

DEEPPAKES AND THE LAW OF ARMED CONFLICT

Summer Crockett

INTRODUCTION

General Smith, field commander for the United States (U.S.) armed forces, listened intently to the broadcast their intelligence teams were receiving. The video showed the leader of an insurgent group they were currently fighting offer to surrender. The footage went on to detail a time and place where the U.S. officers could meet for the acceptance of their surrender. After the video was authenticated with the tools the military had access to and lengthy discussions by the commanders, General Smith and a small contingent of his men went to the designated meeting place. Upon arrival, they were ambushed by a terrorist cell that had been active in the area recently. The insurgents they were supposedly meeting never arrived. After the conflict was over and having suffered many casualties, the terrorist group claimed responsibility for the surrender video declaring that they had created the fake content to lure the U.S. forces into the open. Deepfake technology had been used to deceive the U.S. commander.

While such use of deepfakes in armed conflicts has not yet arisen, this technology is being used in several other meaningful settings with the potential to have significant impacts in the law of armed conflict (LOAC). It has been said that “the ultimate good desired is better reached by free trade in ideas—that the best test of truth is the power of thought to get itself accepted in the competition of the market.”¹ This may have been a valid theory when fake news, information, or communications could be discovered in relatively easy ways. Now, however, technology is developing methods through which even the most advanced computer systems cannot tell the difference between what is real and what is fake,² relying on a *market of ideas*³ theory may not be in the best interest of society.

Today deepfakes, a clever manipulation of technology that presents someone doing or saying something they have never done or said, are

¹ *Abrams v. United States*, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).

² See James Johnson, *The AI-cyber nexus: implications for military escalation, deterrence and strategic stability*, J. OF CYBER POL’Y (Dec. 9, 2019), <https://www.tandfonline.com/doi/full/10.1080/23738871.2019.1701693>.

³ Jill Gordon, *John Stuart Mill and the “Marketplace of Ideas,”* 23 SOC. THEORY AND PRAC. 235 (1997) (explaining that ideas can be analogized to a market economy where there is an exchange and process through which the best ideas will succeed).

becoming ever more present.⁴ Simple deepfakes have been used for some time through the doctoring of photos.⁵ Now, with technological advances, deepfakes have been used to show Obama swearing at Trump,⁶ Nancy Pelosi in a drunken stupor,⁷ and Russia has been accused of using deepfakes during the 2016 U.S. elections.⁸ Doctored footage has also been promulgated in an attempt to influence how British citizens feel about the influx of migrants and refugees into their country.⁹ While not all deepfakes are bad or harmful,¹⁰

⁴ See Robert Chesney & Danielle Citron, *Disinformation on Steroids: The Threat of Deep Fakes*, COUNCIL ON FOREIGN RELATIONS (Oct. 16, 2018), <https://www.cfr.org/report/deep-fake-disinformation-steroids>; Jared Schroeder, *Free Expression Rationales and the Problem of Deepfakes Within the E.U. and U.S. Legal Systems* 1, 7-8 (December 13, 2019). Available at SSRN: <https://ssrn.com/abstract=3503617>.

⁵ David L. Strauss, *Doctored Photos – The Art of the Altered Image*, TIME (June 13, 2011), <https://time.com/3778075/doctored-photos-the-art-of-the-altered-image/>.

⁶ Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 Va. J. L. & Tech. 1, 5 (2020); Kaylee Fagan, *A Viral Video that Appeared to show Obama Calling Trump a ‘Dips—’ Shows a Disturbing New Trend Called ‘Deepfakes,’* BUS. INSIDER (Apr. 17, 2018), <https://www.businessinsider.com/obama-deepfake-video-insulting-trump-2018-4>; See also Schroeder, *supra* note 4.

⁷ Maheen Sadiq, *Real v Fake: Debunking the ‘Drunk’ Nancy Pelosi Footage-Video*, THE GUARDIAN (May 24, 2019), <https://www.theguardian.com/us-news/video/2019/may/24/real-v-fake-debunking-the-drunk-nancy-pelosi-footage-video>.

⁸ Brown, *supra* note 6, at 8 n. 12; See also Mark Mazzetti, *The Plot to Subvert an Election: Unraveling the Russia Story so Far*, N.Y. TIMES (Sep. 20, 2018), <https://www.nytimes.com/interactive/2018/09/20/us/politics/russia-interference-election-trump-clinton.html> (reviewing Russia’s known meddling in U.S. elections); Alfred Ng, *Deepfakes, Disinformation Among Global Threats Cited At Senate Hearing*, CNET (Jan. 29, 2019), <https://www.cnet.com/news/deepfakes-disinformation-among-global-threats-cited-at-senate-hearing/>.

⁹ *Revealed: How Leave.EU faked Migrant Footage*, 4 NEWS (April 16, 2019), <https://www.channel4.com/news/revealed-how-leave-eu-faked-migrant-footage>.

¹⁰ Deepfakes have been used in the film industry for several years. One of the more recent famous ones has been Disney’s use of it in the Star Wars films. From recreating the iconic Peter Cushing to adding dialogue and scenes of Carrie Fisher, Hollywood is no stranger to these techniques. (See Chesney & Citron, *supra* note 4.) In fact, the deepfake technology is almost to the point where advertisements on the web may include images of you and your friends instead of the models that are customarily seen. (Gaurav Oberoi, *Exploring DeepFakes*, MEDIUM (Mar. 5, 2018), <https://goberoi.com/exploring-deepfakes-20c9947c22d9>.)

they can cause confirmation bias issues¹¹, and they tend to spread more quickly than regular truthful representations.¹²

A topic of great concern is how deepfakes can and will influence the LOAC. Questions of morality and ethics are certainly raised, but because the technology is so new there is no consensus as to what should be done about it. Deepfakes are likely to become sources of reliable false information that cause the quick spread of misinformation. The information will be reliable by the fact that deepfakes are often spread by trusted sources—or can be found and verified on several different platforms—and the information presented is consistently the same.¹³ The information will be false because no matter how reliable the deepfake source may appear, the information will still be equivocally false.

In order to handle this arising technology, a few countries have banned deepfakes.¹⁴ Even some areas in the United States have placed restrictions on the emerging technology.¹⁵ While the European Court of

¹¹ During the 2008 elections, rumors that candidate Obama was not actually born in the United States ran rampant through the U.S. There is still a percentage of the population today that believes this information despite irreputable proof to the contrary. While this example is more a spread of disinformation, deepfakes have been known to serve confirmation bias purposes as well. Brown, *supra* note 6, at 10; Stephan Lewandowsky, et al., *Misinformation and Its Correction: Continued Influence and Successful Debiasing*, 13 PSYCHOL. SCI. IN PUB. INT 106, (2012).

¹² “Even when researchers controlled for differences between accounts originating rumors, falsehoods were 70 percent more likely to get retweeted than accurate news.” Robert Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1767 (2019).

¹³ Benedict Carey, *How Fiction Becomes Fact on Social Media*, N.Y. TIMES (Oct. 20, 2017), <https://www.nytimes.com/2017/10/20/health/social-media-fake-news.html>.

¹⁴ Malaysia instituted laws prohibiting deepfakes in part because there is no effective technology to identify and stop the spread of deepfakes before damage has already been done. (Will Knight, *The US Military is Funding an Effort to Catch Deepfakes and Other AI Trickery*, MIT TECH. REV. (May 23, 2018),

<https://www.technologyreview.com/s/611146/the-us-military-is-funding-an-effort-to-catch-deepfakes-and-other-ai-trickery/>.) Singapore has also banned the creation and

disseminating of deepfakes and instituted high monetary penalties for breaking the prohibition. (James Griffiths, *Singapore 'Fake News' Law Comes Into Force, Offenders Face Fines and Prison Time*, CNN (Oct. 2, 2019),

<https://www.cnn.com/2019/10/02/asia/singapore-fake-news-internet-censorship-intl-hnk/index.html>); Casey Newton, *Singapore's Fake News Law Should be a Warning to American Lawmakers: How Misinformation Laws Get Used Against the People Who Advocate for Them*, THE VERGE (Dec. 3, 2019),

<https://www.theverge.com/interface/2019/12/3/20991422/singapore-fake-news-law-censorship-politics-usa>; *Singapore clamps down on fake news with new law*, DW NEWS (May 8, 2019), <https://p.dw.com/p/3IBgm>.) See Brown, *supra* note 6, at 32.

¹⁵ Being one of the first to ban certain types of deepfake creation, Texas enacted criminal standards to be applied to deepfake creation that was false, deceptive, or made to

Human Rights (ECtHR) and the United States Congress, because of their obligation to sustain their Bill of Rights, are both concerned about the free speech implications of banning or permitting deepfakes.¹⁶ The tactical advantages of using such technologies in a LOAC setting make their ban less desirable and may even have the adverse effect of encouraging their greater development and further use.

The strategic advantages created by the use of deepfakes in warfare are undeniable. While using deepfakes in acts of terrorism or perfidy would be prohibited by the LOAC, most uses of deepfakes are too advantageous to ignore. Deepfakes will be used in armed conflict settings, and there are most likely to be significant detrimental effects on civilian populations. Per the constant care provision of Additional Protocol I (AP I),¹⁷ these effects should and can be mitigated through the creation of precautions and using technology to protect the perception of truth for civilians.

Section I will give an in-depth discussion about deepfakes and how they are generated. It will discuss the public concerns surrounding deepfakes, as these can often be analogized to LOAC settings. This section will then review the difficulty of disputing deepfake content in a meaningful way, how cultures with a lack of civilian technology knowledge are impacted, and the free speech concerns that also arise with the use of deepfakes.

Section II will specifically discuss how deepfakes impact LOAC. The history of war and the legal use of disinformation will be reviewed. The importance of the constant care provision in relation to civilians will be discussed, and the benefits and disadvantages of using deepfakes in armed conflicts will also be analyzed.

Section III will review areas of the LOAC that are underdeveloped in the area of handling deepfake technology. Several avenues for establishing law regulations or guidelines will be presented as well.

Section IV will conclude and provide a suggested direction for the next steps the LOAC should take in order to handle the increased usage of deepfakes in military operations.

specifically influence the outcome of a presidential election. (Brown, *supra* note 6, at 45.) AB 730 and AB 602. AB 730 are two bills enacted by California that criminalize deepfakes that are meant to sway voters around election time and also supplies a private right of action to the victims of deepfakes of a sexual nature. (Brown, *supra* note 6, at 45-6; Helena Skinner, *French Charity Publishes Deepfake of Trump Saying 'AIDS is Over,'* EURONEWS (Sept. 10, 2019), <https://www.euronews.com/2019/10/09/french-charity-publishes-deepfake-of-trump-saying-aids-is-over>.)

¹⁶ See Chesney & Citron, *supra* note 12, at 1788-9; *New York Times v. Sullivan*, 447 U.S. 557, 592 (1980) (Rehnquist, C.J., dissenting); Schroeder, *supra* note 4, at 13.

¹⁷ Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), art. 57, Dec. 7, 1978, 1125 U.N.T.S. 3 (hereinafter ADDITIONAL PROTOCOL I).

I. DEEPFAKES

Disinformation and the means by which it is promulgated has been taken to an even deeper level through the emergence of deepfakes.¹⁸ While deepfake technology originally was used to place the facial images of famous people on the bodies of porn stars,¹⁹ the technology is coming to a point of extreme effectiveness in that its use is now being employed in much more productive and potentially dangerous avenues. This section will review the technology used in the creation of deepfakes and the concerns that arise when deepfakes are dispensed.

A. Deepfake Technology

The technology used to create deepfakes is called Generative Adversarial Networks (GANs).²⁰ These GANs or “neural networks” create the fake clip and evaluate its quality.²¹

[T]he generator, draws on a dataset to produce a sample that mimics the dataset. The other network, the discriminator, assesses the degree to which the generator succeeded. In an iterative fashion, the assessments from the discriminator inform the assessments of the generator. The result far exceeds the speed, scale, and nuance of what human reviewers could achieve.²²

This system of machine learning is quickly producing deepfakes at a quality and quantity that would not be achievable by human might alone.²³ Realistic voice duplication applied to deepfake video content is possible and almost undetectable now as well.²⁴ It is the GANs process which makes these

¹⁸ Donie O’Sullivan, *When Seeing is no Longer Believing, Inside the Pentagon’s Race Against Deepfake Videos*, CNN BUS. (Jan. 2019), <https://www.cnn.com/interactive/2019/01/business/pentagons-race-against-deepfakes/>.

¹⁹ Schroeder, *supra* note 4, at 6; See Skinner, *supra* note 14.

²⁰ Chesney & Citron, *supra* note 12, at 1760; Chesney & Citron, *supra* note 4; Schroeder, *supra* note 4, at 7.

²¹ Chesney & Citron, *supra* note 12, at 1760 (2019); Schroeder, *supra* note 4, at 7; See Knight, *supra* note 13 (stating that GANs can create “stunningly realistic artificial imagery”); Skinner, *supra* note 14.

²² Chesney & Citron, *supra* note 12, at 1760; See Skinner, *supra* note 14.

²³ Chesney & Citron, *supra* note 12, at 1760.

²⁴ Brown, *supra* note 6, at 6; Vincent, *supra* note 23. James Vincent, *The AI-generated Joe Rogan fake has to be heard to be believed*, THE VERGE (May 17, 2019), <https://www.theverge.com/2019/5/17/18629024/joe-rogan-ai-fake-voice-clone-deepfake-dessa>.

deepfakes difficult to distinguish from reality because it integrates the input with the desired changes and “regenerate[s] the original input from this encoding” producing a convincing and hard to detect fake.²⁵ With the rate at which this technology improves, it is very likely that deepfake creation technology, that produces very believable content, will soon readily be available to the public.²⁶ The technology is becoming so realistic that it is very difficult for computer programs, let alone people, to identify what is real and what is a deepfake.²⁷ Offering, creating, and promulgating alternative realities to large sections of the world’s population has never been easier or more believable.²⁸

B. Concerns About Deepfakes

Nina I. Brown, an Assistant Professor and scholar in the field of deepfakes, stated that “in the deepfake arms race, there is more at stake than the threat of technological deception it is the reality that deepfakes empower people to choose their truth.”²⁹ This idea is frightening because it implies that people will be making decisions and creating policy on potentially false information that appears too reliable to dismiss.³⁰ The overreaching effects of this are that countries, communities, presidents, generals, LOAC decision-makers and more may all have different perceived truths and falsehoods that spring from believable deepfake content.³¹ An added complication to this is

²⁵ Oberoi, *supra* note 10.

²⁶ Such technology, in a small way, is accessible now through apps such as SnapChat, FaceApp – AI Face Editor, and similar types of app filters. See Vincent, *supra* note 23; Kevin Roose, *Here Come the Fake Videos, Too*, N. Y. TIMES (Mar. 4, 2018), <https://www.nytimes.com/2018/03/04/technology/fake-videos-deepfakes.html>.

²⁷ Chesney & Citron, *supra* note 4; O’Sullivan, *supra* note 17; Knight, *supra* note 13 (explaining that the design of GANs is to “outwit an AI system” making them even harder to detect); See Kevin Stankiewicz, *‘Perfectly Real’ Deepfakes will Arrive in 6 Months to a Year, Technology Pioneer Hao Li Says*, CNBC (Sep. 20, 2019), <https://www.cnn.com/2019/09/20/hao-li-perfectly-real-deepfakes-will-arrive-in-6-months-to-a-year.html>; Oscar Schwartz, *You Thought Fake News was Bad? Deepfakes are Where Truth Goes to Die*, THE GUARDIAN (Nov. 12, 2018), <https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth> (discussing how many people believed and were angry about a video depicting Trump telling Belgium how to handle climate change and later discovered the video was a deepfake).

²⁸ See Skinner, *supra* note 14.

²⁹ Brown, *supra* note 6, at 26; See Chesney & Citron, *supra* note 12, at 1814.

³⁰ Schroeder, *supra* note 4, at 8-10 (explaining that video content is more believable because people feel like they have become first-hand witnesses to an event).

³¹ Schroeder, *supra* note 4, at 10.

that refuting deepfakes in a timely manner is nearly impossible,³² so the world leaders and military commanders leading LOAC maneuvers may all have reliable false information that causes more disasters instead of resolving them.

As politics and war are connected, deepfakes become another weapon in the arsenal of dissemination of falsity. Deepfakes can impact elections through changing public perception of people in political and governmental positions.³³ The proper deepfake placed at the right time can ruin the career of any public figure—military or otherwise.³⁴ Some may argue that that is the way of politics and government and how institutions are run, including the military. However, deepfakes can be a weapon to remove the perfect person from a position in which they could turn the balance of a country and its policies. A deepfake could cause a general to be removed before an important battle or turn public opinion against the military. Deepfakes could potentially be used to create evidence egregious enough to get a president who is against active armed conflict involvements to be impeached and leave the vice president, who is much more aggressive in their war policies, to become commander and chief.

Tied closely with politically motivated deepfakes is the issue of inciting public outrage. Deepfakes can be employed to cause the public to withdraw support from military decision makers.³⁵ In a government system like the United States, this can be problematic because the congress, which controls military funding, is greatly influenced by popular public opinion due to the democratic system of electing congressional members.³⁶ Some dissenters believe that the public will be capable of discerning between real content and deepfakes,³⁷ so the scenario *supra* would not be of great concern. However, deepfakes are often more likely to go viral due to their

³² See Chesney & Citron, *supra* note 12, at 1759-60, 1814; Ng, *supra* note 8 (stating that as deepfake technology improves detection becomes harder).

³³ Brown, *supra* note 6, at 9; Knight, *supra* note 13; Daniel Thomas, *Deepfakes: A Threat to Democracy or Just a Bit of Fun?*, BBC NEWS (Jan. 23, 2020), https://www.bbc.com/news/business-51204954?intlink_from_url=https://www.bbc.com/news/technology&link_location=live-reporting-story; Schroeder, *supra* note 4, at 2; See also Andrew Guess, Brendan Nyhan & Jason Reifler, *Selective Exposure to Misinformation: Evidence from the Consumption of Fake News During the 2016 U.S. Presidential Campaign*, EUR. RES. COUNCIL (Jan. 9, 2018), <https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf> (last visited Mar. 14, 2020) (stating that more than a quarter of U.S. citizens visited a fake news site during the 2016 presidential elections).

³⁴ Chesney & Citron, *supra* note 12, at 1771.

³⁵ Chesney & Citron, *supra* note 12, at 1776.

³⁶ U.S. CONST. art 1, § 8, cl. 11-2.

³⁷ See Oberoi, *supra* note 10

sensationalism,³⁸ and once the content is in the public sphere of ideas it is nearly impossible to dislodge the lies because of the psychology involved in the process.³⁹

While some of these examples may not be conducted during armed conflicts and thereby not implicate the use of AP I 57.1⁴⁰, there is, however, the possibility that if a deepfake is used outside the theater of operations—such as the example of impeaching a president—then 57.1 would still apply because the effects would be felt by the civilian population who would bare the brunt of those consequences. Hence the need for precautions and regulations to be established to protect civilians from the use of deepfakes in armed conflicts.

C. *Difficulty of Disputing*

Researchers are finding that while people approach traditional news sites with a healthy skepticism, this is lacking when they read posts on social media platforms.⁴¹ Psychology Professor Colleen Seifert from University of Michigan explained that “people have a benevolent view of Facebook, for instance, as a curator, but in fact it does have a motive of its own. What it’s actually doing is keeping your eyes on the site. It’s curating news and information that will keep you watching.”⁴² One of the reasons misinformation like deepfakes spreads so easily is because the platforms they are shared on actually are coded to spread them quickly because they tend to keep people on their site and the information spreads at a speed that outstrips fact checkers.⁴³ A second reason why misinformation and deepfakes can spread so quickly and are believed is because they are often spread through a person’s friends on social media platforms.⁴⁴ People are much less skeptical when information is passed onto them by a personal friend.⁴⁵ A third reason that these types of deepfakes can be nearly impossible to refute is that they

³⁸ Brown, *supra* note 6, at 21-2, 36, 55; Schroeder, *supra* note 4, at 2, 29 (explaining that deepfakes are “more visceral and traditionally more credible form of false . . . information”); See Robinson Meyer, *The Grim Conclusion of the Largest-Ever Study of Fake News*, THE ATLANTIC (Mar. 8, 2018), <https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/>.

³⁹ Brown, *supra* note 6, at 16; Carey, *supra* note 13; Chesney & Citron, *supra* note 4; O’Sullivan, *supra* note 18.

⁴⁰ ADDITIONAL PROTOCOL I, *supra* note 17.

⁴¹ Carey, *supra* note 13.

⁴² *Id.*

⁴³ *Id.*

⁴⁴ *Id.*

⁴⁵ *Id.*; Meyer, *supra* note 36.

are shared multiple times.⁴⁶ Repetition serves as a strong reinforcer to influence people's brains into giving a story more credibility because they have heard it more than once.⁴⁷

D. Lack of Civilian Technological Competence

A further complication to this is that deepfakes are especially dangerous in developing countries.⁴⁸ Citizens in these countries are less used to this type of technology and unprepared to handle or contemplate the effects.⁴⁹ They will most likely struggle to realize that what is being seen with their eyes may be a clever deepfake to sow unrest, confusion, propaganda, turn them against their military, or incite them against an incoming military group—effectively removing civilian protections from individuals who otherwise would qualify for them.⁵⁰

E. Compromising the Freedom of Speech

Another difficulty with deepfakes becoming more prevalent is that countries like the United States and those in the European Union take seriously individuals' right to free speech and the protection of information both true and false.⁵¹ This is one of the reasons that social media platforms are rather liberal in the regulation of their content.⁵² This is concerning because most people in the world tend to get their news and information through social media websites.⁵³ With more than a third of the world's population using social media, the spread of deepfakes can be quick and disconcerting.⁵⁴ Since social media platforms usually allow for a free market

⁴⁶ Carey, *supra* note 37.

⁴⁷ Carey, *supra* note 37.

⁴⁸ Thomas, *supra* note 32.

⁴⁹ *Id.*

⁵⁰ *Id.*; See also Geneva Convention Relative to the Treatment of Civilian Persons in Time of War, Aug. 12, 1949, T.I.A.S. 3365; Geneva Convention (III), Relative to the Treatment of Prisoners of War, Aug. 12, 1949, T.I.A.S. 3364; 1907 Regulations Annexed to Convention IV Respecting the Laws and Customs of War on Land; ADDITIONAL PROTOCOL I, *supra* note 16, at art. 5.

⁵¹ Schroeder, *supra* note 4, at 4.

⁵² Communications Decency Act 47 U.S.C. § 230 (1996); See Brown, *supra* note 6, at 42-3; Chesney & Citron, *supra* note 4.

⁵³ Elise Shearer & Jeffery Gottfried, *News Across Social Media Platforms 2017*, PEW RES. CTR. (Sep. 7, 2017), <https://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/>; See Chesney & Citron, *supra* note 4.

⁵⁴ Esteban Ortiz-Ospina, *The Rise of Social Media*, OUR WORLD IN DATA (Sep. 18, 2019), <https://ourworldindata.org/rise-of-social-media>; See Brown, *supra* note 6, at 15 (explaining that the easiest way to disseminate disinformation is through social media platforms); J.

of ideas approach, there is little to no regulation of what disinformation is promulgated.⁵⁵

While victims of deepfakes can seek recourse through current tort statutes in the U.S. to seek recovery from the creators of deepfakes,⁵⁶ this only confounds the problem because disputing the information already disseminated is very difficult.⁵⁷ Some people in the United States still believe the misinformation that former President Obama was not born in the U.S.A.—a popular rumor spread during his campaigning for the position of president.⁵⁸ Years may pass, but if the misinformation sparks reliance on a psychological bias, or the deepfake is too well constructed, it could be impossible to repair the harm caused by their lies. And, even if the technology to detect deepfakes improves, it is the rapid distribution of deepfakes through the internet and social media platforms that causes more problems.⁵⁹ There is a strong argument that social media platforms should more closely monitor, screen, or try to detect deepfakes so that they are not promulgated to the public.⁶⁰ Some proponents of this argue that social media platforms are in the perfect position to detect deepfakes and prohibit them because deepfakes “pass through their systems” when they are posted.⁶¹ The proponents for such an argument may not grasp the complexity of the new generation of deepfakes.⁶² And while the United States congress is pressuring social media platforms to take steps to address this problem,⁶³ the technology used to create deepfakes is far ahead of detection technology making it an uphill and most likely losing battle to identify and prevent their spread.⁶⁴ The dissemination of deepfakes impacts the LOAC because of the sheer scale, influence, and pervasiveness of social media and the psychology of humans

Clement, *Number of Global Social Media Users 2010-2021*, STATISTA (Aug. 14, 2019), <https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/>.

⁵⁵ Brown, *supra* note 6, at 16, 42-3; See Chesney & Citron, *supra* note 4.

⁵⁶ *Id.*

⁵⁷ See Chesney & Citron, *supra* note 12, at 1771; Carey, *supra* note 37.

⁵⁸ Julia Glum, *Some Republicans Still Think Obama was Born in Kenya as Trump Resurrects Birther Conspiracy Theory*, NEWSWEEK (Dec. 11, 2017), <https://www.newsweek.com/trump-birther-obama-poll-republicans-kenya-744195>.

⁵⁹ Brown, *supra* note 6, at 21, citing Lewandowsky, *supra* note 11, at 110 (stating that false information on twitter spreads faster and more broadly than truth).

⁶⁰ Chesney & Citron, *supra* note 4. See Brown, *supra* note 6, at 28, 42 (suggesting social media platforms voluntarily screen for deepfakes).

⁶¹ Skinner, *supra* note 14.

⁶² Dan Robitzski, *DARPA Spent \$68 Million on Technology to Spot Deepfakes*, FUTURISM (Nov. 19, 2018), <https://futurism.com/darpa-68-million-technology-deepfakes> (explaining that forensic detection of deepfakes is going to take time and is not currently effective at recognizing deepfakes).

⁶³ Chesney & Citron, *supra* note 4.

⁶⁴ Brown, *supra* note 6, at 25; Robitzski, *supra* note 59.

who tend to believe false news because of their natural biases.⁶⁵ It would be all too easy for one party to the conflict to post their strategic deepfake on a social media platform. That would make the deepfake available for not only the opposing side in the armed conflict to view but also for numberless civilians. The harm that this might cause because civilians may rely on the false information and place themselves in dangerous areas of conflict, or any other countless examples, necessitates the creation of precautions to protect civilians.

The technology used to create deepfakes is intricate and improving daily. Soon it may be impossible for people or technology to determine what is a deepfake and what is real creating the problem of an unstoppable spread of reliable false information. The harmful potential of deepfakes on civilians and their safety is concerning and must be addressed.

II. DEEPAKE AND DISINFORMATION IN THE LOAC

While the above section has reviewed deepfakes and outlined some of the potential associated harms or difficulties that the LOAC may face because of them, this section will give a more in-depth analysis of the history, dangers and benefits of deepfakes in armed conflicts. It will also review the potential duties of States imposed through AP I and the constant care provision as it relates to civilians, which states that “in the conduct of military operations, constant care shall be taken to spare the civilian population, civilians and civilian objects.”⁶⁶ The key language in this statement is that it appears to apply to all military operations, not just attacks.

A. *History and Dangers that Deepfakes Pose in the LOAC*

The security agencies in the United States and other countries have clearly identified technology as one of their biggest concerns when it comes to safety, security, and warfare.⁶⁷ The chairman of the U.S.A’s Senate’s

⁶⁵ Carey, *supra* note 37; Meyer, *supra* note 36; Oritz-Ospina, *supra* note 51; See Brown, *supra* note 6, at 15 (explaining that the easiest way to disseminated disinformation is through social media platforms); Clement, *supra* note 51.

⁶⁶ ADDITIONAL PROTOCOL I, *supra* note 16, at art. 57(1).

⁶⁷ Ng, *supra* note 8; See Peter Dutton, *Cybersecurity is a National Priority for Australia*, THE STRATEGIST (Nov. 18, 2019), <https://www.aspistrategist.org.au/cybersecurity-is-a-national-priority-for-australia/>; Zolan Kanno-Youngs, *Homeland Security Chief Cites Top Threat to U.S. (It’s Not the Border)*, N.Y. TIMES (Mar. 18, 2019), <https://www.nytimes.com/2019/03/18/us/politics/homeland-security-cyberthreats.html> (explaining U.S.’s top concern is cyber threats); *National Security Threats*, CENTRE FOR THE PROTECTION OF NATIONAL INFRASTRUCTURE, <https://www.cpni.gov.uk/national->

intelligence committee expressed this emphasis well when he stated, “We’re now living in yet another new age: the time characterized by hybrid warfare, weaponized disinformation, all within the context of a world producing more data than mankind has ever seen.”⁶⁸ The seriousness of deepfakes, and their impact on national security and LOAC⁶⁹ appears to be clearly noted and is being addressed by the United States as they have invested over sixty million dollars in technology specifically designed to identify deepfakes.⁷⁰ A forensics expert from the University of Notre Dame noted that this investment is part of a technological arms race.⁷¹ Deepfakes are a top priority in the context of the LOAC because a cleverly designed deepfake released at the right moment, on the right platform, could be the spark to ignite violence and stir civil unrest in an already unstable war torn area.⁷² Robert Chesney and Danielle Citron, professors and researchers in this field of law, further explained the concerns that are being considered as deepfakes become more prevalent in the LOAC when they said:

Soldiers could be shown murdering innocent civilians in a war zone, precipitating waves of violence and even strategic harms to a war effort. Falsified video appearing to show a Muslim man at a local mosque celebrating the Islamic State could stoke distrust of, or even violence against, that community. A fake video might portray an Israeli official doing or saying something so inflammatory as to cause riots in neighboring countries, potentially disrupting diplomatic ties or sparking a wave of violence. False audio might convincingly depict U.S. officials privately “admitting” a plan to commit an outrage overseas, timed to disrupt an important diplomatic initiative.⁷³

It is true that deepfakes can be used to spread hate, lies, and panic. While the use of terror against civilians is strictly prohibited by the LOAC⁷⁴, disinformation has been a tactic used in armed conflicts since ancient times. The Trojan war is made famous by the use of a large, fake, wooden horse which was used to spread disinformation among the enemy.⁷⁵ During World

security-threats (last visited Mar. 17, 2020) (listing Cyber security as one of the United Kingdom’s top concerns).

⁶⁸ Ng, *supra* note 8.

⁶⁹ Chesney & Citron, *supra* note 4.

⁷⁰ Robitzski, *supra* note 59.

⁷¹ Knight, *supra* note 13.

⁷² Brown, *supra* note 6, at 7.

⁷³ Chesney & Citron, *supra* note 12, at 1776.

⁷⁴ ADDITIONAL PROTOCOL 1, Art. 51.2.

⁷⁵ *Trojan War*, HISTORY (Sep. 11, 2019), <https://www.history.com/topics/ancient-history/trojan-war>.

War II disinformation, and early types of deepfakes were used to confuse the Axis forces so that they would not know where D-Day was to take place.⁷⁶ Britain also used similar tactics to wage psychological warfare on the Nazi's and the areas that they had conquered.⁷⁷ What this history suggests is that the LOAC does not recognize disinformation as a breach or a war crime, even if the information is readily accessible to civilians. Rather trickery, ruse, and deception are lauded in history and recognized as legal and strategic methods of warfare.⁷⁸ In fact, not using disinformation would probably be a tactical error given how common it has been in armed conflicts of the past.

B. Lawful and Unlawful Uses of Deepfakes in Armed Conflicts

Deepfakes and misinformation can be tools and weapons in an armed conflict. They are pervasive and effective in the history of armed conflicts. In AP I it states:

Ruses of war are not prohibited. Such ruses are acts which 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. The following are examples of such ruses: the use of camouflage, decoys, mock operations and misinformation.⁷⁹

Due to the historical background of the LOAC accepting misinformation as a suitable tool in war, deepfakes are a legal method of warfare. They will be used in the forum of armed conflicts because they are too strategically beneficial to be ignored. The potential problem with the use of deepfakes in an armed conflict setting, however, arises because the technology is making misinformation more realistic, more persuasive, more permeating, and more accessible to the civilian population more now than ever before. In order to comply with the LOAC, especially the constant care directive, States may need to regulate their use of deepfakes, or even require that deepfakes not be used in certain armed conflict situations.

⁷⁶ Chesney & Citron, *supra* note 12, at 1783.

⁷⁷ Lord Ashcroft, *Man-Eating Sharks on the Prowl in the Channel and a Propaganda Radio Station Wooing the Nazis with Porn: Lord Ashcroft Says Donald Trump Didn't Invent Fake News... Britain Used it to Beat Adolf Hitler!*, DAILY MAIL (Feb. 29, 2020), <https://www.dailymail.co.uk/news/article-8060637/Donald-Trump-didnt-invent-fake-news-Britain-used-beat-Adolf-Hitler-says-Lord-Ashcroft.html>.

⁷⁸ See ADDITIONAL PROTOCOL I, *supra* note 16.

⁷⁹ ADDITIONAL PROTOCOL I, *supra* note 16, at art. 37(2).

Deepfakes can also be a useful tool when military forces are operating in a densely populated area, which has become more common in modern warfare,⁸⁰ because deepfakes can be used to promote recruitment to terrorist groups or turn the normal populace against an occupying force.⁸¹ As this technology continues to improve, it will be an invaluable tool in information operations because it can be used by insurgents to paint themselves or their leaders in a particular light, or make it appear that a person is in a place at a certain time when they were in fact not there.⁸²

Strategically deepfakes could potentially become one of the most useful tools in the arsenal of LOAC weaponry. Tactically speaking, deepfakes have limitless benefits that can alter some of the ways war is currently waged. Deepfakes can be used to turn non-government organizations (NGOs) or the International Committee of the Red Cross (ICRC) against one party or another, which in turn would probably cause global public opinion to shift against them, making it more difficult to ascertain weapons, supplies, and information.⁸³ As a tool of psychological manipulation, a deepfake could make it appear that a leader of a group is meeting with their enemies to broker a deal, turning their civilian basis of support against them.⁸⁴ Also, technology that might be used to create deepfakes could become an important target during an airstrike, or a catastrophic virus or malware because of the potential damage that their creation can cause.⁸⁵

If anything, the fact that the United States Defense Advanced Research Projects Agency's (DARPA) is spending time and resources on technology to detect and handle deepfakes demonstrates that the military considers this technology to be a great threat to peace and security.⁸⁶ With the increased speed at which deepfake technology improves there is apprehension among researches that soon humans and potentially artificial intelligence will be unable to detect or recognize what is a deepfake and what is not.⁸⁷

It is important to also realize that this technology can be used to cause panic among civilians by creating video that claims a nuclear, or other attack is immanent, while in actuality the fake clip was only meant to incite

⁸⁰ Chesney & Citron, *supra* note 12, at 1783 (citing the example of the United States' conflict in Afghanistan).

⁸¹ *Id.*

⁸² *Id.* at 1808-9.

⁸³ Chesney & Citron, *supra* note 4.

⁸⁴ Chesney & Citron, *supra* note 12, at 1809.

⁸⁵ *Id.*

⁸⁶ O'Sullivan, *supra* note 17; See Knight, *supra* note 13; Robitzski, *supra* note 59.

⁸⁷ Knight, *supra* note 13.

hysteria.⁸⁸ However, using deepfakes to create panic among the general population—which would be the perfect screen for infiltration, taking over a city by distracting the local troops, or divert occupying forces to policing instead of protecting the borders—clearly falls into the categories of terror and perfidy, both of which are strictly prohibited by the LOAC.⁸⁹

While deepfake technology appears to have a historical basis in the LOAC and is beneficial in several strategic ways, it may need to be heavily regulated, because while most uses of deepfakes in armed conflicts are legally allowable there is a concern that civilians may be unable to trust news sources and discern truth from the lies and misinformation. AP I may require that deepfakes be regulated so as to limit their impact on civilians.

C. Impact of Deepfakes on Civilians

As seen *supra*, deepfakes spread quickly through social media and their misinformation tends to linger for years because of their believability and the psychological confirmation bias they rely upon.⁹⁰ This has detrimental effects on civilians because it destroys the social contract of society.⁹¹ It inhibits civilian's ability to trust news sources, or to discern what is being used as a deepfake in an armed conflict context, or one that is an authentic message from their own military advising or appraising them of different situations. For example, State A creates a deepfake broadcast showing the leader of militant group B surrendering and announces that all fighting is to cease immediately. This broadcast would lure the militants out of their places of security, making them easier targets for State A. However, civilians who also see the deepfake, would probably leave their areas of security as well, placing more of them in danger, and within the range of State A's pending attack. This would likely violate the obligations established in AP I.⁹² While the use of deepfakes in most military contexts is permitted by

⁸⁸ Chesney & Citron, *supra* note 12, at 1808-9.

⁸⁹ ADDITIONAL PROTOCOL 1, Art. 37, 51.2.

⁹⁰ Carey, *supra* note 37; Meyer, *supra* note 36; Oritz-Ospina, *supra* note 51; See Brown, *supra* note 6, at 15 (explaining that the easiest way to disseminated disinformation is through social media platforms); Clement, *supra* note 51; *Fake News and Why it Matters*, SPEAKING OF PSYCHOLOGY (Aug. 21, 2019) (downloaded using iTunes).

⁹¹ *Social Contract Theory*, ETHICS UNWRAPPED, <https://ethicsunwrapped.utexas.edu/glossary/social-contract-theory> (last visited Apr. 1, 2020) (explaining that social contract is the idea “that people live together in society in accordance with an agreement that establishes moral and political rules of behavior”); See Alexander Yap, Lisa Snyder & Sherrie Drye, *The Information War in the Digital Society: A Conceptual Framework for a Comprehensive Solution to Fake News*, 3 ACAD. OF SOC. SCI. J.1214 (2018).

⁹² ADDITIONAL PROTOCOL I, *supra* note 16, at art. 57.

the LOAC, it may be necessary to forbid or greatly regulate them in order to avoid substantial harm to civilians and follow the provision to take constant care of civilians during military operations.

III. POTENTIAL CHANGES IN THE LOAC TO ADJUST TO USAGE OF DEEPFAKES

As seen in the prior sections, deepfakes can cause several problems. This section will review some areas of the LOAC that have yet to develop sufficiently in this area to interact with or handle deepfake technology and provide the necessary constant care to civilians. Several avenues for pursuing law remedies will also be presented.

A. Areas of the LOAC That Are Not Adequate to Handle Deepfake Technology

A deepfake may be created and disseminated by a third party who is not acting in the conflict.⁹³ The question then becomes if such a deployment of a cyber tool would constitute as an armed attack.⁹⁴ Unfortunately, the law is unclear on this point.⁹⁵ However, there is evidence that such a step by a third party would at least qualify as a *jus ad bellum* armed attack.⁹⁶

This example becomes more complex if the third party simply allows State A to post a deepfake—using the third party’s band width—as part of their attack against State B. There is no international law consensus of this in the LOAC. The Tallinn Manual states that “[u]sing a public, international and openly accessible network such as the internet for military purposes does not violate the law of neutrality. This is so even if it, or components thereof, is located in neutral territory.”⁹⁷ While this has yet to be codified as international law, recent state practice may support the idea that dispensing deepfake materials through a third-party neutral’s territory may not disqualify the third party’s neutrality.⁹⁸ However, further state practice will be needed to establish a LOAC’s application of the cyber space usage in dispensing deepfakes.

⁹³ Chesney & Citron, *supra* note 12, at 1810; Johnson, *supra* note 2.

⁹⁴ “‘Attacks’” means acts of violence against the adversary, whether in offence or in defence.” Additional Protocol 1, art. 49.

⁹⁵ TALLINN MANUAL 2.0 ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS, 342 (Columbia University Press eds., 2nd ed. 2017) [hereinafter TALLINN 2.0].

⁹⁶ U.S. DEP’T OF DEF., DEPARTMENT OF DEFENSE LAW OF WAR MANUAL para 16.1.1 [hereinafter DOD MANUAL].

⁹⁷ TALLINN 2.0, *supra* note 88 at 556.

⁹⁸ *Id.*

Two inter-related areas of the LOAC that are not adequately established in relation to cyber technology and the creating or dispensing of deepfakes are the principles of distinction and indiscriminate attacks.⁹⁹ The LOAC does recognize psychological warfare and propaganda distribution through several types of platforms as legitimate military practices even if the intended audience are civilians.¹⁰⁰ The problem with deepfakes is that they are typically disseminated on such broad platforms that there is no discrimination and large numbers of civilians could be impacted by the deceitful content. The argument in the LOAC that most likely protects the use of deepfakes against distinction and indiscriminate attacks is the idea that deepfakes may not rise to a level of severity in which they are considered to be an “attack.” An attack is defined by AP I Article 49 as “acts of violence against the adversary, whether in offence or in defense.”¹⁰¹ If the LOAC standard for deepfakes was limited to an actual attack, regulation may be quite simple. For example if a deepfake were to incite riots, looting, or lead civilians to attack their own forces, then a deepfake may qualify as an attack. However, according to AP I Article 57, as quoted *supra*,¹⁰² there is the provision of constant care of civilians in presumably all military actions not simply attacks. Therefore, deepfakes probably will not qualify as an attack, but would qualify as a military action and will probably need regulation in order to protect civilians.

B. Alternatives in Handling Deepfakes in the Context of the LOAC

From the discussion *supra*, it is obvious that deepfakes could have the potential to be an incredible and potentially harmful tool in armed conflicts. They are a source of false information that can change the psychological perspective of leaders, soldiers, militants, and civilians.¹⁰³ How the LOAC approaches handling deepfakes now can have ripple effects for centuries of warfare. Solutions need to be created in order to help protect civilians and ensure that constant care and proportionality are followed.¹⁰⁴ Discouraging creation of harmful deepfakes is nearly impossible, as the typical international law remedy of naming and shaming appears to leave little

⁹⁹ ADDITIONAL PROTOCOL I, *supra* note 16, at art 49.

¹⁰⁰ TALLINN 2.0, *supra*, note 88, at 421

¹⁰¹ ADDITIONAL PROTOCOL I, *supra* note 16, at art 49.

¹⁰² *Id.* at art 57

¹⁰³ Jerry Taylor, *How to Change Public Opinion*, NISKANEN CTR. (Feb. 23, 2015), <https://www.niskanencenter.org/how-to-change-public-opinion/>; Gregory Elder, *Intelligence in War: It Can be Decisive*, Cent. Intelligence Agency, https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/csi-studies/studies/vol50no2/html_files/Intelligence_War_2.htm (last visited Apr. 9, 2020).

¹⁰⁴ See TALLINN 2.0, *supra* note 88, at 470; DoD MANUAL, *supra* note 89, para. 16.5.1.1.

impression on the creator and is a pitiful form of punishment, as demonstrated by Russia's condemned actions during the United States' 2016 elections where they have received minimal consequences.¹⁰⁵ This section will explore a few options on how the LOAC may handle deepfakes in this quickly changing forum of conflict in order to protect civilians.

1. Banning

Some countries, as mentioned *supra*,¹⁰⁶ have opted for banning deepfakes. The reasons for this ban vary from the harm they cause to the trouble, expense, and impossibility of discovering them before their harmful message has disseminated to the masses. While these reasons are valid, banning deepfakes in armed conflicts does not comport with its standing in customary international law.¹⁰⁷ Disinformation, ruse, and misinformation have been used for hundreds of years in war.¹⁰⁸ The problem that arises here is that the historical uses of misinformation were usually somewhat limited to the military. Deepfakes, on the other hand, will spread quickly over social media platforms and will be nearly impossible to limit to the military. So, if international law wanted to ban deepfakes in armed conflicts to prevent civilian harm, customary international law would need to change, and states would have to clearly state that they see it as a breach of the LOAC under the constant care provision of AP I, or a treaty could be drafted that clearly states that deepfakes are not permitted in armed conflicts.

The benefit to banning deepfakes in armed conflicts through treaties or customary international law are two-fold. Firstly, it protects the social contract of society. When people in the world cannot trust what they are seeing in the media to be factual, if it is being presented as such, then it will break down the trust within society. With that fail of trust, consequences such as no reliable informational sources, evidence might be unreasonably questioned, or vital warnings for civilian safety may be ignored. Secondly, there is precedence in the LOAC that banning methods of war that are unreasonably destructive or detrimental are permissible and advantageous. The 1925 Gas Protocol, the 1972 Biological Weapons convention, and the 1993 Chemical Weapons Convention¹⁰⁹ were established to prevent the

¹⁰⁵ Chesney & Citron, *supra* note 4. While this is not an armed conflict setting, it is a typical example that "naming and shaming" does not impact well-developed and powerful nations as much as it does on smaller developing countries.

¹⁰⁶ Note, *supra* note 13.

¹⁰⁷ Note, *supra* note 72-3.

¹⁰⁸ Note, *supra* note 71-3.

¹⁰⁹ 1925 Protocol for the Prohibition of the Use in War of Asphyxiating, Poisonous or Other Gases, and of Bacteriological Methods of Warfare, Documents on the Law of War, Feb. 8, 1928; 1972 Convention on the Prohibition of the Development, Production and

substantially harmful effects of chemical and biological weapons on all people as well as ensure that the constant care provisions were fulfilled once that standard was established.¹¹⁰ While deepfakes do not instantly do physical damage to people, the psychological effects of deteriorating trust and the reality people perceive may be enough to warrant new law in the LOAC to ensure the continued functionality of society.

Another option would be to use a partial ban. In a partial ban, parties could use deepfakes in military communications or operations, but the parties would not be allowed to release their deepfake on platforms that are easily accessible to civilian populations. This option would protect the civilian interests while still allowing some of the strategic capabilities of deepfakes by introducing an element of discrimination. However, a further concern would be that people in the military could share a deepfake over social media with their civilian friends causing the societal negative impacts discussed *supra*. Banning then becomes one of the safest ways to protect civilians and the social contract of the world.

There are good arguments for banning deepfakes from armed conflicts and war. However, another problem with banning deepfakes is that it may not solve the problem; if anything, it may exacerbate the situation because banning deepfakes could decrease funding from programs that are striving to develop effective methods to detect and combat deepfakes. Banning may be an unwise option because it would be an overbroad regulation prohibiting States from using the time-honored practices of deception. On the other hand, the societal repercussions of permitting the use of deepfakes unfiltered and unrestricted may cause unforeseen disintegrations of important social functions.

2. Platforms Filter

A second proposal that has been suggested in order to decrease a significant amount of the damage caused by deepfakes is to enact domestic laws or policies that social media platforms must create filters, detection technology, or flagging capabilities to mark deepfakes.¹¹¹ The argument is that because most deepfakes are spread through these platforms it would be

Stockpiling of Bacteriological (Biological) and Toxin Weapons and on Their Destruction, Mar. 26, 1975, U.N.T.S. 1015; 1993 Convention on the Prohibition of the Development, Production and Stockpiling of Chemical Weapons and on Their Destruction, available at https://treaties.un.org/pages/ViewDetails.aspx?src=TREATY&mtdsg_no=XXVI-3&chapter=26 (last visited Apr. 9, 2020).

¹¹⁰ ADDITIONAL PROTOCOL I, *supra* note 16, at art. 57.

¹¹¹ Note, *supra* note 57.

easy and efficient for the platforms to be the ones to monitor their content.¹¹² However, this is burdensome and compounded by the fact mentioned *supra* that detecting deepfakes is incredibly difficult. The top militaries in the world are struggling to keep up with the technologies used to produce deepfakes and expecting media platforms to be able to detect them may be unreasonable.¹¹³ Another disadvantage to pursuing this course is that perhaps some of the most dangerous deepfakes may not be dispensed through social media platforms, instead being used to target specific demographics or groups,¹¹⁴ and perhaps when platforms can identify deepfakes the promulgators of deepfakes will simply move to different platforms.

While it does seem reasonable at the outset to hold social media platforms more culpable for the content they have on their websites, such a standard would be nearly impossible to uphold and enforce without completely shutting down those platforms due to the sheer volume of content that is uploaded every minute of every day.¹¹⁵

3. Watermarking

Another option would be that real content—especially related to important news, political figures, or footage showing soldiers performing an act—should be filmed using a watermark.¹¹⁶ This requirement could easily become the standard through domestic law avenues or could quickly become customary international law by instituting it as a common practice. Watermarking video content when it is created can “produc[e] immutable

¹¹² *Id.*

¹¹³ Robitzski, *supra* note 59; J. Clement, *Hours of Video Uploaded to YouTube Every Minute May 2019*, STATISTA (Aug. 9, 2019), <https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/> (stating that there is more than 300 hours of content uploaded every minute, about 30,000 hours of content every hour, making filtering almost impossible).

¹¹⁴ Chesney & Citron, *supra* note 4 (arguing that deepfakes may be sent to target audiences as a form of “reputational sabotage.”)

¹¹⁵ Clement, *supra* note 106; See 37 *Mind Blowing YouTube Facts, Figures, and Statistics – 2020*, MerchDope (Feb. 26, 2020), <https://merchdope.com/youtube-stats/>; See also Dave Chaffey, *Global Social Media Research Summary 2019*, SMART INSIGHTS (Feb. 12, 2019), <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/>; Simon Kemp, *Digital Trends 2019: Every Single Stat You Need to Know About the Internet*, The Next Web (Jan. 30, 2019), <https://thenextweb.com/contributors/2019/01/30/digital-trends-2019-every-single-stat-you-need-to-know-about-the-internet/>.

¹¹⁶ Chesney & Citron, *supra* note 4; *Outsmarting Deep Fakes: Researchers Devise An AI-Driven Imaging System That Protects Authenticity*, TANDON SCHOOL OF ENGINEERING (May 29, 2019), <https://engineering.nyu.edu/news/outsmarting-deep-fakes-researchers-devise-ai-driven-imaging-system-protects-authenticity>.

metadata that marks location, time, and place and attests that the material was not tampered with.”¹¹⁷ While this option appears to be one of the more solid paths to prevent deep fakes, a foreseeable concern is the ability of learning technology to copy and recreate a watermark. Also, the potential costs of equipment that can watermark videos may be disproportionately high for the companies that would need to use them. However, for the moment, this may be the best way to authenticate footage thereby reducing the potential harms civilians may experience from deepfakes.

4. Alibis

Using a person’s electronic device to set up a viable alibi for soldiers, high profile political professionals, and government leaders may also be a way to mitigate the effects of deepfakes.¹¹⁸ Several companies, who are unaffiliated with governments, currently track the location of millions of phone users by pinging where people are by their phone signals.¹¹⁹ This does have privacy violation concerns, but it may have the interesting side effect of proving where a person is at a given time and what they may be doing. By using this type of data, deepfakes may be used in armed conflict situations because governments can quickly refute deepfake content by confirming the location of their soldiers or leaders at the time the supposed recording occurred because of the soldier or leader’s electronic device’s pinged location. This, however, would require the constant tracking and monitoring the movements of people. In an armed conflict setting, this may not be the best option because enemy specialist could hack the computer systems that store the locating data and get the exact location of all your military personnel. Another problem with relying on this type of regulation is that rebutting the deepfake in a timely manner will still be difficult, and substantial harm could occur during the interim from release to rebuttal.¹²⁰

Using technology to track the movements of military personnel may be an option, but not a safe one. Another problem could be that if the deepfake only uses people in the video who are on the other side, tracking their movements to verify this video actually was filmed by them would be complicated.

¹¹⁷ Chesney & Citron, *supra* note 4.

¹¹⁸ Chesney & Citron, *supra* note 4.

¹¹⁹ Stuart A. Thompson & Charlie Warzel, *Twelve Million Phones, One Dataset, Zero Privacy*, N.Y. TIMES (Dec. 19, 2019), <https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>.

¹²⁰ Carey, *supra* note 37.

5. Authenticated Platforms

A fifth suggested course to handle deepfakes, so they can still be used strategically but cause minimal damage to civilians and nations, is to have platforms that are authenticated. This would mean that the platform would be verified and that all content uploaded there would be as true and accurate as possible. It would most likely need to be set up through domestic law. This way civilians, and people seeking for verified, true and accurate information would have a place they could go to and rely on the content to not be fake.

This is a viable option; however, there are some elements that are concerning. A single authenticated platform may be dangerous because civilians would trust the content posted thereon, thereby making the platform subject to an increased number of cyberattacks. If an insurgent group could hack onto an authenticated platform and disseminate a deepfake, the damage done there might have greater effects than deepfakes posted on unauthenticated social media platforms. Another difficulty would be that deepfakes may lose their effectiveness for military operations because spreading misinformation could be made more difficult if people are looking to one or several specific websites to find what information is true. A third concern would be deciding who would run the authenticated platforms. Official news websites may vie for such a position; however, current news casters also struggle to identify fakes and, at times, spread them accidentally.¹²¹ Governments could be in charge of the authenticated platforms; however, this may raise concerns with the EU and the United States' constitution because it would be giving inordinate amount of power to governments, and it would closely parallel the way Hitler was able to disseminate Nazi ideology,¹²² which would cause concern to many civilians. But, the best argument for authenticated platforms would be that they could potentially recirculate a free market exchange of ideas because the information therein would have been previously authenticated.

C. The Best Options to Protect Civilians

The most persuasive options for the LOAC to use in approaching deepfakes are probably banning them, use watermarking technology, or using

¹²¹ John Bowers, Time Hwang, & Jonathan Zittrain, *What Should Newsrooms do About Deepfakes? These Three Things, for Starters*, NIEMANLAB (Nov. 20, 2019) <https://www.niemanlab.org/2019/11/what-should-newsrooms-do-about-deepfakes-these-three-things-for-starters/> (stating that journalists must double check where they get their information in order to prevent the spread of deepfakes).

¹²² *Nazi Propaganda and Censorship*, Holocaust Encyclopedia, <https://encyclopedia.ushmm.org/content/en/article/nazi-propaganda-and-censorship> (last visited Apr. 4, 2020).

an authenticated platform. There is strong historical legal precedence in the LOAC that suggests deepfakes may need to be banned. Banning this technology through international or domestic law would ensure that the social contract is upheld, and civilians are protected from the side effects of having no trustworthy source of information and news. However, with such an aggressive approach to deepfakes, the strategic value of them may outweigh the potential harmful impact on civilians. The law would also have to address if deepfakes are to be banned only in LOAC contexts, or if they need to be applied in a broader way through domestic law regulations. This may be a poor option as there are many benefits from deepfake technology that can help civilians with normal daily living.¹²³

While watermarking is a potential solution, the fast progression of technology makes the potential that watermarked videos can be easily faked a real possibility. The authenticated platform option does have a myriad of benefits and may be the preferable option. The main issues with the authenticated platform would be the need for excellent cyber security, and a clear discussion of who is running the platform and what is allowed to be posted.¹²⁴

IV. CONCLUSION

With the emergence of deepfake technology, several areas of the law will need to adapt. Deepfakes are a form of disinformation that present false video content that is believable, and, with the progression of technology, these deepfakes are close to being indiscernible from actual footage. While deepfakes are certainly lawful in military operations, and currently permitted under the LOAC, there is a concern that they may detrimentally impact civilians.¹²⁵ In order to fulfill the obligations in AP I to take constant care of civilians during military operations, States may need to establish regulations that will protect civilians from the harmful effects of deepfakes.

The most promising options for regulating deepfakes would be through banning certain types of deepfakes, limiting where and how they can be

¹²³ Brown, *supra* note 6, at 33 (describing how deepfakes can be used in a variety of ways including healthcare and entertainment); Chesney & Citron, *supra* note 12, at 1769-71 (explaining that deepfakes can be used as art, in education, and to assist those with disabilities); Lyrebird, AI, DESCRIPT, <https://www.descript.com/lyrebird-ai> (last visited July 10, 2019) (a website that allows people who have lost their voice through illness to use past recordings of their voice to say new things); See *United States v. Alvarez*, 567 U.S. 709, 729 (2012) (explaining that in order to preserve First Amendment rights speech of all kind must be permitted).

¹²⁴ Schroeder, *supra* note 4, at 14 (stating that democratic societies are concerned with importance of free speech and permitting a diversity of opinions and views).

¹²⁵ See Note, *supra* note 77.

disseminated during military operations, or use authenticated platforms with certifiably true information and news for the safety of civilians. Ultimately, the goal would be to ensure that all necessary care is taken in preserving and protecting civilians according to the standards established by customary international law.