TRANSATLANTIC WORKING GROUP

Actors, Behaviors, Content: A Disinformation ABC

Highlighting Three Vectors of Viral Deception to Guide Industry & Regulatory Responses†

Camille François
Graphika and Berkman Klein Center for Internet & Society at Harvard University[1]

September 20, 2019

Contents

Introduction

As the historic phenomenon of propaganda[2] unfolds today in a variety of social-media manifestations, a plethora of terms has emerged to describe its different forms and their implications for society: "fake news," online disinformation, online misinformation, viral deception, etc.[3] The speed and scale at which disinformation is now able to spread online has led to mounting pressure on regulators around the globe to address the phenomenon, yet its multifaceted nature makes it a difficult problem to regulate. Effective remedies must take into account the different vectors of contemporary disinformation and consider the multiplicity of stakeholders, tradeoffs in different approaches, disciplines, and regulatory bodies able to meaningfully contribute to responses.

† One in a series: A working paper of the Transatlantic High Level Working Group on Content Moderation Online and Freedom of Expression. Read about the TWG: https://www.ivir.nl/twg/.

Major technology platforms have invested in better responses to disinformation, notably by adapting their community guidelines or terms of service. Observed through the lens of platform enforcement, "disinformation" breaks down into a number of different violations manifest on different products which are enforced by distinct teams. This points to a key concern with regard to the current industry responses to viral deception: while disinformation actors exploit the whole information ecosystem in campaigns that leverage different products and platforms, technology companies' responses are mostly siloed within individual platforms (if not siloed by individual products!).

This concise "ABC" framework doesn't aim to propose one definition or framework to rule them all, but rather seeks to lay out three key vectors characteristic of viral deception[4] in order to guide regulatory and industry remedies. Manipulative **actors**, deceptive **behaviors**, harmful **content**: each vector presents different characteristics, difficulties, and implications. Unfortunately, they are also often intertwined in disinformation campaigns, suggesting that effective and long-term approaches will need to address these different vectors with appropriate remedies.

This "ABC" also seeks to reconcile approaches throughout applicable disciplines (e.g., cybersecurity, consumer protection, content moderation) and stakeholders. While the public debate in the U.S. has been largely concerned with **actors** (who is a Russian troll online?), the technology industry has invested in better regulating **behavior** (which accounts engage in coordinated and inauthentic behavior?) while governments have been most preoccupied with **content** (what is acceptable to post on social media?).

"A" is for Manipulative Actors

"On the Internet, nobody knows you're a ~~dog~~ Russian military operative."[5]

The Russian disinformation campaign targeting the U.S. 2016 presidential election[6] has brought to the public's attention how keen certain government actors were to leverage social media to manipulate and influence audiences at home and abroad by engaging in information operations. It has also painfully brought to light the lack of government and industry preparedness and proactivity in the face of these threats. The cybersecurity sector, which bears the brunt of detecting these threat actors and preventing their nefarious activities, had been most focused on protecting physical networks and not enough on detecting those actors on social media networks. Facebook's April 2017 white paper on the issue of information operations (which also marks the first in-depth acknowledgement of this problem by a large technology platform) makes this point clearly and acknowledges that the Facebook cybersecurity team had to expand its scope to appropriately respond to this threat: "We have had to expand our security focus from traditional abusive behavior, such as account hacking, malware, spam and financial scams, to include more subtle and insidious forms of misuse, including attempts to manipulate civic discourse and deceive people."[7]

Manipulative actors, by definition, engage *knowingly* and with clear intent in viral deception campaigns. Their campaigns are *covert*, designed to obfuscate the identity and intent of the actor orchestrating them. Throughout the technology industry, detection and enforcement of this vector of viral deception campaigns rely on the cat-and-mouse game of a) identifying threat actors willing and able

to covertly manipulate public discourse and b) keeping those actors from leveraging social media to do so,[8] as they refine their strategies to evade detection.

Because this detection practice has its roots in the cybersecurity realm, terms of service and community guidelines do not always address these issues, or provide a clear basis to support detection and enforcement efforts against manipulative actors. Precedents in this area include platform rules laying out specific actors who are prevented from using the services (e.g.,. Foreign Terrorist Organizations[9]), but it is worth noting that no major platform to date has included language in its terms of service explicitly prohibiting governments from covertly using its services to conduct influence campaigns.[10] Setting an industry precedent, in August 2019, following investigations disclosing that Chinese State-controlled media leveraged Twitter advertising to promote content critical of pro-democracy protests in Hong Kong, Twitter announced that it would no longer allow "State-controlled media" to use its advertising products.[11] The state-controlled media entities can continue to remain "organic users" (meaning normal and/or verified accounts on the Twitter platform), but their ability to use ads to reach users who are not already following them is now restricted. In doing so, Twitter will likely face difficulty determining which entities are "taxpayer funded entities" and "independent public broadcasters" allowed to use the advertising services vs. "state-controlled media (...) financially or editorially controlled by the state" prohibited from doing so. States have also used a variety of techniques to conceal their direct involvement in seemingly independent online media properties: the Kremlin-controlled Baltnews network[12] and the Iranian-controlled IUVM[13] network are good illustrations.

Note that this problem has little to do with "banning" anonymity or pseudonymity online: both serve important purposes in protecting vulnerable voices and enabling them to participate in critical conversations.[14] Banning anonymity/pseudonymity would prevent such participation while doing little to prevent sophisticated and well-funded actors from exploiting this vector. The deceptive actors we are concerned with here are well-funded military and intelligence apparatus or campaign apparatus, not "somebody sitting on their bed that weighs 400 pounds," as President Trump famously characterized the anonymous troll. Clint Watts describes these figures as "Advanced Persistent Manipulators,"[15] a moniker that stresses the parallels and overlaps between the actors engaged in information operations[16] and hacking:[17]

> Similar to the challenge APT[18] actors have posed to information and cyber security professionals, social media companies now face malign actors that can be labeled as Advanced Persistent Manipulators (APMs) on their platforms. These APMs pursue their targets and seek their objectives persistently and will not be stopped by account shutdowns and platform timeouts.... They have sufficient resources and talent to sustain their campaigns, and the most sophisticated and troublesome ones can create or acquire the most sophisticated technology.[19]

Since 2017, we have seen multiple examples of viral deception campaigns whose *primary* vector is a deceptive actor. Notable examples include false persona "Guccifer 2.0"[20] used by the GRU, false identities tying back to the Islamic Republic of Iran Broadcasting and operating on multiple platforms,[21] and Facebook's December 2018 takedown of accounts in Bangladesh that were found to be misrepresenting their true identity and attempting to mislead voters ahead of the elections.[22]

Governments also have a role to play in detecting and mitigating harms caused by manipulative actors online, although defining the contours of government action in this space remains a largely unexplored policy question. Around the U.S. 2018 midterms elections, for instance, the U.S. government led actions to detect and share relevant information on manipulative actors with the technology sector[23] and to disrupt and deter these actors from engaging in information operations.[24]

"B" is for Deceptive Behavior

"On the Internet, nobody knows you're a ~~dog~~ bot army."

Deceptive behavior is a fundamental vector of disinformation campaigns: it encompasses the variety of techniques viral deception actors may use to enhance and exaggerate the reach, virality and impact of their campaigns. Those techniques run from automated tools (e.g., bot armies used to amplify the reach and effect of a message) to manual trickery (e.g., paid engagement, troll farms). At the end of the day, deceptive behaviors have a clear goal: to enable a small number of actors to have the *perceived impact* that a greater number of actors would have if the campaign were organic.[25]

Interestingly, while there are significant differences in the various disinformation definitions and terms of service applicable to the issue among technology companies, the focus on *deceptive behavior* appears to be a clear convergence point throughout the technology industry.

Google's definition of disinformation, as laid out in its February 2019 White Paper on "How Google Fights Disinformation," points to those deceptive behaviors as a core vector of how disinformation affects Google's platforms:

> We refer to [...] deliberate efforts to deceive and mislead using the speed, scale, and technologies of the open web as "disinformation."[26]

In Facebook's case, deceptive behavior is mostly defined through the "Coordinated Inauthentic Behavior"[27] policy, which has led to numerous takedowns since it was implemented in 2018.[28] While Facebook has shared records and data points regarding the content and accounts taken down for their participation in "coordinated and inauthentic behavior," enforcement in this realm remains opaque throughout the major technology companies.

While the detection and mitigation techniques in this area can be similar to spam detection, an area generally opaque for the public and regulators and not subject to much public scrutiny, the free speech implications of taking down *content* and social media *accounts* (especially political content during election cycles) justify much higher scrutiny of these practices. Relevant questions to technology platforms in this area include:

- Applicable rules: Which are the applicable policies set forth by the platform to address deceptive behaviors on their products?
- Enforcement: What enforcement options are available to the platforms to take action against accounts and content that violate the rules on deceptive behavior? Platforms generally acknowledge a range of options from content demotion to account suspension, although those enforcement options are rarely spelled out for users or made clear for users affected.

- Detection and prioritization: Which teams are effectively in charge of detecting deceptive behaviors, how much of this detection relies on machine learning classifiers (and which ones?), and how does prioritization of potential issues and focus areas work at the platform level?
- Transparency: How will affected users (including good faith actors mistakenly engaging in deceptive behaviors, consumers of information spread by deceptive behavior, bad faith actors seeking to best understand what telltale signs trigger enforcement, etc.) be notified when action is taken against content or accounts? Can those decisions be appealed, and if so, how? Will the platform share transparency metrics regarding its enforcement of rules relative to distortive behavior, both at the annual and the aggregate level (through the existing mechanism of Transparency Reports) and through press releases published when enforcement happens?
- Product vulnerabilities and changes: When deceptive behaviors exploit vulnerabilities in platforms and products, what changes are made to address them?[29]

The industry's lack of proactivity in tackling some of these campaigns and growing public anxiety about disinformation have led regulators to craft frameworks to specifically address deceptive behavior. California's "Bot Law," for instance, is a clear attempt to regulate deceptive behavior on social media:

> It shall be unlawful for any person to use a bot to communicate or interact with another person in California online, with the intent to mislead the other person about its artificial identity for the purpose of knowingly deceiving the person about the content of the communication in order to incentivize a purchase or sale of goods or services in a commercial transaction or to influence a vote in an election. A person using a bot shall not be liable under this section if the person discloses that it is a bot. The disclosure required by this section shall be clear, conspicuous, and reasonably designed to inform persons with whom the bot communicates or interacts that it is a bot.[30]

The "Manipulative Actor" and "Deceptive Behavior" vectors are particularly challenging to address through effective regulatory frameworks because of the dramatic asymmetry of information between the platforms targeted by these campaigns and the rest of the world. While open-source investigation techniques and a few available tools allow others to scrutinize online activity for campaigns run by a manipulative actor or using deceptive techniques, it is undeniable that platforms have much more visibility into those issues than external researchers and stakeholders. Some platforms' community standards or terms of service either indirectly prevent the type of external research that may lead to detecting and exposing distortive behaviors (e.g., when existing and important safeguards also prevent researchers from collecting the data they'd need to analyze distortive behaviors) or directly seek to prevent it (e.g., with rules explicitly preventing the use of data in order to perform detection of deceptive behavior).

Finally, some of the platforms' own systems may actually enhance those deceptive behaviors by disinformation actors: algorithmic reinforcement is a core concern in this area.[31] While anecdotal evidence suggests machine learning based recommendations systems may easily be gamed into promoting campaigns "boosted" by adversarial distortive behavior, the difficulties discussed above

with regard to external research have prevented more systematic examinations of these issues throughout the various platforms.

"C" is for Harmful Content

"On the Internet, nobody knows you're a ~~dog~~ deepfake."

Finally, it is sometimes the case that the content of posts and messages justifies classifying a campaign as an instance of viral deception. Content is the most visible vector of the three: while it is difficult for an observer to attribute messages to a manipulative actor or to observe behavioral patterns across a campaign, every user can see and form an opinion on the content of social media posts. This is likely why regulators have focused on content aspects when regulating disinformation.

This vector calls for detection and enforcement strategies in the realm of content moderation[32]. Unfortunately, regulatory and legal frameworks often struggle to properly define categories of "harmful content" they seek to regulate (see ongoing debates about the definitions of "violent extremism," "hate speech," "terrorist content," etc.) or to properly take into account that a lot of the speech they consider to be "harmful" is protected under human rights law. Governments' appetite to regulate viral deception through the content lens risk further eroding protections to freedom of expression online.

The intersection of harmful content and disinformation campaigns can manifest in several ways:

- Entire categories of content can be deemed "harmful" because they belong to the realm of viral deception, e.g., health misinformation.[33]

Technology platforms have so far mostly proposed to address the categories of content deemed most "harmful" for their disinformation nature by adding context for users alongside the content, such as "flags" or "fact-checking" content. Some platforms though have taken a more radical route by banning entire categories of disinformation content from their services.

Photo-sharing platform Pinterest, for instance, takes action against harmful medical information shared on its platform. Its "Health Misinformation" policy reads:

> "Pinterest's misinformation policy prohibits things like promotion of false cures for terminal or chronic illnesses and anti-vaccination advice. Because of this, you're not allowed to save content that includes advice where there may be immediate and detrimental effects on a Pinner's health or on public safety."[34]

- The content of a campaign itself (not its diffusion mechanism) can be manipulated to deceive users and therefore belong to the realm of "disinformation" (e.g., use of manipulated media on the range from "deepfakes" to "cheap fakes"[35]).

- "Harmful content" can be promoted by deceptive actors or by campaigns leveraging distortive behaviors (e.g., "troll farms amplifying harassment campaigns").

It should indeed be noted that viral deception campaigns whose primary vector is a deceptive actor or distortive behavior can participate in amplifying other types of harmful content categories, such as hate speech, harassment, and violent extremism.

Conclusion and recommendations

Viral deception campaigns spread across platforms and through three core vectors: manipulative actors (A), deceptive behavior (B) and harmful content (C). As such, they represent a complex and multifaceted problem for policy makers and regulators to address. This "ABC" framework therefore offers a few modest recommendations for policy makers and regulators navigating this maze:

- Each dimension matters. Regulatory efforts focused on viral deception tend to exaggerate the role of harmful content: balanced approaches will consider how manipulative actors (both foreign and domestic) and deceptive behaviors contribute to the problem.
- Each dimension comes with its own set of challenges, tradeoffs, and policy implications. Specific disciplines may be necessary and/or best suited to address each of them. For instance, cybersecurity (and threat intelligence in particular) is a core component of how manipulative actors get detected; how the resulting signals get shared across the industry and with the relevant parties (researchers, public institutions) is a key policy question. Consumer protection frameworks (and stakeholders) may be ideally situated to help regulate deceptive behavior issues. Policies and regulatory frameworks that center around one type of remedy only (such as content takedowns*)* are insufficient.
- On a final (and related) note, Manipulative Actors (A) and Deceptive Behaviors (B) are dimensions on which the information asymmetry between the technology platforms on which this activity unfolds and the rest of the stakeholders in the debate is immense. How to ensure that the public, media, and policy stakeholders are able to meaningfully analyze both the issues and potential impacts of remedies in place is a fundamental question in this space.

Appendix: Examples of Disinformation Campaigns Spanning the Three Vectors

- A Disinformation Campaign in the Philippines (Facebook)

On March 28, 2019, Facebook removed 200 pages, groups and accounts engaged in "coordinated inauthentic behavior" on Facebook and Instagram in the Philippines. Facebook's press release[36] highlights the manipulative actor along with the deceptive behavior elements of the campaign:

> We're taking down these Pages and accounts based on their behavior, not the content they posted. In this case, the people behind this activity coordinated with one another and used fake accounts to misrepresent themselves, and that was the basis for our action.

Follow-up analysis highlights that the content taken down by Facebook in this campaign did contain "harmful content," notably in the form of hate speech and manipulated media (Photoshopped images of politicians in wheelchairs enticing viewers to question the health of candidates).[37]

- The Russian Internet Research Agency's "Columbia Chemical" Campaign (Twitter)

On September 11, 2014, a set of seemingly uncoordinated Twitter accounts engaged in disseminating news of a chemical incident and toxic fumes in the city of St. Mary Parish in Louisiana. Along with the social media campaign, videos of the "incident" were uploaded and officials and media were contacted by available channels with an alarming messaging – "Take shelter!" – and links to a dedicated website (www[.]columbiachemical[.]com).[38]

It wasn't long until officials realized that the campaign, with its false images of the incident and alarming messages, constituted harmful content – "a hoax," as it was initially described. It was later made clear that the accounts used to spread the content were coordinated to give the impression of a mounting local panic, using distortive behavior to create the illusion of a spontaneous wave of local panic.

It took a few more years for the major technology platforms and the U.S. Government to provide a final attribution on those accounts, confirming that the Internet Research Agency troll farm in Saint Petersburg was indeed the actor operating the accounts.[39]

Notes

[1] Camille François works on cyber conflict and digital rights online. She is Chief Innovation Officer of Graphika, where she leads the company's work to detect and mitigate disinformation, media manipulation and harassment in partnership with major technology platforms, human rights groups and universities around the world. She also is an affiliate of Harvard University's Berkman Klein Center for Internet & Society. An earlier version of this paper was presented as the second meeting of the Transatlantic Working Group on Content Moderation Online and Freedom of Expression, in Santa Monica, Calif., on May 9-12, 2019. The author thanks her colleagues in the working group for their engagement with this work during the sessions. Their feedback and encouragement greatly benefited this final paper.

[2] See for instance: Tworek, Heidi JS. News from Germany: The Competition to Control World Communications, 1900–1945. Harvard University Press, 2019.

[3] For a thoughtful typology of the different aspects of the phenomenon, see for instance Claire Wardle and Hossein Derakhshan's "Information Disorder" framework: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

[4] Viral deception here is used as an umbrella term for the multiple facets of contemporary disinformation online, see Jamieson, Kathleen Hall. *Cyberwar: How Russian Hackers and Trolls Helped Elect a President What We Don't, Can't, and Do Know*. Oxford University Press, 2018.

[5] I hope readers will forgive this 2019 edit to the famous cartoon published by Peter Steiner in the New Yorker on July 1993.

[6] See the Mueller Report: https://www.justice.gov/storage/report.pdf

[7] Jen Weedon, William Nuland and Alex Stamos, "Information Operations and Facebook", v.1: https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf

[8] For an example of a takedown solely motivated by the actor behind the content, see Facebook's "The IRA Has No Place On Facebook" post on April 3, 2018: "We removed this latest set of Pages and accounts solely because they were controlled by the IRA — not based on the content." https://newsroom.fb.com/news/2018/04/authenticity-matters/

[9] See for instance Microsoft's "*Approach to Terrorist Content*" statement (published May 20, 2016), which notes that "there is no universally accepted definition of terrorist content" and that Microsoft relies on organizations listed in the Consolidated United Nations Security Council Sanctions List to define and take action against terrorist content posted on its platforms: https://blogs.microsoft.com/on-the-issues/2016/05/20/microsofts-approach-terrorist-content-online/

[10] There are, however, multiple policies that indirectly cover aspects of these campaigns. An example with Google Ads' "Misrepresentation Policy": https://support.google.com/adspolicy/answer/6020955?hl=en

[11] In 2017, Twitter had similarly banned Russian State-controlled media Russia Today and Sputnik from using their advertising products (https://blog.twitter.com/en_us/topics/company/2017/Announcement-RT-and-Sputnik-Advertising.html). The August 2019 policy extends this ad-hoc remediation done in the wake of the investigation

regarding the Kremlin's election interference efforts on social media to all of "state-controlled" media: https://blog.twitter.com/en_us/topics/company/2019/advertising_policies_on_state_media.html

[12] See the Aug. 29, 2019 BuzzFeed investigation, "This Is How Russian Propaganda Actually Works in the 21st Century": https://www.buzzfeednews.com/article/holgerroonemaa/russia-propaganda-baltics-baltnews

[13] See for instance DFRLab's "In Depth: Iranian Propaganda Network Goes Down," March 26, 2019, https://medium.com/dfrlab/takedown-details-of-the-iranian-propaganda-network-d1fad32fdf30

[14] For an examination of how manipulative actors use "pseudoanonymity" to "impersonate marginalized, underrepresented, and vulnerable groups to either malign, disrupt or exaggerate their cause," see Friedberg and Donovan's piece in the MIT JODS: https://jods.mitpress.mit.edu/pub/2gnso48a

[15] Clint Watts, "Advanced Persistent Manipulators", Feb. 12, 2019: https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/

[16] For a global inventory of actors organized for social media manipulation, see: Bradshaw, Samantha, and Philip Howard. "Troops, trolls and troublemakers: A global inventory of organized social media manipulation." (2017).

[17] See also "False Leaks: A Look at Recent Information Operations Designed To Disseminate Hacked Material," Camille Francois, CYBERWARCON 2018. Video: https://www.youtube.com/watch?v=P8iXN8j4gMk

[18] APT here refers to Advanced Persistent Threat, a term commonly used in the threat intelligence industry to describe State-sponsored and state-affiliated groups engaged in hacking operations. See: https://en.wikipedia.org/wiki/Advanced_persistent_threat

[19] Clint Watts, "Advanced Persistent Manipulators," Feb. 12, 2019: https://securingdemocracy.gmfus.org/advanced-persistent-manipulators-part-one-the-threat-to-the-social-media-industry/

[20] Guccifer is a social media persona who claimed to be the hacker who hacked the Democratic National Committee in 2016, and who used this deceptive identity to engage WikiLeaks and the media. The account was in reality operated by Russian military intelligence: https://en.wikipedia.org/wiki/Guccifer_2.0

[21] See for instance Google's Kent Walker update on action taken against IRIB and broader State-Sponsored activity on Google's products: https://blog.google/technology/safety-security/update-state-sponsored-activity/

[22] https://newsroom.fb.com/news/2018/12/take-down-in-bangladesh/

[23] See for instance reporting by the Associated Press, "Facebook blocks 115 accounts ahead of US midterm elections", Nov. 6, 2018, https://www.apnews.com/19aabf8ba7b6466b859f4d0afd9e59be. The AP reports: "Facebook acted after being tipped off Sunday by U.S. law enforcement officials. Authorities notified the company about recently discovered online activity "they believe may be linked to foreign entities."

[24] See Ellen Nakashima's reporting in the Washington Post, "U.S. Cyber Command operation disrupted Internet access of Russian troll factory on day of 2018 midterms", Feb, 26, 2019, https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html

[25] I am borrowing here from a definition my colleagues and I have used to frame detection techniques. See Francois, Barash, Kelly: https://osf.io/aj9yz/

[26] "How Google Fights Disinformation," Feb. 2019, available at: https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf

[27] See "Coordinated Inauthentic Behavior Explained," https://newsroom.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/

[28] A blog post entitled "Removing Bad Actors On Facebook", from July 2018, seems to be the first public reference to "coordinated and inauthentic behavior": https://newsroom.fb.com/news/2018/07/removing-bad-actors-on-facebook/

[29] An example of a product change directly motivated by a platform's need to tackle distortive behaviors on its products can be found in the January 2019 YouTube announcement: ""To that end, we'll begin reducing recommendations of borderline content and content that could misinform users in harmful ways": https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html

[30] https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1001

[31] See for instance former YouTube engineer Guillaume Chaslot's project regarding algorithmic reinforcement of fringe and harmful views on YouTube: https://algotransparency.org/methodology.html

[32] For an in-depth discussion of the various issues plaguing the content moderation industry, see Roberts, Sarah T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, 2019. or Gillespie, Tarleton. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[33] This is a good place for a quick reminder of the differences between misinformation and disinformation. Dictionary.com, which made "misinformation" the word of the year in 2018, defines it as "false information that is spread, regardless of whether there is intent to mislead." It describes disinformation as "deliberately misleading or biased information; manipulated narrative or facts; propaganda".

[34] https://help.pinterest.com/en/article/health-misinformation

[35] See Britt Paris and Joan Donovan, "Deep Fakes and Cheap Fakes", published Sept. 18th 2019 by the Data & Society Research Institute, https://datasociety.net/output/deepfakes-and-cheap-fakes/
[36] https://newsroom.fb.com/news/2019/03/cib-from-the-philippines/
[37] https://medium.com/graphika-team/archives-facebook-finds-coordinated-and-inauthentic-behavior-in-the-philippines-suspends-a-set-d02f41f527df
[38] Adrian Chen's 2015 account in The New York Times Magazine is the first public account of this campaign: https://www.nytimes.com/2015/06/07/magazine/the-agency.html
[39] See for instance the reports commissioned by the Senate Select Intelligence Committee regarding the IRA's online activity targeting the USA: https://comprop.oii.ox.ac.uk/research/ira-political-polarization/