COMPUTATIONAL PROPAGANDA

IF YOU MAKE IT TREND, YOU MAKE IT TRUE

RENEE DiRESTA

There are invisible rulers who control the destinies of millions.
— Edward Bernays, *Propaganda*

The pioneering public relations consultant Edward Bernays's words are nearly a century old, but today, in an era of rampant misinformation and insidious disinformation campaigns online, they seem startlingly apt. The rulers Bernays was talking about were public relations specialists, and at the time *propaganda* was not a pejorative. But when we consider this statement in the context of the current information ecosystem, replete with manipulative narratives spread by bots and human operators alike, it's somewhat jarring. Now the "invisible rulers" are the people who control the algorithms that populate the feeds of two billion users, and the strategists who are most adept at gaming them.

In Bernays's time, *propaganda* was used somewhat interchangeably with *public relations*. The concept has evolved since then; modern and postwar propaganda can perhaps be defined in the simplest terms as *information with an agenda*. The information is not objective, but it also isn't necessarily false — in fact, to be most

effective, propaganda is often anchored in a partial truth. Regardless of whether it's true or false, propaganda has a consistent aim: to influence the target to feel a certain way or form a certain opinion about a concept or entity. Propaganda is most often associated with governments, but activist groups, companies, and the media also produce it.

Propaganda has been a tool of rulers since the time of the ancient Greeks and Romans; the term itself comes from an administrative body of the Catholic Church that was dedicated to "propagating" the Catholic faith in non-Catholic countries. Kings used it, religious leaders used it, and even the American Founding Fathers used it to shape opinions and influence societies. In fact, Bernays believed that propaganda was essential to the functioning of a democracy:

> The conscious and intelligent manipulation of the organized habits and opinions of the masses is an important element in democratic society. Those who manipulate this unseen mechanism of society constitute an invisible government which is the true ruling power of our country. We are governed, our minds are molded, our tastes formed, our ideas suggested, largely by men we have never heard of. This is a logical result of the way in which our democratic society is organized. Vast numbers of human beings must cooperate in this manner if they are to live together as a smoothly functioning society.

This manipulation took the form of informing people, persuading people, or integrating people with people. "Of course," Bernays noted, "the means and methods of accomplishing these ends have changed as society has changed."

The methods changed rather dramatically decades after Bernays's death; the internet was born, and it transformed and upended the stranglehold that rulers and elites traditionally had on the flows of information. And as the internet itself changed, with the emergence of a handful of social platforms that serve two billion people, a series of unintended consequences stemming from design choices and business models democratized propaganda and helped it evolve into what Phil Howard and Sam Woolley of the University of Oxford called *computational* propaganda: "the use of algorithms, automation, and human curation to pur-

posefully distribute misleading information over social media networks."

Most people have heard of, and experienced, *misinformation*: it's false information, and on the early internet it often took the form of email chains, perhaps from your gullible uncle, telling you that if you forwarded a message to ten people, Bill Gates would send you money. This kind of misinformation is the raison d'être of the hoax-buster Snopes, the site that explains what's going on when things are wrong on the internet. But if you turn on the news today, in the United States, in Europe, or in the United Kingdom, you will probably be hearing and reading about *disinformation* — perhaps in the context of Brexit, perhaps in the context of the 2016 U.S. presidential election. The false information that made the rounds in these situations was more nefarious than the original fare on Snopes; it was disseminated on purpose. Misinformation and disinformation are both, at their core, misleading or inaccurate information; what separates them is intent. Misinformation is the inadvertent sharing of false information; the sharer didn't intend to mislead people and genuinely believed the story. Disinformation, by contrast, is the deliberate creation and sharing of information known to be false. It's a malign narrative that is spread deliberately, with the explicit aim of causing confusion or leading the recipient to believe a lie. Computational propaganda is a suite of tools or tactics used in modern disinformation campaigns that take place online. These include automated social media accounts that spread the message and the algorithmic gaming of social media platforms to disseminate it. These tools facilitate the disinformation campaign's ultimate goal — media manipulation that pushes the false information into mass awareness.

As with propaganda, disinformation is not a new phenomenon: it was widely used during the Cold War by the KGB to sow distrust and cause confusion. But today's computationally driven disinformation campaigns are a form of information warfare. They are deployed in their new, high-velocity, well-targeted digital form to interfere in elections and break up societies. The narrative is often *laundered*: it first appears on an "alternative" site or blog, or as a post on a message board — something outside of traditional, more carefully vetted press — and is pushed into the social ecosystem using deliberate, coordinated tactics designed to game social net-

work algorithms. Determined manipulators work to ensure that content will jump from one social platform to the next, blanketing the ecosystem, touching a wide, receptive audience, and eventually trending on a prominent platform and – they hope – getting picked up by mainstream media.

As Lenin purportedly put it, "A lie told often enough becomes the truth." In the era of computational propaganda, we can update that aphorism: "If you make it trend, you make it true."

The early creators of the internet valued open participation and the free flow of information; early platforms such as Geocities and Blogger democratized access to creation tools. Anyone could have a site – for free! – in which he or she could create content and share a point of view with the world. In his seminal essay "A Declaration of the Independence of Cyberspace," the internet pioneer John Perry Barlow described this new, open platform as "a world where anyone, anywhere may express his or her beliefs," and where "we have no elected government." Zero-cost publishing eliminated the old editorial gatekeepers and gave anyone a voice. It was such a significant structural and cultural change that these newly empowered creators eventually came to have a name: the fifth estate.

As a result of the decentralized nature of Web 1.0, blogs were hard to discover, and there wasn't much to do except read them and occasionally leave a comment. Online communities began to thrive but were generally relegated to disparate message boards or specialized chat services. In the early days of the web, it was difficult to find information, let alone reach millions of people or make content trend.

But with the advent of social networks in the early 2000s, online audiences and activity gradually consolidated onto a handful of platforms. The social web was exciting; it was fun to be on the same platform as friends and family, and the ability to easily discover new people who shared the same interests was compelling. The once decentralized web became centralized. Facebook emerged, originally as a place for Ivy Leaguers to check out their college classmates; over the next decade, it became a platform for two billion people to connect with friends and family. It evolved into a source of news, a marketplace for online yard sales, a place to find interest groups, a place to share photos, a place to engage

with brands, a messenger tool, a venue for live video — a singular hub for spending time online. It expanded farther with the acquisition of Instagram and WhatsApp. At the same time, Google was evolving, becoming not only the web's dominant search engine but a major email platform, a social network, a messenger client, a photo-sharing platform, a source of productivity tools, financial tools, self-driving cars, internet-enabled balloons, and — with the acquisition of YouTube — a user-created video-sharing behemoth. Much smaller but still extremely popular platforms Reddit and Twitter also emerged for sharing and discussing news or quickly disseminating photos of cats and other memes.

These companies had to make money, and each chose the advertising business to monetize its growing audience. Advertising has been the business model that powers the internet since the earliest days of banners, popups, and garish blue hyperlinks; it's sometimes called "the original sin of the internet." As anyone who has seen an ad knows, advertising is most effective when it's targeted and engaging, and the social platforms were uniquely suited both to target and to engage. As users spent time on these platforms they sent out signals based on whom they followed, what they clicked on, what they Liked, and what they read. A person who checks in at high-end restaurants and travels frequently? Probably wealthy. One who searches for baby-related content or products? Possibly pregnant. The platforms processed each action and engagement into a profile, using behavior both on- and off-platform as signal. Then they leveraged that data both to inform ad targeting for their customers (the advertisers) and to decide what content to serve their users. As people spent more and more time on the social platforms, and as the platforms offered an ever-larger set of features from which to derive signal, the sophistication of user profiles increased. This model came to be called "surveillance capitalism" by its critics.

The design choices of the platforms were driven largely by the incentive structures of their ad-based business model: to serve people ads, those people must be active on the platform. To keep people active on the platform, the platforms needed to show them engaging content. The big social networks — Twitter, Facebook, YouTube — are competitive not because they contain similar features but because they are competing for their users' time and

attention; they are all attention brokers. Keeping people engaged is critical to operating a successful business.

Advertisers and platforms alike drove the arms race of engagement tactics, pioneering clickbait and testing ever more innovative ways of targeting audiences to interact with ads. Through the power of analytics, the platforms and advertisers came to learn that content with high emotional resonance – including anger or outrage – performed better. Platforms and advertisers looked at where people's eyes went as they scanned a screen and identified the kind of content they were most likely to engage with: prominent images, auto-played videos. Eventually the ads became indistinguishable in form from the rest of the content on the platforms, and the views–clicks–actions tracking became increasingly more refined. Lookalike audiences appeared, based on an algorithm that ran correlations between people with overlapping interests and proclivities; they eliminated the need for advertisers to create their own demographic or interest-based targeting criteria.

Anyone who wanted to target an audience, whether broad or niche, with a message could do so – quickly, easily, and inexpensively. And easy targeting of massive numbers of people consolidated on a handful of social platforms was not just a boon for marketers – it was also a boon for propagandists. The platforms, especially Facebook, offered reach. With inexpensive ads, a talented propagandist could pay for engagements and direct people to follow a Facebook Page or join a Facebook Group, enabling the propagandist to develop a long-term relationship.

To keep users engaged on site (to serve them ads), the platforms built design prompts to drive participation. They created friendly open text fields that simply asked "What's on your mind?" or "What's happening?" The barriers to participation decreased until they were frictionless, and every user became a content creator. It wasn't necessary to write a long blog post; a user could post a status update on Facebook or "microblog" on Twitter – or, as smartphones became ubiquitous, just snap a photo and add a filter on Instagram. Millions of people shared pictures of sandwiches and tweeted gripes about flight delays. And during election season, they tweeted and posted and vlogged about politics.

As user numbers and participation grew and content volume exploded, the primary information feeds on each site – News Feed

on Facebook, Feed on Twitter — became highly curated and personalized. Timelines were no longer reverse-chronological; it became impossible to scroll to the "end" of a feed. Indeed, there was no longer an "end" — feeds had become bottomless. We had moved into something new: a period of information glut, or at least content glut.

Managing a vast amount of content required new tools to help surface — and spread — the best of it. Virality engines became a core feature of social platforms: Twitter has the retweet button and hashtags that cluster tweets into topics and help users find and engage in conversations they care about. Facebook has the Share and Like buttons, which push shared content from one person's social graph into others'. The term *virality* is an allusion to epidemiology — information spreads throughout one cluster of people, and then expands outward to others, who spread it among their own clusters, and so on. Compelling content keeps people on the platform; seeing one's content go viral keeps people creating it. Anyone who can gather enough momentum from sharing, Likes, retweets, and other message-amplification features can spread a message across the platforms' large standing audiences for free. Anyone — including propagandists.

Virality engines are incredibly powerful; like most of the technology underpinning computational propaganda, they can be used both to inform and to harm. It's dual-use technology: social networking tool and information-war weapon. The internet has given people the ability to speak truth to power, to call attention to grievances, and to share moments of joy. But when manipulators use these tools to push a malign narrative that goes viral, it can spread far and fast — and it's extremely difficult to debunk it after the fact.

There are long-term benefits to achieving virality: the content often continues to be algorithmically amplified by the platforms themselves. To help solve the information-glut problem, the platforms have to act as curators. Three curatorial functions in particular have proliferated across the web: search, trending, and recommendation engines. And each of these has become a battleground for motivated propagandists.

Search seems straightforward; a user searches for a keyword or asks a question, and the platform returns the most relevant con-

tent that answers the query. But search has become a way of pushing an agenda. The signals that Google and others use to rank content can be manipulated, based on the techniques broadly associated with the term *search engine optimization* (SEO). SEO is practiced by legitimate businesses trying to be among the first listed in response to keyword searches relevant to their offering. SEO is also employed by groups who want to push a particular message. If a keyword does not bring up many results, then the only content that will turn up in search is the content produced by that group. Michael Golebiewski and danah boyd of Data & Society call this a data void. It ensures that someone curious about a unique term will see precisely the message that the propagandist wishes to present. This can have serious implications when the keyword is related to extremism or conspiratorial misinformation. As an example, searching for "Vitamin K shot" (a routine health intervention for newborns) returns almost entirely anti-vaccine propaganda; anti-vaccine conspiracists write prolific quantities of content about that keyword, actively selling the myth that the shot is harmful, causes cancer, causes SIDS. Searches for the phrase are sparse because medical authorities are not producing counter-content or fighting the SEO battle in response.

Search algorithms try to help people find answers; trending algorithms show users what large groups of people are talking about or reading *right now*. These algorithms draw signals from common keywords or URLs or hashtags being shared at the same time. These signals indicate that something important or interesting is happening. Trending piques curiosity; it's a draw for users who want to engage with others about a topic that is being actively discussed in real time. It is also a draw for propagandists, who manipulate the algorithm in two primary ways. The first is by flooding an active hashtag with propaganda to reach the people who are actively participating. The propagandist can push an agenda or shift, derail, or co-opt the conversation. The second technique is to create a trending topic from scratch, via coordinated manipulation using bots (automated accounts that pretend to be human). Leveraging automated accounts or fake personas to spread a message and start it trending creates the illusion that large numbers of people feel a certain way about a topic. This is sometimes called "manufactured consensus." Trending is a power-

ful tool for manipulating the media. If mainstream reporters who spend time on Twitter notice the trend and cover it without first verifying its legitimacy, the message may land in a newspaper with a huge readership or on a television program. If the reporters see and debunk such a trend, they are often still inadvertently reinforcing it. And if they ignore it, propagandists can still push a narrative to conspiratorial or media-suspicious communities claiming that the mainstream media is ignoring the truth. If something has trended, the propagandist has won the battle.

Recommendation engines are the third type of curatorial algorithm that enables computational propaganda and disinformation to spread. They suggest content that a user is likely to find compelling. In its most basic form, this is done via "content-based filtering" – the user is shown more content from a site or related to a topic that he or she has already engaged with and likes, such as Amazon recommending a new book about gardening to people who have previously bought gardening books. The other version of recommendations is based on what's known as "collaborative filtering" – a type of algorithm informed by what *people similar to a given user* like. This is why, for example, Pinterest may show someone who's pinned a lot of fitness content new pins featuring recipes for healthy dishes; people who are fitness focused are often also diet conscious. It can feel a bit like serendipity when it's done well, but when it is misused, collaborative filtering sends people who are prone to certain conspiracy theories – for example, that the attacks of 9/11 were an inside job – content related to other conspiracies, such as Pizzagate (the inane theory that Hillary Clinton and other Democrats operated a secret child-trafficking ring out of the basement of a pizza parlor in Washington, D.C.). This conspiracy cross-pollination happens on social platforms because the best predictor of belief in a conspiracy theory is belief in another conspiracy theory. The recommendation engine isn't wrong; the person who believes the conspiracy theory about 9/11 is likely to be interested in Pizzagate. But researchers share a deep concern that recommendation engines on social networks may be suggesting increasingly extreme content to users who are most prone to be radicalized.

These curatorial algorithms are designed to process simple social signals; they are not equipped to assess factual accuracy, and

there is no underlying ethics that enables them to recognize the negative consequences of serving up extremist propaganda. Algorithms do not function in the same way as human editors, for better or worse. They can be manipulated – perhaps not more than people can, but certainly differently from the way people are. This was brought into stark relief when Facebook eliminated human editors from an oversight role on its trending topics list after allegations that the humans were biased against conservative media. The subsequent algorithm-only trending feature immediately displayed a false story that Megyn Kelly had been fired from Fox News. The story was inaccurate, the domain (thelastlineofdefense.org) was dubious – but the article had gone viral on Facebook, and enough users were sharing it that it qualified for trending.

The original belief that the web would democratize access to all the world's information and give everyone a voice seems quaint today. Incentive structures, design decisions, and technology have delivered a manipulatable system that is being gamed by propagandists. It privileges popularity over factual accuracy. We still tell ourselves that high-quality, accurate content will rise to the top. But social algorithms are designed to amplify what people are talking about, and popularity is both easy to feign and often irrelevant when it comes to certain types of information – say, information about science or health.

One of the first groups to demonstrate how effectively the system worked for bad actors was ISIS. From 2012 to 2016, ISIS built a virtual, online caliphate. They posted videos to YouTube that were then cross-posted to Facebook and Twitter for dissemination. They created thousands of Twitter accounts, human fanboys as well as bots, leveraging the platform for one-on-one recruiting of sympathizers as well as to create trending hashtags and gloat about terrorist attacks. The journalist Nick Bilton described them as "the world's deadliest tech startup," noting their dominance at the precise type of digital marketing techniques that the social platforms were designed for:

> Ghost Security Group, a counterterrorism organization, has noted in the past that ISIS utilizes almost every social app imaginable to communicate and share its propaganda, in-

> cluding mainstays like Twitter and Facebook; encrypted chat apps such as Telegram, Surespot, and Threema; and messaging platforms including Kik and WhatsApp. The terror group shares videos of beheadings on YouTube and even more gruesome clips on LiveLeak. They use the remarkably secure Apple iMessage to communicate. They preach to their disciples across the world using Internet radio stations. When a terror attack takes place, they use Twitter to claim responsibility and their followers subsequently cheer with favorites and retweets.

ISIS was transparent about who they were and what they were doing; there were no covert tactics, no laundering of narratives, no pretending to be Americans. They were running a pure propaganda campaign, pushing the narrative that they were defenders of Islam, mighty warriors who had established a utopian caliphate, formidable foes against Western imperialist invaders, and — using images of jihadis holding kittens — warm and friendly family men. They attempted to instill fear in their enemies using videos of beheadings; they rallied potential recruits with promises of glory in high-production video game–style shorts, and they targeted women recruits with stories of honorable lives as the wives of brave fighters. Nicholas Glavin of the U.S. Naval War College summarized the themes in ISIS messaging: "brutality, mercy, victimhood, belonging, war and utopianism." High emotional resonance, manipulative persuasion. But what was distinct about ISIS was that they ran their online outreach operation like a sophisticated digital marketing campaign. ISIS was building a brand, right down to the prominent iconography of the black flag, and disseminating it computationally for maximum reach and organic spread. When researchers conducted a census of their activity on Twitter in December 2014, ISIS appeared to be operating between 46,000 and 70,000 Twitter accounts; by contrast, Twitter would later go on to identify a mere 3,814 accounts affiliated with the Russian Internet Research Agency's extensive operation in the 2016 election.

ISIS was perhaps the first highly visible adversary in what has come to be an ongoing information war, though the Russian Internet Research Agency's operation began around the same time.

The Russians ran a much more subtle type of computational propaganda campaign, but they used a similar playbook facilitated by the same infrastructure: the social network ecosystem. The Internet Research Agency's disinformation campaign appeared on all major platforms. There were dozens of Facebook pages and Instagram accounts, targeting both right-wing and left-wing interests, most of which pushed memes (highly shareable) and occasionally news stories with a bit of editorializing. Facebook events also featured prominently, with the aim of taking the operation to the streets. There were Twitter bots for amplification, and Twitter personas for spreading narratives and engaging with media — there were even some Twitter accounts that pretended to *be* media, with handles that sounded like local news stations. There were YouTube channels with original content, including one that featured two millennial black men having conversations in a style popular within vlogger culture. Reddit, Tumblr, and Medium all found evidence of Russian activity; the Internet Research Agency used Twitter's Vine video app and created meme boards on Pinterest. Music apps and games were created or co-opted; the popular game Pokémon Go even featured in the operation. There were propaganda blogs, such as BlackMattersUS, created to host original written content, many of which closely resembled citizen journalism efforts. WhiteHouse.gov petitions were created or co-opted.

Topics ran the gamut, with most of the propaganda focusing on social issues that had high emotional resonance: LGBTQ rights, religious culture, Second Amendment rights, secession, veterans affairs, and Native American affairs, among others. But the majority of the content, a prevalent theme across all platforms, was race, and material intended to stoke racial tension. The propaganda co-opted Black Lives Matter messaging and black culture (especially black women's culture) on the left; it created and amplified highly charged Blue Lives Matter and "southern culture" (Confederacy-related content) on the right. The percentage of explicitly political propaganda that mentioned the candidates in the 2016 election was low, but it, too, incorporated the themes of the culture wars rather than attacking policy planks, and used the rhetoric of ridicule and disgust. Although the investigation into the operation is ongoing, the content that researchers have uncovered was entirely negative toward the candidacy of Hillary Clinton on both the right

and the left (where the narratives turned to voter suppression); in the content on the right, the majority of the content was pro Donald Trump. This does not indicate coordination with the Trump campaign; that investigation is also ongoing. But it does show that determined actors can run self-directed propaganda campaigns and achieve incredible reach — Facebook alone estimated that 146 million users engaged with the content.

One of the reasons the reach was so high was that the propaganda was so shareable. And this was because one of the things that Russia did remarkably well in its quest to fracture American societal cohesion was leverage memes.

It is impossible to talk about modern-day propaganda on the social web without delving into memes. Most people familiar with the term think of a meme as a picture, often of a cat, with pithy writing in big white letters. But memes are more than that. The term was coined in 1976 by Richard Dawkins in *The Selfish Gene*. "We need a name for the new replicator, a noun that conveys the idea of a unit of cultural transmission, or a unit of *imitation*," he wrote. Memes are cultural genes; they are part of the body of a society, transmitted from person to person, able to mutate. Memes can be pictures, icons, lyrics, catchphrases — anything that individuals would immediately recognize and that people can piggyback on and apply to other scenarios. "Winter is Coming" — a phrase from the popular fantasy series *Game of Thrones* — is now used by thousands of people to describe any situation of impending doom, ranging from the serious (a favored political candidate losing) to the sardonic (a visit from a disliked mother-in-law). Another popular example is "This is fine" — a meme featuring a visual of a dog calmly sitting at a table drinking a cup of coffee while the room around him burns.

Memes seem like a novelty, but the defense industry and DARPA (Defense Advanced Research Projects Agency) — which created ARPANET, the forerunner of the internet — have studied them for years: "a meme is information which propagates, has impact, and persists." Memes are popular; people enjoy creating and sharing them. They feel authentic, and their messages resonate. DARPA researchers believe that memes can change individual and group values and behavior. They fit our information consumption in-

frastructure: big image, limited text, capable of being understood thoroughly with minimal effort. Person-to-person transmission and social virality tools enable them to spread easily and jump from group to group, evolving and changing as they do. They solidify cultural bonds and in-group ideological identity, and turn big ideas into emotional snippets. They have sticking power and virality potential, and they now become part of the cultural zeitgeist at internet speed. Memes are the propaganda of the digital age.

One piece of evidence supporting the power of memes is the fact that the more authoritarian governments ban memes; Russia banned the meme of Putin made to look like a gay clown. China banned a meme of president Xi Jinping reimagined as Winnie the Pooh. Ridicule is one of the most potent forces for breaking a powerful brand or cutting down a symbol of authority, and meme culture is extraordinarily adept at ridicule. Jeff Giesea, a consultant who helped build support for President Donald Trump during the 2016 campaign, put it simply in an article for NATO StratCom: "Cyber warfare is about taking control of data. Memetic warfare is about taking control of the dialogue, narrative, and psychological space. It's about denigrating, disrupting, and subverting the enemy's effort to do the same." Giesea's work has focused on countering ISIS propaganda as well – countering the computational with the computational. "Trolling, it might be said, is the social media equivalent of guerrilla warfare, and memes are its currency of propaganda."

And that gets at one of the key challenges today: computational propaganda and influence operations are inexpensive, they appear to be effective, and the tech platforms that built the infrastructure are having a very difficult time stopping them. As terrorist networks, state actors, domestic ideologues, conspiracy theorists, politicians, and everyday internet users able to create and distribute propaganda, we have yet to answer the key question: How can we push back against this ubiquitous online manipulation?

In any effort to find solutions to the epidemic of propaganda on social media, we must stay attuned to the social dynamics and human biases that the manipulators insidiously prey on. Russian propagandists, for example, exploited societal fractures such as race because these tensions are a very real problem. We also need

to take into account the problem of human biases intersecting with technological design, and the problem of finite attention. For most of human history, people lived in an environment of information scarcity. We now live in an era of information glut. There is simply too much content, so the platforms decide what we will see in the time we allocate to them. As the economist Herbert Simon put it, "A wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it." The platforms also recognize that people find comfort in seeing things that conform to their own worldview. People have an emotional response to angry content and hateful content, and the unsophisticated yet phenomenally astute algorithms serve up content that meets these base needs. Computational propaganda spreads, in part, because of the enthusiasm of the people who are targeted to receive it, and who then feel a deep desire to pass the message along. The crowd that spreads the message is digital now, gathered in Facebook Groups and Twitter hashtags instead of in a stone-and-asphalt public square, but the human needs, the motivations, the drives remain the same. Facts are largely irrelevant; a recent study of Twitter from 2006 to 2017 tracked 126,000 rumors spread by 3 million people. "False news reached more people than the truth; the top 1% of false news cascades diffused to between 1000 and 100,000 people, whereas the truth rarely diffused to more than 1000 people. Falsehood also diffused faster than the truth." The researchers who published this conclusion, Soroush Vosoughi, Deb Roy, and Sinan Aral, noted that false news "was more novel than true news," and suggested that novelty might be one reason people shared it.

The power to influence opinions lies with those who can most widely and effectively disseminate a message. If you control — or effectively game — the algorithms that decide dissemination, you control the messages people see. And if you control the messages, you control the people reading them. As Bernays put it, "In almost every act of our lives, whether in the sphere of politics or business, in our social conduct or our ethical thinking, we are dominated by the relatively small number of persons . . . who understand the mental processes and social patterns of the masses. It is they who pull the wires that control the public mind, who harness old social

forces and contrive new ways to bind and guide the world." In the age of computational propaganda, it is they who exploit societal fissures and influence the vote.

The owners of the platforms that created the algorithms are at a loss. They struggle to define their responsibilities, they struggle with the subtle distinctions among propaganda, disinformation, and misinformation, and they struggle with a desire to remain neutral and maintain a commitment to the principle of free speech. The decision of whether and how to moderate content is complex; First Amendment protections don't apply to users who post content on a private platform, and legally the platforms have their own First Amendment right to moderate content. However, the platforms have traditionally tried to maintain as light a touch as possible. This is partially out of a commitment to the ideal of free expression, but also due to the desire to avoid the controversy and allegations of censorship and bias that inevitably follow prominent takedowns.

Ironically, the platforms powering the information ecosystem of a free and democratic society are more vulnerable to manipulation than the censored platforms that operate in authoritarian regimes. Our commitment to free speech has rendered us hesitant to take down disinformation and propaganda until it is conclusively and concretely identified as such beyond a reasonable doubt. That hesitation gives narratives the opportunity to take hold, and gives propagandists an opportunity to reach the vulnerable.

As the infrastructures for disseminating information have changed over the centuries, with the advent of the printing press, the radio, and the television, people have adapted – but in doing so they established new norms and new regulations to protect consumers and mitigate or neutralize newly enabled societal harms. Conversations about norms and regulations for mitigating the effects of computational propaganda are taking place in earnest now, as the 2018 U.S. elections loom. Some people are taking a demand-side approach to the problem, stressing more media literacy and trying to educate people about how to avoid falling victim to disinformation. Organizations like Common Sense Media have developed "News Literacy 101" curricula, and there are even games that explain and combat misinformation. Activist groups are lobbying the platforms to take responsibility for reducing the

supply of disinformation, both publicly and in private, via back-channel conversations with the engineers and decision makers who work at the companies. And the regulators are getting actively involved, considering legislation on a variety of fronts: to break up the platforms on antitrust grounds, to improve privacy and restrict data gathering, to require disclosure of automated accounts, and to mandate ad transparency.

One of the primary challenges to a solution is that this is a systems-level problem, and even the platforms have power only over their own walled garden. Around the same time that ISIS began to amass their audience and Russia was secretly launching its campaign initiative, researchers began to observe that conspiracy theorists, fringe ideologues, and anti-government activists were also gaining a prominent share of voice in the social ecosystem, one that seemed disproportionate to their representation in the offline world. Anyone with a message and a botnet could manipulate the system in fairly predictable ways. These coordinated campaigns became so common that they were fairly formulaic – except, it seemed, for the platforms that were in the best position to intervene. This is because although the problem is systemic, each platform, as noted, has visibility primarily into its own data, and very little awareness of what is happening elsewhere. No one oversees the health of the *ecosystem* in which two billion people get their news and information; outside of the United States, the disinformation and propaganda problem in other countries has led to genocide (Myanmar), murders (India), and state-sponsored trolling.

The stakes are high: we are living in an era in which trust in the media is at an all-time low. Trust in authority figures is also at an all-time low. And now, as the vulnerabilities in the social ecosystem have become more apparent, we are seeing not just an erosion of trust in the tech platforms but also a rising suspicion of fellow participants. A quick glance into hot-topic political Twitter threads reveals people accusing one another of being bots simply because they disagree.

There has always been propaganda. But it has not previously been algorithmically amplified and deliberately targeted to reach precisely the people who are most vulnerable. It has never before been so easy to produce or so inexpensive to spread. The idealistic

vision of citizen journalism and "the fifth estate" has manifested as an information ecosystem in which polarized partisan extremists, state intelligence services, and terrorists push self-serving propaganda and attack the very idea of objectivity. A sitting president advances the mantra of "alternative facts" and retweets conspiracy theorists to his hundreds of millions of followers.

These technologies will continue to evolve: disinformation campaign content will soon include manufactured video and audio. We can see it coming but are not equipped to prevent it. What will happen when video makes us distrust what we see with our own eyes? If democracy is predicated on an informed citizenry, then the increasing pervasiveness of computational propaganda is a fundamental problem. Through a series of unintended consequences, algorithms have inadvertently become the invisible rulers that control the destinies of millions. Now we have to decide what we are going to do about that.