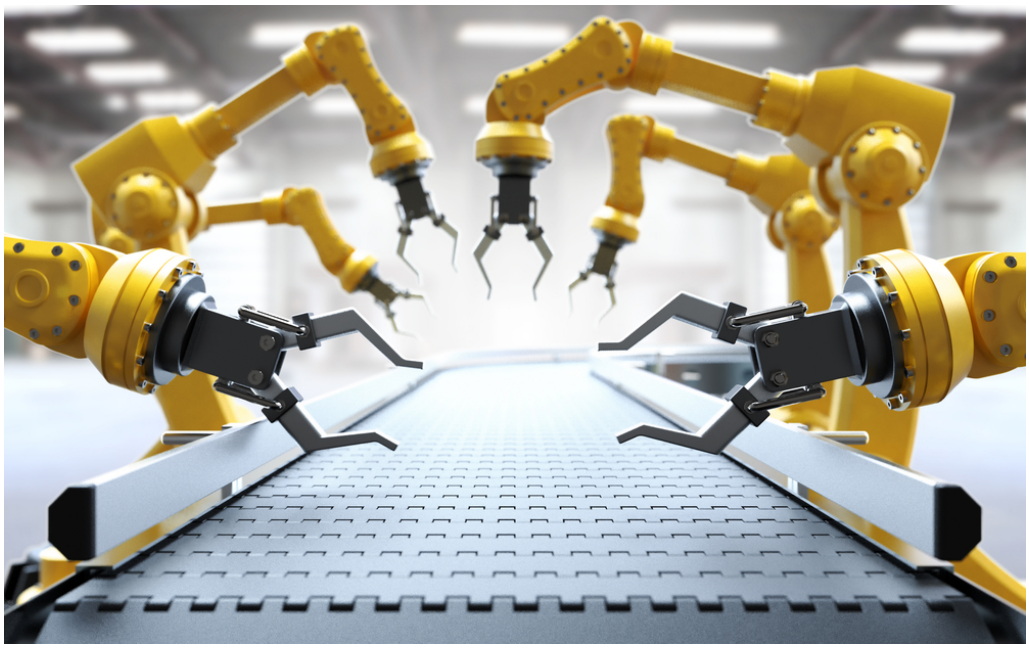
National Endowment for Democracy
Supporting freedom around the world

forum
International Forum for Democratic Studies

THE BIG QUESTION: HOW WILL 'DEEPFAKES' AND EMERGING TECHNOLOGY TRANSFORM DISINFORMATION?



Disinformation, the intentional use of false or misleading information for political purposes, is increasingly recognized as a threat to democracy worldwide. Many observers argue that this challenge has been exacerbated by social media and a declining environment for independent news outlets. Now, new advances in technology—including but not limited to "deepfakes" and other forms of synthetic media—threaten to supercharge the disinformation crisis.

The International Forum for Democratic Studies asked five leading experts about the role that deepfakes and other emerging applications may play in the disinformation landscape. (Their answers have been edited for length and clarity, and do not necessarily reflect the views of the National Endowment for Democracy.)

***Nic Dias** is a Senior Research Fellow for First Draft. He recently graduated from the Columbia Journalism School and specialized in data and computation. Follow him on Twitter @niccdias.*

Any robust defense against malicious, newsworthy "deepfakes" and other AI-generated synthetic media is going to have to involve journalists. Their purpose—to seek the truth

on behalf of the public—is best aligned to this task. Sophisticated, algorithmically generated fabrications aside though, journalists continued to be fooled by comparably simple cases of photoshopping and imposter social media accounts. A study from ICFJ not long ago found that, while 71 percent of journalists use social media to find new stories, only 11 percent use social media verification tools. The first step, then, in developing a robust defense against malicious, AI-generated synthetic media is to establish standards and norms of verification in newsrooms. Put simply, we need to get journalists to get in the habit of checking images and videos in same way that they now check text-based facts.

As is clear from a recent report by First Draft and WITNESS, the means of detecting deepfakes is necessarily computational, though not necessarily automatic. Given this fact, the social platforms will need to be involved in the development of tools to detect these manipulations. They have the resources to fund development of these tools, and are the only actors remotely in a position to effectively scan for distortions. Considerable innovations in reverse video search (the ability to search for other appearances of a video online) have already been made by several start-ups. An effective reverse video search tool alone would go a long way toward helping to identify deepfakes, as well as other kinds of misuses of video.

Renée DiResta *is the Director of Research at New Knowledge, and Head of Policy at nonprofit Data for Democracy. She investigates the spread of malign narratives across social networks, and assists policymakers in understanding and responding to the problem. She has advised Congress, the State Department, and other academic, civic, and business organizations, and has studied disinformation and computational propaganda in the context of pseudoscience conspiracies, terrorism, and state-sponsored information warfare. Follow her on Twitter @noupside.*

Influence operation tactics are constantly evolving. As the platforms develop new features, and as new technologies emerge, adversaries immediately evaluate their potential and exploit them—this is an arms race. Disinformation campaigns exploit modern information infrastructure to manipulate a population. Malign actors blanket the social ecosystem with synchronized propaganda to create media mirages that surround targeted groups.

One of the most challenging shifts in our ability to clear or prevent that mirage will occur when automated accounts—bots—begin to do a passable job of presenting themselves as human when chatting with people. We know that this coming, because there are significant and valuable commercial applications for chatbot technology. We've seen the engagement and impact that a handful of human-run sockpuppet (or fraudulent personality) accounts and a handful of fully automated amplifier bots can generate; in the case of Russia's interference in the 2016 US election, a few thousand human-managed aliases tweeted millions of times and attracted significant audiences.

The amplifier bots are presently primitive, and easy to detect. Sophisticated chatbot technology will reduce the cost of operations, and has the potential to eliminate the need for direct human involvement in running fake online accounts. This will enable sophisticated adversaries to run orders of magnitude more accounts—ones that are

virtually indistinguishable from real people—and expands the availability of the tactic to less-resourced adversaries. These bots will be able to process cues and engage with the individuals they are targeting. The mirage will get far more personalized and targeted... and as a result, more persuasive.

Sam Gregory *helps people use the power of the moving image and participatory technologies to create human rights change. He is Program Director of WITNESS and he also teaches the first graduate level course at Harvard on harnessing the power of new visual and participatory technologies for human rights change. Follow him on Twitter @SamGregory.*

The most serious ramification of deepfakes and other forms of synthetic media is that they further damage people's trust in our shared information sphere and contribute to the move of our default response from trust to mistrust. This could result from either widespread actual usage of deepfakes or widespread rhetorical usage by public figures who call 'deepfakes' on news they don't like or exercise 'plausible deniability' on compromising images and audio. Our current tendency to sensationalize deepfakes contributes to this problem.

Scenarios for the usage of synthetic media (including the ability to plausibly manipulate facial expressions in video, synthesize someone's voice, or make subtle removal edits to a video) include 'floods of falsehood' created via computational propaganda and individualized microtargeting, which could target discrete individuals with fake audio and overwhelm fact-finding and verification through sheer volume of manipulated content. At the other extreme, a timely 'credible doppelganger' of a real person shared in closed messaging apps could incite violence, or a subtle edit of a photo or video could challenge many fact-finding approaches. Right now we are ill-prepared to identify manipulated content, and most forensic approaches do not work at scale. Meanwhile, fact-checking is only just catching up to audio and video.

To counter this threat, we need to ask how it combines with the reduced barrier to entry for existing threats. We need to fight malicious deepfakes within the context of existing trends in misinformation and disinformation such as computational propaganda, coordinated organizing by networked malicious actors, the attention economy, and financial and political pressures on media outlets.

In our ongoing work on solutions around malicious deepfakes, WITNESS held a first cross-disciplinary expert convening which produced a survey of potential solution areas and ongoing work. More information can be found here.

Lisa-Maria Neudert *is a researcher at the Computational Propaganda project, where her work is located at the nexus of political communication, technology studies, and governance. Her previous research has focused on propaganda, social bots, and fake news—in its relation to the evolving digital media ecosystem. Follow her on Twitter @lmneudert.*

Political bots—automated software scripts on social media—have been at the center of digital influence campaigns. They distort public discourse with meaningless chatter,

amplify extremist viewpoints, and suppress minority voices with hate speech. Facing widespread evidence of bot activity in elections all over the world, social media platforms eventually took action, removing millions of bot accounts in 2018 alone. What's more, concerns about political bots emerged on the public agenda with user manuals for bot-spotting popping up across major publications in the US—increasing awareness and media literacy.

As a result, it may appear that democratic societies are finally gaining the upper hand in the cat-and-mouse game of bot detection. While looming threats from deep fakes, psychographics, and propaganda on "dark social" channels may appear more precarious, the next generation of bots is preparing for attack. This time around political bots will leave repetitive, automated tasks behind, and instead become intelligent: the rapid advances in artificial intelligence and natural language processing that help Amazon's Alexa and Apple's Siri get smarter are also teaching propaganda bots how to talk. To drive innovation, toolkits and platforms for natural language processing are open to third party developers; but bad actors also have access and could again leverage sophisticated technologies to manipulate democratic processes. When computational propaganda becomes conversational, bots will become more human and harder to detect than ever.

Sam Woolley *is the research director of the Digital Intelligence Lab at the Institute for the Future, a research associate at the Oxford Internet Institute, and an associate member of Green Templeton College at the University of Oxford. He specializes in the study of automation/AI, political communication, and information warfare. He is also a co-founder and former research director of the Computational Propaganda (ComProp) research team at the University of Oxford and the University of Washington. Follow him on Twitter @samuelwoolley.*

The spread of propaganda using digital media across multiple elections and security crises in multiple countries has left the world in little doubt as to the rise and repercussions of disinformation. In Myanmar, online rumors about the Muslim Rohingya minority fueled the murders of tens of thousands. In India, disinformation campaigns on Facebook led to harassment of women and attacks upon journalists. Throughout Europe, far-right parties used false online stories about refugee violence against women to propel isolationism and fear.

In the future, what emergent technologies will be used to spread computational propaganda? Tools such as artificial intelligence, automated voice systems, machine learning, deepfakes, interactive memes, virtual reality, and augmented reality will make digital disinformation more effective and harder to combat. Deepfakes are getting the lion's share of attention in relationship to the digital disinformation problem because they can be used to emulate real pictures and video. But future computational propagandists will not only target the visual. Forthcoming campaigns will likely harness our other senses: realistic sounding AI voices represent a new future for push polling and VR will allow for multisensory propaganda experiences.

But the future of computational propaganda, while frightening, is also manageable. We may not be able to alter how the internet was used to challenge democracy during

moments in years past, but we can follow signals to prevent manipulation in the future. With simulation, we can bring people closer together by putting them in another person's shoes so they can see, hear and—to a degree—feel what difference is like. We can use online video to challenge people's mental models and harness new social media platforms to encourage an omnivorous information diet. What has become a deep skepticism about the news media can be redirected into a skepticism about polarizing content and falsehood.