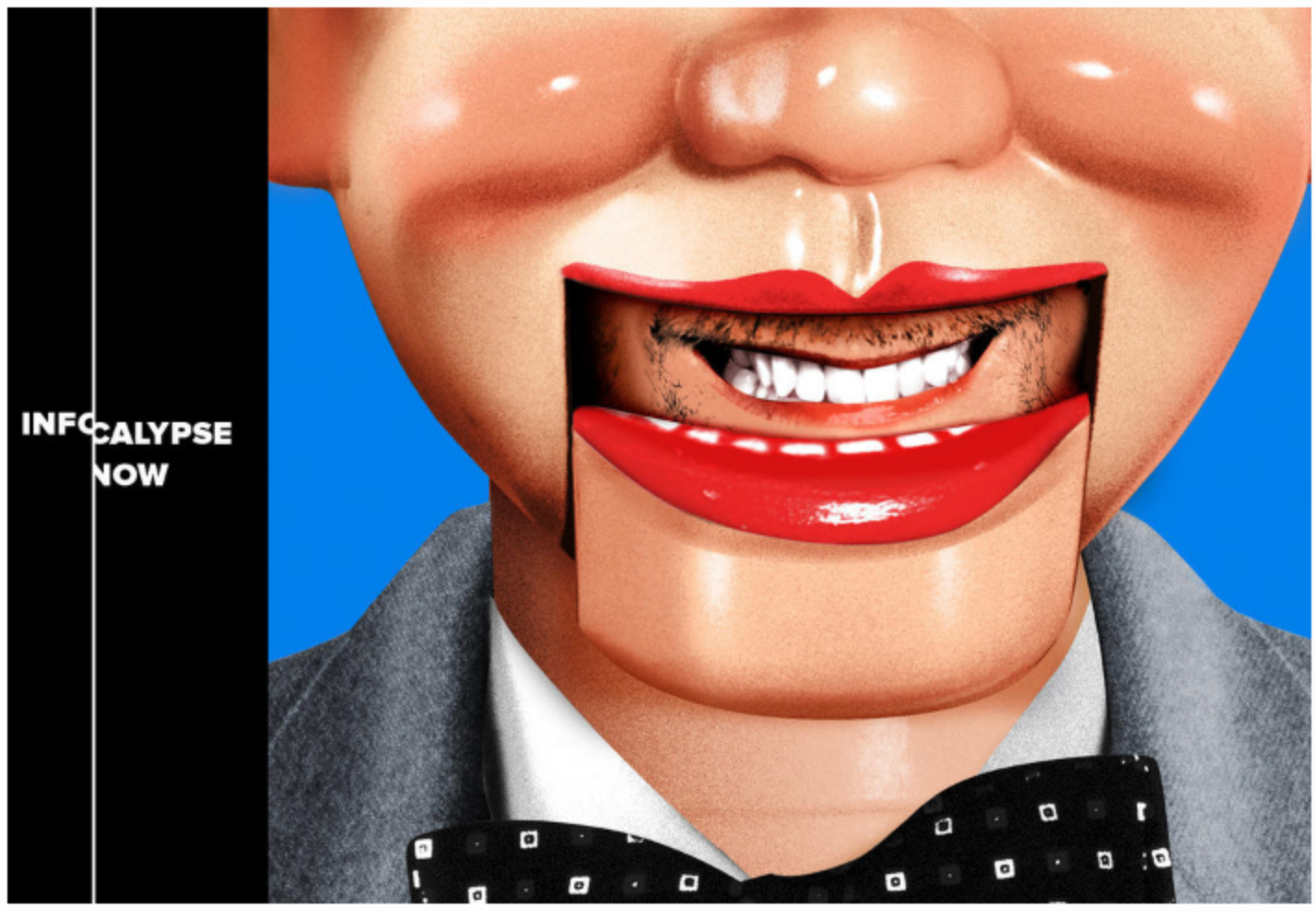
He Predicted The 2016 Fake News Crisis. Now He's Worried About An Information Apocalypse.

By **Charlie Warzel**, www.buzzfeed.com
février 11, 2018

Afficher l'original

ai bots fake news info wars + 1 autre(s)



In mid-2016, Aviv Ovadya realized there was something fundamentally wrong with the internet — so wrong that he abandoned his work and sounded an alarm. A few weeks before the 2016 election, he presented his concerns to technologists in San Francisco's Bay Area and warned of an impending crisis of misinformation in a presentation he titled "Infocalypse."

The web and the information ecosystem that had developed around it was wildly unhealthy, Ovadya argued. The incentives that governed its biggest platforms were calibrated to reward information that was often misleading and polarizing, or both. Platforms like Facebook, Twitter, and Google prioritized clicks, shares, ads, and money over quality of information, and Ovadya couldn't shake the feeling that it was all building toward something bad — a kind of critical threshold of addictive and toxic misinformation. The presentation was largely ignored by employees from the Big Tech platforms — including a few from Facebook who would later go on to drive the company's NewsFeed integrity effort.

"At the time, it felt like we were in a car careening out of control and it wasn't just that everyone was saying, 'we'll be fine' — it's that they didn't even see the car," he said.

Ovadya saw early what many — including lawmakers, journalists, and Big Tech CEOs — wouldn't grasp until months later: Our platformed and algorithmically optimized world is vulnerable — to propaganda, to misinformation, to dark targeted advertising from foreign governments

— so much so that it threatens to undermine a cornerstone of human discourse: the credibility of fact.

But it's what he sees coming next that will really scare the shit out of you.

"Alarmism can be good — you should be alarmist about this stuff," Ovadya said one January afternoon before calmly outlining a deeply unsettling projection about the next two decades of fake news, artificial intelligence-assisted misinformation campaigns, and propaganda. "We are so screwed it's beyond what most of us can imagine," he said. "We were utterly screwed a year and a half ago and we're even more screwed now. And depending how far you look into the future it just gets worse."

That future, according to Ovadya, will arrive with a slew of slick, easy-to-use, and eventually seamless technological tools for manipulating perception and falsifying reality, for which terms have already been coined — "reality apathy," "automated laser phishing," and "human puppets."

Which is why Ovadya, an MIT grad with engineering stints at tech companies like Quora, dropped everything in early 2016 to try to prevent what he saw as a Big Tech-enabled information crisis. "One day something just clicked," he said of his awakening. It became clear to him that, if somebody were to exploit our attention economy and use the platforms that undergird it to distort the truth, there were no real checks and balances to stop it. "I realized if these systems were going to go out of control, there'd be nothing to reign them in and it was going to get bad, and quick," he said.

Today Ovadya and a cohort of loosely affiliated researchers and academics are anxiously looking ahead — toward a future that is alarmingly dystopian. They're running war game-style disaster scenarios based on technologies that have begun to pop up and the outcomes are typically disheartening.

For Ovadya — now the chief technologist for the University of Michigan's Center for Social Media Responsibility and a Knight News innovation fellow at the Tow Center for Digital Journalism at Columbia — the shock and ongoing anxiety over Russian Facebook ads and Twitter bots pales in comparison to the greater threat: Technologies that can be used to enhance and distort what is real are evolving faster than our ability to understand and control or mitigate it. The stakes are high and the possible consequences more disastrous than foreign meddling in an election — an undermining or upending of core civilizational institutions, an "infocalypse." And Ovadya says that this one is just as plausible as the last one — and worse.

Worse because of our ever-expanding computational prowess; worse because of ongoing advancements in artificial intelligence and machine learning that can blur the lines between fact and fiction; worse because those things could usher in a future where, as Ovadya observes, anyone

could make it "appear as if anything has happened, regardless of whether or not it did."

And much in the way that foreign-sponsored, targeted misinformation campaigns didn't feel like a plausible near-term threat until we realized that it was already happening, Ovadya cautions that fast-developing tools powered by artificial intelligence, machine learning, and augmented reality tech could be hijacked and used by bad actors to imitate humans and wage an information war.

And we're closer than one might think to a potential "Infocalypse." Already available tools for audio and video manipulation have begun to look like a potential fake news Manhattan Project. In the murky corners of the internet, people have begun using machine learning algorithms and open-source software to easily create pornographic videos that realistically superimpose the faces of celebrities — or anyone for that matter — on the adult actors' bodies. At institutions like Stanford, technologists have built programs that that combine and mix recorded video footage with real-time face tracking to manipulate video. Similarly, at the University of Washington computer scientists successfully built a program capable of "turning audio clips into a realistic, lip-synced video of the person speaking those words." As proof of concept, both the teams manipulated broadcast video to make world leaders appear to say things they never actually said.

As these tools become democratized and widespread, Ovadya notes that the worst case scenarios could be extremely destabilizing.

There's "diplomacy manipulation," in which a malicious actor uses advanced technology to "create the belief that an event has occurred" to influence geopolitics. Imagine, for example, a machine-learning algorithm (which analyzes gobs of data in order to teach itself to perform a particular function) fed on hundreds of hours of footage of Donald Trump or North Korean dictator Kim Jong Un, which could then spit out a near-perfect — and virtually impossible to distinguish from reality — audio or video clip of the leader declaring nuclear or biological war. "It doesn't have to be perfect — just good enough to make the enemy think something happened that it provokes a knee-jerk and reckless response of retaliation."

Another scenario, which Ovadya dubs "polity simulation," is a dystopian combination of political botnets and astroturfing, where political movements are manipulated by fake grassroots campaigns. In Ovadya's envisioning, increasingly believable AI-powered bots will be able to effectively compete with real humans for legislator and regulator attention because it will be too difficult to tell the difference. Building upon previous iterations, where public discourse is manipulated, it may soon be possible to directly jam congressional switchboards with heartfelt, believable algorithmically-generated pleas. Similarly, Senators' inboxes could be flooded with messages from constituents that were cobbled together by machine-learning programs working off stitched-together content culled from text, audio, and social media profiles.

Then there's automated laser phishing, a tactic Ovadya notes security researchers are already whispering about. Essentially, it's using AI to scan things, like our social media presences, and craft false but believable messages from people we know. The game changer, according to Ovadya, is that something like laser phishing would allow bad actors to target anyone and to create a believable imitation of them using publicly available data.

"Previously one would have needed to have a human to mimic a voice or come up with an authentic fake conversation — in this version you could just press a button using open source software," Ovadya said. "That's where it becomes novel — when anyone can do it because it's trivial. Then it's a whole different ball game."

Imagine, he suggests, phishing messages that aren't just a confusing link you might click, but a personalized message with context. "Not just an email, but an email from a friend that you've been anxiously waiting for for a while," he said. "And because it would be so easy to create things that are fake you'd become overwhelmed. If every bit of spam you receive looked identical to emails from real people you knew, each one with its own motivation trying to convince you of something, you'd just end up saying, 'okay, I'm going to ignore my inbox.'"

That can lead to something Ovadya calls "reality apathy": Beset by a torrent of constant misinformation, people simply start to give up. Ovadya is quick to remind us that this is common in areas where information is poor and thus assumed to be incorrect. The big difference, Ovadya notes, is the adoption of apathy to a developed society like ours. The outcome, he fears, is not good. "People stop paying attention to news and that fundamental level of informedness required for functional democracy becomes unstable."

Ovadya (and other researchers) see laser phishing as an inevitability. "It's a threat for sure, but even worse — I don't think there's a solution right now," he said. "There's internet scale infrastructure stuff that needs to be built to stop this if it starts."

Beyond all this, there are other long-range nightmare scenarios that Ovadya describes as "far-fetched," but they're not so far-fetched that he's willing to rule them out. And they are frightening. "Human puppets," for example — a black market version of a social media marketplace with people instead of bots. "It's essentially a mature future cross border market for manipulatable humans," he said.

Ovadya's premonitions are particularly terrifying given the ease with which our democracy has already been manipulated by the most rudimentary, blunt-force misinformation techniques. The scamming, deception, and obfuscation that's coming is nothing new; it's just more sophisticated, much harder to detect, and working in tandem with other technological forces that are not only currently unknown but likely unpredictable.

For those paying close attention to developments in artificial intelligence and machine learning, none of this feels like much of a stretch. Software currently in development at the chip manufacturer Nvidia can already convincingly generate hyperrealistic photos of objects, people, and even some landscapes by scouring tens of thousands of images. Adobe also recently piloted two projects — Voco and Cloak — the first a "Photoshop for audio," the second a tool that can seamlessly remove objects (and people!) from video in a matter of clicks.

In some cases, the technology is so good that it's startled even its creators. Ian Goodfellow, a Google Brain research scientist who helped code the first "generative adversarial network" (GAN), which is a neural network capable of learning without human supervision, cautioned that AI could set news consumption back roughly 100 years. At an MIT Technology Review conference in November last year, he told an audience that GANs have both "imagination and introspection" and "can tell how well the generator is doing without relying on human feedback." And that, while the creative possibilities for the machines is boundless, the innovation, when applied to the way we consume information, would likely "clos[e] some of the doors that our generation has been used to having open."

In that light, scenarios like Ovadya's polity simulation feel genuinely plausible. This summer, more than one million fake bot accounts flooded the FCC's open comments system to "amplify the call to repeal net neutrality protections." Researchers concluded that automated comments — some using natural language processing to appear real — obscured legitimate comments, undermining the authenticity of the entire open comments system. Ovadya nods to the FCC example as well as the recent bot-amplified #releasethememo campaign as a blunt version of what's to come. "It can just get so much worse," he said.

Arguably, this sort of erosion of authenticity and the integrity of official statements altogether is the most sinister and worrying of these future threats. "Whether it's AI, peculiar Amazon manipulation hacks, or fake political activism — these technological underpinnings [lead] to the increasing erosion of trust," computational propaganda researcher Renee DiResta said of the future threat. "It makes it possible to cast aspersions on whether videos — or advocacy for that matter — are real." DiResta pointed out Donald Trump's recent denial that it was his voice on the infamous *Access Hollywood* tape, citing experts who told him it's possible it was digitally faked. "You don't need to create the fake video for this tech to have a serious impact. You just point to the fact that the tech exists and you can impugn the integrity of the stuff that's real."

It's why researchers and technologists like DiResta — who spent years of her spare time advising the Obama administration, and now members of the Senate Intelligence Committee, against disinformation campaigns from trolls — and Ovadya (though they work separately) are beginning to talk more about the looming threats. Last week, the NYC Media Lab, which helps the city's companies and academics collaborate, announced a plan to bring together technologists and researchers in June to "explore

worst case scenarios" for the future of news and tech. The event, which they've named Fake News Horror Show, is billed as "a science fair of terrifying propaganda tools — some real and some imagined, but all based on plausible technologies."

"In the next two, three, four years we're going to have to plan for hobbyist propagandists who can make a fortune by creating highly realistic, photo realistic simulations," Justin Hendrix, the executive director of NYC Media Lab, told BuzzFeed News. "And should those attempts work, and people come to suspect that there's no underlying reality to media artifacts of any kind, then we're in a really difficult place. It'll only take a couple of big hoaxes to really convince the public that nothing's real."

Given the early dismissals of the efficacy of misinformation — like Facebook CEO Mark Zuckerberg's now-infamous statement that it was "crazy" that fake news on his site played a crucial role in the 2016 election — the first step for researchers like Ovadya is a daunting one: Convince the greater public, as well as lawmakers, university technologists, and tech companies, that a reality-distorting information apocalypse is not only plausible, but close at hand.

A senior federal employee explicitly tasked with investigating information warfare told BuzzFeed News that even he's not certain how many government agencies are preparing for scenarios like the ones Ovadya and others describe. "We're less on our back feet than we were a year ago," he said, before noting that that's not nearly good enough. "I think about it from the sense of the enlightenment — which was all about the search for truth," the employee told BuzzFeed News. "I think what you're seeing now is an attack on the enlightenment — and enlightenment documents like the Constitution — by adversaries trying to create a post-truth society. And that's a direct threat to the foundations of our current civilization."

That's a terrifying thought — more so because forecasting this kind of stuff is so tricky. Computational propaganda is far more qualitative than quantitative — a climate scientist can point to explicit data showing rising temperatures, whereas it's virtually impossible to build a trustworthy prediction model mapping the future impact of yet-to-be-perfected technology.

For technologists like the federal employee, the only viable way forward is to urge caution, to weigh the moral and ethical implications of the tools being built and, in so doing, avoid the Frankensteinian moment when the creature turns to you and asks, "Did you ever consider the consequences of your actions?"

"I'm from the free and open source culture — the goal isn't to stop technology but ensure we're in an equilibria that's positive for people. So I'm not just shouting 'this is going to happen,' but instead saying, 'consider it seriously, examine the implications," Ovadya told BuzzFeed News. "The thing I say is, 'trust that this isn't not going to happen.'"

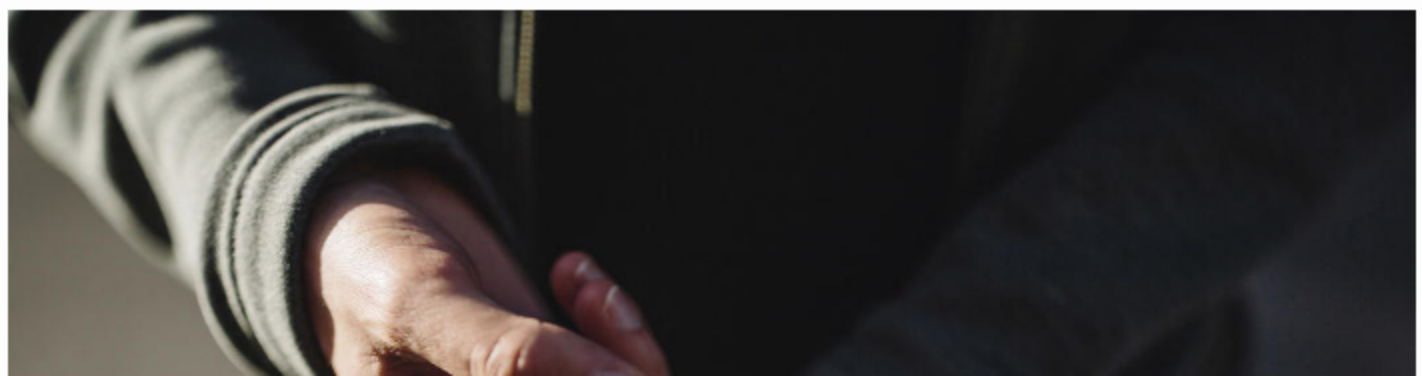
Hardly an encouraging pronouncement. That said, Ovadya does admit to a bit of optimism. There's more interest in the computational propaganda space then ever before, and those who were previously slow to take threats seriously are now more receptive to warnings. "In the beginning it was really bleak — few listened," he said. "But the last few months have been really promising. Some of the checks and balances are beginning to fall into place." Similarly, there are solutions to be found — like cryptographic verification of images and audio, which could help distinguish what's real and what's manipulated.

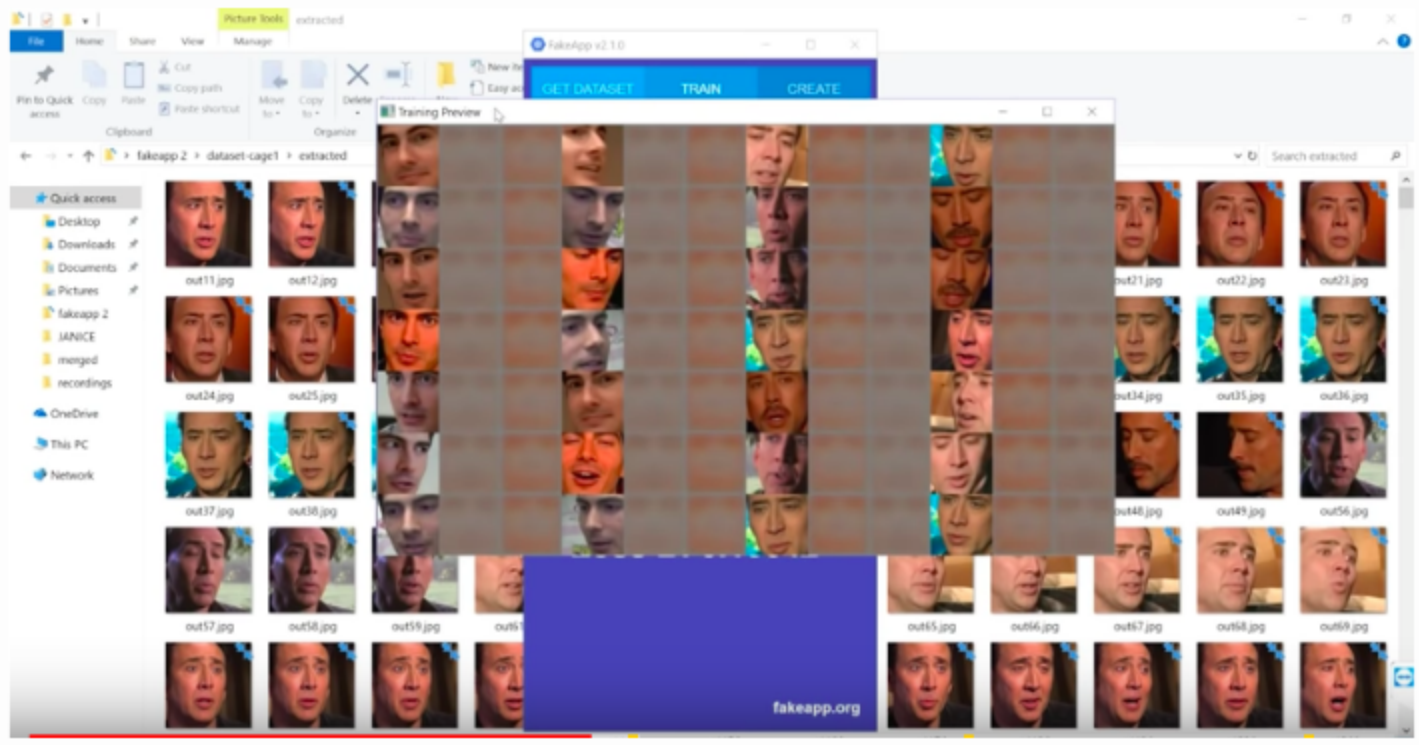
Still, Ovadya and others warn that the next few years could be rocky. Despite some pledges for reform, he feels the platforms are still governed by the wrong, sensationalist incentives, where clickbait and lower-quality content is rewarded with more attention. "That's a hard nut to crack in general, and when you combine it with a system like Facebook, which is a content accelerator, it becomes very dangerous."

Just how far out we are from that danger remains to be seen. Asked about the warning signs he's keeping an eye out for, Ovadya paused. "I'm not sure, really. Unfortunately, a lot of the warning signs have already happened." ●

Outside Your Bubble is a BuzzFeed News effort to bring you a diversity of thought and opinion from around the internet. If you don't see your viewpoint represented, contact the editor at bubble@buzzfeed.com. Click here to join our Facebook group and discuss stories that have changed how you see the world, or here for more on Outside Your Bubble.

And if you want to read more about the future of the internet's information wars, subscribe to Infowarzel, a BuzzFeed News newsletter by the author of this piece, Charlie Warzel.





For more BuzzFeed content, download our **mobile app**, sign up for **newsletters**, or follow us on **Facebook**, **Twitter** or **Pinterest**.