instagram

Founded in 2010, Instagram is an application (app) that enables users to share, comment, and explore user-generated photos and video. The app also supports the editing of this content through digital filters, which modify the appearance of the images. The platform was acquired by Facebook in 2012.

Snapchat

Founded in 2011, Snapchat is a messaging platform that allows users to share images and video with text captions that are only available

for a short period of time before disappearing. The platform also supports sharing, discovery, and collaboration around "stories," i.e., collections of images that make up short-form storylines.

Platforms are limited in four important respects that may result in them being persistently vulnerable to disinformation campaigns and hinder an effective response to them.

- *Ideological Constraints:* Many platforms have taken a relatively hands-off approach toward the veracity of content flowing through their services, driven by a core belief in free expression and the "wisdom of the crowd." This attitude is evident in entrepreneurs like Twitter co-founder Evan Williams, who has explained that a core idea of the platform was that "once everybody could speak freely and exchange information and ideas, the world is automatically going to be a better place."[2] Facebook co-founder Mark Zuckerberg likewise reflects this thinking and recently emphasized his view that it is imperative that the company "be extremely cautious about becoming arbiters of truth ourselves."[3] These ideological constraints have also informed the design of platforms like Reddit, which curate content primarily based on the "upvotes" and "downvotes" of its users.

- *Technical Constraints:* Designing an effective response to online disinformation campaigns requires a solution that can quickly be applied by a machine to millions or billions of pieces of content uploaded by users to platforms each day. However, false or misleading information can take many forms, and users can have many different views on what is "true." This makes designing a generalized filter for disinformation challenging on a technical level.

- *Legal Constraints:* Regulations such as the DMCA and CDA 230 create incentives for platforms to largely avoid filtering or intervening in content hosted on their services, simultaneously shielding them from liability for that content. Though this has largely made user-generated platforms viable, it has also made platforms less active than they might otherwise be in combatting activity such as harassment, hate speech, and disinformation.[4]

- *Financial Constraints:* Many online platforms rely on advertising as a core source of revenue. This encourages platform designs that feature easy account creation and work to deepen user engagement and time on the site. These business interests may create incentives to avoid making changes that would inhibit the ease of joining a service for the purpose of excluding bots, or aggressively targeting disinformation that may nonetheless generate significant attention and sharing activity.

---

2 David Streitfeld, "'The Internet Is Broken': @ev Is Trying to Salvage It," *New York Times,* May 20, 2017, https://www.nytimes.com/2017/05/20/technology/evan-williams-medium-twitter-internet.html?mcubz=1.

3 Mark Zuckerberg, Status Update - November 12, 2016, Facebook, https://www.facebook.com/zuck/posts/10103253901916271.

4 Cf. Sarah Jeong, *The Internet of Garbage,* (Forbes Signature Series: 2015, EPUB).

Known Perpetrators

Disinformation campaigns are driven by a range of perpetrators with different motivations. There are three important categories of actors that have invested prominently in these activities:

Political Actors

- Disinformation can be politically valuable, organized by state and non-state actors looking to manipulate discourse around targeted political leaders and institutions.
- With their substantial resources, these campaigns can be multifaceted, creating original content distributed through state-owned media and third-party intermediaries to spread false information. They can also leverage paid operatives and swarms of bots to help propagate disinformation on a user-to-user level and to suppress opposition.

> Example: Advanced Persistent Threat (APT) 28 and APT 29 are two major cyber-espionage actors identified by US intelligence as having played a role in the 2016 disinformation campaign. Both are associated with Russian intelligence and have shown methods consistent with the sophisticated capabilities of nation-state actors.[5]

Commercial Interests

- The monetization of the Internet through advertising has also produced a financial motive for creating disinformation, which is widely shared through the Internet and drives traffic to a website.
- Commercial campaigns are focused on the maximization of traffic around a fixed set of web properties and lack the resources of state-owned media to help produce and propagate sophisticated false content. As a result, many of these sites copy or only slightly modify content drawn from elsewhere on the web.

> Example: Residents of Veles, Macedonia discovered that content about Donald Trump increased the pageviews on their websites, leading to increased ad revenue for them. Over 150 pro-Trump domains were found to be registered to individuals in Veles alone. Stories that maximize traffic are exploited, regardless of their value as news.[6]

5 NCCIC / FBI, *GRIZZLY STEPPE - Russian Malicious Cyber Activity,* December 29, 2016, https://www.us-cert.gov/sites/default/files/publications/JAR_16-20296A_GRIZZLY%20STEPPE-2016-1229.pdf.

6 Samanth Subramanian, "Inside the Macedonian Fake-News Complex," *Wired*, February 15, 2017, https://www.wired.com/2017/02/veles-macedonia-fake-news/.

Trolls

- Disinformation campaigns can emerge purely from a motivation to seek entertainment or notoriety online. "Trolling"—bullying activity aimed at provoking anger in targets for the amusement of the perpetrators—has been a long-standing feature of online culture.
- Trolling campaigns are coordinated through informal, usually anonymous groups of users online and can deploy a range of techniques from the fabrication of misleading documents and other information, to coordinated, direct attempts to engage and mislead other users online.

  Example: The proliferation of racist and misogynist cartoon images apparently supporting Donald Trump's presidential campaign resulted in a number of traditional media investigations. Trolls fed journalists stories intended to provoke outrage, and then offered other journalists the story of how they fooled the first journalists. The resulting media frenzy increased the visibility of the offensive memes it sought to critique.[7]

- These groups do not always operate independently of one another, and they are not always easily distinguishable. Russian state actors were also connected to efforts to mobilize informal online troll communities to spread disinformation during the election season.[8] Consequently, what researchers observe is a multilayered ecosystem of disinformation efforts emerging online.

Emerging Threats

As disinformation campaigns are exposed to the public and actions are taken to discourage these activities, perpetrators are forced to advance the "state of the art." There are three key trends and technologies that likely will significantly increase the influence of these campaigns going forward:

Cyber Attacks

- Disinformation campaigns increasingly exploit hacking to further their ends. This serves the purpose of helping to distract and disrupt the ability of targets to resist these campaigns, as well as bolstering the credibility of certain channels spreading the stolen information. It also produces opportunities to uncover embarrassing or other discrediting information on targets.

7 Jesse Singal, "How Internet Trolls Won the 2016 Presidential Election," *New York*, September 16, 2016, http://nymag.com/selectall/2016/09/how-internet-trolls-won-the-2016-presidential-election.html.

8 Office of the Director of National Intelligence, Background to "Assessing Russian Activities and Intentions in Recent US Elections": The Analytic Process and Cyber Incident Attribution," January 6, 2017, https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Example: In 2017, hackers were able to compromise several news sites within Qatar to spread a disinformation narrative discrediting the country's emir and encouraging the regional blockade of the country.[9]

The Evolution of AI and Machine Learning

- Breakthroughs in AI and the subfield of machine learning are generating many new methods for believably simulating human behavior with machines. These methods are published openly, and the technical tools to deploy them are increasingly available as the cost of high-powered computing decreases. This will likely encourage key perpetrators to experiment with these techniques to make their disinformation easier to believe and harder to debunk going forward.

Example: Face2Face, a recent demonstration from Stanford University, shows that machine learning can be used to create believable fabricated video from footage of existing individuals, such as political figures and other key leaders.[10]

Metadata

- Perpetrators benefit from low-cost access to precise metrics around the success of the false information that they produce. Using off-the-shelf services designed for marketing, disinformation campaigns can access granular information about the audience and adjust strategies in near-real time to more effectively spread disinformation.
- Leveraging metadata, academic researchers have also increasingly come to understand the complex processes by which content is shared, gains credibility, and eventually goes "viral" through a community of users. Future campaigns might use this knowledge to effectively time the distribution of messages to key influencers within a network, creating mass cascades of activity around their content.

Example: One recent paper develops detailed models for how hoaxes persist within Wikipedia and are spread throughout the web.[11] Such information could be used to help malicious actors produce disinformation that more reliably avoids detection and spreads more effectively.

9 Patrick Wintour, "Russian hackers to blame for sparking Qatar crisis, FBI inquiry finds," *Guardian*, June 7, 2017, https://www.theguardian.com/world/2017/jun/07/russian-hackers-qatar-crisis-fbi-inquiry-saudi-arabia-uae.

10 Justus Thies, et al., "Face2Face: Real-Time Face Capture and Reenactment," 2017, http://www.graphics.stanford.edu/~niessner/thies2016face.html.

11 Srijan Kumar, et al., "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes," MobiCom '98 Proceedings of the 4th annual ACM/IEEE international conference on Mobile computing and networking, pp 233-241, Dallas, Texas, USA—October 25-30, 1998, PDF available at http://dl.acm.org/citation.cfm?id=2883085.

Policy Recommendations

Congress can take a number of proactive steps to help combat the threat from disinformation campaigns and empower existing efforts currently underway:

- *Better Fact-Checking Tools:* Funding should be put toward the creation of collaborative online platforms for fact-checking digital media, enabling journalists and citizens to more quickly work together to reject or authenticate disinformation circulating through the web.[12]
- *Media Literacy Surge:* Investments and partnerships can be made in media literacy campaigns to educate the public on best practices in evaluating the quality of information online and encourage them to play a role in actively challenging disinformation where it exists.
- *Investigating "Warning Labels":* Studies should be initiated to investigate the potential effectiveness of "warning labels" that signal to users of online platforms when the veracity of a given piece of content has been called into question by journalistic outlets.[13] This work would also explore the feasibility of creating robust, machine-readable signals of information quality that could be integrated into ranking algorithms and other systems.
- *Bolstering Journalism*: Government can convene relevant players and encourage the development of partnerships that ensure reliable financing of investigative journalism from the public and private sector.
- *Public Alert Systems:* Legislation could be passed to require online platforms to provide data on significant campaigns of disinformation to the public in real time. This would provide researchers, journalists, and the public an increased awareness of the activity and an ability to assess it.[14]

12 Existing efforts include projects like Check. For further information, please see: Meedan, *Check*, https://meedan.com/en/check/.

13 An early example of such "disputed" labels are being prototyped by Facebook. Jon Constine, "Facebook now flags and down-ranks fake news with help from outside fact checkers," *TechCrunch*, December 15, 2016, https://techcrunch.com/2016/12/15/facebook-now-flags-and-down-ranks-fake-news-with-help-from-outside-fact-checkers/.

14 Further discussion of how increased transparency may help to combat disinformation: Tim Hwang and Sam Woolley, "The Most Important Lesson From the Dust-Up Over Trump's Fake Twitter Followers," *Slate*, June 2, 2017, http://www.slate.com/articles/technology/future_tense/2017/06/the_lesson_of_the_dust_up_over_trump_s_fake_twitter_followers.html.

The Atlantic Council is a nonpartisan organization that promotes constructive US leadership and engagement in international affairs based on the central role of the Atlantic community in meeting today's global challenges.

The Konrad-Adenauer-Stiftung is a German political foundation and international think tank with strong ties to the political party of German chancellor Angela Merkel. The Konrad-Adenauer-Stiftung is working worldwide to strengthen democracy and the rule of law.
On the basis of Christian democratic values, the Konrad-Adenauer-Stiftung is a reliable partner and promoter of German-American friendship, a strong Atlantic partnership, and intensive cooperation between the United States and the European Union.

Atlantic Council

1030 15th Street, NW, 12th Floor,
Washington, DC 20005

(202) 463-7226, www.AtlanticCouncil.org