Foreign Policy
at BROOKINGS

JANUARY 2023

DEEPFAKES AND INTERNATIONAL CONFLICT

Daniel L. Byman, Chongyang Gao, Chris Meserole,
and V.S. Subrahmanian

EXECUTIVE SUMMARY

Deceit and media manipulation have always been a part of wartime communications, but never before has it been possible for nearly any actor in a conflict to generate realistic audio, video, and text of their opponent's political officials and military leaders. As artificial intelligence (AI) grows more sophisticated and the cost of computing continues to drop, the challenge deepfakes pose to online information environments during armed conflict will only grow.

To navigate that challenge, security officials and policymakers need a far greater understanding of how the technology works and the myriad ways it can be used in international armed conflict. Deepfakes can be leveraged for a wide range of purposes, including falsifying orders from military leaders, sowing confusion among the public and armed forces, and lending legitimacy to wars and uprisings. While these tactics can and often will fail, their potential to impact an adversary's communications and messaging mean that security and intelligence officials will inevitably use them in a wide range of operations.

For policymakers and officials in democratic states, deepfakes pose a particularly difficult challenge. Given the importance of a trusted information environment to democratic societies, democratic governments should generally be wary of deepfakes, which threaten to undermine that trust. Yet security and intelligence officials in the United States and other democracies will nonetheless have strong incentives to deploy deepfakes against their adversaries, particularly in the context of armed conflict. As a result, the U.S. and its democratic allies should consider developing a code of conduct for deepfake use by governments, drawing on existing international norms and precedents.

Further, the U.S. should also consider establishing something like a "Deepfakes Equities Process," loosely modeled on similar processes for cybersecurity, to determine when the benefits of leveraging deepfake technology against high-profile targets outweighs the risks. By incorporating the viewpoints of stakeholders across a wide range of government offices and agencies, such an inclusive, deliberative process is the best way to ensure deepfakes are used responsibly.

INTRODUCTION

On March 2, 2022, shortly after Russia had launched a full-scale invasion of neighboring Ukraine, a video message showing Ukrainian President Volodymyr Zelenskyy briefly appeared on the news website Ukraine 24.[1] Dressed in his iconic olive shirt, Zelenskyy's tone and attire matched his other messages of the time. Yet the message itself was altogether different: Rather than urging Ukrainians to carry on their fight, Zelenskyy instead implored them to lay down their arms and surrender. Not surprisingly, the video then quickly spread on VKontakte, Telegram, and other social media platforms, where it was picked up and reported on by global media.[2]

Zelenskyy's office immediately disavowed its authenticity, noting that it was exactly the kind of "deepfake" they had warned about before the war. Nonetheless, as the first high-profile use of a deepfake during an armed conflict, the incident marked a turning point in information operations. Deceit and media manipulation have always been a part of wartime communications, but never before has it been possible for nearly any actor in a conflict to generate realistic audio, video, and text of their opponent's political officials and military leaders. As artificial intelligence (AI) grows more sophisticated and the cost of computing continues to drop, the challenge deepfakes pose to online information environments will only grow. Policymakers and government officials will need to develop robust systems for monitoring and authenticating both public and private messages in real time, while also evaluating when — if at all — to leverage the technology themselves.

This is particularly true when it comes to military and intelligence operations. Going forward, militaries and security services will need to assume that rival state and nonstate actors alike will have access to deepfake capabilities that can generate compelling audio and video of any state official, leader, or soldier. As a result, they will need to develop the kind of robust authentication mechanisms and "pre-bunking" strategies that Ukraine has already pioneered. Moreover, they will need to understand how deepfake technology adds further complexity to the communications challenges that militaries and insurgent groups already face. Democratic governments will need to develop strategies for how to operate in such an environment without undermining the integrity of their communications or key values and norms.

This policy brief offers an overview of how deepfake technologies will impact security and intelligence operations. Although public attention has fixated on the use of deepfakes in influence operations and propaganda campaigns, the technology is likely to be used far more widely, including in targeted military and intelligence operations. The potential benefits of targeted deepfakes are significant enough that even some democratic governments are likely to generate and deploy them, despite the risks that deepfakes pose to democratic societies overall. To navigate this complexity, security officials and policymakers need a far greater understanding of how the technology works and the myriad ways it can be used. Ultimately, however, they will also need a consistent process for determining when the benefits of using deepfakes outweigh the risks. The U.S. government should establish a Deepfakes Equities Process to decide when to leverage deepfake technology against high-profile targets — similar to the Vulnerabilities Equities Process established, in part, to decide whether to exploit an adversary's "zero-day" cybersecurity vulnerability, an unknown or unresolved security flaw that the target has zero days to fix.

WHAT IS A DEEPFAKE?

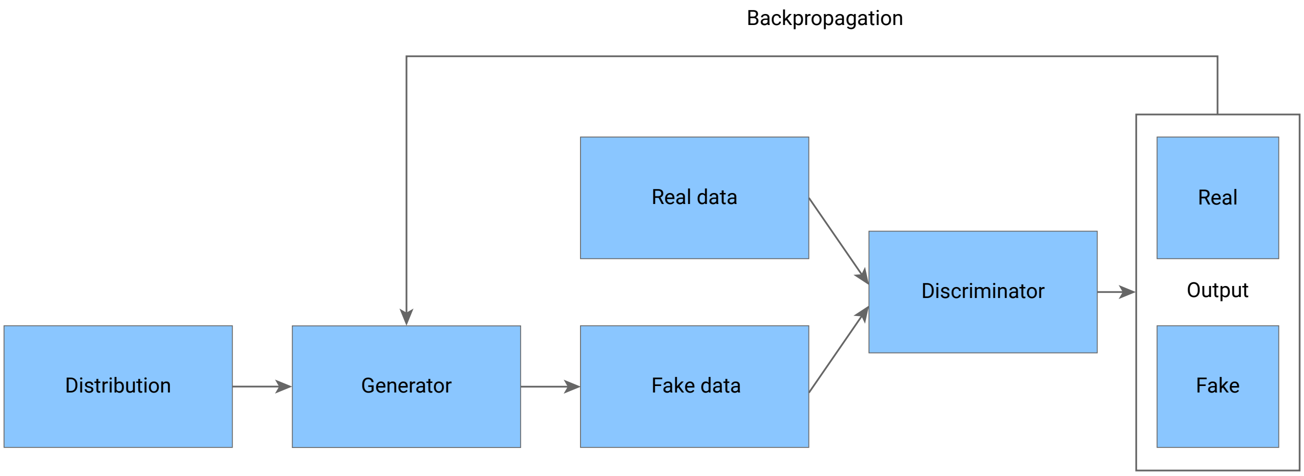
Deepfakes are only possible because of recent breakthroughs in machine learning. Although the field has a long history — the first "machine learning" paper was published in the late 1950s — until recently, its promise far outpaced its performance. Over the past 15 years, however, machine learning has finally fulfilled its initial potential. Driven by the widespread availability of multimodal data (for example, news articles, social media, audio, imagery, and video), as well as the dramatic reduction in costs of high-performance central processing unit (CPU) and graphics processing unit (GPU) computing clusters, machine learning techniques are now ubiquitous. The techniques are used to detect credit card fraud, power autonomous vehicles, and curate social media feeds among many other applications.

At the core of the recent machine learning revolution are deep neural networks, often referred to as "deep learning." With enough data and computing power, deep learning models can be enormously powerful, including for generating realistic images, audio, and text. Indeed, they are so effective that the "deep" in deep learning is what gave rise to the term "deepfake."

To generate deepfake video, the specific deep learning algorithm typically used is called a GAN, short for generative adversarial network. Though there are many variations of GANs, they all use deep neural networks and follow the same simple architecture (see figure 1).

FIGURE 1

Architecture of a GAN



For computer vision, or the field of AI that enables computers to interpret and react to visual images, a GAN consists of two key components: a Generator algorithm that tries to generate a fake image and a Discriminator algorithm that tries to distinguish between real and fake images. Imagine that someone wants to generate synthetic images of the White House using a set of real images. At the onset, the Generator will create an image that is essentially random noise and looks nothing like the White House. The Discriminator algorithm will quickly and easily identify the Generator's image as fake compared to the real images. The Generator now learns that it must generate an image that is different from the previously generated image. Based on this feedback, the Generator algorithm can then generate a better fake in a bid to fool the Discriminator. As the process is repeated, over time, the Generator will generate better and better images and the Discriminator will become worse and worse at separating real images from fake images. The GAN converges when a stable state is reached, such that neither the Discriminator's performance nor the Generator's performance continues to improve over time.

Although GANs were first developed in 2014, they have matured rapidly as computing power, available data, and algorithms have all improved. GANs now exist for any form of digital content, from static images and text to streams of audio and video. Further, the algorithms underlying GANs are often released with open-source licenses, meaning that anyone online can download and train them. The main constraint is not technical expertise, but rather having the necessary training data (which governments can usually gather in a matter of days) and computing power (usually a CPU/GPU cluster) to develop a compelling deepfake of a specific person or target image.

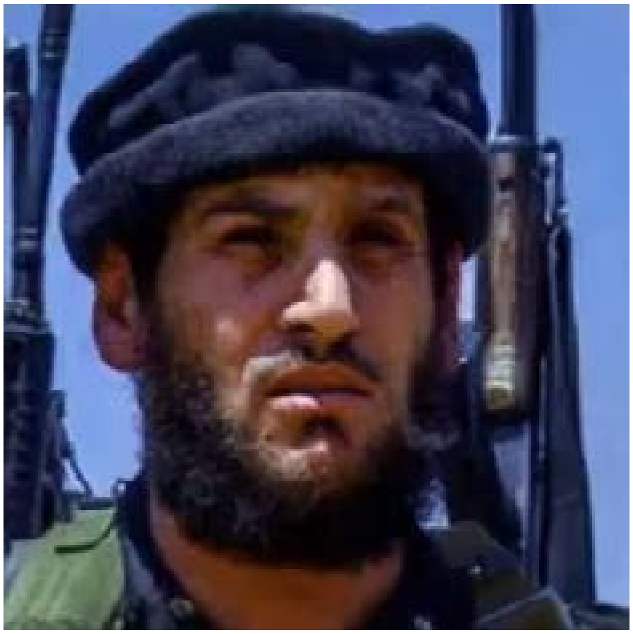
For example, consider TREAD, short for Terrorism Reduction with Artificial intelligence Deepfakes. Developed at Northwestern University, TREAD enables qualified researchers to generate deepfake videos to better understand the technology's future use in conflict and security contexts. For instance, TREAD was used to develop a fake video of Abu Mohammad al-Adnani, a now deceased terrorist of the Islamic State, saying exactly the same thing as Syrian President Bashar al-Assad says in a real video (see figure 2 or see the video here.). TREAD generated the fake through a two-phase process: training and operational use.

In the training phase, a trainer (Assad in figure 2) records a 15-to-20-minute video. The trainer is an individual who puts words in the mouth of a target. For instance, a counter-terrorism officer might want to generate deepfakes of a currently wanted terrorist to discredit them or discourage their followers from using violence. An audio recording of the target and a good headshot are also needed. These two inputs along with the video of the trainer are fed into TREAD, which learns a complex GAN-based model in order to generate a fake video of the target. This training phase can take several hours, but it only needs to be performed once for a given trainer-target combination.

Once the model has been trained, the operational use phase can begin. At this stage, the same trainer records another video that puts words into the mouth of the target. The video could be just a few seconds long or may drag on for hours. The trained model from the training phase uses this input to generate a deepfake video of the target saying the same thing with highly realistic facial and other expressions, as well as highly realistic audio in the target's own voice. The method is independent of the language used and hence usable worldwide.

FIGURE 2

Sample deepfake video of Mohammad al-Adnani saying something that Bashar al-Assad said



Sources for the video: Assad video, https://www.youtube.com/watch?v=bT7GHd2c-Hs; Adnani headshot, https://www.bbc.com/news/world-middle-east-37234207; Adnani audio, https://archive.org/details/MohammadAlAdnaniSpeech

TREAD's deepfake generation capabilities were put together using publicly available, open-source GAN-based code to generate fake audio and fake imagery. TREAD seamlessly combined these off-the-shelf tools to generate the desired deepfakes. Anyone with a reasonable background in machine learning can — with some systematic work and the right hardware (typically a CPU/GPU computing cluster that may cost no more than $10,000) — generate deepfake videos at scale by building models similar to TREAD. The intelligence agencies of virtually any country, which certainly includes U.S. adversaries, can do so with little difficulty.

As deepfakes have become more successful in fooling humans, scientists have started developing techniques for the automated detection of deepfakes. Simple techniques include looking for mistakes made by the deepfake developers (for example, when they forget to delete or reset GPS coordinates recorded in the metadata associated with an image or video). Somewhat more sophisticated techniques include examining the technical properties of the image (for example, the border regions that separate the human face from the image background) and of the camera used (for example, the typical camera signatures that exist in a video). Deepfake detection methods can also include looking for inconsistencies between landmark points in faces — such as the tip of a nose or the center of the eye — in real versus deepfake images and videos and between these landmarks and the spoken audio. Some of these methods work well today.

When a new deepfake detector is deployed, deepfake creators are likely to replace the Discriminator in a GAN (see figure 1) with a version of that deepfake detector and retrain their Generator to evade it. The result will be a cat-and-mouse game similar to that seen with malware: When cybersecurity firms discover a new kind of malware and develop signatures to detect it, malware developers make "tweaks" to evade the detector. The detect-evade-detect-evade cycle plays out over time.

The TREAD research illustrates the importance of deepfake detection methods going forward. The ease with which deepfakes can be developed for specific individuals and targets, as well as their rapid improvement — most recently through a form of AI known as stable diffusion — point toward a world in which all states and nonstate actors will have the capacity to deploy deepfakes in their security and intelligence operations. Security officials and policymakers will need to prepare accordingly.[3]

POSSIBLE USES IN INTERNATIONAL CONFLICT

As machine learning grows more powerful and cost-effective, the challenge that manipulated media poses to democracy will grow more pronounced. At a high level, authoritarian regimes are likely to leverage cheaply produced fake media of democratic presidents and leaders in their influence operations abroad. For example, during the 2016 U.S. presidential election, Russia pushed an array of false content on Facebook, Twitter, and other platforms to undermine Hillary Clinton's presidential campaign and to polarize American society in general. Because of the speed of social media, information — real or false — often spreads quickly, leading to information cascades that, because of the seeming ubiquity of a fake fact, can lead to its broad acceptance. Indeed, efforts to deny misinformation often result in it being spread more widely. Studies have additionally shown that false information spreads faster than true information. Worse, the better and more easily deepfakes can be produced, the easier it will be for any piece of authentic content or information to be dismissed as inauthentic, ultimately making the notion of democratic deliberation untenable. Widespread use of deepfakes can tarnish a democratic society's image and undermine the legitimacy of democratic governance.

To date, most discussions around deepfakes have focused on threats to democracy and society. Yet the easy generation of manipulated media will also exacerbate age-old challenges in conflict environments. Disinformation efforts have long been used to confuse, discredit, and undermine enemies during war. Octavian used poetry and coins to spread the impression that Mark Anthony was a philandering drunk.[4] Irish rebels claimed George II was ill to undermine his image as a strong leader.[5] During the Cold War, the Czechoslovak secret service put out false information claiming that leading Western politicians in West Germany had collaborated with the Nazis. In addition, the Soviet Union's KGB put out, or leaked, false information asserting that U.S. intelligence was linked to the assassination of U.S. President John F. Kennedy, that the Jimmy Carter administration had endorsed South Africa's apartheid government, and that the U.S. government invented the AIDS virus to destroy nonwhites (which was then widely believed). U.S. intelligence disinformation campaigns were less effective, but among other cases, the agency was successful in publishing false editions of communist publications, duplicating their format but enclosing more subversive content.[6]

Today, efforts that use deepfakes will make these kinds of disinformation efforts much more powerful. As the aforementioned fake Zelenskyy video suggests, deepfakes are now being used in international conflicts, and their role is only likely to grow in the coming years.[7] States and nonstate actors, particularly illiberal ones, can use deepfakes for many purposes. The following list of applications is only limited by the creativity of those designing the deepfakes.

Legitimizing war and uprisings. Countries have long used false information and staged outrage as pretexts for war. Before Germany invaded Poland in 1939, German SS officers dressed in Polish uniforms seized a radio station and broadcast a message condemning Germany — a "false flag" to justify a Germany invasion. In 2017, several Gulf states began a confrontation with Qatar after a computer hack led to the spread of fake quotations from Qatar's emir.[8] Today, more realistic audio and video deepfakes could "reveal" plans for an invasion (that must be preempted) or other nefarious intentions.[9] For instance, a deepfake might foster or legitimate an insurgency; Robert Chesney and Danielle Citron give the example of a fake video showing a U.S. general burning a Koran.[10]

Falsifying orders. Both audio and visual content can be created and put in the mouths of commanders. Russia's video of Zelenskyy had him instructing Ukrainian soldiers to lay down their arms and surrender to invading Russian forces. Some false orders might include video of senior leaders telling soldiers to lay down their arms, retreat, launch nonexistent chemical weapons, or otherwise encourage mass surrenders, to dislodge well-defended troops, or to make the forces vulnerable.[11]

Sowing confusion. When civilians and soldiers are instructed to ignore leaders' instructions as potential fakes, they may also inadvertently ignore legitimate orders — creating confusion at a dangerous time. Disinformation operations are most effective when truth and fiction are blended together so that people cannot discern the difference. In Gabon, a possible deepfake of the country's president giving a stilted, expressionless address led his critics to question the leader's ability to rule, and the military attempted a coup.[12] In this case, it was not clear whether the address was a deepfake (the president may simply have been ill), and it might have just been uncertainty that led different actors to draw their own conclusions about whether the president remained fit to rule. This confusion is especially likely when the deepfake reinforces existing cognitive biases that make it hard to believe uncomfortable facts, such as evidence that a popular leader is engaging in bad behavior.[13] Similarly, news organizations may hesitate to report on breaking news, fearing that they may be fooled by a deepfake. Indeed, as Chesney and Citron point out, the possibility of deepfakes creates a "liar's dividend," allowing political leaders to dispute the authenticity of their own genuine misbehavior.[14] In essence, this inadvertent confusion is a mirror image of false orders: instead of false instructions being followed, legitimate ones are discarded.

Sowing confusion can also be done at a tactical level during a conflict. Imagine that a deepfake showed an image of enemy soldiers entering a city or raising a flag in a captured city. This could lead defenders to believe that their positions are not defendable and thus lead them to flee.

Although open democracies are vulnerable to disinformation campaigns, so too are authoritarian regimes, as Henry Farrell, Abraham Newman, and Jeremy Wallace contend.[15] AI algorithms are sometimes based on biased data that reflect a regime's political prejudices. Indeed, the primary disinformation targets of regimes are often their own people. Such regimes lack institutional feedback mechanisms and thus might "find themselves in the throes of an AI-fueled spiral of delusion."[16] Other regimes might try to take advantage of such spirals, feeding additional bad information into the system.

Dividing the ranks. Divided armies usually fight poorly, and as a result, maintaining unit cohesion and esprit de corps is an important part of training and leadership.[17] Adversaries can generate content that shows top political or military leaders voicing racist remarks, expressing disdain for their soldiers and political bosses, laughing at the dead and wounded, or otherwise discrediting them. Such deepfakes may also show military leaders questioning the authority of their superiors or describe an ongoing battle or war as a losing proposition, leading to a lack of will to fight among the rank and file.

Undermining popular support. Armies do not fight alone. They need recruits and financial backing and, perhaps most importantly, high morale; they want to be seen as fighting for something. For insurgents, popular support is also vital to ensure that they have adequate food and safe passage (to evade militarily superior government forces). Deepfakes might show military forces committing human rights abuses, favoring one community over another, fleeing as cowards rather than fighting bravely, looting and stealing from the local community, or otherwise betraying the cause and the people they claim to be defending.

Polarizing societies. Russia has already tried to use fake news to polarize American society, playing up tension over Black Lives Matter rallies and other protests. Deepfakes could add to this tension by showing white police officers shouting racial slurs while gunning down unarmed Black men.[18] Such efforts may increase divisions within a military and lead to decreased confidence in political leaders.

Dividing allies. Allies have different security priorities and domestic political concerns, and false content can play up these differences. For example, during the Cold War, the KGB put out convincing, but false, "leaked" official U.S. reports that called for using nuclear weapons on the territory of members of the North Atlantic Treaty Organization, creating widespread anger.[19] Deepfakes offer a potentially convincing way to show leaders making disdainful, uncaring comments about allies and allied casualties; laughing at a sensitive issue for an ally (for example, the flow of refugees or energy shortages); or behaving in other ways that would fray a bilateral or multilateral relationship.

Discrediting leaders. Deepfakes can be used to discredit leaders. A possible deepfake was used in Myanmar to show a former government minister saying he bribed the country's former leader, Aung San Suu Kyi, whom the military regime sees as an enemy.[20] The manipulated video of a supposedly "drunk" Nancy Pelosi, speaker of the U.S. House of Representatives, could have been done more masterfully via deepfakes.[21] Deepfake videos could show a leader saying racist, callous, or insensitive things; sneering at casualties or political partners; or acting in other ways that offend important countries, peoples, and constituencies.

For now, the greatest danger is from states that have considerable technological capacity (like Russia) or can hire it (like Saudi Arabia). However, as technology improves and becomes more accessible, smaller states, nonstate actors, and even individuals could begin using deepfakes.[22] Adversaries might create deepfakes of U.S. soldiers shooting civilians or of U.S. leaders discussing plans to seize territory, empower religious rivals, or bolster terrorist narratives. Quality will vary, but even less professional efforts may sow confusion or undermine government policies.

Good disinformation campaigns play on the likely reaction of the target. During the Cold War, U.S. intelligence sent falsified letters that were supposedly from the East German trade ministry to several of its customers around the world, noting that East Germany could no longer accept orders due to economic problems and, even more troubling, that the trade was not producing enough propaganda value. The East Germans predictably denounced the letter as a forgery but had to send out an explanation to all its customers, unsure of which ones had received the letter — thereby damaging their prestige broadly and confusing many more customers.[23] A series of deepfakes could be planned, with a denial of one deepfake leading to additional "revelations" that further discredit a leader.

Although the focus above is on how countries might use deepfakes against the United States and its allies, deepfakes might be used against U.S. adversaries and other threats in order to protect the international order. A deepfake might show a Russian general ordering troops to withdraw from a besieged city in Ukraine, allowing vital humanitarian relief to enter the city. Or it might show a terrorist leader making critical statements about rivals, thus splitting an overall movement. The list is long.

Of course, deepfakes will often fail, and even when they are viewed as genuine, they will rarely be magic policy bullets for the United States or for its adversaries. As the Zelenskyy video shows, some deepfakes may be clumsy and easily dismissed. Even better produced ones may have only limited impact: Thomas Rid's work on the history of influence campaigns shows that influence operations are often hit or miss.[24] Also, as audiences' awareness of deepfakes and disinformation in general grows, so too will their skepticism.

Yet that skepticism, in turn, raises its own policy problem. As audiences are warned to doubt even realistic-looking video and audio, they may doubt fake and real communications alike. And that general skepticism — whether in specific instances (soldiers who fear that orders may be a deep fake doubt or delay following legitimate ones) or in broader perceptions instances (for example, when people are not sure a leader's actual speech is true) — can make diplomacy and military operations harder and poison politics in general.

POLICY RESPONSE AND RECOMMENDATIONS

The sheer range of potential use cases of deep fakes poses a daunting challenge to policymakers and security analysts in the United States and other democracies.

On the defensive side, guarding against deep fakes will be far from straightforward. At a technical level, although it's still possible to design and train algorithms today that can identify deep fake images, videos, and texts, in the long-term such an approach is unlikely to work – any advances in the algorithmic detection of deep fakes can be baked into the next generation of algorithms used to generate them. Eventually we may reach an endpoint where detection becomes infeasible, or too computationally intensive to carry out quickly and at scale.

Instead, defending against deep fakes will require robust forms of authenticating and verifying digital content, and greater digital literacy and critical reasoning among the public at large. For the security and intelligence enterprise, it will also require systems capable of assuring the provenance and chain of custody of a given piece of audio, video, or text.

Deep fakes pose a challenge for intelligence analysts, journalists, and others trying to parse the truth in real-time. The appearance of important, seemingly accurate video relevant to a crisis or challenge is hard to ignore, but analysts must be cautious. Single-source information must be treated with even more suspicion. Slowing down and verifying information is even more vital given the likelihood of deception, but less careful analysts will often fill any resulting information void, leading the deep fake to be more widely believed. Similarly, journalists might emulate intelligence products that discuss "confidence levels" with regard to judgments or otherwise make levels of uncertainty clearer.[25]

When real-time operations are subject to deep fakes, different approaches might be used. Information from a separate source, such as verification codes, for example, might be necessary to show an order is legitimate.

By contrast, comparatively little policy attention has been given to when and how democratic officials should use deepfakes themselves. On the one hand, the lack of attention is understandable. Democracies have a vested interest in preserving the integrity of both domestic and global information environments and should be reluctant to adopt tactics that risk undermining public faith and trust in the capacity for shared conceptions of truth. This is especially true in contexts of armed conflict, where basic news and information can be highly contested and democratic governments need to preserve the trust of their populations; a botched manipulation of one news story might discredit government efforts to address hundreds of other pieces of news. For this reason, democratic officials should be reluctant to rely on deepfakes as part of a public information operation. Even if their adversaries are leveraging deepfakes in a computational propaganda campaign, they should refrain from responding in kind; when it comes to deepfake propaganda, fighting fire with fire will burn democratic societies far more because their publics are accustomed to a better information environment and their political systems depend on an informed public. Many of the uses of deepfakes laid out above — especially dividing allies, undermining popular support, and polarizing societies — are not ones that democratic governments and officials should exploit.

Democratic governments should also be wary of creating and deploying deepfakes in intelligence operations despite their advantages. Even without using deepfakes, the use of deception tactics poses significant risks. Consider how the United States and other democracies have used public health programs as cover for their operatives. For example, to verify that Osama bin Laden lived at his compound in Abbottabad, Pakistan, the United States reportedly exploited a local house-by-house vaccination program for polio.[26] Despite the risks that using the program as cover posed to popular trust in future public health campaigns — a risk that has since borne out[27] — the U.S. government nonetheless appeared to have decided that the benefits and importance of confirming the location of a high-value target outweighed the risks. If a democratic government is willing exploit a polio eradication program for a high-value target, they may also be willing to bear other ethical risks associated with deepfakes.

Given that democratic governments will almost certainly consider generating and distributing deepfake content, they should establish robust oversight and accountability mechanisms to govern its generation. One approach would be to develop an international agreement or code of conduct on the use of deepfakes by governments, perhaps under the auspices of the United Nations. States that have already started using deepfakes might be unenthusiastic, even they may want to see some limits on deepfakes that could disrupt industry or health care systems. Fortunately, there are precedents to draw from. One example, though flawed and imperfect, is the 2001 Convention on Cybercrime (also known as the Budapest Convention) established by the Council of Europe; it has protocols for nation states to exchange valuable information in cybercrime investigations.

A useful model for deepfakes is the Vulnerabilities Equities Process, which the U.S. government developed to manage its response to the discovery of zero-day cybersecurity vulnerabilities. Since these vulnerabilities pose perhaps the most serious cyber-related risks, the government needed to develop a case-by-case approach to weigh the benefits and risks of exploiting a given vulnerability and of leaving the vulnerability undisclosed. For instance, a zero-day vulnerability in a Cisco router or a Safari browser could expose sensitive personal data of millions of American citizens and companies. In such cases, the U.S. government may prefer to disclose the vulnerability to the vendor involved so that it can be quickly fixed, rather than leave the vulnerability undisclosed so that it can be exploited in an operation against an adversary. The Vulnerabilities Equities Process aims to ensure that vulnerabilities are leveraged only when there is a compelling reason to do so and when the upsides trump the potential downsides, including a loss of faith and trust in the security of modern hardware and software.

The United States and other democratic governments should consider establishing a Deepfakes Equities Process since there are legitimate concerns about the potential impact of deepfakes on the information environment and, by extension, the diplomatic, political, and military environments. The decision to generate and use deepfakes should not be taken lightly and not without careful consideration of the trade-offs. The use of deepfakes, particularly those designed to attack high-value targets in a conflict setting, will affect a wide range of government offices and agencies. Each stakeholder should have the opportunity to offer input, as needed and as appropriate. Establishing such a broad-based, deliberative process is the best route to ensuring that democratic governments use deepfakes responsibly.

APPENDIX

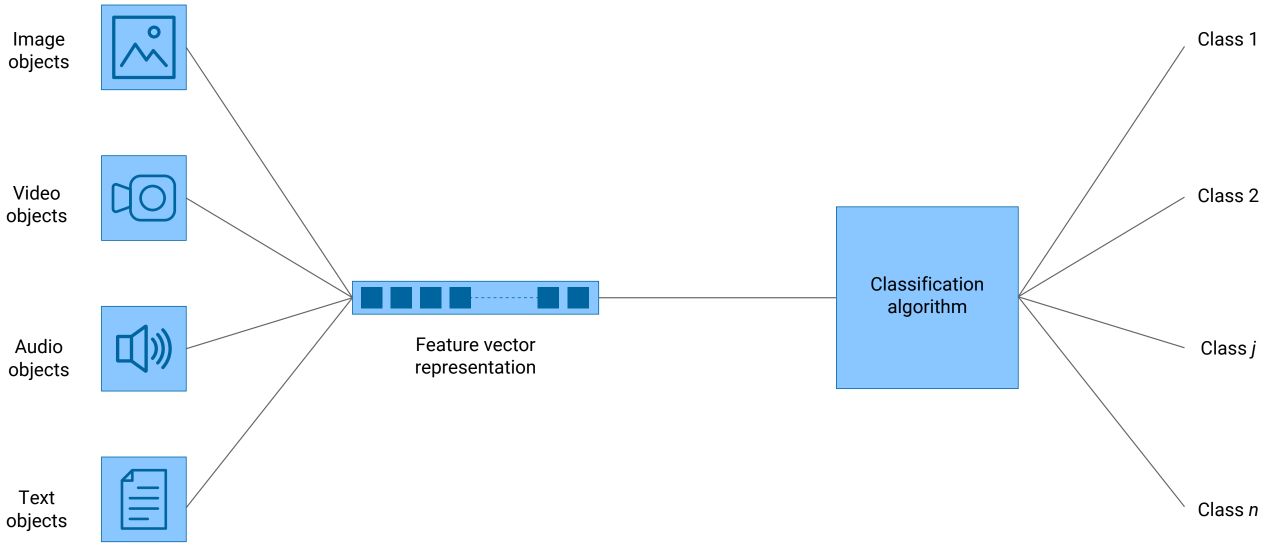
Deepfakes can be produced by a wide array of machine learning algorithms, including generative adversarial networks. But all leading algorithms typically rely on classification algorithms and deep learning.

Classification algorithms

Figure 3 shows a very high-level view of a traditional classification process. For instance, suppose an analyst wants to determine whether a piece of text denotes an honest or fake online review of a product on a website like Amazon. In this case, the input is a "text object" consisting of the text of the review, along with some data about the author, perhaps about other reviews written by the same author, and about the time at which the review was posted. From this text object, a "feature extraction" algorithm extracts some "features" that may or may not be intelligible to humans. In the case of review text, human-understandable features might include the sentiment score of the text denoting how positive or negative the review is on a numeric scale,[28] information about the percentage of shoppers who found the author's past reviews helpful,[29] whether the author was previously reported for any suspicious activity, and more. These kinds of features are combined together to form a feature vector which will form the input to a machine learning "classification algorithm" which will take numerous such feature vectors as input, one for each "text object" and classify it into one of many possible classes. For instance, we might have just 2 classes (let's call them 1 and 2) corresponding to honest reviews versus fake reviews. In general, a classification algorithm may take the feature vector representations of many different objects and classify them into multiple classes (e.g. class 1 for honest reviews, class 2 for reviews that may be honest or not and need to be examined by a human moderator, and class 3 for a fake review). In this case, Figure 3 would classify the feature vectors into n=3 classes number 1, 2, and 3 respectively.

FIGURE 3

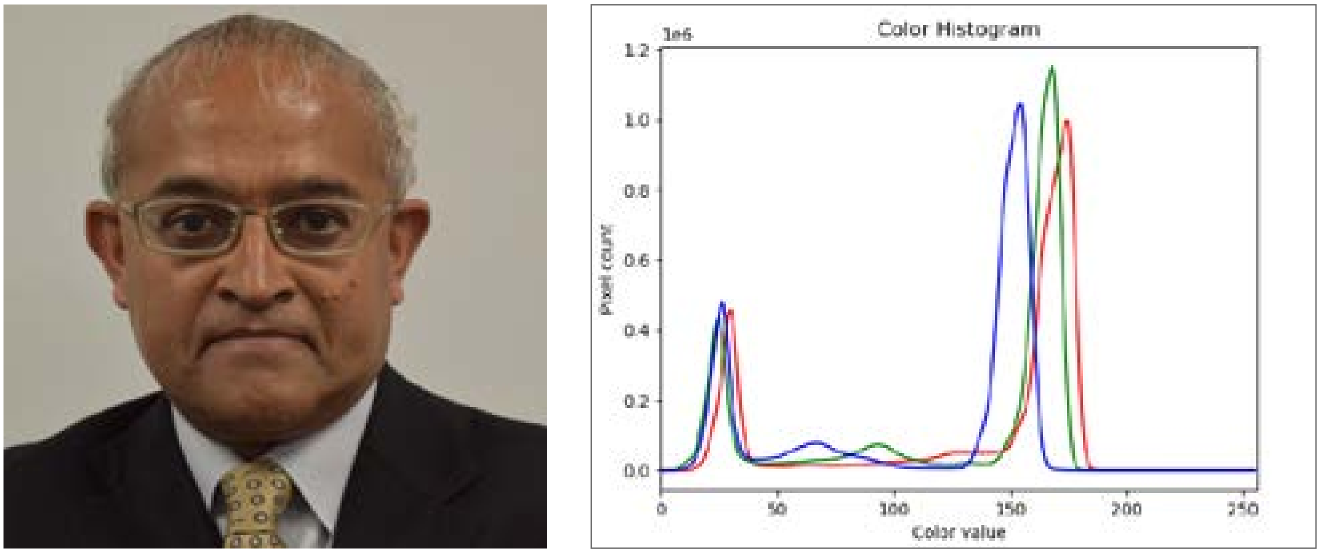
Classification process



A self-driving car application might classify images into two output classes: a "pedestrian present" class (1) or a "no pedestrian present" class (2). In this case, machine learning classifiers may use "low-level" information about the image in order to come up with features automatically that might not make a lot of sense to human users. Such features might include color histograms showing how different pixels' red/green/blue color channels are distributed, i.e. what number or percentage of pixels are red, green, and blue respectively.[30] Figure 4 shows the color histogram associated with an image of author V.S. Subrahmanian. The x-axis of the figure shows the number of pixels in the image that are in a particular shade of red, green, and blue (drawn from a list of 256 shades for each). Such color histograms generate part of the feature vectors associated with images.

FIGURE 4

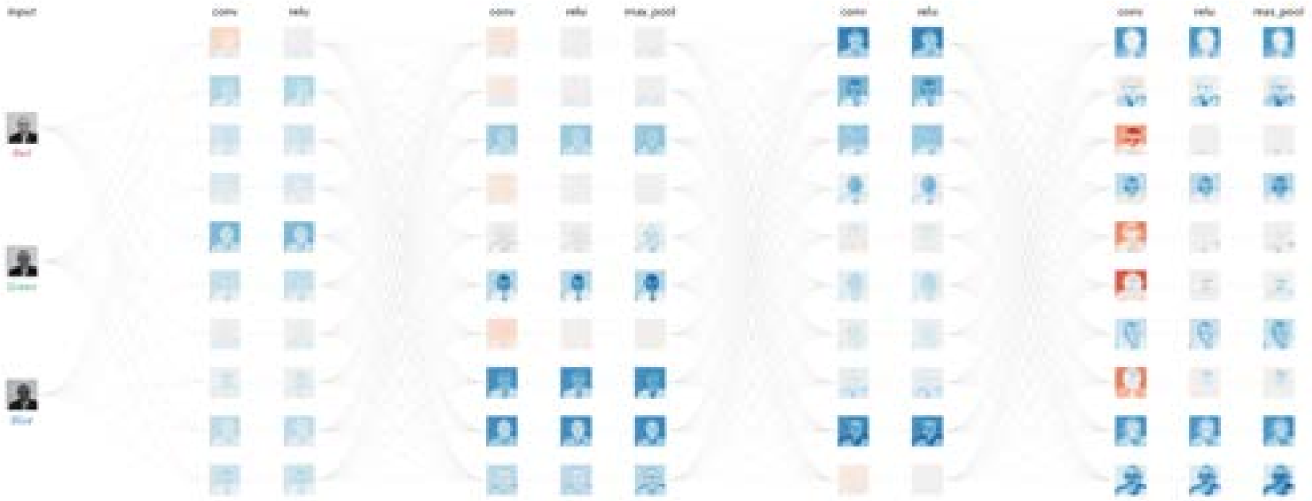
Color histogram associated with an image of author V.S. Subrahmanian



Classification algorithms may also coarsen an image so that small differences between adjacent pixels are harmonized. For instance, figure 5 shows a coarsening of the above image of author V.S. Subrahmanian along each of the red, green, and blue channels at different points during the coarsening using a convolutional neural network (CNN for short). Going from left to right, the image is getting refined into a less coarse representation of the original bin Laden image. This figure was generated using an off-the-shelf CNN (https://poloclub.github.io/cnn-explainer/).

FIGURE 5

Coarsening of the V.S. Subrahmanian image using a convolutional neural network



Source: The above coarsening of the image of V.S. Subrahmanian was generated using the CNN software code at: https://poloclub.github.io/cnn-explainer/

Once these features are created — either through definition by a human or automatically — and the feature vector for a given data object (such as the text object or image object discussed above) is fully assembled, the classifier generates a solution to classify the feature vector and hence the associated object, e.g. to classify a review as honest or fake.

However, prior to operational use, the classification algorithm (or simply classifier) needs to be trained. Training involves feeding the classifier a bunch of data objects for which the class to which the object belongs (e.g. whether a review is real or fake) is known in advance, e.g. through prior investigation or analysis. The training process generates one or more classification rules. When a trained classifier is put to operational use, the learned rule is applied to a new feature vector. If the new feature vector satisfies the rule, the classification algorithm classifies the new feature vector (and its associated data object) into one of the classes being considered. For instance, if we return to the self-driving car example, the classification algorithm may classify a new image captured by the car's camera as belonging to class 1 ("pedestrian present"), otherwise it classifies it as belonging to class 2 ("no pedestrian present"). Good classifiers try to minimize misclassification error (in other words, feature vectors belonging to one class should not be classified as belonging to the other class). Standard metric systems to measure performance of a classifier include F1-Score and Area Under a Receiver Operating

Characteristic Curve (AUC) which is a number between 0 and 1. In both cases, the closer the number is to 1, the better the classifier is performing at predicting the class to which an input object belongs.

Though our discussion has primarily focused on binary classification where the classifier classifies data objects into two classes, most classification algorithms can be used to classify data objects into multiple classes.

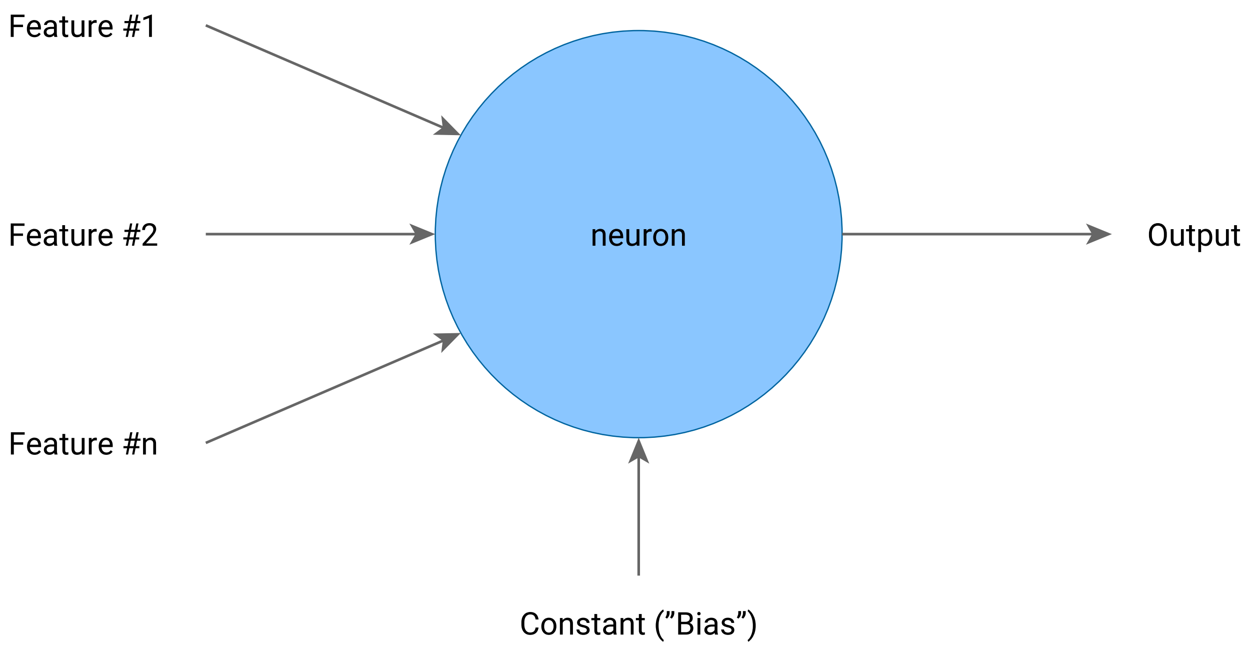
Deep learning

The term "deep learning" usually refers to classification using deep neural networks. A neural network takes some data objects or features (usually numeric) as input, computes a function using those inputs, and then checks to see if the resulting output number exceeds a given threshold or not. If the output number exceeds a threshold, then the neural network says that the input object belongs to class 1; otherwise, it says the input object belongs to class 0.

Figure 6 shows the working of a single neuron (depicted by a circle) in a neural network. The neuron can be fed a feature vector as input (for example, the features associated with a given image). Each input has an associated weight, and in addition, there is another numeric value called a "bias." Assuming that the weights and bias are known in advance, the neuron computes some function and generates a numeric value as output by using the feature values, the feature weights, and the bias. Sometimes, this numeric value is converted to a 0 or a 1 depending on whether the value exceeds or is below a threshold.

FIGURE 6

Working of a single neuron



A deep neural network (DNN) can be thought of as a whole bunch of neurons on steroids. A DNN consists of a set of “layers” that are the vertically arranged neurons shown in figure 7. Initially, all the edge weights are assigned some values (which will later change).

In the “forward propagation” phase of training, the input feature vectors (input layer) are typically fed into each neuron in the first hidden layer as shown in figure 7. Each neuron performs its computation, and the resulting outputs are usually fed as inputs to each neuron in the second hidden layer. This layer does the same computation, and this process repeats for subsequent layers. The last hidden layer (layer m in figure 7) generates one output for each neuron in that layer, and these outputs are merged into a single output value. Suppose this output value is between 0 or 1 (and could be exactly 0 or exactly 1 as well); this final output is then compared with the ground truth associated with that training sample.

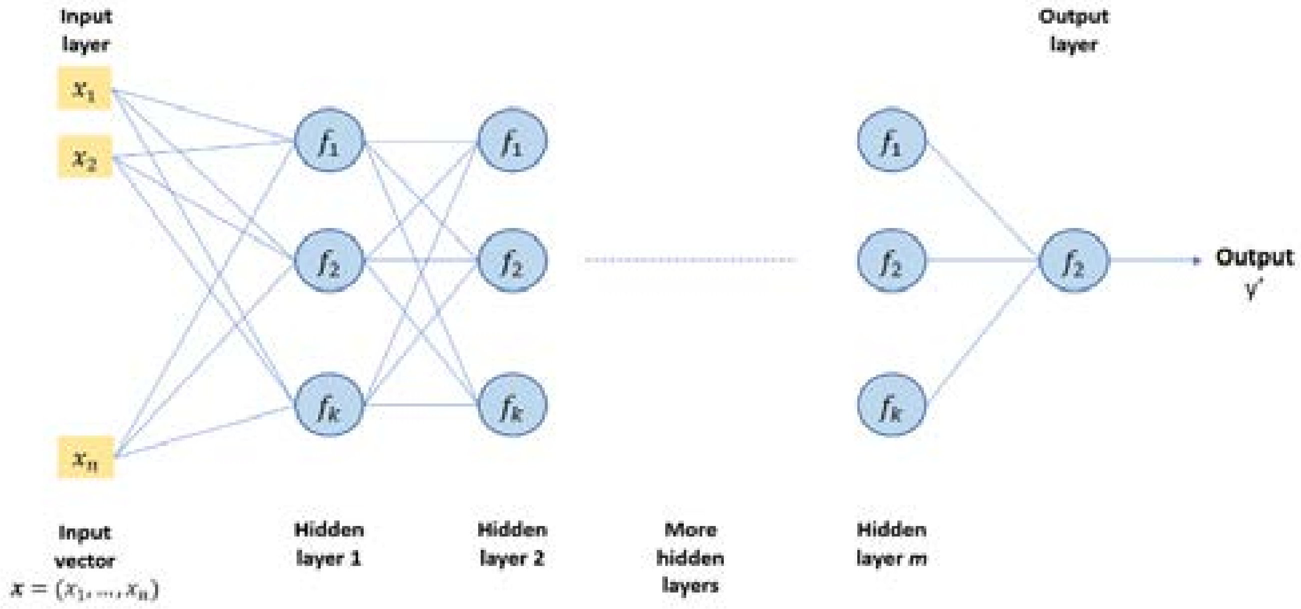
The most important part of a DNN is the “backward propagation” phase. Suppose the ground truth for the most recent sample is a 1; an analyst would then want the output generated in the most recent forward propagation phase above to be close to 1. To achieve this, backward propagation walks backward from the right side of figure 7 to the left side. The weight of each edge is adjusted a very, very tiny bit so that the output of the DNN, had it previously used these weights, would have caused the last value produced as output to be slightly closer to 1 than before. But this has to be done extremely carefully so that good classification results for previously considered training samples are not considered.

This process of forward and backward propagation is repeated for every training sample. At this stage, the training process ends, and all the weights, biases, and/or thresholds in the DNN have been learned.

When put to operational use (in other words, after training), the feature vector of a new sample is fed into the DNN as input and generates an output that is the prediction of the DNN for that feature vector.

FIGURE 7

A deep neural network (weights along edges are not shown)



DNNs are extremely powerful because they offer a huge amount of flexibility. DNNs with lots of layers (for example, 3 versus 30) can be used, and whatever functions are desired can be implemented within a neuron. Different numbers of inputs can be piped into different neurons within hidden layers. All of this allows DNNs to compute a huge range of complex functions. The deeper the networks (in terms of number of layers), the smaller the error (usually) on training samples but the greater the problems with overfitting the network to the training data; and these problems can potentially lead to greater error when the trained DNN is used operationally.

Both GANs and more recent innovations like stable diffusion rely on deep learning and machine learning classification to produce photorealistic images and videos.

REFERENCES

1 The Telegraph, "Deepfake video of Volodymyr Zelensky surrending surfaces on social media," YouTube, March 17, 2022, https://www.youtube.com/watch?v=X17y-rEV5sl4.

2 Tom Simonite, "A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be," *WIRED*, March 17, 2022, https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.

3 For example, the text-to-image breakthroughs that stable diffusion has made possible will almost certainly be modified for video soon, too.

4 Izabella Kaminska, "A lesson in fake news from the info-wars of ancient Rome," *Financial Times*, January 17, 2017, https://www.ft.com/content/aaf2bb08-dca2-11e6-86ac-f253db7791c6.

5 "A brief history of fake news," BBC News, December 4, 2020, https://www.bbc.co.uk/bitesize/articles/zwcgn9q.

6 For an excellent review, see Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare* (New York: Farrar, Straus and Giroux, 2020).

7 Tom Simonite, "A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be."

8 Saif Ahmed Al Thani, "Letters: Quotes were falsely attributed to the emir of Qatar and its foreign minister," *The Guardian*, May 30, 2017, https://www.theguardian.com/world/2017/may/30/quotes-were-falsely-attributed-to-the-emir-of-qatar-and-its-foreign-minister.

9 Robert Chesney and Danielle Citron, "Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?," Lawfare, February 21, 2018, https://www.lawfareblog.com/deepfakes-looming-crisis-national-security-democracy-and-privacy.

10 Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics," *Foreign Affairs*, January/February 2019, https://www.foreignaffairs.com/articles/world/2018-12-11/deepfakes-and-new-disinformation-war?cid=otr-authors-january_february_2019-121118.

11 Greg Allen and Taniel Chan, "Artificial Intelligence and National Security," (Cambridge: Belfer Center for Science and International Affairs, July 2017), https://www.belfercenter.org/publication/artificial-intelligence-and-national-security.

12 Ali Breland, "The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink," *Mother Jones*, March 15, 2019, https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

13 Robert Chesney and Danielle Citron, "Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?"

14 Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics."

15 Henry Farrell, Abraham Newman, and Jeremy Wallace, "Spirals of Delusion," *Foreign Affairs*, September/October 2022, https://www.foreignaffairs.com/world/spirals-delusion-artificial-intelligence-decision-making.

[16] Ibid.

[17] Jason Lyall, *Divided Armies: Inequality and Battlefield Performance in Modern War* (Princeton University Press, 2020).

[18] Robert Chesney and Danielle Citron, "Deepfakes: A Looming Crisis for National Security, Democracy and Privacy?"

[19] Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare*, 120-128.

[20] Tom Simonite, "A Zelensky Deepfake Was Quickly Defeated. The Next One Might Not Be."

[21] "Fact check: 'Drunk' Nancy Pelosi video is manipulated," Reuters, August 3, 2020, https://www.reuters.com/article/uk-factcheck-nancypelosi-manipulated/fact-check-drunk-nancy-pelosi-video-is-manipulated-idUSKCN24Z2BI.

[22] Robert Chesney and Danielle Citron, "Deepfakes and the New Disinformation War: The Coming Age of Post-Truth Geopolitics."

[23] Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare*, 81.

[24] Thomas Rid, *Active Measures: The Secret History of Disinformation and Political Warfare*.

[25] "Intelligence Community Directive 203: Analytic Standards," Office of the Director of National Intelligence, January 2, 2015, https://irp.fas.org/dni/icd/icd-203.pdf.

[26] Saeed Shah, "CIA organised fake vaccination drive to get Osama bin Laden's family DNA," *The Guardian*, July 11, 201, https://www.theguardian.com/world/2011/jul/11/cia-fake-vaccinations-osama-bin-ladens-dna.

[27] See, for example, Monica Martinez-Bravo and Andreas Stegmann, "In Vaccines We Trust? The Effects of the CIA's Vaccine Ruse on Immunization in Pakistan," Research Briefs in Economic Policy, no. 276, CATO Institute, November 10, 2021, https://www.cato.org/sites/cato.org/files/2021-11/RB-276.pdf; and Jackie Northam, "How The CIA's Hunt For Bin Laden Impacted Public Health Campaigns In Pakistan," All Things Considered, NPR News, Islamabad, September 6, 2021, https://www.npr.org/2021/09/06/1034631928/the-cias-hunt-for-bin-laden-has-had-lasting-repercussions-for-ngos-in-pakistan.

[28] V.S. Subrahmanian and Diego Reforgiato, "AVA: Adjective-Verb-Adverb Combinations for Sentiment Analysis," *IEEE Intelligent Systems* 23, no. 4 (2008): 43-50, https://ieeexplore.ieee.org/document/4580544.

[29] Jiahua Du, Jia Rong, Sandra Michalska, Hua Wang, and Yanchun Zhang, "Feature selection for helpfulness prediction of online product reviews: An empirical study," *PloS One* 14, no. 12 (2019): e0226902, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0226902.

[30] Mark Grundland and Neil A. Dodgson, "Color histogram specification by histogram warping," *SPIE Proceedings* 5667 (2005), https://www.spiedigitallibrary.org/conference-proceedings-of-spie/5667/1/Color-histogram-specification-by-histogram-warping/10.1117/12.596953.short?SSO=1 .

ABOUT THE AUTHORS

Daniel L. Byman is a professor at Georgetown University and a senior fellow at the Brookings Institution.

Chongyang Gao is a Ph.D. student in computer science at Northwestern University's McCormick School of Engineering and a research assistant at Northwestern's Roberta Buffett Institute for Global Affairs.

Chris Meserole is a fellow in Foreigsn Policy at the Brookings Institution and director of research for the Brookings Artificial Intelligence and Emerging Technology Initiative.

V.S. Subrahmanian is the Walter P. Murphy Professor of Computer Science at Northwestern University's McCormick School of Engineering and a faculty fellow at the Northwestern Roberta Buffett Institute for Global Affairs. He also heads the Northwestern Security and AI Lab.

ACKNOWLEDGMENTS

The authors would like to thank Derek Belle for his assistance, Lori Merritt for editing, and Rachel Slattery for layout.

DISCLAIMER

The Brookings Institution is a nonprofit organization devoted to independent research and policy solutions. Its mission is to conduct high-quality, independent research and, based on that research, to provide innovative, practical recommendations for policymakers and the public. The conclusions and recommendations of any Brookings publication are solely those of its author(s), and do not reflect the views of the Institution, its management, or its other scholars.

BROOKINGS

The Brookings Institution
1775 Massachusetts Ave., NW
Washington, D.C. 20036
brookings.edu