MARCH 2024

Government Use of Deepfakes

The Questions to Ask

AUTHORS
Daniel Byman
Daniel W. Linna Jr.
V. S. Subrahmanian

A Report of the CSIS Transnational Threats Project

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

MARCH 2024

Government Use of Deepfakes

The Questions to Ask

AUTHORS
Daniel Byman
Daniel W. Linna Jr.
V. S. Subrahmanian

A Report of the CSIS Transnational Threats Project

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

About CSIS

The Center for Strategic and International Studies (CSIS) is a bipartisan, nonprofit policy research organization dedicated to advancing practical ideas to address the world's greatest challenges.

Thomas J. Pritzker was named chairman of the CSIS Board of Trustees in 2015, succeeding former U.S. senator Sam Nunn (D-GA). Founded in 1962, CSIS is led by John J. Hamre, who has served as president and chief executive officer since 2000.

CSIS's purpose is to define the future of national security. We are guided by a distinct set of values—nonpartisanship, independent thought, innovative thinking, cross-disciplinary scholarship, integrity and professionalism, and talent development. CSIS's values work in concert toward the goal of making real-world impact.

CSIS scholars bring their policy expertise, judgment, and robust networks to their research, analysis, and recommendations. We organize conferences, publish, lecture, and make media appearances that aim to increase the knowledge, awareness, and salience of policy issues with relevant stakeholders and the interested public.

CSIS has impact when our research helps to inform the decisionmaking of key policymakers and the thinking of key influencers. We work toward a vision of a safer and more prosperous world.

CSIS does not take specific policy positions; accordingly, all views expressed herein should be understood to be solely those of the author(s).

Center for Strategic & International Studies
1616 Rhode Island Avenue, NW
Washington, DC 20036
202-887-0200 | www.csis.org

Acknowledgments

The authors would like to thank Di Cooke and Ines Oulamine for their comments on previous versions of this article. The authors also thank Sihan Feng, Northwestern Pritzker School of Law student, for conducting legal research for this article and helping with legal analysis of the hypotheticals.

This report was made possible by general funding to CSIS and the Buffett Institute for Global Affairs, the McCormick School of Engineering, and the Pritzker School of Law at Northwestern University.. No direct sponsorship contributed to this report.

Contents

Introduction

In March 2022, a fake video of Ukrainian president Volodymyr Zelensky telling his soldiers to lay down their arms was posted on a Ukrainian website. The video also appeared on Facebook, Twitter, YouTube, and a host of Russian channels.[1] A similar video shared on the social platform Telegram surfaced in November 2023 in which the head of Ukraine's armed forces, General Valery Zaluzhny, is heard making a similar speech.[2] Other fake videos of Zelensky show him sporting a swastika and appearing in a Pride parade.[3] Widely believed to have been created and disseminated by Russia, these fakes reflect the use of advanced information technology methods to covertly influence warfare and shape the broader information environment in Russia's favor.

This state of affairs is only likely to worsen with the rapid proliferation of deepfakes, or artificial images generated by artificial intelligence (AI), as well as manipulable digital media in general. As AI has improved, deepfakes have gone from primitive to highly realistic, and they will only get harder to distinguish. (A representative list of some popular tools for generating synthetic artifacts is provided in the appendix.)

This proliferation of AI provides an unparalleled opportunity for state actors to use deepfakes for national security purposes. In addition to the Zelensky and Zaluzhny videos, there have been numerous uses of deepfake technology in the context of politics and international conflict. In May 2023, deepfake videos claimed to show the Kurdistan Workers' Party (PKK) terror group endorsing Kemal Kilicdaroglu, the main opposition candidate for the Turkish presidency. Another deepfake against an opposition candidate, this time allegedly pornographic, led him to step down from the presidential race.[4] In January 2024, a robocall that faked President Joe Biden's voice advised New

Hampshire residents not to vote in the presidential primary.[5] There are reports that the Venezuelan government used deepfake technology (specifically the Synthesia tool shown in the appendix) to generate deepfakes of news anchors portraying the Venezuelan economy in glowing terms.[6]

The Synthesia product was also reportedly used to generate videos supportive of the 2022 coup in Burkina Faso.[7] Initial versions allowed users to use their own visual likeness or an avatar to generate a video in the voice of the individual portrayed. Recent updates can convert "text-based sources [inputs] into full-fledged synthetic videos in a matter of minutes."[8] However, unscrupulous individuals may use similar technology to generate highly realistic videos of nonconsenting individuals. It is therefore only a matter of time before politicians, military leaders, and others are falsely accused, through weaponized audio and video, of participating in acts they never carried out. These deepfakes may include made-up speeches, fake murders, and other provocative acts.[9]

On the positive side, deepfakes have been used for various legitimate purposes, labeled here as "beneficial deepfakes." One example is the set of deepfakes of soccer superstar David Beckham, who was portrayed supporting an antimalaria message in which he speaks in a host of foreign languages; the visual content is real, but the audio is not. These messages, created by Synthesia, were presumably generated with his consent.[10] An Indian politician had previously used a similar idea to create deepfake videos of his own political speeches in 20 languages using his own voice, enabling him to share his political message more effectively.[11]

Will the lure of deepfakes, whether beneficial or malign, prove irresistible to democratic governments? If the past is a guide, the answer seems to be a resounding yes. Photoshop and similar tools have long allowed the easy editing of images, and deepfakes bring this phenomenon to a new level.[12] September 2022 saw reports of the U.S. military using fake Twitter, Facebook, and Instagram accounts to promote a pro-Western message in Central Asia and the Middle East.[13] In February 2023, leaked documents revealed the U.S. military may be actively seeking to run covert overseas influence campaigns that use deepfakes to "generate messages and influence operations via non-traditional channels."[14]

Will the lure of deepfakes, whether beneficial or malign, prove irresistible to democratic governments? If the past is a guide, the answer seems to be a resounding yes.

It will not be long before major democracies, including the United States, start or at least consider using deepfakes to achieve their ends, if they have not already done so. This paper examines hypothetical cases in which deepfakes might be used and argues that deepfakes should not be used without a clearly articulated set of guardrails that consider both the benefits and the risks of a proposed government-run deepfake-enabled operation.

There is need for an unbiased evaluation of the trade-off between the short-term benefits and long-term risks of a hypothetical deepfake campaign by a democratic nation. Moreover, a transparent

and publicly articulated process is required for performing such evaluations, even if the details of a specific deepfake or deepfake campaign are classified. To these ends, this paper focuses on answering two fundamental questions:

- **What** are the right questions to ask when a government agency contemplates using deepfakes to further its mission?
- **Who** should be responsible for asking these questions and who should approve or reject a request to use a deepfake or perform a deepfake campaign? What processes and governance mechanisms should these people use when contemplating deepfake use in military and intelligence settings?

To answer these questions, this paper presents five hypothetical security-related scenarios a democratic government might confront. The authors conducted interviews with six leaders who have expert knowledge in both AI technology and national defense. These include a retired U.S. general, a retired U.S. intelligence agency chief, a leading EU think tank employee working in disinformation and technology, a senior tech company executive, a retired general from Asia, and a former senior lawyer at the White House and National Security Council (NSC). The authors engaged in a discussion with each of these experts on the what and who questions listed above. This report summarizes the authors' methods and findings.

This paper argues democratic governments should consider several issues before they deploy deepfakes. First, how effective is the deepfake likely to be, especially if other methods are available? Second, will the deepfake be visible to a narrow audience (e.g., just the inner circle of a foreign terrorist organization) or a broad one (such as an entire country or the citizens in the government's own country)? Third, could the deepfake harm innocent civilians, distorting their views on important issues? Fourth, does the proposed deepfake use comply with applicable international law? Fifth, is the deepfake portraying a truly prominent and influential person, such as a president or religious or social leader, or is it portraying a less important person? Sixth, is the deepfake part of a tit-for-tat process or to protect a country's people from immediate harm, or is there a lesser goal? Finally, how likely is it that a deepfake will be traced back to the home country, with potentially profound repercussions for a government's relationship with its own people? To answer these questions, the authors recommend a deepfake equities process that brings together a range of stakeholders to determine whether the deepfake should be used, with the initial presumption being that, in general, deepfakes should not be used.

Potential Risks for Democratic Governments

A previous report by two of the authors on this topic details the potential uses of deepfakes in conflict settings as well as the costs and risks.[15] Several risks are worth noting for democratic governments in particular. The loss of trust that the population of a democratic nation has in both offline and online media and news sources will be the first casualty, making it easier for adversaries to sow disinformation that divides a target population and for antigovernment domestic actors to use the deepfake to degrade public trust in the government. This loss of trust, in turn, may allow such adversaries to influence elections, reduce support for the target nation's military and

diplomatic activities, encourage riots and internal conflict, sway stock markets, and more. As democracies depend on open communication and informed voters, anything that interferes with these concepts—such as deepfakes—should be viewed with concern.

As democracies depend on open communication and informed voters, anything that interferes with these concepts—such as deepfakes—should be viewed with concern.

A second major long-term risk of the use of deepfakes is the loss of credibility of the government that uses them. Once a government can be credibly accused of using lies as an instrument of state policy, the value of any statements they make could be diminished for years, if not decades, at home as well as abroad.

Third, adversaries may benefit from the "liar's dividend," wherein they can explain away real evidence of corruption or abuse by claiming the information is fake.[16]

Methodology

The authors created five scenarios in which a democratic government may be tempted to intervene. These five scenarios were shared with six leaders in the national security sector from three regions: the United States, the European Union, and Asia. The leaders represented four types of stakeholders: intelligence, military, industry, and nongovernmental organizations with an interest in the topic.[17]

Election Scenario

Country X is rich in oil, minerals, and other natural resources and has a government that is not friendly to Western countries. The leader of Country X is an unsavory character responsible for human rights abuses and linked to international terrorist groups. In the aftermath of recent elections, Country X is experiencing political instability, with many protests and moderate and pro-Western leaders credibly contending the election was stolen from them. In the midst of ongoing instability, intelligence agencies propose to use deepfakes showing the leader of Country X thanking a Western diplomat for sending him money, receiving a payoff from a wealthy national of Country X, laughing at the deaths of citizens of Country X in protests, and ordering ballot boxes to be stuffed with fraudulent ballots. The intelligence agencies hope this will discredit him and contribute to his downfall.

This scenario is intended to address a common foreign policy challenge: an adversarial government that is also repressive to its own people. Such governments often threaten U.S. and allied interests and harm their own people, though rarely do they pose a grave threat.

Genocide Scenario

Country Z is run by a small cabal whose leader (L1) is planning a genocide of a minority group that constitutes about 15 percent of the population. Unbeknownst to L1, his intelligence chief (L2) has had discussions with Western governments about the future of the country, including turning against his boss, possibly to save himself or to gain power. He wants your government's help overthrowing the leader, preventing a genocide, and becoming the leader of a government that will be better and more closely aligned with Western principles. The time has come to overthrow L1. To do so, your country's intelligence agency has generated deepfake videos of L1 boasting about the amount of money he has stolen from the treasury and stating that most members of the majority group are stupid and will have to follow him even though the minority group poses no threat. Western intelligence agencies hope that this deepfake will reduce L1's support among the population of Country Z and enable L2 to gain enough momentum to take over and avert genocide.

This scenario is designed to test responses in one of the worst situations imaginable: the deliberate mass killing of civilians. Historically, the United States and other Western governments have not intervened effectively in such situations.[18] Deepfakes offer one means of intervention that leaders might turn to in such a crisis.

Invasion Scenario

Country R is mobilizing its military forces and preparing to invade a small neighbor. Western intelligence learns Country R is preparing deepfake videos that show that its neighbor's troops fired first at Country R's forces and that the small state's leaders were planning to oppress ethnic citizens living in Country R. Your government considers releasing a fake video of Country R's leader boasting about creating fake videos and how easily he can fool his own people into believing what he wants. The hope is that the fake video will preemptively discredit Country R's fake videos and decrease the popularity of Country R's leader.

In this situation, the deepfake is being used in response to other fake videos—an attempt to fight fire with fire—in a situation involving potential war and human rights abuses.

Stock Market Scenario

A criminal hacker group has released a deepfake video of a corporate billionaire saying his company's latest research has not panned out and that such efforts are not likely to come to fruition for the next 5–10 years. The company stock crashes, along with related stocks, and the hackers reap huge rewards. The billionaire is politically connected, and he reaches out to his government (Western Country X), demanding it help prevent this foreign attack from destabilizing the market and ensure other companies (especially his) are safe. Desperate to stabilize the market and prevent contagion, Country X decides to release a deepfake video of a known critic of the company saying the report is false and the technology is sound.

To determine if government use of deepfakes is appropriate outside a traditional security context, this scenario explores a different situation: a private citizen and a company seeking assistance. In addition, the information in this deepfake is at least partially true.

Intellectual Property Theft Scenario

Company C is a world leader in the pharmaceutical industry. It invests massive amounts of money in the design of new drugs for a variety of diseases. Company C has been previously hacked and is fearful the designs for a new blockbuster drug will be stolen by adverse nation-states. It designs deepfake materials (text documents, images, videos, PowerPoints) that are close enough to the real design to appear highly credible, even to experts, but are flawed enough that the designs do not work. The idea is that an adversary who steals a fake design will execute the design and waste an immense amount of time and money developing a product that does not work.

This last scenario explores another commercial use of deepfakes. In this case, the deepfake is defensive and may never be seen by anyone other than its creators.

Questions to Ask If and When Governments Contemplate Using Deepfakes

All the experts interviewed expressed deep concerns about governments using deepfakes. They were united in their view that deepfakes should be used only in limited settings and under authorization from appropriate authorities. However, there were differences in their opinions about when deepfakes should be used, which scenarios were appropriate, how they should be used, and who should authorize their use.

Drawing on the interviews, Figure 1 summarizes the key questions that should be asked when a government contemplates the use of a deepfake. These questions address (1) the likely efficacy of the deepfake, (2) its audience, (3) the potential harms, (4) the legal implications, (5) the nature of the target, (6) the goal of the deepfake, and (7) the traceability of the deepfake back to the originating democratic government.

1. **Efficacy**

What is the intended purpose of the deepfake? Will using the deepfake achieve the desired purpose? Are there other ways to achieve the intended goal that do not involve deception?

Deepfakes may fail to achieve their goals or may make only a marginal difference. As an example, consider the election scenario. A senior intelligence official might believe that generating a deepfake of the unsavory leader will lead to his downfall. However, the reality might not be so simple. The leader may (correctly) argue the video is a deepfake created to tarnish his image. He might additionally invoke the specter of a U.S. or other foreign takeover of the country and argue that the deepfake is an attempt to grab control of the country's mineral wealth, creating a nationalistic backlash against the United States in favor of the dictator.

Figure 1. Deepfake Use by Governments: The Questions to Consider

Source: Authors' analysis.

Or consider the genocide scenario, where the leader of Country Z is planning to kill many members of the country's ethnic minority. In the long run, it may be wiser to disclose true evidence of the leader's plans, as well as evidence pointing to prior acts of ethnic cleansing he might have carried out at a smaller scale. An influence campaign that puts out honest information, such as via the country's cell phone network, might be less easily refutable by the leader.

However, there may be reasons to use a deepfake to make accurate but inaccessible acts more apparent. In the context of this scenario, a retired non-U.S. military leader argued,

> If the deepfake is replicating something where it is difficult to get videos of actual facts and you make a video that is based on the truth to avoid something worse from happening, in those cases I would say it is justified as long as it is based on the truth.

The interviewee added,

> Anything based on falsity is not likely to be effective. If you base your narratives on the truth and use deepfakes to protect the truth, it will be effective. Otherwise, it will eventually be exposed and be ineffective and lead to loss of credibility.

Thus, at least during times of extreme danger, interviewees expressed some support for the use of deepfakes when they are merely representations of the truth, even if the video is itself fake. But also consider that how to represent "the truth" may not be so straightforward, particularly in contentious situations.

In the intellectual property (IP) theft scenario, the deepfake is intended to have an effect only after a criminal act is performed (i.e., the theft of the IP). Only after the criminal act happens does the deepfake impose costs on the IP thief. Such costs include inducing delays in the IP thieves' projects, inducing additional uncertainty in the thieves' minds about whether the stolen material is real or fake, and frustrating adversaries. Also, there is less risk of the deepfake reaching a broader audience that then accepts it as true. All but one interviewee was enthusiastic about the use of deepfakes in this scenario.

In the case of the stock market scenario, a former senior White House official stated the critic portrayed in the deepfake would "immediately jump up and deny" they ever made the statement. Such a quick and emphatic denial might blunt any desired outcome of the deepfake and could cause volatility that makes the situation worse, further undermining the credibility of those who endorse the deepfake.

2. Audience

Who is the target audience for the deepfake? Is it a domestic audience? A highly focused audience? Or is the deepfake expected to reach a large number of people?

A key question when contemplating the use of a deepfake is whether it can be disseminated so that only a small, focused group of people end up seeing it. Multiple experts pointed out that once a deepfake is in the public domain, it may spread rapidly and uncontrollably. One retired U.S. military officer said, "The idea that you will produce some deepfake material and then it will only play in the specific context for the specific audience is really difficult. You can't control it."

However, for example, if an intelligence agency were to generate a deepfake of two senior Russian security officials, such as Foreign Intelligence Service (SVR) chief Sergey Naryshkin and General Valery Gerasimov, discussing something potentially treasonous and then strategically leak it to Russian president Vladimir Putin, the audience for such a deepfake would be very small. The deepfake would not deceive the entire Russian population or other large audiences. In this regard, a former senior White House attorney said, "Deepfakes should not leak into the bloodstream of the internet. . . . [If there are going to be] lots of eyeballs on a deepfake, don't go there."

A key question when contemplating the use of a deepfake is whether it can be disseminated so that only a small, focused group of people end up seeing it.

Nevertheless, it is difficult to anticipate the size of the audience affected by a deepfake. In the above example, only a few people might initially view the deepfake, but if they act on this "intelligence," believing it to be true, they might take actions that affect millions of people. One interviewee expressed concern that even if such use of a deepfake were successful, the truth might leak months or years into the future. For example, a prominent news outlet such as the *New York Times* might write a piece titled "How the Biden Administration Fooled Putin" six or twelve months after the

deepfake was used. Such a story might severely compromise real evidence the United States offers in future national security situations. Given that clandestine U.S. operations and collection methods regularly leak, it is plausible, even likely, that reports of deepfake use would leak too. In addition, as deepfake detection technology improves, a deepfake used today might be uncovered in the future.

All interviewees agreed that governments should not try to use deepfakes to influence their own population. As a retired non-U.S. official at the level of general stated, "If we are using cognitive means against anybody, you should not misinfluence your own population."

However, there was some support for using deepfakes generated in a foreign language, making it unlikely the majority of the U.S. population would understand the content of the deepfake, though the use of subtitles or automated translation tools could easily change this. Further, the images themselves might be easily understood even without audio translation. One corporate leader with deep expertise in AI said, "Is there a risk to the domestic population if a deepfake used in a foreign country uses a foreign language? Yes, and [it] still needs guardrails if it gets to the U.S."

If a government deploys a deepfake overseas, even in a foreign language, is there still a risk it will affect the domestic population? If a deepfake directed at an overseas population emphasizes rising crime in that country or other social problems, it might influence travel blogs, social media commentary, or other information that would affect whether Americans travel to a part of the world. It might even influence U.S. immigration policy and attitudes toward the country in question.

Another important question to consider is, Who are the possible beneficiaries of the deepfake's use? The country's citizens as a whole? A specific individual, company, or industrial sector? In the case of the stock market scenario, the direct beneficiaries are limited. All interview subjects were united in the view that the government should not be in the business of helping individuals or private corporations foster lies to the population, even if the request for help comes due to an attack by a foreign state. In the stock market scenario, the billionaire could, as a citizen, ask the Federal Bureau of Investigation (FBI) (or the equivalent for other democratic governments) for help in investigating the origins and perpetrators of the deepfake that targeted him rather than ask the government to run deepfakes to correct the market, which would deceive the American people and others.

3. Harms

Who might be harmed by the deepfake? What is the probability the deepfake will harm innocent civilians?

All subjects interviewed were deeply concerned a deepfake might harm innocent civilians. According to one interview subject, "There should be no unintended harm to civilians."

In the election, genocide, and invasion scenarios, the goal of the deepfake is to destabilize an abhorrent leader. The main people likely to be directly harmed are the despicable leader and the leader's close supporters. However, if the deepfake leads to civil unrest (e.g., if people rise up against the leader), then protesters may be injured or killed. Then the leader might simply be replaced by another leader with similar or more extreme goals. The use of the deepfake might lead to the deaths of innocent civilians but not avert the threat.

A shared goal in the election, genocide, and invasion scenarios is to discredit or destabilize the leader. But what if the leader escalates the scenario out of fear, paranoia, perceived self-defense, or something else and cracks down on the population, a neighboring country, or the country perpetrating the deepfake? It is difficult to predict how an adversary would respond to the deepfake, and some responses could foster tremendous harm.

Additionally, in the case of the IP theft scenario, harms may occur to innocent people. For instance, suppose a Chinese entity steals the design documents for a hypersonic missile from a U.S. company, and suppose that company had generated 99 fake versions of that design document by using a large language model or a system specifically focused on generating believable fake technical documents to deter IP theft.[19] Further, suppose the entity responsible for the theft has analyzed all 100 versions of the document (1 real, 99 fake) and decided a specific version (v) is the real one. The assessment that v is real is incorrect (i.e., the deception by the victim company was successful). The IP thief executes on the design, but during testing, the missile blows up, killing six employees of the Chinese entity. In this case, the victim of the IP theft might argue it is not responsible as the deadly outcome was a direct consequence of a criminal act perpetrated by the Chinese entity and, hence, the Chinese entity should be held responsible for its misdeeds. All but one of the interviewees accepted this argument.

4. Legal

Will the contemplated use of a deepfake under the circumstances violate international law?

International law has a role to play when states use deepfakes in their interactions with other states. But how international law regulates cyberspace generally, much less deepfakes specifically, lacks clarity. States, experts, and scholars have debated for more than a decade how international law applies to activities in cyberspace. While areas of agreement have emerged, several published positions reveal the many divisions and uncertainties.

Determining how international law applies to deepfakes raises new issues beyond those that scholars focus on when analyzing cyberattacks. Peter B. M. J. Pijpers, when analyzing influence operations, describes cyberspace as consisting of three dimensions: physical, virtual, and cognitive.[20] International law, like most law, focuses on the physical world. For example, if foreign actors enter another state and surreptitiously destroy election ballots before they are counted, this would be a clear violation of the state's sovereignty and the principle of nonintervention. With the advent of the internet, however, a state can affect activities in another state through a virtual dimension without entering another state's physical world. These virtual activities can result in direct physical impacts—for example, hacking that causes the destruction of computers and other physical devices. These virtual activities can also cause virtual damage, such as manipulating voting tallies on a machine or in a database, which might be considered analogous to physical effects when analyzed under traditional international law.

Deepfakes, which are deployed in the virtual dimension but are focused on having an impact in the cognitive dimension of cyberspace, present new challenges for international law. Given the recent emergence of deepfakes, it is not surprising that the body of literature on this topic is small.

Nevertheless, for the purposes of this paper, the authors can outline at a high level the principles and rules most likely to apply and that states should consider when deepfakes are used in the national security context.

Deepfakes, which are deployed in the virtual dimension but are focused on having an impact in the cognitive dimension of cyberspace, present new challenges for international law.

Before analyzing the hypotheticals presented, it is important to recognize that a state's use of deepfakes will involve far more than simply creating the deepfake. For example, states will have specific goals, and consideration of the context will play an important role in the legal analysis. States will also decide how to deploy the deepfake or campaign of deepfakes using specific social media channels or other methods of publishing and disseminating content. States might hack social media accounts, news organizations, or other organizations, including in the targeted state, to disseminate content. Or states might deploy individuals in the targeted state to carry out these or related activities. These examples illustrate some of the context that should be considered when a state uses deepfakes. Considering the context will greatly inform the analysis of whether the activities related to a state's use of deepfakes complies with applicable international law.

To begin, it will be useful for states to consider whether the use of deepfakes involves use of force, the threat of force, intervention in the domestic affairs of another state, or violation of the sovereignty of another state.[21] When engaged in armed conflict, states cannot engage in perfidy, which applies to "acts inviting the confidence of an adversary to lead him to believe that he is entitled to, or is obliged to accord, protection under the rules of international law applicable in armed conflict, with intent to betray that confidence."[22] Michael N. Schmitt in the *Tallinn Manual 2.0*, a compilation of expert views on the application of international law to cyber operations, provides an example of perfidy: "Consider the case of a perfidious email inviting the enemy to a meeting with a representative of the International Committee of the Red Cross, but which is actually intended to lead enemy forces into an ambush."[23]

Contrast perfidy with a "ruse" of war, which is not prohibited. According to the Protocol Additional to the Geneva Conventions (Protocol I), ruses are acts that are "intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary with respect to protection under that law."[24] Protocol I provides traditional examples of ruses such as "use of camouflage, decoys, mock operations and misinformation."[25] The *Tallinn Manual 2.0* provides several examples of ruses, including simulating nonexistent forces, transmitting false information showing operations beginning, feigned cyberattacks, bogus orders purportedly issued by the enemy, and transmitting false intelligence information intended for interception.[26]

According to Protocol I, states must also consider that "in the conduct of military operations, constant care shall be taken to spare the civilian population, civilians and civilian objects."[27] The *Tallinn Manual 2.0* states, "The term 'spare' refers to the broad general duty to 'respect' the civilian population, that is, to consider deleterious effects of military operations on civilians."[28]

To determine the propriety of using deepfakes, a state may have considered the prior activities of the target state, which may have precipitated a response. For example, proportional countermeasures may be appropriate even when not engaged in armed conflict.[29]

Outside of armed conflict settings, the principle of sovereignty prohibits acts that "interfere with, or usurp, an inherently governmental act or cause such effects on the territory." Protected governmental acts include elections, crisis management, and national security.[30] The following analysis of the scenarios begins with sovereignty as it "is a foundational principle of international law."[31]

International law scholars generally do not view cyber espionage as a violation of sovereignty, and some contend this even when such activities take place within another state's territory or cause territorial effects.[32] Eric Talbot Jensen writes that these scholars "for instance . . . are of the view that remote cyber activities that violate domestic law on espionage would not, in themselves, violate international law." [33] Some states argue that sovereignty is not a rule but a foundational principle on which the use-of-force and nonintervention principles are based, meaning that noncoercive cyber operations and deepfakes would not violate international law.[34]

While often discussed together, sovereignty and nonintervention must be differentiated, as sovereignty focuses on territorial control and governmental acts and does not require coercion.[35] Nonintervention is seen to emanate from sovereignty.[36] Nonintervention "prohibits States from engaging in coercive interference, directly or indirectly, with the domestic affairs of another State."[37] Case law illustrates the application of nonintervention as a customary rule.[38] The International Court of Justice has defined the protected affairs of a state, known as the *domaine réservé*, as "the choice of a political, economic, social, and cultural system, and the formulation of foreign policy." [39] According to Pijpers, "The *domaine réservé* . . . is the area 'in which each State is permitted, by the principle of State sovereignty to decide freely.'"[40]

Nonintervention requires coercion, but coercion lacks a universal standard and is not well defined. Robert Jennings and Arthur Watts state, "To constitute intervention, the interference must be forcible or dictatorial, or otherwise coercive; in effect depriving the State intervened against of control over the matter in question."[41] Schmitt defines coercive action as "intended to cause the State to do something, such as take a decision that it would otherwise not take, or not to engage in an activity in which it would otherwise engage."[42]

According to Schmitt, the *Tallinn Manual 2.0* says that states should also consider due diligence, or their obligation "to ensure that their territory is not used as a location from which cyber operations having serious adverse consequences for the target State are launched."[43] But this would not require

a state to monitor or prevent cyber operations, only to stop them when the state has notice.[44] Many states, however, reject the notion that due diligence is required.[45]

Turning to the hypotheticals, how might international law principles and rules apply to a country's use of deepfakes? A country using deepfakes would intend to cause a specific outcome and would undertake action to disseminate the deepfakes, providing many additional facts to inform the legal analysis. Even without these additional facts in hand, one could consider the international law principles and rules that might apply.

In the election scenario, targeting an election—an inherently governmental function—implicates the principle of sovereignty. A certain magnitude of state action is required to violate another state's sovereignty, but neither physical damage nor loss of functionality is required when an inherent government function is involved. Additionally, propaganda aimed to influence an election result, which could possibly describe the activity in the election scenario, is usually allowed.[46] Likewise, for the nonintervention principle to apply, the use of the deepfake would need to be coercive, which seems unlikely based on the facts described. Thus, the country's use of a deepfake in this scenario might not run afoul of international law, depending on the country's intent and other facts that would emerge.

In the genocide scenario, the clear human rights violations may justify the use of deepfakes, notwithstanding principles of sovereignty and nonintervention, particularly when neither leader seems to enjoy legitimacy.[47]

In the invasion scenario, the contemplated deepfake would not have a direct impact in the target country and is likely to be seen as propaganda that does not implicate the principles of sovereignty or nonintervention. If the deepfake leads to an indirect coercive impact in the target state and the state that used the deepfake intends to cause this coercive impact, a majority of experts in the *Tallinn Manual 2.0* take the position this would violate the principle of nonintervention.[48] In the scenario for this paper, the authors have not ascribed any such intent to the state using the deepfake. In the invasion scenario, there may also be a humanitarian justification for intervening.[49] The deepfake could also be justified as a countermeasure, though experts are split on whether a non-injured state can undertake countermeasures on behalf of another state.[50]

For the stock market scenario, it is unclear whether other states will be affected by the deepfake and thus whether and how international law should apply. To the extent it is relevant, cyberattacks that result in economic damages may give rise to a countermeasure claim.[51]

For the IP theft scenario, if only private actors are affected, international law likely has no role. In this instance, the harm occurred only because a counterparty committed the wrongful act of stealing Company C's IP. In an extreme case, if the home country is aware of Company C's actions and there is a high likelihood of harm to another country, under the due diligence doctrine the home country might be obligated to take appropriate action to stop these acts. Even then, many states do not accept the due diligence doctrine.

5. Target

Who or what is the intended target of the deepfake? Is it the president of a country or a prominent person such as a political or religious leader? Is it a living person?

Consider the invasion scenario. Suppose an intelligence agency wants to create a deepfake of President Putin saying something false. The bar for approving such a deepfake should be the highest possible. In the United States, it might require approval from the president, while in some EU countries, the approving authority might be a prime minister, especially if the deepfake were to be disseminated widely in Russia and beyond. To quote one interviewee, a retired U.S. military leader at the level of general or higher, "Deepfakes of a national leader would need presidential approval under a finding from the White House."

Another U.S. military leader at the level of general or higher stated that U.S. presidents may be deeply reluctant to approve the use of deepfakes to target the leader of a foreign nation because of the precedent it would set: "A POTUS tends to think a lot about precedent—are we creating the right precedents? Subordinates tend to be more focused on objectives." The same individual went on to say, "The slippery slope argument is very strong, and the costs may be higher than we anticipate."

In both the election and genocide scenarios, one corporate leader said, "If a human rights abuser uses a deepfake, I can see doing a lot of other deepfakes to discredit the first (e.g., 100 variations) to educate people." Simply put, if the leaders of the two countries involved in these scenarios use deepfakes, then targeting them with deepfakes might be considered an appropriate response.

However, if the deepfake portrays a fictional version of a real event occurring on the ground (e.g., a genocide) for which there is compelling evidence, then disseminating a deepfake image or video representing the event may be fair because (1) the event is real and compelling evidence (perhaps nonvisual) exists; (2) there may not be real, compelling imagery or video because anyone capturing such imagery would be killed or placed in harm's way; and (3) no single world leader is being explicitly portrayed in the deepfake. Thus, some of the interviewees were comfortable with disseminating a deepfake image of a burial pit with bodies to illustrate what is really happening in a country at war. Nevertheless, such use has risks. For instance, how does the government creating the deepfake know its portrayal of the burial pits is consistent with reality? Will the use of a deepfake backfire, convincing people that the real burial pits are nonexistent because a deepfake was used? Considerable effort would be needed to ensure the deepfake accurately represents, and is believed by the public to be, the reported reality. In addition, both civil society and investigative journalist groups such as Bellingcat often offer their own highly credible reports, which is vital in democratic systems.

Another alternative is to release the deepfake publicly but clearly label it a deepfake. In essence, the deepfake would be a dramatization of real events, which frequently occurs in documentaries and films.

Whether the subject of a deepfake is living or dead also should matter. A dead person cannot lose power or money or otherwise suffer harm. On the other hand, consider that others close to the

dead person may be harmed. The deniability of the deepfake is also affected, as the person in question cannot simply say the video is fake.

6. Goal

What is the goal of the deepfake? Is it intended to protect the nation's citizens from immediate harm? Is it a tit-for-tat response to deepfake use by another government? Is it intended to educate a given population about deepfakes and suggest they be wary of any content they see?

The interviewees largely agreed it was acceptable for governments to use deepfakes under three circumstances: an immediate threat, a tit-for-tat response, or education and discrediting.

Immediate threat. At least one U.S. military leader felt using a deepfake in response to an immediate threat was potentially acceptable as long as no other measures were available to defeat the threat. However, he cautioned,

> The deepfake is still not something I would do unless there is an imminent threat because there are all these other tools you can use rather than try to put words in the mouths of these leaders, and the potential for boomerang is still too high.

Thus, an immediate threat might constitute one set of circumstances where deepfakes could be used. For instance, in the case of the genocide scenario, there may be members of the minority group who are U.S. citizens living in the country where the potential genocide would occur. As in the case of the Rwandan genocide of 1994, a U.S. president may be deeply reluctant to deploy U.S. troops to avert the genocide, and the presence of UN peacekeepers may not be useful. In such cases, it might make sense to use deepfakes to avert the genocide.

In some cases, the deepfake does not necessarily need to be believable in the long term. In the genocide scenario (or in cases of military deception), the deepfake could be circulated among the target state's leaders to misguide them into leaving a certain area exposed, allowing the at-risk population to be evacuated. However, the use of deepfakes in the invasion scenario may not make sense if the potential for retaliatory deepfakes or a kinetic response by the country in question is high.

Tit for tat. Would it be acceptable for the United States to use deepfakes in retaliation for deepfake use by a foreign government? One non-U.S. military leader saw nothing wrong in doing so: "Tit-for-tat countering may be OK. . . . If an adversary is trying to create strategic effects in the victim nation, then the way international law is framed, in the interim, I would say the same logic should apply."

A corporate leader likewise supported tit-for-tat use of deepfakes. He advocated a "no-first-use policy" but said "it is OK to respond in kind to use by a bad guy." He went on to say that he was "OK with using deepfakes defensively," and in such a case the actors "must flood the internet when such deepfakes are deployed, as one message doesn't work well."

Thus, according to both these experts, in the case of the Zelensky video referenced earlier, it would have been acceptable for Ukraine to target Putin not just with one retaliatory deepfake but with a large number, presumably as a deterrent for future deepfake use by Russia. Appropriate authorities must answer the question of what constitutes a number that is large but not so large

that it is escalatory. Such an effort, however, might poison the internet even further, perhaps to the detriment of billions of ordinary users.

Perhaps the best response to a foreign deepfake may be other cyber means. A former senior White House attorney and NSC member stated that if Russia is spreading deepfakes inside the United States, the United States should "get out in front of Russia's deepfakes and punch down the story." The interviewee was also supportive of "shut[ting] down Russia's efforts to create/spread the deepfakes using cyber tools."

Education/discrediting. Can a deepfake be used to show how convincing deepfakes can be? Can AI-generated deepfakes be used to educate a population and make them understand they should not trust everything they see? One industry expert stated the use of deepfakes to educate the population about deepfakes is perfectly acceptable. This interviewee, who has in-depth knowledge about YouTube's policies, remarked,

> Educational use—don't believe what you see—is OK. For example, YouTube approved a deepfake of "[the] pope in [a] puff jacket" if it was posted for educational purposes but would block it if it was used as an ad or in a way intended to create harm.

Of course, when deepfakes are used for educational purposes, they should be clearly marked as deepfakes.

7. Traceability

What is the probability of the deepfake being traced back to the creator? What will the blowback be in the event it is attributed to the government that created it?

When a government uses deepfakes in pursuit of its goals, the interviewees felt several questions must be asked: Will the deepfake be detected by other actors? How severe will the repercussions of the deepfake likely be, if they are publicly attributed to the country that created and distributed it? Will it implicate third parties (e.g., if Israel uses a deepfake, will it implicate the United States)?

The experts had deep concerns about detectability and blowback. One U.S. military leader at the level of general or higher stated the importance of intelligence tradecraft: "If we do this, can they trace it back to us? Is there something inherently identifying that is traceable back to us?"

Another U.S. military leader said the amount of blowback "should be linked to the approval process all the way up to the NSC." Simply put, a deepfake, especially one of a foreign leader, may need very high levels of approval up to the highest levels of the U.S. government. He went on to posit a hypothetical scenario: "Suppose a Silicon Valley firm releases a deepfake of Xi Jinping. What about his reaction? They would blame the USA. This is coming." This suggests the need for a more compelling set of guardrails against corporate and individual use of deepfakes to target foreign states or their leaders.

In contrast, a former senior White House and NSC lawyer stated, "If a covert action finding had previously been approved, then the CIA alone would decide if deepfakes should be used." Speaking on the same topic, a non-U.S. military leader said, "Who created the deepfake doesn't matter. It

is about who is endorsing and authenticating it." Thus, in the case of a Silicon Valley company releasing a deepfake, the military leader suggested the U.S. government should not endorse the deepfake or distribute it in any way, even if it represents a reality. Perhaps, in this case, the U.S. government's immediate action should be to disavow the corporate deepfake and immediately take action to hold those responsible to account to the fullest extent of the law.

Who Should Manage a Deepfake Equities Process?

The above questions are important to ask—but equally important is who is asking them. Democratic countries should create a deepfake equities commission (DEC) to weigh the answers to these questions. The following is a discussion of how this might work in the United States.

The DEC should be a White House–run interagency working group chaired by an NSC senior director. It should include one representative from each of the Departments of Defense, Justice, State, and Treasury as well as a representative from the Office of the Director of National Intelligence to ensure that different perspectives are brought together. In addition to these national security and legal perspectives, it is necessary to bring in representatives who can weigh the domestic effects—something outside the traditional national security realm—which may require bringing in an appropriately cleared representative from civil society to represent the public interest.

Because of the novelty of deepfakes and their potentially broad effects on the American people, any recommendation for deepfake use by the U.S. government should be approved by the president. Should deepfake use become more common, it is plausible there may be established categories for use that are preapproved, but for the initial stages and possibly beyond, their importance and the precedent they might set require the highest level of approval. Similarly, if deepfakes are used as part of an ongoing conflict, authority may be delegated to the secretary of defense or another high-ranking subordinate until clear use categories are established.

Other countries will have different institutional arrangements, but the same basic principles should hold. Different parts of the government representing different commercial and security perspectives should be brought together, as should individuals representing the informational health of the polity. Because of the novelty and potential impact of deepfakes, the top elected official, such as the prime minister in countries with a parliamentary system of government, should approve their use.

When a situation arises where use of a deepfake might be considered, the relevant agency should present the use case, with particular detail not only on the immediate effects but also on the potential blowback to the home country's people, allies, and the credibility of the government should the deepfake use be discovered. For example, if the goal is to discredit a foreign military official, the Department of Defense might propose a deepfake. The scenario would not only outline the content of the deepfake and its anticipated effects on the adversary's military but also how it might shape public opinion inadvertently in the United States (or the home country contemplating deepfake use) and in allied countries. Officials from other agencies would weigh in on its legality, diplomatic consequences, and other ramifications. The designated official representing U.S.

public opinion would also weigh how dangerous the information would be if it infected domestic discussion. Other officials might review its necessity, as there may be more traditional cyber or other operations that have a more established precedent that could achieve the same intended effects. Finally, all involved officials would discuss the implications for other policy goals if the United States were associated with deliberately promulgating false information.

Conclusion

Deepfakes are a new information tool, and it is important to think about their use and consequences now rather than wait to cobble together a policy following haphazard usage by U.S. government agencies or allied entities. This process is particularly important as deepfakes have profound consequences for the long-term credibility of any government and can shape and potentially worsen domestic political debates.

Although it is tempting to simply declare that a government will never use deepfakes, their potential power and reach make them attractive tools. Governments, however, may want to act—at times, for the best reasons—without thinking through the long-term implications of deepfake use. Thus, deepfakes must have a set of rules and criteria to guide their use and ensure governments are considering all relevant factors and long-term as well as short-term perspectives. For these rules to work in practice, governments must draw on a range of perspectives, including those beyond the traditional national security realm, when discussing deepfake use. Although the authors have presented some initial criteria, rich debate and observations on the use of deepfakes in practice by other governments will inform how to weigh these factors and what to add to ensure robust practice.

About the Authors

Daniel Byman is a senior fellow with the Transnational Threats Project at the Center for Strategic and International Studies (CSIS). He is also a professor at Georgetown University's School of Foreign Service and director of the Security Studies Program. He is the foreign policy editor for Lawfare and a part-time senior adviser to the Department of State on the International Security Advisory Board. In addition to serving as the vice dean for the School of Foreign Service at Georgetown, he was a senior fellow at the Center for Middle East Policy at the Brookings Institution and a professional staff member with both the National Commission on Terrorist Attacks on the United States (9-11 Commission) and the Joint 9/11 Inquiry Staff of the House and Senate Intelligence Committees. He formerly served as research director of the Center for Middle East Public Policy at the RAND Corporation and as a Middle East analyst for the U.S. intelligence community. Dr. Byman is a leading researcher and has written widely on a range of topics related to terrorism, insurgency, intelligence, social media, artificial intelligence, and the Middle East. He is the author of nine books, including *Road Warriors: Foreign Fighters in the Armies of Jihad* (Oxford, 2019), *Al Qaeda, the Islamic State, and the Global Jihadist Movement: What Everyone Needs Know* (Oxford, 2015), and *A High Price: The Triumphs and Failures of Israeli Counterterrorism* (Oxford, 2011). He is the author or coauthor of almost 200 academic and policy articles, monographs, and book chapters as well as numerous opinion pieces in the *New York Times, Wall Street Journal, Washington Post*, and other leading journals. Dr. Byman is a graduate of Amherst College and received his PhD in political science from the Massachusetts Institute of Technology.

Daniel W. Linna Jr. has a joint appointment at the Northwestern Pritzker School of Law and McCormick School of Engineering as a senior lecturer and the director of law and technology initiatives. Dan's teaching and research focus on innovation and technology, including the use of artificial intelligence and data analytics to improve legal-services delivery as well as the law, regulation, and governance of computational technologies. Dan is also an affiliated faculty member at CodeX – The Stanford Center for Legal Informatics. Dan received his BA from the University of Michigan, received a second BA and an MA in public policy and administration from Michigan State University, and graduated magna cum laude, Order of the Coif, from the University of Michigan Law School. Dan began his legal career with a one-year judicial clerkship for U.S. Court of Appeals judge James L. Ryan. After his clerkship, he joined Honigman Miller Schwartz and Cohn, where he was elected equity partner in 2013. Before law school, Dan was an information technology manager, developer, and consultant.

V. S. Subrahmanian is the Walter P. Murphy Professor of Computer Science and a faculty fellow in the Buffett Institute for Global Affairs at Northwestern University. He is an expert on probabilistic and machine learning based methods to analyze text/geospatial/relational/social network data, learn behavioral models from the data, forecast actions, and influence behaviors. His models have been used to forecast terror attacks and terror network evolution, to reduce poaching, to identify

bad actors on social media, to forecast systemic banking crises, to maximize airline profits, and more. He has written nine books, edited ten, and published over 300 articles. He was named to ISIHighlyCited.com which lists the top-most cited computer scientists of all time. He has received numerous awards. He is an elected fellow of the American Association for the Advancement of Science and the Association for the Advancement of Artificial Intelligence. His work has been featured in numerous outlets such as the *Baltimore Sun*, *The Economist, Science, Nature*, the *Washington Post*, and American Public Media. He serves on the editorial boards of numerous journals including *Science*, on the boards of directors of the Development Gateway Foundation (set up by the World Bank) and SentiMetrix, Inc., and on the research advisory board of Tata Consultancy Services. He previously served on DARPA's Executive Advisory Council on Advanced Logistics and as an ad hoc member of the U.S. Air Force Science Advisory Board (2001).

Appendix

A deepfake of a computational artifact (e.g., an image, a piece of text, or an audio or video clip) is a synthetic version of the artifact that is generated using advanced AI techniques. Machine learning classifiers have been trained for well over a decade to distinguish between real and fake artifacts. Since 2017, new generative AI techniques have created increasingly hard-to-detect fakes.

One popular technique, generative adversarial networks (GANs), considers the generation of fakes in terms of a game: A generator module in the GAN generates fake artifacts, while a discriminator module (a machine learning classifier) in the GAN tries to detect which of a set of artifacts (both real and fake) are fake. The discriminator reports the results, which the generator then uses as feedback to generate better fakes. This iterative process (generate-discriminate-feedback, generate-discriminate-feedback, and so on) continues until an equilibrium is reached and neither the generator nor the discriminator can substantially improve at its task in subsequent iterations. Numerous organizations worldwide have adapted and improved GANs and other AI techniques, such as Stable Diffusion, providing ready access to a huge number of AI tools capable of generating deepfakes for a wide variety of multimodal artifacts.[52]

Table 1 presents a range of artifacts and generation tools that could be used to generate different kinds of deepfake content.

Table 1: Well-Known Synthetic Computational Artifact Generation Tools

Artifact type	Name	URL
Image		
	Dall-E 2	https://openai.com/dall-e-2
	MidJourney	https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F
	DreamStudio	https://dreamstudio.ai/
	Firefly	https://www.adobe.com/sensei/generative-ai/firefly.html
	Runway	https://runwayml.com/
Audio		
	Resemble	https://www.resemble.ai/
	VoiceBox	https://about.fb.com/news/2023/06/introducing-voicebox-ai-for-speech-generation/
	Murf	https://murf.ai/?gclid=CjwKCAjw_aemBhBLEiwAT98FMhdLcF_XJvpbi3lBxl-UneiR24QWvjG1LFP6zJSYNu57wyw_aUaMwRoCNp8QAvD_BwE
	Play	https://play.ht/
	Speechify	https://speechify.com
Video		
	Synthesia	https://www.synthesia.io/home
	Pictory	https://pictory.ai
	Lumiere	https://lumiere-video.github.io/
	Stability*	https://stability.ai/

* Also generates audio.

Source: Authors' compilation.

Endnotes

1 Bobby Allyn, "Deepfake Video of Zelenskyy Could Be 'Tip of the Iceberg' in Info War, Experts Warn," NPR, March 16, 2022, https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia. The video can be found here: "Debunking a deepfake video of Zelensky telling Ukrainians to surrender," FRANCE 24 English, YouTube video, March 17, 2022, https://www.youtube.com/watch?v=2tgqX5WVhr0&t=56s.

2 The Zaluzhny video can be found at ЦЕНТР ПРОТИДІЇ ДЕЗІНФОРМАЦІЇ [Center for Anti-Misinformation], "Ворожі TG-канали координовано поширюють діпфейки відеозвернення Головнокомандувача ЗСУ Валерія Залужного," [Enemy TG-channels coordinately spread deepfakes of the video address of the Commander-in-Chief of the Armed Forces of Ukraine Valery Zaluzhny], Telegram post, November 7, 2023, 4:08 p.m., https://t.me/CenterCounteringDisinformation/7744.

3 Nur Ibrahim, "9 Doctored Pics and Deepfakes of Volodymyr Zelenskyy," Snopes, March 4, 2023, https://www.snopes.com/list/volodymyr-zelenskyy-fact-checks/.

4 Demetrios Ioannou, "Deepfakes, Cheapfakes, and Twitter Censorship Mar Turkey's Elections," *Wired*, May 26, 2023, https://www.wired.com/story/deepfakes-cheapfakes-and-twitter-censorship-mar-turkeys-elections/.

5 Em Steck and Andrew Kaczynski, "Fake Joe Biden robocall urges New Hampshire voters not to vote in Tuesday's Democratic primary," CNN, January 22, 2024, https://www.cnn.com/2024/01/22/politics/fake-joe-biden-robocall/index.html.

6 Jeronimo Gonzalez, "AI Avatars Are Spreading Pro-Venezuela Propaganda," Semafor, February 21, 2023, https://www.semafor.com/article/02/21/2023/venezuela-uses-ai-avatars-to-disseminate-propaganda.

7 Sophia Smith Galer, "Someone Made AI Videos of 'Americans' Backing a Military Coup in West Africa," Vice, January 27, 2023, https://www.vice.com/en/article/v7vw3a/ai-generated-video-burkino-faso-coup.

8 Shubham Sharma, "Synthesia Launches LLM-Powered Assistant to Turn Any Text File or Link into AI Video," Venture Beat, January 31 2024, https://venturebeat.com/ai/synthesia-launches-llm-powered-assistant-to-turn-any-text-file-or-link-into-ai-video/.

9 Outside of national security, deepfakes disproportionately target women. Many applications superimpose the faces of celebrities and other unsuspecting victims onto the faces of porn actresses, leading to fake sex videos that appear to show the victim engaging in pornographic acts. Deepfake audios have been used to extort money from parents who believe their child has been kidnapped after listening to a deepfake audio message from the child. Sophie Maddocks and Hailey Reissman, "What Is Deepfake Porn and Why Is It Thriving in the Age of AI?," Annenberg School for Communication at the University of Pennsylvania, July 13, 2023, https://www.asc.upenn.edu/news-events/news/what-deepfake-porn-and-why-it-thriving-age-ai; and Faith Karimi, "'Mom, These Bad Men Have Me': She Believes Scammers Cloned Her Daughter's Voice in a Fake Kidnapping," CNN, April 29, 2023, https://www.cnn.com/2023/04/29/us/ai-scam-calls-kidnapping-cec/index.html.

10 Leander Sodji, "How We Made David Beckham Speak 9 Languages," Synthesia, October 26, 2023, https://www.synthesia.io/post/david-beckham.

11 Charlotte Jee, "An Indian Politician Is Using Deepfake Technology to Win New Voters," MIT Technology Review, February 19, 2020, https://www.technologyreview.com/2020/02/19/868173/an-indian-politician-is-using-deepfakes-to-try-and-win-voters/.

12 Matthew Fecteau, "The Deep Fakes Are Coming," War Room, April 23, 2021, https://warroom.armywarcollege.edu/articles/deep-fakes/.

13 Stephen Silver, "Defense Department to Investigate Military-Run Fake Social Media Accounts," National Interest, September 20, 2022, https://nationalinterest.org/blog/techland-when-great-power-competition-meets-digital-world/defense-department-investigate.

14 Sam Biddle, "U.S. Special Forces Want to Use Deepfakes for Psy-Ops," The Intercept, March 6, 2023, https://theintercept.com/2023/03/06/pentagon-socom-deepfake-propaganda/.

15 Daniel L. Byman, Chongyang Gao, Chris Meserole, and V. S. Subrahmanian, "Deepfakes and International Conflict," Brookings Institution, January 2023, https://www.brookings.edu/articles/deepfakes-and-international-conflict/.

16 Robert Chesney and Danielle Keats Citron, "21st Century-Style Truth Decay: Deep Fakes and the Challenge for Privacy, Free Expression, and National Security," *Maryland Law Review* 78, no. 4 (2019): 882, https://digitalcommons.law.umaryland.edu/mlr/vol78/iss4/5/.

17 Each expert had ample time both to review these five scenarios and preview the types of questions the authors were going to ask them. Interviews with the experts lasted approximately 60 minutes and typically included an opening statement from the expert, followed by a period of question and answer and open discussion. All experts were encouraged to share their views as expansively as they wished. In order to protect the anonymity of the experts, the authors did not record the interviews but did take detailed notes. The interviews were held during the April-July 2023 period.

18 Samantha Power, *"A Problem from Hell": America and the Age of Genocide* (Basic Books, 2013).

19 Stella Biderman et al., "Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling," *Proceedings of Machine Learning Research* 202 (2023): 2397–2430, https://proceedings.mlr.press/v202/biderman23a.html; and Almas Abdibayev et al., "Using Word Embeddings to Deter Intellectual Property Theft through Automated Generation of Fake Documents," *ACM Transactions on Management Information Systems* 12, no. 2 (2021): 1–22, Article 13, https://doi.org/10.1145/3418289.

20 Peter B. M. J. Pijpers, *Influence Operations in Cyberspace and the Applicability of International Law*

(Edward Elgar, 2023), 227.

21 Ibid., 147.

22 "Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I)," *American Journal of International Law* 72, no. 2 (April 1978): 457–502, https://doi.org/10.2307/2200004.

23 Michael N. Schmitt, *Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations* (Cambridge University Press, 2017), 491, https://doi.org/10.1017/9781316822524.

24 "Protocol Additional to the Geneva Conventions," Art. 37(2).

25 Ibid.

26 Schmitt, *Tallinn Manual 2.0*, 495.

27 "Protocol Additional to the Geneva Conventions," Art. 57(1).

28 Schmitt, *Tallinn Manual 2.0*, 476.

29 Ibid., 111.

30 Marko Milanović and Michael N. Schmitt, "Cyber Attacks and Cyber (Mis)Information Operations during a Pandemic," *Journal of National Security Law & Policy* 11 (January 2020): 255, https://doi.org/10.2139/ssrn.3612019.

31 Schmitt, *Tallinn Manual 2.0*, 1.

32 Eric Talbot Jensen, "The Tallinn Manual 2.0: Highlights and Insights," *Georgetown Journal of International Law* 48 (March 2017): 742, https://papers.ssrn.com/sol3/Delivery.cfm/SSRN_ID2932110_code812464.pdf?abstractid=2932110&mirid=1.

33 Michael N. Schmitt, "'Virtual' Disenfranchisement: Cyber Election Meddling in the Grey Zones of International Law," *Chicago Journal of International Law* 19, no. 1 (August 2018): 30, https://centaur.reading.ac.uk/89663/.

34 Jensen, "The Tallinn Manual 2.0: Highlights and Insights," 742.

35 Pijpers, *Influence Operations*, 147.

36 Ibid.

37 Hitoshi Nasu, "Deepfake Technology in the Age of Information Warfare," Lieber Institute West Point, March 1, 2022, https://lieber.westpoint.edu/deepfake-technology-age-information-warfare/.

38 Pijpers, *Influence Operations*, 148.

39 Military and Paramilitary Activities in and against Nicaragua (Nicaragua v. United States of America). Merits, Judgment. I.C.J. Reports 1986, p. 108.

40 Ibid. See Pijpers, *Influence Operations*, 151.

41 Robert Jennings and Arthur Watts, *Oppenheim's International Law*, 9th ed. (Oxford University Press, 2008), 1:430–49.

42 Schmitt, "'Virtual' Disenfranchisement," 51. See also Ido Kilovaty, "The International Law of Cyber Intervention," in *Research Handbook on International Law and Cyberspace*, ed. Nicholas Tsagourias and Russell Buchan (Edward Elgar, 2021), 97.

43 Schmitt, "'Virtual' Disenfranchisement," 53 (citing Tallinn Manual 2.0).

44 Jensen, "The Tallinn Manual 2.0: Highlights and Insights," 744–45.

45 Schmitt, *Tallinn Manual 2.0*, 31.

46 Ibid., 26.

47 Ibid., 324. A minority of *Tallinn Manual* experts would support intervention in a humanitarian crisis.

48 Ibid., 320.

49 Ibid., 324.

50 Ibid., 132.

51 Ibid., 111.

52 Antonia Creswell et al., "Generative Adversarial Networks: An Overview," *IEEE Signal Processing Magazine* 35, no. 1 (January 2018): 53–65, https://doi.org/10.1109/MSP.2017.2765202; and Robin Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models," in *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Piscataway, NJ: Institute of Electrical and Electronics Engineers, 2022), 10684–95, https://doi.org/10.48550/arXiv.2112.10752.

COVER PHOTO JACKIE NIAM/ADOBE STOCK

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

1616 Rhode Island Avenue NW
Washington, DC 20036
202 887 0200 | **www.csis.org**