G | M | F The German Marshall Fund of the United States
STRENGTHENING TRANSATLANTIC COOPERATION

Policy Paper

August 2019 | No.22

AI STARTUPS AND THE FIGHT AGAINST ONLINE DISINFORMATION

ANYA SCHIFFRIN, WITH INTRODUCTION BY ELLEN P. GOODMAN

Please direct inquiries to:

The German Marshall Fund of the United States
1744 R Street, NW
Washington, DC 20009
T 1 202 683 2650
F 1 202 265 1662
E info@gmfus.org

This publication can be downloaded for free at http://www.gmfus.org/listings/research/type/publication.
The views expressed in GMF publications and commentary are the views of the authors alone.

Cover photo credit: vs148 (Shutterstock)

About GMF

The German Marshall Fund of the United States (GMF) strengthens transatlantic cooperation on regional, national, and global challenges and opportunities in the spirit of the Marshall Plan. GMF contributes research and analysis and convenes leaders on transatlantic issues relevant to policymakers. GMF offers rising leaders opportunities to develop their skills and networks through transatlantic exchange, and supports civil society in the Balkans and Black Sea regions by fostering democratic initiatives, rule of law, and regional cooperation. Founded in 1972 as a non-partisan, non-profit organization through a gift from Germany as a permanent memorial to Marshall Plan assistance, GMF maintains a strong presence on both sides of the Atlantic. In addition to its headquarters in Washington, DC, GMF has offices in Berlin, Paris, Brussels, Belgrade, Ankara, Bucharest, and Warsaw. GMF also has smaller representations in Bratislava, Turin, and Stockholm.

About the Author(s)

Anya Schiffrin is the director of the technology, media, and communications specialization at Columbia University's School of International and Public Affairs, and a PhD candidate at the University of Navarra, Spain. Her most recent books are *African Muckraking: 75 Years of African Investigative Journalism* (Jacana 2017) and *Global Muckraking: 100 Years of Investigative Reporting from Around the World* (New Press, 2014).

Ellen P. Goodman is a non-resident senior fellow for the Digital Innovation and Democracy Initiative at the German Marshall Fund. She is also a Rutgers Law School professor. She co-directs and co-founded the Rutgers Institute for Information Policy and Law.

Acknowledgments

This paper was made possible with support from the Nasdaq Educational Foundation. Thanks to John Batelle, Justin Hendrix, and Raju Narisetti for introductions to the firms profiled. I am also grateful for the research provided by Chloe Oldham and edits suggested by Hollie Russon Gilman, Gordon Crovitz, Alejandro Romero, and others who commented on drafts. Above all, thanks to our interviewees for their time and to Dean Merit Janow and the committee that allocated the research funding.

G | M | F The German Marshall Fund of the United States
STRENGTHENING TRANSATLANTIC COOPERATION

COLUMBIA | SIPA
School of International and Public Affairs

Executive Summary

On both sides of the Atlantic, governments, foundations, and companies are looking at how to solve the problem of online dis/misinformation. Some emphasize the demand side of the problem, believing it important to focus on consumer behavior and the use of media literacy and fact-checking. Some focus on legal remedies such as platform-liability and hate-speech laws as well as privacy protections. Meanwhile, others try to raise the quality of journalism and support local news in the hope that creating more reliable content will be a counterweight to the dis/misinformation found online.

In short, there are myriad solutions aimed at addressing the problem of online dis/misinformation. This study looks at one kind of fix: the small companies in the information ecosystem that use natural language processing as well as human intelligence to identify and, in some cases, block false or inflammatory content online. There are impediments to the success of this entrepreneurial approach, including the fact that disinformation detection by algorithms is complicated, it is hard to scale, and that it is unclear whether the platforms have an incentive to adopt such technology. It is very likely that platforms such as Facebook or Twitter—which already screen, block, and remove fake accounts and content—will copy the technology or will buy out the small firms for the skills of their staff and for their products in order to gain access to the AI needed for further screening.

This paper looks at thirteen such companies, most of which are building solutions to identify false information online through a combination of people and natural language processing. Nascent and not yet widespread, these businesses are seeking to find new commercial applications for their products and, in some cases, hoping to entice the social media platforms to buy them out.

Ultimately, there is disagreement among the companies surveyed as to whether natural language processing alone will be able to identify all forms of dis/misinformation online, the regulatory and policy approaches that could complement their activities, and if the industry will be able to survive at all within an ever-changing information ecosystem.

AI Startups and the Fight against Online Disinformation

INTRODUCTION

ELLEN P. GOODMAN

Growing citizen and government dissatisfaction with digital platforms is creating new pressure on content moderation. In response, social media platforms have begun to take action. Facebook is developing a "supreme court" to adjudicate content removal decisions; Twitter is initiating new policies to flag content that runs afoul of its community standards; and Google is calling for "smart regulation" of digital information platforms to prevent harm. At the same time, governments around the world are introducing or proposing new legislation to impose more liability on platforms for promoting harmful content and disinformation, to mandate more aggressive content moderation, or merely to require more transparency.

All of this activity around moderation assumes or insists on the ability of platforms to detect and deal with harmful content. Content analysis and fact-checking remains an area of innovation and experimentation. In this paper, Anya Schiffrin shows how start-ups and small companies are tackling this problem. Media policy in Western democracies favors competitive environments and a diverse ecosystem for the production of news.

In diversity and experimentation, there are more opportunities for a robust exchange of views and the production of accountability journalism. The same might be true for disinformation detection: the more vibrant and diverse the ecosystem for flagging content, the more accurate and reliable the information will be.

The interviews with practitioners on which this paper's findings are based raise doubts about the viability of small-scale entrepreneurship in this space. The research suggests that successful entities in the disinformation solutions space will be acquired by the major platforms, or that those platforms will develop their own models of content-flagging. As a result, it is unrealistic to expect there to be a competitive and diverse ecosystem of content moderation in a world of dominant platforms.

This research is significant for the policy discussion in that proposals calling directly or indirectly for more platform content moderation must consider where the technology and methods for doing so come from and who will likely own them.

AI Startups and the Fight against Online Disinformation

ANYA SCHIFFRIN

Many solutions are being tested to combat the alarming spread of dis/misinformation online. Governments are imposing regulations as to what can and cannot be said online; foundations are funding fact-checking initiatives; and efforts are being made to build trust in the media. Journalists are building relationships with their audiences and trying to strengthen local news reporting in order to provide societies with credible and useful information. Journalists are also tracking the people and governments creating and spreading false news and propaganda. Facebook, Google, and Twitter claim to be tackling the problem of removing dangerous and illegal content. However, since these platforms profit from the spread of dis/misinformation and have been slow to respond to warnings about the dangerous effects on society, many lack trust in the platforms ability to successfully combat the spread of harmful content online. For example, Facebook has blocked tools promoting transparency of advertising on its site,[1] and it has also consistently refused to release data to researchers seeking to measure the impact of exposure to disinformation.

In this fragmented universe of solutions, which is characterized by a lack of comprehensive platform regulation, several small private-sector companies have proposed ways of solving the problem. This paper looks at what solutions they are developing and outlines some characteristics of the growing sector. This paper also tackles the question of whether the solutions proposed by these small companies will be effective in the absence of widespread adoption by the giant platforms.

The research in this paper is based on more than twenty in-depth interviews, carried out between December 2018 and February 2019, and also offers short profiles on thirteen of the firms. These companies were identified by asking experts within the field of content moderation for suggestions as well as by reading press reports. Most of the interviews were administered over the phone, but a few were conducted in person. The companies studied were asked about their technology and how it works, their business models, annual revenues, and their plans to scale, as well as their thoughts about government regulation.

Trying to Fill a Regulatory Gap

For the most part, the firms surveyed did not set out to crack the problem of dis/misinformation online. Many were involved in other activities when they came across dis/misinformation and decided to do something about it, noted Alejandro Romero of Alto Data Analytics, a data analysis firm based in Madrid with a presence in Brazil, the United Kingdom, and the United States. "All the companies working in this space were working on something else and the disinformation they saw had an impact on what they were doing. They saw a business opportunity and thought this could be an interesting learning. I've not seen a company that started up just to fight disinformation," he said.

1. Jeremy Merrill and Ariana Tobin, "Facebook Moves to Block Ad Transparency Tools — Including Ours," ProPublica, January 28, 2019.

Disinformation

Disinformation is false information spread deliberately to deceive. The word is a loan translation of the Russian *dezinformatsiya*, derived from the title of a KGB "black" propaganda department.* Different scholars have provided taxonomies of the different kinds of misinformation and disinformation (including, for example, satire and false context as well as imposter content and fabricated content) as well as of who the different actors and targets are.**

* Garth Jowett and Victoria O'Donnell. "What Is Propaganda, and How Does It Differ From Persuasion?", Propaganda and Persuasion, Sage Publications, 2005.

** Claire Wardle and Hossein Derakshan, "Information Disorder: toward an interdisciplinary framework for research and policymaking", Council of Europe, September 2017

The start-up executives interviewed assumed that widespread regulation of online hate speech and of platforms was not imminent, and that consumer and corporate demand for their products would continue. They noted that Facebook, YouTube, and Twitter had no incentive to change a business model that was profit from generating outrage and engagement. In this scenario, without regulation, the platforms would continue to allow, or even encourage, the circulation of falsehoods online.

The firms surveyed exist because of a lack of action by social media companies that have no financial incentive to fix the problem and because government regulation has not yet been passed. Further, the regulation that does exist—in the case of the United States, Section 230 of the Communications Decency Act—protects the companies from being liable for what is put on their platforms. Many interviewees argued that Facebook, YouTube, and Twitter have been extremely lax and irresponsible in allowing hate speech and disinformation to contaminate their networks. Some note that, technically, it is not a hard problem to fix but that the incentive for Facebook, Twitter, or YouTube to do so is nonexistent. "If the platforms try to tackle the problem internally it will be a huge revenue loss for them so they do not have an incentive to do it," said Sagar Kaul, one of the founders of MetaFact, a fact-checking platform based in India.

According to a leading technology journalist, Julia Angwin, "online disinformation is a lot like the spam problem. And it could probably be solved the way we solved the spam problem. We solved spam with a combination of legal and technical measures. First, Congress passed a law (the CAN-SPAM Act)[2] that imposed fines on the senders of spam. Faced with liability, the email industry then set up a 'blacklist' of known spammers that they all agreed to block. Similarly, if there was some legal or financial cost to the platforms, they would likely set up a 'blacklist' of disinformation outlets to block. But they currently have no incentive to do so." Facebook's Mark Zuckerberg has called for global regulation,[3] but observers note that his comments are belied by the amount of time and money the company and other tech giants spend lobbying to avoid regulation.

According to the publisher John Batelle, who has extensive experience launching and investing in media and technology companies, "The platforms are extremely good at making soothing noises. They are silver-tongued and very good at what they do. Their response is: 'We are hiring 20,000 people. We are all over this problem, we have a community that will flag false news.' Facebook's point of view is: 'We got this. We acknowledge the problem. We are in the twelve-step program.' But they have not admitted the first step: 'There is a power beyond ourselves.' Facebook has never acknowledged this problem is bigger than them."

Some firms try to work with advertisers to pressure the platforms to remove dis/misinformation. In the absence of regulation or financial incentives, there is no reason for the platforms to do so. Therefore,

2 Controlling the Assault of Non-Solicited Pornography and Marketing Act, 2003.

3 Mark Zuckerberg, "The Internet needs new rules. Let's start in these four areas" Washington Post, March 29, 2019.

John Batelle on How the Funding Landscape Works

Publisher and serial entrepreneur John Battelle explained that some "angels" are willing to invest amounts up to $1–2 million into start-ups of the kind surveyed, but that it is unlikely that large investors will put in lots of money because they would want to see major returns. He said: "It is not hard to get some money at this stage but raising a seed round is not proof of much. A sophisticated, later-stage investor would say 'I am never going to put a lot of money in this field because it is controlled by the big companies.' If Twitter or Facebook were looking for new ideas—for example, how to identify fake news using images—then they could acquire [an early-stage company by] paying half a million or so per engineer. Or, they will just copy the technology. There is no reason for a VC [venture capitalist] to put $10 million into a company with that profile. Institutional VCs know that, should the large platforms identify one of these seed-stage companies as doing anything useful, they will either copy it or acquire it for not much money. They certainly will not depend on a third-party technology. VCs who want to put a lot of money to work will not win in this scenario."

According to Batelle, "These investors do not want to put in money at a $2 million valuation, only to get out at a $3 million valuation. This means the innovation space has become a desert, blocked by the large platforms which have a monopsony over demand for acquisitions….a functioning market should have a flourishing ecosystem of innovation. What we have here is a market failure because of an oligarchy."

Without acquisition by the big platforms, Battelle added, the start-ups dealing with disinformation will not scale. "Entrepreneurs often say 'This time is different,' but then again, that is what founders of new companies always say. They have a new technology, and it may or may not appeal to the platforms. And it may appeal to the other parts of the market that matter, such as consumers who could put an extension into their browsers. These are people who are motivated not to put fake news in their lives. But that is a long shot. You cannot build a company based on a web browser extension."

many of the start-ups hope that reputational risk and naming-and-shaming will prove effective. They are working with advertisers to see if the latter can push the platforms to take more action.

Assuming that respectable brands will not want to run their advertisements next to unsavory content, some of the interviewees expect that reputable corporate advertisers will push for change and force platforms to crack down on hate speech online as well as on dis/misinformation. These start-ups and some other groups, like the United For News coalition, hope to persuade advertisers to push the platforms to do a better job of screening advertisements.

Many of those interviewed describe how advertisers could be compelled to pressure the platforms. Additionally, some journalists advocate for advertisers to work directly with media outlets rather than with programmatic buying through algorithms and third parties. Not only does this ensure more transparency through direct relationships, but also supports quality media outlets.

Regarding United For News, John Battelle said: "I am a huge fan of the philosophy. Direct media buying supports what I think is the most important part of publishing—the direct relationship between publisher, audience and marketer. Programmatic advertising has stripped away the context of an experience with the audience. Advertisements go into unknown places, out of context. Direct advertising will re-establish the connection with the audience. The problem is that programmatic advertising is cheap for advertisers and cheap for publishers to implement, so it is irresistible—especially for the advertisers, who control the ecosystem."

Advertisers would themselves be damaged by government bans on microtargeting or the spread of regulation aimed at protecting privacy. Therefore, pushing the platforms to take action would protect both sectors (platforms and brand advertisers) from regulations they do not want. Groups like United for News also hope to protect the quality of journalism. By persuading advertisers to work with outlets directly they hope that advertising revenue will go to high quality journalism sites rather than automatically appearing on random web pages or next to dis/misinformation. For their part, publishers hope that, even though it is more time consuming, and therefore expensive, reputable brands can be persuaded to advertise on high-quality news sites rather than risk an algorithm dropping their advertisements alongside false or shady content.

While some of the firms in this field have been around for years, others are nascent and small. Many of their founders got help from friends and family as well as small start-up grants. The companies are now busy trying to develop and launch their tools or scale in size. For example, Metafact was started with funding from Hanyang University in partnership with the Seoul metropolitan government, while the Global Disinformation Index relied on foundations. Others have core business activities and/or ties to intelligence and government agencies, cyber security, or fraud detection. A few, like Brandwatch, have a core activity of monitoring the web for corporate clients.

Some, like Vett News and NewsGuard, consider fighting disinformation to be a part of their core activity. NewsGuard has raised $6 million for their news rating system and launched in the United States and in key markets in Europe. Microsoft has licensed NewsGuard globally, including for its Edge mobile browser.

Some firms, such as Truepic, hope for commercial applications for their technology. Truepic verifies photographs taken with its technology and assigns them a unique number stored in the cloud. This will not only help governments, human rights groups, and media organizations that need to verify, for example, pictures of atrocities or human rights violations. This could also be extremely useful for insurance companies that are vetting claims and rooting out fraud. Some firms, such as AdVerif.ai, give pro bono help to media outlets.

There may be a way for some of these firms to generate revenue from the technology they have developed, but only a few are likely to scale. According to NYC Media Lab's Justin Hendrix, "There are probably a handful of scalable ones that use machine learning and natural language processing. Extracting information from video and images is difficult and something you can profit from. Companies that figure out how to do that will make some money and the rest of the firms will have a hard time scaling."

Problem Areas

There is some disagreement among the founders of the start-ups surveyed as to how much of the screening of dis/misinformation can be automated and how much cannot be done without people. Can artificial intelligence (AI) and natural language processing identify all or most of the bad stuff? Interviewees were split on this question. Some argued that people are an essential part of the process because so many of the sites were designed to mislead and look like real news sites. As a result, it is almost impossible for a computer to recognize all the different characteristics of these kinds of sites. Danny Rogers from the Global Disinformation Index explained: "It is more pernicious because 60 percent of RT is high-quality journalism and it looks different from the low-quality sites put up quickly to get eyeballs."

Others interviewed are more optimistic about the possibilities that detecting and even suppressing false/disinformation can be automated nearly completely using AI and natural language processing. All have admitted that there was some uncertainty as to how this could be done as well as limits on how effective it would be. Those interviewed also noted

Danny Rogers on Whether Technology Can Solve the Problem

New York University's Danny Rogers has a background in intelligence and the dark web, which he uses in his work on combatting fraud and identity theft. Developing the Global Disinformation Index is a labor of love and he hopes it will be used by advertisers who want to avoid placing their ads next to disinformation, as well as by platforms and other tech companies to help de-platform disinformation efforts. "I do not think automated natural language processing is scalable. Computers are not going to be able to distinguish content that is designed to fool humans," he said.

Rogers distinguished between two kinds of disinformation and said they need to be analyzed differently. At the top of the food chain are "highly organized threat actors like state-run operations or commercial ones like Cambridge Analytica." At the bottom are decentralized purveyors such as trolls, 4chan, and clickbait purveyors.

One kind of disinformation is high-quality and comes from fully fledged media operations such as Breitbart or RT. "High quality sites will be largely impossible to differentiate. A computer will have a really hard time computing the difference between Breitbart and CNN. Breitbart is very nuanced and you have to look at it from a journalistic perspective, not a computational one. The 'junky' misinformation has lots of hallmarks such as spelling errors, recycled material; it is often presented on a WordPress template. All of this has signatures that you can identify."

One unresolved question is how to define the problem so as to get bipartisan consensus. "Using the word 'disinformation' makes conservatives wary of a liberal plot to silence the media. But it is not two equal sides: it is irrationality versus enlightenment thinking," said Rogers.

According to him, "Facebook's job is to get people to click on links. They do not want to combat this. They run affiliate marketing conferences teaching people how to do this. They got away with it for years until the 2016 elections. Facebook is siphoning off all the ad revenues of these clicks. Twitter has no incentive from a business model perspective to kick off the bots because their stock price is directly tied to their user count. The bots make them look bigger and more popular than they are."

that tech solutions would take seven to ten years to implement and so would not be any faster than waiting for comprehensive government regulations. "Private-sector solutions are all seven to ten years down the road. I do not see one coming up in one to two years," said Joe Ellis from Vidrovr.

Another problem is that much of the dis/misinformation is put on small fake sites that are new and change constantly, so it may be impossible to control. Organizations like Mediabiasfactcheck.com cannot keep up with the new sites constantly appearing, said Alejandro Romero. Further, these sites inject their false information into Facebook and Twitter where it gets circulated and becomes impossible to track its impact unless these networks provide more transparency on advertising campaigns.

The human factor also comes into play as it is hard to see how different tech solutions can be applied all over the world in places with different values. "How do you handle a global platform operating in different places with different values? Even if the technology existed, the application in different contexts would be a nightmare," said Romero.

According to Marie Frenay, a member of the office of the European Commission's Vice-President and Commissioner for Digital Single Market Andrus

Ansip, "There are very promising research projects and start-ups which explore the potential of AI to detect disinformation, identify patterns. We need to continue investing in this area. At the same time, it is also clear that human expertise is needed. I see technologies as tools that can assist disinformation experts in their work. It is about complementarity. As disinformation is conducted more and more subtly and covertly, making it harder to detect and attribute, we need the best of human brains and machines to address it."

The Paradox of Ratings Systems

Others worry about the paradox of rating systems—whether the rating systems can be gamed, politicized, or corrupted in the same way the credit rating agency Moody's was before the 1997 financial crisis. At that time, the integrity of Moody's business model came under fire because, in addition to providing ratings, it also sold services to countries wanting to improve their ratings. It was also notorious for giving high ratings to countries that collapsed shortly after. The problems with the credit rating agency model also apply to the rating agencies in the disinformation/news sector, in that ratings can become politicized and highly contested. The judgment and legitimacy of the 'rating actor' can come into question based on the rating methodology and perceptions about how the rating agency is funded and managed.

Julia Angwin said: "A better model would be peer accreditation, where journalists band together to enforce a set of standards on their industry and only include outlets that meet the criteria."

The "industry standard" model is also being attempted by the Paris-based Reporters Without Borders, which is working with the European Standards Authority and an international coalition of journalists in order to come up with a list of credible media outlets that follow internationally agreed-upon standards. Making such lists transparent opens up the rating entities to criticism. The Reporters Without Borders list has already been criticized for including Russia-funded RT France.

The technology journalist Will Oremus has also criticized NewsGuard for giving a positive "trustworthy" rating to FoxNews.com and an extremely negative rating for Al Jazeera. He speculates that in the future "the ratings authorities [could] become *too* powerful."[4]

"Making decisions about what misinformation to suppress or promote almost has to be done anonymously because if people know who is behind the effort they may not trust it. It is strange but people almost seem to trust Facebook more than they would trust another group. The second you know who is behind the effort people will start arguing about whether the group is qualified to pass judgment," Reg Chua, chief editorial operating officer of Thomson Reuters said. "It is not that I am in favor of secret cabals deciding what we read; but more that all having a group vet information does is move the debate about who to trust upstream—from the news source to the vetters."

Newsguard co-founder Gordon Crovitz points out, however, that research on this question shows the opposite: consumers are willing to trust journalists to rate other journalists so long as they operate in a transparent manner rather than Silicon Valley's non-disclosed algorithms.[5]

One criticism of the companies trying to use AI to look at dis/misinformation is that they ignore its underlying causes and sources. For Alejandro Romero, "the social networks are the last building block" in a process that begins with entities that find vulnerabilities in society and then target them, stoking fears about subjects such as immigration in order to affect the integrity of elections and political decision-making.

Views about Regulation

Interviewees expressed mixed views as to whether government regulation would be a good idea. Some

4 Will Oremus, "Just Trust Us," Slate, January 25, 2019.

5 For more information, see Newsguard's website. Accessed on June 17

entrepreneurs such as Joe Ellis and Gordon Crovitz from Newsguard said they believed in free-market approaches. Mark Little from Kinzen said: "I am afraid of regulation that does not solve the problem but will make the perception of elite control of the media worse." Ellis also seemed wary of regulation: "The disinformation question is really hard. I do not know how to solve it. The best way to try to solve it is to give as much power to the actual consumer as possible so that search involves user intentions."

Others who are more open to the idea of regulation acknowledged that they did not have the detailed policy knowledge to understand the best approach to regulating disinformation. They noted the regulatory differences between the Untied States and European context and the relative pros and cons of different regulatory approaches.

Reducing the scope of Section 230 would fix the problem straight away, according to Eric Feinberg of GIPEC and Julia Angwin, as platforms would be held liable for illegal content. Germany's NetzDG law, implemented at the start of 2018 and opposed by many internet rights groups on freedom-of-expression grounds, made platforms liable for defamatory content and hate speech. Under this new law, if platforms such as Facebook and Twitter fail to remove such illegal content, they could be fined up to €50 million.

Other ideas included:

- Removal of programmatic ads networks.
- Legislation for a greater number of human fact checkers.
- Twitter promoting fact-checked content rather than paid ads.
- Strong privacy regulations to help prevent exploitation, identity theft, and microtargeting.

Frantisek Vrabel from Semantic Visions in the Czech Republic suggested: "A ban on microtargeting would work. We strongly recommend [the] European Commission to regulate Facebook, regulate algorithms so that they do not microtarget based on creating small information bubbles that fragment, atomize and divide our societies." As all interviewees noted, the threat of regulation would be an intrinsic part of getting the platforms to police themselves more. "There is a carrot-and-stick approach. Big tech's incentives are that advertisers are pushing that; but the stick is even more powerful and governments are pushing for that," said Or Levi from AdVerif.ai

Conclusion

In a world of fragmented solutions to the problem of online dis/misinformation, the small start-ups using AI and natural language processing are a niche to watch. It is clear that the entrepreneurs interviewed see a business opportunity in using both people and natural language processing to identify and possibly remove dis/misinformation online. However, these entrepreneurs are also aware of the limitations of this approach. First, they note that AI is not yet able to identify all of the myriad forms of dis/misinformation contaminating the information ecosystem. Second, even if it were possible to use the technology at scale, there is little evidence that Facebook, Google, and Twitter would use it—one of many reasons why regulation of the platforms is essential. Third, these tech-based solutions do not address the larger economic, social, and political reasons that dis/misinformation spreads. An entire ecosystem, including programmatic advertising and the anonymity of domain registrants, contributes to creating the ideal conditions for dis/misinformation.

In the words of Alejandro Romero, "Platforms will not take on active defense of truth-telling institutions [and] even if the platforms wanted to fix the problem, they can only have an important but limited impact in the disinformation landscape. They are a contributor to a massively organized disinformation system. But the digital ecosystem is broken and the possibilities of gaming the system are endless."

Annex. Company Profiles

AdVerif.ai

Or Levi: "Terrorism and violence are traditionally what companies are trying to eliminate]. Our focus is more fake news [that is] more challenging for technology to detect." The business model is for advertising and a blacklist for publishers and advertisers to protect their brand from being associated with fake content and screen fake contact.

What is it? Developing tools that use natural language processing to see if something is fake or suspect. Creating a blacklist for publishers and advertisers who want to protect their brand from being associated with fake content. And an API to help screen fake contact.

Funding: Bootstrap, raising funding.

Staff size: Three.

Launch date: Company began in 2017, beta tool launched in 2019.

Alto Data Analytics

Alejandro Romero: "Alto was not created to research disinformation. We just found it."

What is it? Alto provides actionable insights based on public data: "We help our clients understand the world faster, better". The company provides AI- and machine learning-based software that harvests, indexes, analyzes, and visualizes public data that allow businesses to understand insights. Alto also offers services based on the help from a team of experts.

Funding: Privately, no external funding/VC funding.

Staff size: Over 100.

Launch date: 2012.

Brandwatch

Paul Siegel: "A true story looks the same as a false story. Both are a collection of sentences." "There is a distinction between language and truth that is hard to make."

What is it? U.K.-based company that monitors the web and sells clients' data on the online public perception of their brand.

Funding: Privately held, VC funded as well as private clients.

Staff size: Over 550.

Launch date: 2006.

Future plans: Sees a market for selling services to government, government contractors, and enterprises that are targeted by disinformation.

Factmata

Dhruv Ghulati: "I think legislation will drive the platforms to do things and is needed urgently. It is a shame that it has come to regulation and that they have not taken it on properly."

What is it? Software company that provides AI tools that detect specific types of disinformation. For businesses, it monitors their brand online to eliminate undesirable information about them.

Funding: Has raised $1.8 million from seed funding.

Staff size: 10.

Launch date: 2017.

Future plans: It is developing browser extensions and an app for individuals.

GIPEC

Eric Feinberg: Feinberg was in advertising technology and has a patent on what he calls "anticipatory intelligence." He was angered when he began finding unsavory content online including Islamic State posts calling for attacks on U.S. troops. His software looks for words like "caliphate," beginning with hashtags and then trails it through the web. "My systems dig through all accounts using it." Now Feinberg has a faux account and so the algorithms push pro-Jihadi content to him as well as to Islamic State sympathizers.

Feinberg notes that he is "not going after the top, it is the peer-to-peer, the sympathizers….You've got Islamic State, radical jihadists getting radicalized in Indonesia, Bangladesh, Philippines….Facebook's algorithm has connected the world for radical jihad."

What is it? Feinberg's tool monitors where clients' ads are going online and generally helps its clients protect their brands online. Combines their research and reports with entities such as the Digital Citizens Alliance (DAC), a nonprofit 501(c)(6) organization. Its Internal Revenue Service form does not indicate its donors, simply stating that DCA's revenue comes from "program services."

Funding: Bootstrap, looking for capital.

Staff size: Unknown.

Launch date: 2015.

Future plans: Hopes to be licensed, funded and to work with social media companies to reduce extremist content on their platforms.

Global Disinformation Index

Danny Rogers: "Right now we have the most brand-unsafe environment in the history of advertising. It is the Wild West. Platforms have no incentive to actually secure themselves. We are trying to catalyze grassroots support and get the advertising buyers to have a say."

What is it? A U.K. non-profit trying to make an "AI-powered classifier which can identify junk domains automatically" and would then work with programmatic ad networks so that they have a "dynamic blacklist of sites thereby choking off funding for disinformation networks." Rogers said: "We want the Global Disinformation Index to be the ones to take on the risk. We have no skin in the game and can provide transparent, neutral ratings that platforms and the brand safety community can use."

"The goal is to have a couple of products. One a self-updating blacklist of junky open web sites that are worth blocking in the ad exchange. This can be used by the ad tech community to allow them to block ad buy on junky sites. No one company wants to take a stand or say 'this is good or bad.' So we want to be neutral and transparent and be the risk-absorbing entity," added Rogers.

Funding: $50,000 grant from Knight Foundation. Now has $1 million in seed money. Other funders include USAID and Luminate.

Staff size: Three co-founders.

Launch date: 2018.

Future plans: The company is part of the Reporters Without Borders "technical advisory committee" and working with the European Standards

Organization to get consensus-based standards developed for media outlets who opt in. Eventually, this could lead to certification of outlets that meet certain standards of transparency and other criteria.

MetaFact

Sagar Kaul: "MetaFact is creating disinformation defense solution for newsrooms, brands, and organizations. By leveraging next-gen technology like advanced Al, to analyze pattern and bucketing data sets, they help newsrooms to understand if a certain discourse around a particular topic is genuine, or is a targeted campaign trail orchestrated to change public opinion or inflict financial damage. Detecting bots that spread false claims so profiling them is of paramount importance. Profiling human-run bot-like accounts is tougher, yet achievable with a claims-first approach. By being able to detect a claim as soon as it's uploaded online our tool is able to track the interaction of bot accounts and influencers before any other tool is able to detect it as a threat. By using our claim first approach we can proactively detect, monitor, and defend brands from disinformation attacks before they gain momentum and inflict financial losses."

What is it: A company that builds AI-based disinformation detection and defense solutions.

Funding: Bootstrap, friends, and family. Obtained a grant for $20,000 from Hanyang University in South Korea. Also just completed an accelerator program in Ireland through the National Digital Research Centre, sponsored by the Irish government and Enterprise Ireland. Metafact was the first startup selected from outside of Ireland and received €30,000. Metafact is working on the IBM Watson platform having been selected for the IBM Global Entrepreneur program that provides IBM Cloud credits.

Staff size: 5.

Launch date: Prototype was launched in April 2018. A "minimum viable product" website for the tool will be launched by the end of July 2019.

Future plans: Launching the tool for media houses, businesses and corporations. Making the tool available in other languages. Working on Deep Fakes. Developing an AI-enabled media literacy app for kids. According to Sagar Kaul, "Media literacy has to play an important part. We're developing a mobile app that limits what kids can see on their phones and at the same time help them understand echo chambering and how disinformation is spread. AI will be an integral part of the app. It will understand the needs of the kids and based on that the app can give recommendations for reading material."

NewsGuard

What is it? Ratings system that assigns red or green ratings and explanatory "nutrition label" write-ups to thousands of news sites around the world, based on nine basic criteria of journalistic practice relating to credibility and transparency. Available through Microsoft and also through its browser extension for Chrome, Safari, Edge, and Firefox. Newsguard has partnered with more than 200 public libraries in the United States to provide a news-literacy tool for library computer users. It also has a brand-safety product called BrandGuard that provides marketers with a whitelist of generally reliable news websites safe for programmatic advertising and a blacklist of sites that misinform.

Founded by journalists Steven Brill (founder of American Lawyer and CourtTV) and Gordon Crovitz (former publisher of *The Wall Street Journal* and an early investor in *Business Insider*).

Funding: Raised $6 million with the Publicis Groupe advertising agency as the lead investor. It charges the digital platforms to grant their users access to the ratings and nutrition labels instead of charging the actual publications being rated.

Staff size: A total of fifty, including forty journalists covering the United States, the United Kingdom, Germany, France, and Italy.

Launch date: 2018 in the United States in 2018; 2019 in France, Germany, Italy, and the United Kingdom.

Semantic Visions

Frantisek Vrabel: "We do not focus on analysis of online social networks, but we focus on online news. In our experience the disinformation and propaganda start on news sites and blogs."

What is it? A large, speedy database and a web mining system that are used for risk assessment and monitoring. It has roots in the defense industry and uses open-source intelligence.

Funding: Work for corporate clients (real-time risk detection solution integrated into SAP Ariba business commerce platform risk) pays for the work on disinformation. Recently won a $250,000 grant from the U.S. Department of State's Global Engagement Center to help fund the development of cutting-edge new technology to combat disinformation online.[1]

Staff size: 25.

Launch date: 2005.

Truepic

Mounir Ibrahim: "We want to make sure anyone in the world with a smart phone has the ability to capture an image and prove its contents are real."

What is it? Image verification technology. Truepic has several products. Users can download the free app and whenever they take a picture the system will log the time, date, location and pixelation, and assign it an encrypted code that will be stored in the cloud. Truepic also has developed a remote inspections platform (known as Truepic Vision) for clients in insurance, banking, and lending.

Funding: Not profitable but generating revenue, raised $8 million in 2018.

Staff size: 30.

Launch date: 2014.

Future plans: Have commercial applications.

Vett News

Paul Glader: "People would like a tool that when they are reading an article would allow them to understand if it is fake or not, good or not, opinion or not, validation or confirmation." Paul Glader got interested in verification of content after he wrote a very widely circulated article about what to look for.[2]

What is it? Like NewsGuard, Vett News currently provides a Chrome extension that rates news sites based on their reliability: green for trustworthy and red for unreliable.

Funding: Bootstrap, not yet ready to raise funding.

Staff size: About 5.

Launch date: 2017.

Future plans: Get a browser into schools and libraries and the ad tech market, and be bought by Facebook or Twitter.

1 For more information see U.K. Department for Digital, Culture, Media and Sports, "Semantic Visions wins $250,000 Tech Challenge to Combat Disinformation," March 8, 2019.

2 Paul Glader, "10 journalism brands where you find real facts rather than alternative facts" Forbes, Jan 2017

Vidrovr

Joe Ellis: "A lot of the video people watch is not used or found from a search perspective. [Our company wants] to infuse intent from the domain in the region." Instead of passively receiving information, Ellis hopes that Vidrovr can help audiences as well as video producers to become more active about what they see.

"You are at the behest of what the algorithm chooses to show you. Radical transparency will build trust so audiences can know who you are and what you have done. We think that will help combat misinformation and help companies monetize their video, allowing them to be transparent and provide intent into user search," said Ellis who took leave from his Columbia PhD program to run Vidrovr.

"When you start a company, you have a vision for building it and that can get derailed by running out of money and going out of business or you get gobbled up by a large company that thinks what you are doing is really interesting. Google is very good at building search solutions and using machine learning technology to make content available, and [it] has a lot of talent. Can a start-up do better than the big companies? Usually not."

What is it? Software that helps companies and news outlets index, annotate, and search their videos.

Funding: Got tech start start-up money aimed at student projects and then VC money. Raised $1.25 million

Clients: Noting that it cannot disclose most of its customers, Ellis said Vidrovr "work[s] with some of the largest broadcasters in the United States."

Staff size: 7.

Launch date: 2016.

Future plans: Switch video viewing from social sites to an over-the-top (personal), more mobile-based platform.

G | M | F The German Marshall Fund of the United States
STRENGTHENING TRANSATLANTIC COOPERATION

Washington • Ankara • Belgrade • Berlin
Brussels • Bucharest • Paris • Warsaw

www.gmfus.org