VOLUME 3 November 2022 ISSUE 2

JOURNAL ON EMERGING TECHNOLOGIES



## ARTICLES

# DEEPFAKE FIGHT: AI-POWERED DISINFORMATION AND PERFIDY UNDER THE GENEVA CONVENTIONS

*Major D. Nicholas Allen*

# DEEPFAKE FIGHT: AI-POWERED DISINFORMATION AND PERFIDY UNDER THE GENEVA CONVENTIONS*

MAJOR D. NICHOLAS ALLEN**

*All that we are not stares back at what we are.*

*- W.H. Auden*[1]

## INTRODUCTION

On February 24, 2022, missiles began hitting major cities across the country: Kyiv, Kharkiv, Chernihiv.[2] Russian infantry, armor, mechanized fighting vehicles, mobile artillery, aviation, trucks, and supply assets charged over Ukraine's border at every point of the compass except west. The war the world feared for years would happen, that had actually *been* happening but on a smaller, deniable scale, started.[3]

---

* The views, opinions, and assertions provided in this article, notwithstanding those cited, are the views, opinions, and assertions of the author alone. This article does not necessarily reflect the views or positions of the United States Army, the Department of Defense, or the United States government.

**Judge Advocate, United States Army. Presently assigned as Chief of National Security Law, 25th Infantry Division, Schofield Barracks, Hawaii. L.L.M. in Military Law, 2021, The Judge Advocate General's School, United States Army, Charlottesville, Virginia; J.D., 2010, University of Baltimore School of Law; B.A., 2006, University of Florida. Previous assignments include Command Judge Advocate, United States Army Security Assistance Training Management Organization, Fort Bragg, North Carolina, 2018-2020; Defense Counsel, Fort Bragg Trial Defense Service Field Office, Fort Bragg, North Carolina, 2016-2018; Battalion Judge Advocate, 2nd Battalion, 3rd Special Forces Group (Airborne), Fort Bragg, North Carolina, 2014-2016; Trial Counsel, Fort Jackson, South Carolina, 2013-2014; Legal Assistance Attorney, Office of the Staff Judge Advocate, Fort Jackson, South Carolina, 2012-2013. Member of the bar of Maryland. The author wishes to thank the editors and staff of the Notre Dame Journal on Emerging Technologies as well as the myriad mentors, colleagues, and friends who assisted with this article. Most of all the author thanks his wife Anna and his children Jackson and Finley for their boundless love and support.

[1] W.H. AUDEN, THE SEA AND THE MIRROR 204 (1944).

[2] *See e.g.* Madeline Fitzgerald, *Russia Invades Ukraine: A Timeline of the Crisis*, U.S. NEWS & WORLD REP. (Feb. 25, 2022, 5:49 PM), https://www.usnews.com/news/best-countries/slideshows/a-timeline-of-the-russia-ukraine-conflict; John Psaropoulos, *Timeline: The First 100 Days of Russia's War in Ukraine*, AL JAZEERA (Jun. 3, 2022), https://www.aljazeera.com/features/2022/6/3/timeline-the-first-100-days-of-russias-war-in-ukraine.

[3] *Id.* Deniability was a key component of Russia's hybrid military involvement in Ukraine when it invaded the Crimean Peninsula in 2014, doing so with troops sent from its territory and armed with its weapons and equipment but lacking any

But the expected quick Russian victory did not materialize. In the following days the Ukrainian military fought harder and better than Russia had planned for, resulting in thousands of Russian troops killed, hundreds of Russian combat vehicles destroyed, and almost none of Russia's apparent major military objectives achieved.[4] Russian forces also slogged through self-inflicted logistics woes which further degraded Russian forces' abilities to maneuver, caused many Russian crews to abandon their vehicles across Ukraine, and quickly became a point of tremendous embarrassment for Russian military leaders.[5]

In the public relations sphere Russia would be in arguably its deepest hole. Worldwide condemnation of its invasion would feed an enormous sanctions regime,[6] a strengthening among NATO alliances as

---

identifying features or flags. The troops became known pejoratively around the world as "Little Green Men." Russian President Vladimir Putin eventually admitted the obvious shortly after his forces secured the Crimean Peninsula. *See e.g.* Silvia Aloisi & Frank Jack Daniel (eds.), *Timeline: The Events Leading up to Russia's Invasion of Ukraine*, REUTERS (Feb. 28, 2022, 11:03 PM), https://www.reuters.com/world/europe/events-leading-up-russias-invasion-ukraine-2022-02-28/; Vitaly Shevchenko, *"Little Green Men or Russian Invaders?",* BBC (Mar. 11, 2014), https://www.bbc.com/news/world-europe-26532154; Steven Pifer, *Watch Out for Little Green Men*, BROOKINGS (Jul. 7, 2014), https://www.brookings.edu/opinions/watch-out-for-little-green-men/. These same hybrid forces would also aid separatists in the eastern Ukrainian Luhansk and Donetsk regions during years of fighting against the armed forces of Ukraine prior to Russia's all-out invasion in 2022. *Id.*

[4] *See supra* note 2; *see also* Paul D. Shinkman, *Russia Abandons March on Kyiv, Focuses Embattled Troops Instead on Donbas*, U.S. NEWS & WORLD REP. (Mar. 25, 2022 at 3:29 PM), https://www.usnews.com/news/world-report/articles/2022-03-25/russia-abandons-Mar.-on-kyiv-focuses-embattled-troops-instead-on-donbas.

[5] *See supra* note 2; *see also* Anna Ahronheim, *Fuel and Logistics Problems Frustrate Russian Advance*, JERUSALEM POST (Feb. 27, 2022 at 2:39 PM), https://www.jpost.com/international/article-698800; Bonnie Berkowitz & Artur Galocha, *Why the Russian Military is Bogged Down by Logistics in Ukraine,* WASH. POST (Mar. 30, 2022 at 10:17 AM), https://www.washingtonpost.com/world/2022/03/30/russia-military-logistics-supply-chain/; Brad Lendon, *What Images of Russia's Trucks Say About its Military's Struggles in Ukraine*, CNN (Apr. 14, 2022 at 12:06 AM), https://www.cnn.com/2022/04/14/europe/ukraine-war-russia-trucks-logistics-intl-hnk-ml/index.html; Ann Marie Dailey, *What's Behind Russia's Logistical Mess in Ukraine? A US Army Engineer Looks at the Tactical Level*, ATL. COUNCIL (Mar. 21, 2022), https://www.atlanticcouncil.org/blogs/new-atlanticist/whats-behind-russias-logistical-mess-in-ukraine-a-us-army-engineer-looks-at-the-tactical-level/.

[6] *See* Chad P. Bown, *Russia's War on Ukraine: A Sanctions Timeline*, PETERSON INST. FOR INT'L. ECON. (Jul. 1, 2022 at 12:45 PM), https://www.piie.com/blogs/realtime-economic-issues-watch/russias-war-ukraine-sanctions-timeline; *see also List of Sanctions Against Russia After it Invaded Ukraine*, AL JAZEERA (Mar. 3, 2022 at 12:04 PM), https://www.aljazeera.com/news/2022/2/25/list-of-sanctions-on-russia-after-invasion.

well as potential expansion of NATO,[7] and the growth of Ukrainian President Volodymyr Zelenskyy as an international hero figure.[8] Even at home Moscow would have to confront a significant counter swell among the Russian people, leading Moscow to resort to Soviet-style tactics of mass arrests, severe free speech restrictions, and intimidations to suppress the dissent movement.[9]

On March 16, 2022, a new tactic emerged. Ukraine 24, a major television news network in Ukraine, broadcast a quixotic video of Ukrainian President Zelenskyy imploring his troops, not to push to victory, but to surrender.[10] In a motif similar to his daily press briefings and which would have been familiar to his daily viewers, President Zelenskyy appeared behind a podium with short-crop hair, a thin growth of beard, wearing an olive-green shirt, and with presidential symbols in the background. However, instead of his usual remarks encouraging

---

[7] *See supra* note 2; *see also Finland and Sweden Submit Applications to Join NATO*, N. ATL. TREATY ORG. (May 18, 2022 at 9:08 AM), https://www.nato.int/cps/en/natohq/news_195468.htm; A. Wess Mitchell, *Putin's War Backfires as Finland, Sweden Seek to Join NATO,* U.S. INST. OF PEACE (May 26, 2022), https://www.usip.org/publications/2022/05/putins-war-backfires-finland-sweden-seek-join-nato. While Türkiye initially opposed Finland and Sweden's joining NATO, significantly slowing full acceptance, Türkiye has since dropped its opposition by signing a tripartite agreement with Finland and Sweden which now paves the way for the two countries to become NATO's newest member states. George Wright, *Turkey Supports Finland and Sweden NATO Bid*, BBC (Jun. 29, 2022), https://www.bbc.com/news/world-europe-61971858.

[8] *See* Laura King, *Waging War, Wielding Words: Zelenksy's Speeches Have Made Him a Folk Hero*, LOS ANGELES TIMES (Mar. 16, 2022 at 1:28 PM), https://www.latimes.com/world-nation/story/2022-03-16/ukraine-zelensky-speeches-have-made-him-folk-hero; Nidhi Razdan, *Volodymyr Zelensky: From TV Star to War Hero*, NEW DELHI TELEVISION (Mar. 31, 2022 at 6:02 PM), https://www.ndtv.com/world-news/volodymyr-zelensky-from-tv-star-to-war-hero-full-transcript-2840813.

[9] *See* Courtney Subramaniam & Anna Nemtsova, *In Russia Thousands Defy Police Threats to Protest the Invasion of Ukraine. Can it Make a Difference?*, USA TODAY (Mar. 7, 2022 at 12:00 PM), https://www.usatoday.com/story/news/politics/2022/03/04/russia-ukraine-war-protests/9351061002/?gnt-cfr=1; Anton Troianovski & Valeriya Safronova, *Russia Takes Censorship to New Extremes, Stifling War Coverage*, THE NEW YORK TIMES (Mar. 4, 2022), https://www.nytimes.com/2022/03/04/world/europe/russia-censorship-media-crackdown.html; Marko Milanovic, *The Legal Death of Free Speech in Russia*, EUR. J. INT'L. L.: EJIL TALK! (Mar. 8, 2022), https://www.ejiltalk.org/the-legal-death-of-free-speech-in-russia/ (comparing current laws in Russia criminalizing the characterization of the Russian invasion of Ukraine as either an "invasion" or a "war" to similar laws from the Soviet Union).

[10] Bobby Allyn, *Deepfake Video of Zelenskyy Could be 'Tip of the Iceberg' in Info War, Experts Warn*, NPR (Mar. 16, 2022 at 8:26 PM), https://www.npr.org/2022/03/16/1087062648/deepfake-video-zelenskyy-experts-war-manipulation-ukraine-russia; Jane Wakefield, *Deepfake Presidents Used in Russia-Ukraine War*, BBC (Mar. 18, 2022), https://www.bbc.com/news/technology-60780142.

Ukrainians to remain strong and detailing his armed forces' needs to the world, President Zelenskyy claimed instead that "[b]eing the president was not so easy," that "[i]t didn't work out," "[t]here is no tomorrow," and finally "I advise you to lay down your arms and return to your families. It is not worth dying in this war."[11] A chyron also ran at the bottom of the news broadcast claiming that Ukraine had surrendered.[12]

News agencies and social media companies around the world sped to analyze the video and quickly determined that this realistic video was not actually real at all.[13] Instead it was the most recent employment of a still-young technology – a deepfake.

---

[11] Samantha Cole, *Hacked News Channel and Deepfake of Zelenskyy Surrendering is Causing Chaos Online,* VICE (Mar. 16, 2022 at 7:08 AM), https://www.vice.com/en/article/93bmda/hacked-news-channel-and-deepfake-of-zelenskyy-surrendering-is-causing-chaos-online (providing a rare uncommented version of the entire video). While the entire, unaltered video is otherwise difficult to find due to being removed from social media sites or being flagged for false content, a transcript in Ukrainian of the purported remarks is available on the Way Back internet archive. WAYBACK MACH., https://web.archive.org/web/20220316142015/https://u24.ua/ (last visited Jul. 5, 2022)(Ukrainian-to-English translation provided via Google translate and compared to translation provided in Cole, *id.*).

[12] Cole, *supra* note 11.

[13] *Id.*; *see also* James Pearson & Natalia Zinets, *Deepfake Footage Purports to Show Ukrainian President Capitulating*, REUTERS (Mar. 17, 2022 at 2:16 AM), https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/; Joshua Rhett Miller, *Deepfake Video of Zelensky Telling Ukrainians to Surrender Removed from Social Platforms*, THE NEW YORK POST (Mar. 17, 2022 at 12:20 PM), https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/; Tom Simonite, *A Zelensky Deepfake was Quickly Defeated. The Next One Might Not Be.*, WIRED MAG. (Mar. 17, 2022 at 1:30 PM), https://www.wired.com/story/zelensky-deepfake-facebook-twitter-playbook/.


[*Fig. 1.* Side-by-side stills contrasting the deepfake Zelensky video on the left with a genuine video of President Zelensky on the right making remarks at a news conference days prior.][14]

As of the writing of this article the video has had no discernible direct impact on the battlefield or Ukraine's war effort, likely due to its relatively poor quality.[15] But the confusion it sowed, even if temporary, provided immediate and worldwide effects in the information space[16] and demanded priceless time and attention from President Zelenskyy and members of his administration to rebut.

The episode remains a clarion call to those who contemplate the future of media manipulation and digital deception. The evolutionary march of digital deception leads straight to the battlefield, and few capabilities when at their highest potential are better primed to cause confusion and chaos in the battlefield's information space than deepfake technology.

"Deepfake" is the term associated with ultra-realistic video and audio images created not by human actors but by artificial intelligence. Originally associated with salacious pornography videos that depicted

---

[14] Images at Graham Cluley, *Deepfake President Zelensky Calls on Ukraine to Surrender, as TV Station Hacked*, BITDEFENDER (Mar. 17, 2022), https://www.bitdefender.com/blog/hotforsecurity/deepfake-president-zelensky-calls-on-ukraine-to-surrender-as-tv-station-hacked/.
[15] Simonite, *supra* note 13.
[16] Cole, *supra* note 11.

unwitting victims participating in sex acts,[17] people have used the technology to create perceptually perfect fake videos of such figures as President Barack Obama, celebrities like Emma Watson and Nicolas Cage, or even Russian President Vladimir Putin as early as 2018.[18] The technology has manipulated images of weather patterns and even depicted the life cycle of a daisy without needing human input for guidance.[19]

The Zelenskyy deepfake is also not the first time that a deepfake has made a mark during a time of crisis. In 2019, a deepfake-caused crisis instigated an attempted coup in Gabon, which nearly caused a civil war.[20] Supporters of Gabonese President Ali Bongo Ondimba became convinced that, after the President had not been seen for several days, a video purporting to show President Bongo alive, astute, and on the job was not real but instead was a deepfake. In support of this assumption, citizens pointed to differences in the President's demeanor, physical appearance, his apparent inability to use a hand, and even raised skepticism about the video's lighting.[21] Local newspapers had also speculated about deepfake, and on January 7, 2019, military officers from

---

[17] *See* Thanh Thi Nguyen, et al., Deep Learning for Deepfakes Creation and Detection 2 (Jul. 28, 2020, 17:54 UTC), https://arxiv.org/pdf/1909.11573.pdf; Yisroel Mirsky & Wenke Lee, The Creation and Detection of Deepfakes 1-2 (Sep. 13, 2020, 22:44 UTC), https://arxiv.org/pdf/2004.11138.pdf. This derogatory use of deepfake technology has caused significant harm for hundreds if not thousands of victims since its inception. However, this impact is beyond the scope of this article. For a devoted analysis of deepfake technology and its role in revenge pornography or other related victimizing activities, *see, e.g.*, Nina I. Brown, *Deepfakes and the Weaponization of Disinformation*, 23 VA. J.L. & TECH. 1, 7-8 (2020); Danielle Citron, *Sexual Privacy*, 128 YALE L.J. 1870, 1898-1902 (2019) (detailing how nonconsensual deepfake pornography videos violate sexual privacy rights); Rebecca A. Delfino, *Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act*, 88 FORDHAM L. REV. 887, 895-99 (2019) (discussing the ways that deepfake pornography is used, the harm it causes, and the problems with finding recourse in the law for victims); Russell Spivak, *'Deepfakes': The Newest Way to Commit One of the Oldest Crimes*, 3 GEO. L. TECH. REV. 339, 345-48 (2019) (discussing the history of deepfake proliferation from a Reddit user who posted the first deepfake videos to use in nonconsensual pornographic content to comparatively benign modifications of movie and television clips, and describing how private companies financially benefit from evolutions in media manipulation).

[18] *See* Bloomberg Quicktake, *It's Getting Harder to Spot a Deepfake Video* (Sep. 27, 2018), https://www.youtube.com/watch?v=gLoI9hAX9dw/

[19] *Id.*

[20] *See* Sarah Cahlan, *How Misinformation Helped Spark an Attempted Coup in Gabon*, WASH. POST (Feb. 13, 2020, 3:00 AM), https://www.washingtonpost.com/politics/2020/02/13/how-sick-president-suspect-video-helped-sparked-an-attempted-coup-gabon/; Ali Breland, *The Bizarre and Terrifying Case of the 'Deepfake' Video that Helped Bring an African Nation to the Brink*, MOTHER JONES (Mar. 15, 2019), https://www.motherjones.com/politics/2019/03/deepfake-gabon-ali-bongo/.

[21] Breland, *supra* note 20.

the Gabonese armed forces attempted a coup d'état by forcibly seizing a broadcast station and sending messages in an effort to "restore democracy."[22]

While the coup did not succeed[23] and the video was most likely real,[24] the impact of the episode is enough to give skeptics of deepfake manipulation further pause. No actual manipulation was necessary. Deepfake technology's existence alone brought the country to the edge of non-international armed conflict.[25]

With media manipulation at such new heights, international actors must not neglect its technical and legal impact on the battlefield. This Article therefore attempts to assess the current state of deepfake technology, look ahead to its potential future applications in armed conflict, process the ways in which current law contemplates such deception, and distill recommendations for improving governance where needed.

First, the Article will examine the origins of media manipulation and warfare in order to provide context for the later analysis of where deepfake deception fits in today's information arsenal. Second, the Article will detail the current state of deepfake technology. This discussion will explore the technology's structural roots, in both variational autoencoders and the more popular method via generative

---

[22] The Associated Press, *Gabon's Government Quashes Coup Attempt, Killing 2, Officials Say*, CBC (Jan. 7, 2019, 2:00 AM), https://www.cbc.ca/news/world/gabon-coup-attempt-1.4968177.

[23] Two of the officers were killed in a resulting raid and the others captured. *Id.*

[24] The President as it turns out had suffered a stroke and needed treatment, both of which likely explained his differences in mannerisms and appearance. Cahlan, *supra* note 20; *see also* Janosch Delcker, *Welcome to the Age of Uncertainty*, POLITICO (Dec. 17, 2019, 7:50 PM), https://www.politico.eu/article/deepfake-videos-the-future-uncertainty/. In a cryptic follow-up, later analysis of the video concluded both that the video was "likely" real but also could not rule out that it still could have been a deepfake. *Id.*

[25] While threatening to expand into a non-international armed conflict, this episode would not likely qualify as one under the Tadic Factors as the conflict, while involving a clash between government forces and an armed, uniformed, organized non-governmental force, was not "protracted," having started and ended in a day. *See Prosecutor v. Tadić*, Case No. IT-94-1, Decision on Defence Motion for Interlocutory Appeal on Jurisdiction, ¶ 70 (Int'l Crim. Trib. for the Former Yugoslavia Oct. 2, 1995) (finding that an armed conflict exists "whenever there is a resort to armed force between States or protracted armed violence between governmental authorities and organized armed groups or between such groups within a State."). Furthermore, in finding that the conflict in the Balkans qualified as "protracted," the International Criminal Tribunal for the former Yugoslavia observed that conflict between State and non-State forces had existed for years and involved "large-scale violence." *Id.* Neither of those facts presented in Gabon, though nothing about deepfake technology mitigated those possibilities.

adversarial networks, to show how deepfake technology can be complex yet accessible to trends and expected future advances. Third, the Article will detail abilities and limits for detecting deepfake manipulations and will analyze methods for determining attribution for an act of deepfake-derived deception both in a technological sense and a legal sense. Fourth, the Article will discuss the laws that may impact uses of deepfake technology in armed conflict. This discussion will look chiefly through a *jus in bello* lens to confront the conflict that arises when international humanitarian laws which permit misinformation may have to thwart misinformation. Fifth, the Article will distinguish uses of deepfake manipulation that would require enforcement of the laws against perfidy or violations of honor from uses which would qualify as lawful ruse. Finally, the Article will conclude with recommendations on how to improve the governance of deepfake technology even as the technology continues to evolve and its deception capabilities become sharper.

## I. MEDIA MANIPULATION AND WAR

### *A. Genesis and the First Fake*

In 1838, Louis Daguerre captured the first photograph of a human[26] – accomplished almost by accident. Attempting to use his photography process to capture a picture of a Parisian street, he could not capture humans or any other mobile items such as horse carriages because his process required seven minutes of light exposure and seven corresponding minutes of no movement. Apparently unaware that the photograph was happening, nobody on the street had any reason to stand still that long. Nobody except, as luck would have it, a distant man standing at a corner having his shoes shined (the shoe-shiner would be captured as well).[27] This photograph, and other similar tin-plate "daguerreotypes" that followed, were revolutionary, heralded at the time

[26] Adam Withnall, *This is the First Ever Photograph of a Human – and how the Scene it was Taken in Looks Today*, INDEP., (Nov. 5, 2014, 4:45 PM), https://www.independent.co.uk/news/world/world-history/first-ever-photograph-human-and-how-scene-it-was-taken-looks-today-9841706.html; Robert Krulwich, *First Photo of a Human Being Ever?*, NAT'L PUB. RADIO, (Oct. 25, 2010, 10:17 AM), https://www.npr.org/sections/krulwich/2011/03/31/130754296/first-photo-of-a-human-being-ever (comparing the 1838 daguerreotype photograph with an 1848 photograph made in Cincinnati, Ohio).

[27] *See* Withnall, *supra* note 26.

for their "truthful likeness,"[28] and soon Mr. Daguerre would seek official recognition of his direct positive photographic printing process from the French Academy of Sciences.[29]

Mr. Daguerre, however, had a rival. Hippolyte Bayard was a fellow Frenchman who created his own photography process while Louis Daguerre was developing his.[30] Mr. Bayard hoped to beat Mr. Daguerre and achieve recognition from the French Academy of Sciences as the first claimant to the direct positive photographic printing process. When, however, Mr. Daguerre instead submitted his work in the first week of 1839 on what would become known as the daguerreotype process, he beat Mr. Bayard, torpedoing Mr. Bayard's ambitions and relegating him to the status of a follow-behind.[31]

Severely chafed and eager to continue to prove himself, Mr. Bayard chose to pioneer a different kind of first – the first fake photograph. It was morbid. In his 1840 photograph entitled "Self Portrait as a Drowned Man,"[32] Mr. Bayard spliced a self-portrait of his

---

[28] LIBR. OF CONGRESS, THE DAGUERREOTYPE MEDIUM, https://www.loc.gov/collections/daguerreotypes/articles-and-essays/the-daguerreotype-medium/ (last visited Oct. 20, 2020).

[29] *See* LOUISE JACQUES MANDÉ DAGUERRE, HISTORY AND PRACTICE OF PHOTOGENIC DRAWING ON THE TRUE PRINCIPLES OF THE DAGUERRÉOTYPE, WITH THE NEW METHOD OF DIORAMIC PAINTING 1-6 (J.S. Memes, LL.D. trans., Smith, Elder and Co. ed. 1839) (also available online at https://archive.org/details/historyandpract00memegoog/page/n8/mode/2up (Jul. 15, 2008 at 10:12 AM)) (detailing the submission made to the French Academy of Sciences as well as both the acceptance of the submission by the French government and the purchase of the process from Mr. Daguerre); *see also* Randy Alfred, *Aug. 19, 1839: Photography Goes Open Source*, WIRED, (Aug. 19, 2010, 7:00 AM) (discussing Louis Daguerre's advancement of direct positive photography and his efforts to have the process officially acknowledged and shared, resulting in the publication of his work in Aug. of 1839).

[30] Michal Sapir, *The Impossible Photograph: Hippolyte Bayard's "Self-Portrait as a Drowned Man"*, 40 MOD. FICTION STUD., no. 3, 1994, at 619-29. It should also be noted that William Henry Fox Talbot was also simultaneously working in England to develop his own photographic process and that Mr. Talbot, though not within the same professional circles as Mr. Daguerre and Mr. Bayard, was also a contemporary competitor of Mr. Bayard at the French Academy of Sciences that year. However, Mr. Bayard's follow-on actions appear to have been most influenced by his disappointment in his competition against Mr. Daguerre. *Id*; *see also* THE GETTY MUSEUM, HIPPOLYTE BAYARD, http://www.getty.edu/art/collection/artists/1840/hippolyte-bayard-french-1801-1887/ (last visited Oct. 20, 2020).

[31] *Id*.

[32] *Id*. *See also* Sean O'Hagan, *Exposed: Photography's Fabulous Fakes*, THE GUARDIAN (Jan. 31, 2016, 1:00 PM), https://www.theguardian.com/artanddesign/2016/jan/31/exposed-photographys-fabulous-fakes (comparing the Bayard fake suicide photograph to later examples of faked photographic images); Michael Zang, *The First Hoax Photograph Ever Shot*,

face, eyes closed and cheeks lifeless, on to a different self-portrait of his pale upper torso and darkened hands, appearing to show that he had committed suicide.[33] On the back of the picture was even a purported suicide note in which Mr. Bayard wrote "the poor wretch has drowned himself," that "he has been at the morgue for several days, and no-one has recognized him or claimed him," and warning the viewer that "you'd better pass along for fear of offending your sense of smell . . . the face and hands of the gentleman are beginning to decay."[34]

While Mr. Bayard, who had not committed suicide, made the photograph as an expression of protest and not as an attempt to fake his own death,[35] his work has served as a predecessor for media manipulation. From nineteenth century presidential touch-ups and face-swaps,[36] to twentieth century fairies,[37] to historical re-writes,[38] to twenty-

---

PETAPIXEL (Nov. 15, 2012), https://petapixel.com/2012/11/15/the-first-hoax-photograph-ever-shot/.

[33] *See* Sapir, *supra* note 30.

[34] Quotes translated from the original French. *Id.*

[35] Mr. Bayard would actually go on to experience significant professional success and renown in the field of photographic technology, earning several accolades during his lifetime including in 1863 the *Légion d'honneur* – the highest award that can be bestowed in France. However, his fake suicide photograph has dominated his legacy. *See* Getty Museum, *supra* note 30.

[36] *See e.g.* Michael Waters, *The Great Lengths Taken to Make Abraham Lincoln Look Good in Portraits*, ATLAS OBSCURA (Jul. 12, 2017), https://www.atlasobscura.com/articles/abraham-lincoln-photos-edited (discussing efforts to make President Lincoln appear more virulent to the public during his 1860 presidential campaign by splicing a picture of his face on to the more commanding posture of John C. Calhoun).

[37] The "Cottingley Fairies" was a series of photographs taken in 1917 depicting two young girls, Frances Griffiths and Elsie Wright, playing with winged fairies. The girls made the photographs after the younger girl, Frances (then nine years old), had claimed that she actually had played with fairies in her garden but was not believed. The method of the trick was simple – the girls made hand-drawn cutouts of fairies, stuck them in the ground with hatpins, posed with them, and took the pictures. While it's questionable whether they intended for the photographs to be seen as real, their photographs eventually circulated widely among local societies and in the local news. They even grabbed the attention of famed author Sir Arthur Conan Doyle who wrote a book in defense of the photographs' authenticity. Unfortunately for the reputation of all involved, however, Elsie would confess shortly before her death in the 1980s that the photographs were fake. Hazel Gaynor, *Inside the Elaborate Hoax that made British Society Believe in Fairies*, TIME (Aug. 1, 2017, 9:15 AM), https://time.com/4876824/cottingley-fairies-book/; *see also* SIR ARTHUR CONAN DOYLE, COMING OF THE FAIRIES 13, 196 (1922).

[38] Fourandsix Technologies hosts a webpage entitled "Photo Tampering Throughout History" which provides an in-depth image-based historical profile of famous fake or doctored photographs. Several images reside there of political leaders, such as Joseph Stalin and Mao Tse-Tung, 'erasing' or removing disfavored people posing with the political leader from photographs after the individual fell out of favor with the leader. PHOTO TAMPERING THROUGHOUT HISTORY, http://pth.izitru.com/ (last visited Oct. 21, 2020).

first century Instagram,[39] deceptions in visual media have exploded from the product of a gifted few to an output today so large that by some estimates at least half—if not more—of internet content is artificially created by one means or another.[40] Furthermore, editing and doctoring have evolved from being visually distinct to virtually indistinguishable absent a dedicated professional forensic investigation or a happenchance sloppy edit.[41]

### *B. Photo Fraud Goes to War*

Image and audio manipulation have been a part of war ever since photographers first lugged their equipment to the ravaged battlefields of the Crimean War in 1854. British photographer Roger Fenton, widely acknowledged to be the first war photographer for his work during that war, has been accused of staging his photograph "The Valley of the Shadow of Death," taken after the 1854 Battle of Balaclava, by pre-positioning cannonballs to make the shot more dramatic.[42] Scrutiny has also come down upon famed American Civil War photographers Alexander Gardner and Matthew Brady who purportedly captured the human wreckage at Antietam and Gettysburg but who also allegedly

---

[39] Today, Instagram, a photograph sharing platform, is ubiquitous with modern-day photograph fakes and forgeries where an entire cottage industry has bloomed of self-styled influencers earning income in many cases by having photographs of themselves either altered or invented entirely in order to earn followers. *See e.g.* Janine Puhak, *Instagram Influencer Slammed for 'Fake Traveling' Photos*, FOX NEWS (Dec. 19, 2018), https://www.foxnews.com/travel/instagram-star-slammed-for-fake-traveling-photos.

[40] *See* Max Read, *How Much of the Internet is Fake? Turns out, a Lot of It, Actually*, N.Y. MAG. (Dec. 26, 2018), https://nymag.com/intelligencer/2018/12/how-much-of-the-internet-is-fake.html.

[41] In their sweeping examination of deepfake technology implications, Professors Danielle Citron and Robert Chesney explain how digital forensic efforts to detect fake images have become more and more difficult, noting that the "field of digital forensics has been grappling with the challenge of detecting digital alterations for some time." Danielle K. Citron & Robert Chesney, *Deepfakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CAL. L. REV. 1753, 1759 (2019). This increasing difficulty has long been forecast. *See e.g.* Hany Farid, *Digital Forensics: How Experts Uncover Doctored Images*, SCI. AM. (Jun. 1, 2008), https://www.scientificamerican.com/article/digital-image-forensics/ (observing in 2008 that "today anyone with a computer can readily produce fakes that can be very hard to detect").

[42] *See* MUSÉE D'ORSAY, ROGER FENTON: THE VALLEY OF THE SHADOW OF DEATH, https://www.musee-orsay.fr/en/collections/works-in-focus/photography/commentaire_id/the-valley-of-the-shadow-of-death-16457.html?tx_commentaire_pi1%5BpidLi%5D=847&tx_commentaire_pi1%5Bfrom%5D=844&cHash=1613936201 (last visited Oct. 21, 2020) (discussing the nature of the allegation but dismissing it outright).

moved and propped up bodies in an effort to make the destruction of the war appear more gruesome, or their photographs appear more contemporaneous to the fight.[43]

Today's battlefields have been no exception. Aside from the examples from Ukraine and Gabon discussed earlier, China has been accused of creating false and incendiary content when one of its Twitter accounts posted a fabricated image of an Australian soldier slitting the throat of an Afghan child during the later years of Australia's fight in Afghanistan.[44] North Korea and Iran have also both in recent years distributed photographs purporting to demonstrate larger forces of landing craft[45] and missile launchers,[46] respectively, than they actually possessed. Consider also the 2014 case of the Associated Press having to sever ties with an esteemed combat photographer after editors discovered that the photographer had improperly altered images of an anti-Assad regime fighter in Syria.[47]

Now, thanks to the ever-increasing sophistication of artificial intelligence, technological capabilities to create fake content have experienced a bullet-speed rise in complexity and efficacy. As programmers and developers worldwide have competed voraciously to

---

[43] *See* Michael E. Ruane, *Alexander Gardner: The Mysteries of the Civil War's Photographic Giant*, WASH. POST (Dec. 23, 2011), https://www.washingtonpost.com/local/alexander-gardner-the-mysteries-of-the-civil-wars-photographic-giant/2011/12/12/gIQAptHhDP_story.html.

[44] Zhao Lijian (@zlj517), TWITTER (Nov. 29, 2020, 8:02 PM), https://twitter.com/zlj517/status/1333214766806888448. The tweet was sent by Mr. Zhao Lijian, deputy director of the Information Department of the Chines Ministry of Foreign Affairs. The tweet came on the heels of the Brereton Report conducted by the Australian government which detailed, among other things, apparent unlawful killings by its own troops in Afghanistan. The Australian government called the tweet "utterly outrageous" and demanded an apology which the Chinese government refused to provide, causing further strain in the countries' relationship. Kirsty Needham, *Australia Demands Apology from China After Fake Image Posted on Social Media*, REUTERS (Nov. 29, 2020, 9:59 PM), https://www.reuters.com/article/us-australia-china/australia-demands-apology-from-china-after-fake-image-posted-on-social-media-idUSKBN28A07Y.

[45] *See* Alan Taylor, *Is This North Korean Hovercraft-Landing Photo Faked?*, THE ATLANTIC (Mar. 26, 2013), https://www.theatlantic.com/photo/2013/03/is-this-north-korean-hovercraft-landing-photo-faked/100480/; Damien Mcelroy, *North Korea 'Photoshopped' Marine Landings Photograph*, THE TELEGRAPH (Mar. 27, 2013), https://www.telegraph.co.uk/news/worldnews/asia/northkorea/9956422/North-Korea-Photoshopped-marine-landings-photograph.html.

[46] *See* Adam Hadhazy, *Is that Iranian Missile Photo a Fake?*, SCI. AM. (Jul. 10, 2008), https://www.scientificamerican.com/article/is-that-iranian-missile/; David Folkenflik, *On the Smokey Trail of a Faked Missile Photo*, NAT'L PUB. RADIO (Jul. 11, 2008, 1:07 PM), https://www.npr.org/templates/story/story.php?storyId=92454193.

[47] *See* Associated Press, *AP Severs Ties with Photographer who Altered Work*, ASSOCIATED PRESS ONLINE (Jan. 22, 2014), https://www.ap.org/ap-in-the-news/2014/ap-severs-ties-with-photographer-who-altered-work.

improve artificial intelligence, they have made simultaneous advances in how artificial intelligence learns and performs. These advances, as explained below, have the battlefield poised for serious complexities.

## II. THE TECHNOLOGY BEHIND DEEPFAKE

### *A. Defining the Device*

To understand deepfake technology and thereby understand its legal ramifications, it is important to first understand what deepfake technology is not. Deepfake media are not works of total human invention. Unlike the copy-pasting of a missile battery, deepfake media does not necessitate human decisions at all stages.

What distinguishes deepfake media from other variants of falsified images, and what makes their nature so convincing, is that they are self-correcting. Deepfake technology is mathematically engineered from and through artificial intelligence. In particular, deepfake technology is a consequence of machine learning. Machine learning, defined generally as the ability of a computer to solve a problem without being explicitly programmed,[48] can take such primitive forms as a 1642 hand-dialed device that calculated taxes.[49] The earliest modern mathematical models for defining and developing machine learning explored the game of checkers to determine whether an IBM computer could learn from and defeat a human opponent. It did.[50] The next natural step was to see if an IBM computer could learn from and defeat a human opponent at chess. It did.[51]

---

[48] *See* JOHN R. KOZA ET AL., *Automated Design of Both the Topology and Sizing of Analog Electrical Circuits Using Genetic Programming*, *in* ARTIFICIAL INTELLIGENCE IN DESIGN '96 151, 153 (John S. Gero & Fay Sudweeks eds., 1996) (paraphrasing the work of Arthur Lee Samuel, the inventor of modern machine learning applications); *see also* Arthur L. Samuel, *Some Studies in Machine Learning Using the Game of Checkers*, 3 IBM J. 211-29 (1959).

[49] Pascal's Arithmetic Machine, also known as the Pascaline, was an early calculator invented by French mathematician Blaise Pascal in 1642. Designed to help tax collectors like the inventor's father, it required Mr. Pascal to implement several mathematical equations into the Pascaline's design so that the device could produce accurate, arithmetically-derived tax figures with the simple turning of a few dials. *See* Paul A. Freiberger & Michael R. Swaine, *Pascaline*, ENCYCLOPAEDIA BRITANNICA (Apr. 26, 2019), https://www.britannica.com/technology/Pascaline.

[50] Samuel, *supra* note 42; *see also* Bernard Marr, *A Short History of Machine Learning – Every Manager Should Read*, FORBES (Feb. 19, 2016, 2:31 AM), https://www.forbes.com/sites/bernardmarr/2016/02/19/a-short-history-of-machine-learning-every-manager-should-read/#468739a915e7.

[51] Marr, *supra* note 50.

Today, machine learning has grown into several sub-disciplines, each guided in large part by algorithm design and designer intent. For example, logistic regression has helped as early as 1990 to recommend cesarean deliveries based on patient data provided by physicians.[52] Another algorithm, known as Naive Bayes, can help sort desirable emails from spam emails.[53] Algorithm-based programs such as these, however, rely on "representations," that is to say, collections of information provided by human input (whether a computer programmer or a user checking their email inbox)[54] which communicates within the algorithms what right looks like.[55] In other words, machine learning in these contexts continues to require human hand-holding.

While such a fact is not inherently problematic, it has, in some sense, posed a barrier to more advanced machine learning. From this conundrum came deep learning. The concept is cogently explained by researchers at the Massachusetts Institute of Technology who pioneered certain advances in machine learning, explaining:

> "The hierarchy of concepts enables the computer to learn complicated concepts by building them out of simpler ones. If we draw a graph showing how these concepts are built on top of each other, the graph is deep, with many layers. For this reason we call this approach to AI deep learning."[56]

An artificial intelligence designed to build and perfect images based on an algorithmic infrastructure that through multiple efforts generates *its own* representations (as opposed to constantly requiring human inputs) demonstrates deep learning. Amazon's Alexa AI, for example, employs deep learning through Google's proprietary Natural Language Processing program that enables Alexa to swiftly scan virtually all recorded words in the English language in order to improve how it receives and responds to a person's command.[57] This way, if Alexa AI

---

[52] IAN GOODFELLOW ET AL., DEEP LEARNING 3 (2016).
[53] *Id.*
[54] Also known as a "feature." *Id.*
[55] *Id.* at 4.
[56] *Id.* at 2.
[57] *See* Alexandre Gonfalonieri, *How Amazon Alexa Works? Your Guide to Natural Language Processing (AI)*, TOWARDS DATA SCI. (Nov. 21, 2018), https://towardsdatascience.com/how-amazon-alexa-works-your-guide-to-natural-language-processing-ai-7506004709d3; Brian Barrett, *The Year Alexa Grew Up*, Wired (Dec. 19, 2018, 10:00 AM), https://www.wired.com/story/amazon-alexa-2018-machine-learning/ (detailing how Alexa's NLP enables it to find a radio station when a person requests it by a station nickname).

issues an incorrect return the first time, it could issue a correct return on a second or third attempt without any human command to make a different attempt.[58] Thus, when deep learning began to enable artificial intelligence to fabricate media, the term "deepfake" grew from a recognition of the role of deep learning in the creation of otherwise unreal or nontruthful media.[59]

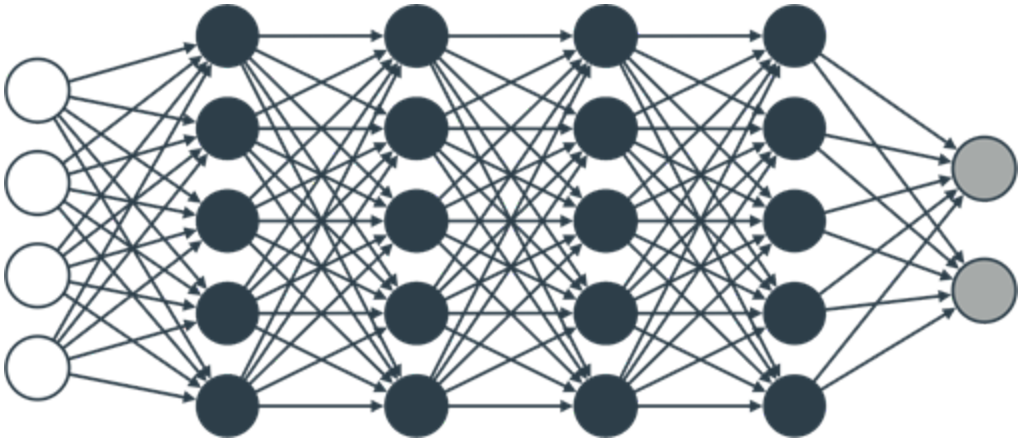
That the algorithmic function generates without human input, much less corrects without human input, is what fundamentally distinguishes deepfake from other methods of fabrication. How this occurs lies in the most basic component of information-gathering–the node –and the most basic component of computer activity.

### *B. Building Blocks*

Merriam-Webster defines a "node" *inter alia* as a point at which other parts originate or center.[60] In the field of computer science, a node is, at its essence, a point of information.[61] A node can be either a device, such as a phone or computer, or a point of information input, such as a year, hair color, or height. A network occurs when two or more nodes become connected.[62] Thus, for example, a computer connected to the internet forms at least one network with the computer being one node and the internet[63] another. Additional computer connections then branch from this original network. Computer scientists sometimes represent clusters of nodes in what are called "trees" due to the fact that nodes will subordinate from a primary node (also called a "parent node") in a fashion that graphically represents a tree.[64] As they grow in complexity and function, producing even rudimentary thought patterns, these tree networks can be described as "neural networks," a nod to the similarly complex and hyper-connected nature of the human brain.[65]

---

[58] *Id.*
[59] *See* Riana Pfefferkorn, *"Deepfakes" in the Courtroom*, 29 B.U. PUB. INT. L. J. 245, 246 (2020) (describing the term "deepfake" as a "portmanteau of 'deep learning' and 'fake'.").
[60] *Node*, MERRIAM-WEBSTER ONLINE DICTIONARY, https://www.merriam-webster.com/dictionary/node (last visited Oct. 22, 2020).
[61] *See* BRIAN HARVEY & MATTHEW WRIGHT, SIMPLY SCHEME: INTRODUCING COMPUTER SCIENCE 299 (2nd ed. 1999); *see also* COMPUTER BUSINESS REVIEW, WHAT IS A NODE?, https://techmonitor.ai/what-is/what-is-a-node-4927877 (last visited Oct. 22, 2020).
[62] *Id.* (*see also* Harvey, *supra* note 61 at 306-07).
[63] Or, more accurately, servers hosting internet content.
[64] *See* Harvey, *supra* note 61 at 297.
[65] *See* Citron, *supra* note 41(citing Larry Hardesty, *Explained: Neural Networks*, MIT NEWS (Apr. 14, 2017), https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 (explaining that the term "neural network" was first coined as far back



[*Fig. 2*. A demonstrative representation from Dr. Luis Serrano of a multi-layered neural network. For example, this particular network features five horizontal layers, left-to-right, not counting the first column of "input" nodes. Note that the four dark columns are "hidden," meaning that a person interacting with this network would see the input (for example, a Google search request for a local restaurant) and the output (a website link to a local restaurant) but would not see the various interconnected networks operating to filter out incorrect returns and find a correct return.][66]

Neural networks are the central infrastructure of artificial intelligence, serving as highways and byways along which machine learning, more complex representation learning, and eventually deep learning, occurs. While heavy research focus on neural networks waned during the first decade of the twenty-first century,[67] intensity of interest renewed with the advent of better computer processing abilities.[68] Then in the second decade, leaps in artificial neural network interaction theory

as 1944 by researchers at MIT)). *See also* GOODFELLOW, *supra* note 52, at 13 (observing that the early efforts to develop neural networks termed these networks "artificial neural networks" directly due to researchers' intent on using said networks to better understand the function of the human brain).

[66] LUIS SERRANO, GROKKING MACHINE LEARNING Ch. 10, fig. 10.1 (2020), https://livebook.manning.com/book/grokking-machine-learning/chapter-10/v-13/1. *See also* Jason Brownlee, *How to Configure the Number of Layers and Nodes in a Neural Network*, MACHINE LEARNING MASTERY (Jul. 27, 2018), https://machinelearningmastery.com/how-to-configure-the-number-of-layers-and-nodes-in-a-neural-network/.

[67] *See* Hardesty, *supra* note 65(describing how interest in neural networks rose and fell repeatedly during the 20th and 21st centuries).

[68] *Id.* (noting that advances in video game performance particularly fueled improvements in computer processing abilities which set the conditions for a neural network resurgence).

set the conditions for the deepfake technology that exists and evolves today.[69]

### *C. Two Tales of Two Networks*

Leaps in artificial neural network interaction theory occurred in the evolution of variational autoencoders (VAEs) and most notably the pioneering development of generative adversarial networks (GANs).[70] Both disciplines use the relationship between two or often more networks to help train the networks to create a desirable output product, but in notably different ways.

#### 1. Variational Autoencoders

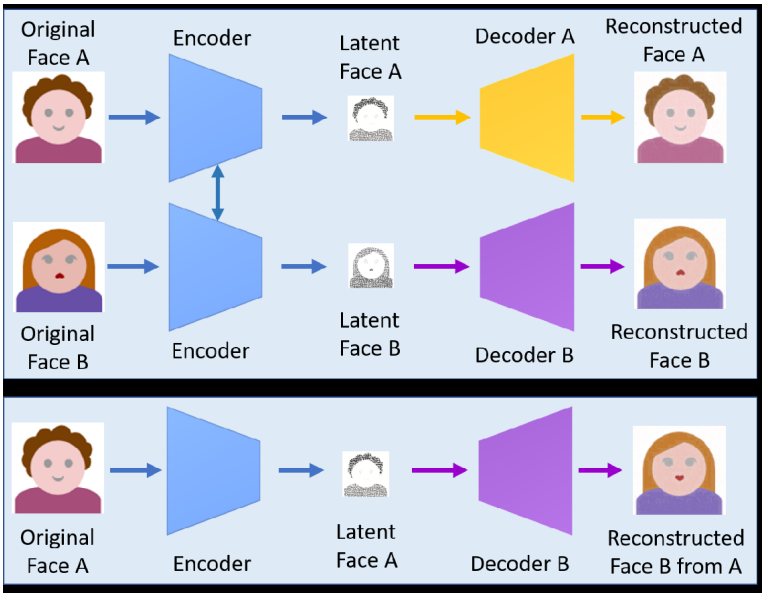
As alluded to in the introduction, the first widely known deepfake synthetic media creation was by a Reddit user who employed autoencoders to conduct a simple face swap to create pornographic content of female celebrities.[71] Today, given that most deepfake content relies on simple changes, such as face swaps, face editing, or face synthesis,[72] developers still often make deepfake content with autoencoders.

Autoencoders focus on two network players, an encoder and a decoder, which interact through an intermediary layer sometimes described as a "bottleneck" layer.[73] The encoder network receives an input, for example in the form of a picture of a person with dark hair (the source image).[74] The encoder identifies, categorizes, and condenses variables about that source image, such as jaw structure, hair color, lighting, etc. into the bottleneck. The decoder then extracts those variables from the bottleneck and recreates the source image. Once the autoencoder has accomplished this initial feat, the encoder then receives

---

[69] *Id.*; see also Michael Woolridge, A Brief History of Artificial Intelligence 139 (2020)(observing "In the second decade of the twenty-first century, AI has attracted more interest than any new technology since the World Wide Web in the 1990s.").
[70] Ian J. Goodfellow et al., Generative Adversarial Nets (Jun. 10, 2014, 6:58 UTC) (Neural Information Processing Systems conference paper), https://arxiv.org/abs/1406.2661; *see also* Citron, *supra* note 41, at 1760.
[71] *See* discussion *supra* note 17.
[72] *See* Mirsky, *supra* note 17 at 3; *see also* Andreas Rössler et al., FaceForensics++: Learning to Detect Manipulated Facial Images 1, 4 (Aug. 26, 2019, 17:59 UTC), https://arxiv.org/pdf/1901.08971.pdf.
[73] *See* Rössler, *supra* note 72at 14; *see also* Ben Dickson, *What are Deepfakes?*, TECHTALKS (Sep. 4, 2020), https://bdtechtalks.com/2020/09/04/what-is-deepfake/.
[74] *See* Nguyen, *supra* note 17 at 2; Rössler, *supra* note 72 at 14; Dickson, *supra* note 73.

a second input, for example a person with light hair (the target who, in the case of a face swap, the developer wants depicted in place of the face from the source image). The encoder, with some degree of guidance from the developer, distills variables about the target image into the bottleneck layer where the source image variables still reside, and the compression of data mitigates margins of error. The decoder extracts variables from both images and then attempts to construct the goal synthetic image as exemplified here:




[*Fig. 3.* Graphical representation of synthetic media creation via an encoder-decoder pair. The goal fake image is at the bottom right.][75]

Autoencoders predate deepfake technology so this advent is not new. What accelerated these neural networks towards deepfake-level capacity were *variational* autoencoders (VAE).[76] Prior autoencoders required users to comb laboriously through sometimes thousands of images in order to find useful variables for decoder use.[77] VAEs, on the other hand, use probabilistic generative modeling, meaning the decoder tries to predict from the information available in the bottleneck layer what the goal hybrid image should be.[78] The result has been described

[75] The image is from Nguyen, *supra* note 17 at 3.
[76] *See* Lars Ruthotto & Eldad Haber, An Introduction to Deep Generative Modeling 22 (Mar. 9, 2021, 02:19 UTC), https://arxiv.org/pdf/2103.05180.pdf.
[77] *See* Dickson, *supra* note 73 (describing the process of selecting images from a video and cropping each one to just portray a face).
[78] Diedrick P. Kingma & Max Welling, An Introduction to Variational Autoencoders 28-30 (Dec. 11, 2019, 17:33 UTC), https://arxiv.org/pdf/1906.02691.pdf (describing how VAE training can develop an "*importance sampling* technique" [emphasis original] to assist with VAE inferences).

as "elegant" and "simple to implement."[79] However, VAEs can still demand a significant amount of time and data,[80] and they suffer from distinct image blurriness[81] which has made other avenues more attractive.

### 2. Generative Adversarial Networks

The creation of GANs, by comparison, has been a game-changer in media manipulation . Employing the analogy of the counterfeiter and the cop [82] imagine a counterfeiter is trying to sneak a counterfeit picture past a cop who is diligently looking out for counterfeit pictures. Being a first attempt, the counterfeiter's first efforts are rudimentary. When the cop obtains the picture, the cop easily determines that the picture is a fake and discards it. The counterfeiter, however, learns that the cop has detected faults in the picture. The counterfeiter determines to avoid those faults, generates a new picture that does not include those faults, and tries again. The process continues, the counterfeiter removing one detected fault from the creation process after another, until the counterfeiter has removed so many faults that the cop can no longer detect the difference between an authentic picture and a fake picture.

Generative adversarial networks operate in the same way. A GAN consists essentially of a pair of neural networks that compete against each other.[83] One network, termed a "generator,"[84] will act as the counterfeiter, generating information that it has manufactured. The other network, termed a "discriminator,"[85] will act as the cop, filtering out information that does not match the parameters set for authenticity. A programmer will build the discriminator network first. In the process, the programmer will define the properties that characterize an authentic

---

[79] *See* Goodfellow, *supra* note 52 at 688.
[80] *See* Matthew Stewart, *GANs vs. Autoencoders: Comparison of Deep Generative Models*, TOWARDSDATASCIENCE (May 12, 2019), https://towardsdatascience.com/gans-vs-autoencoders-comparison-of-deep-generative-models-985cf15936ea. However, databases have proliferated online to facilitate such data collection. CelebFaces Attributes Dataset, for example, contains over 200,000 face images of over 10,000 public figures. *Id.*
[81] *See id.*; Kingma, *supra* note 78 at 32.
[82] This analogy is most commonly associated with Mr. Ian Goodfellow, an often-credited trailblazer of GAN development who also uses the analogy often to illustrate the concept. *See* Ian Goodfellow, *Introduction to GANs, NIPS 2016 | Ian Goodfellow, OpenAI* (Aug. 24, 2017). https://www.youtube.com/watch?v=9JpdAg6uMXs.

[83] *See* Goodfellow, *supra* note 70 at 1.
[84] *Id.*
[85] *Id.*

output—often by numerical valuative factors but sometimes simply by uploading authentic video images of a target individual or typing desired spoken text into a prompt in order to shape the desired synthetic output, the goal of the GAN. By defining conditions for success, the programmer has implicitly also begun defining conditions for failure as data that does not match the goal will eventually become waste, or "noise."[86] The programmer will then start building the generator. Through algorithmic inputs, some of which may be purposefully hidden or "latent,"[87] the programmer essentially sets the goalposts for the generator. The generator, once created, immediately begins to transmit data to the discriminator, and the adversarial back-and-forth starts.

Due to the nature of the exchange and the positions of the dueling networks, the discriminator will almost always lose.[88] In fact, arguably the best-case scenario for a discriminator is that the discriminator network will get to the point where it can accurately estimate that at least 50 percent of the data produced by the generator is noise.[89] The generator cannot produce such success without training. Sophisticated training, therefore, is the hinge-point for the effectiveness of deep learning and deepfake technology.

---

[86] *See generally* Serrano, *supra* note 66; *see also* Luis Serrano, *A Friendly Introduction to Generative Adversarial Networks (GANs)* (May 5, 2020), https://www.youtube.com/watch?v=8L11aMN5KY8.

[87] Diego Gomez Mosquera, *GANs from Scratch 1: A Deep Introduction. With Code in PyTorch and TensorFlow*, AI SOC. (Feb. 1, 2018), https://medium.com/ai-society/gans-from-scratch-1-a-deep-introduction-with-code-in-pytorch-and-tensorflow-cb03cdcdba0f.

[88] *See* Goodfellow, *supra* note 82; Serrano, *supra* note 86.

[89] *See e.g.* Jason Brownlee, *How to Identify and Diagnose GAN Failure Modes*, MACHINE LEARNING MASTERY (Jan. 21, 2021), https://machinelearningmastery.com/practical-guide-to-gan-failure-modes/. Google developers posit that this 50 percent assessment rate occurs at a tipping point which arrives when the generator becomes so good that the discriminator's success appears owed more to chance than calculation. *See GAN Training* in *Generative Adversarial Networks* (Jul. 12, 2019), https://developers.google.com/machine-learning/gan/training (last visited Jun. 25, 2022). If, however, the algorithm continues to generate images yet the discriminator begins to reflect an accuracy rate beyond 50 percent, rendering the accuracy rate artificial, this can indicate error in the discriminator which would unintentionally cause the generator to become less effective. *Id.*

### *D. Supervised, Unsupervised, and Semi-Supervised Training*

#### 1. Supervised Learning – Showing the Machine

Supervised learning is not only the original method of machine training but also the most common—so common actually that we all unwittingly participate in it every day. Supervised learning occurs when algorithms receive labeled or pre-defined information with the intention that the algorithm will use that information to achieve a preconceived target output. IBM describes it as the "use of labeled datasets to train algorithms that to [sic] classify data or predict outcomes accurately."[90] Put more directly, supervised learning involves actions by "an instructor or teacher who shows the machine learning system what to do."[91]

Anyone, however, can be an instructor or teacher for AI. We participate in supervised learning-style AI training whenever we ask an Amazon Alexa device to tell us the weather forecast, tap our brakes in vehicles with automated brake performance-enhancing technology[92], or ask Google Translate to convert a question from English to French.[93] Physicians can assist supervised learning by inputting patient data and treatment techniques into algorithmic-based programs to predict the likely journey of a COVID-19 infection and increase chances of successful recovery.[94] Data analysts use algorithms trained with various supervised learning techniques to improve face-recognition technology and predict stock market fluctuations.[95]

We all train artificial intelligence every day via supervised learning without really knowing it. However, pure supervised learning is really only useful for classification modeling (e.g., telling the difference

---

[90] IBM Cloud Education, *Supervised Learning*, IBM CLOUD LEARN HUB (Aug. 19, 2020), https://www.ibm.com/cloud/learn/supervised-learning#toc-what-is-su-d3nKa9tk.
[91] Goodfellow, *supra* note 70 at 103.
[92] *See* Alyssa Schroer, *Artificial Intelligence in Cars Powers an AI Revolution in the Auto Industry*, BUILTIN (Mar. 25, 2020), https://builtin.com/artificial-intelligence/artificial-intelligence-automotive-industry.
[93] Yonghui Wu et al., *Google's Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation* (Oct. 8, 2016), https://arxiv.org/abs/1609.08144.
[94] *See* The Mount Sinai Hospital, *Developing Machine Learning Models to Predict Critical Illness and Mortality in COVID-19 Patients*, MEDICAL XPRESS (Nov. 10, 2020), https://medicalxpress.com/news/2020-11-machine-critical-illness-mortality-covid-.html.
[95] *See* JEREMY WATT ET AL., MACHINE LEARNING REFINED: FOUNDATIONS, ALGORITHMS, AND APPLICATIONS 1 (2016).

between a cat and a dog) or for regressive/predictive modeling (e.g., predicting the rate of student loan debt expansion over time).[96] Other learning techniques, therefore, become necessary to help sharpen AI.

### 2. Unsupervised Learning – Unbinding the Machine

Unsupervised learning deepens the AI talent pool and, ultimately, sets the conditions for deepfake technology to thrive. Whereas supervised learning occurs when an algorithm works within a set of labeled inputs, unsupervised learning removes the training wheels. In this case, a neural network will instead work with unlabeled inputs. Without the use of labeled inputs to communicate goal expectations, the network instead must identify patterns in order to deliver a goal output.[97]

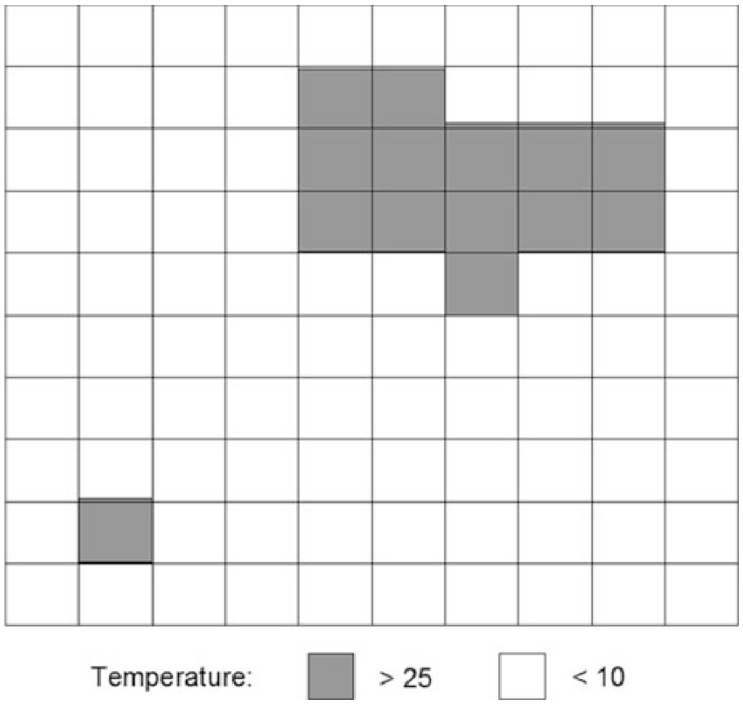
The learning that results is termed "unsupervised" because the human programmer minimizes their influence on the network so that the programmer can test the algorithm's independent ability to learn i.e., to adjust, discern, and identify, mathematically speaking.[98] The kinds of tasks that unsupervised learning tends to accomplish are generally those which group similar kinds of data or information, also known as "clustering."[99] In this way, an algorithm can identify one or several themes in a data group (e.g., pictures of men with dark hair v. men with gray hair v. men with no hair v. men with dark beards v. men with gray beards v. men with no beards) and compartmentalize each piece of data into groups, based on apparent patterns, to present a cluster of results (e.g., all pictures of men with beards) which a person can retrieve by requesting that particular cluster. On a larger scale, clustering assists with everything from data mining to data extraction to data analysis.

By logical and actual extension, unsupervised learning can also accomplish an implied task of clustering, that is to say, identify what data does *not* belong to a data cluster. Known as "anomaly detection"[100] or, in a related context, "denoising,"[101] this task identifies those data points

---

[96] *Id.* at 1-12.
[97] *See* Goodfellow, *supra* note 70 at 103.
[98] *See e.g.* UNSUPERVISED LEARNING ALGORITHMS V (M. Emre Celebi & Kemal Aydin eds., 2016) (observing that unsupervised learning algorithms "automatically discover interesting and useful patterns" in unlabeled data).
[99] *See* Goodfellow, *supra* note 70 at 103; *see also* Tülin İnkaya, Sinan Kayaligil, & Nur Evin Özdemirel, *Swarm Intelligence-Based Clustering Algorithms*, in UNSUPERVISED LEARNING ALGORITHMS 303 (M. Emre Celebi & Kemal Aydin eds., 2016).
[100] *See e.g.* P. Deepak, *Anomaly Detection for Data with Spatial Attributes*, in UNSUPERVISED LEARNING ALGORITHMS 1 (M. Emre Celebi & Kemal Aydin eds., 2016).
[101] Goodfellow, *supra* note 70 at 101, 507 (discussing how denoising autoencoders receive a "corrupted data point as input and [are] trained to predict the original, uncorrupted data point as [their] output.").

or characteristics which do not comport with the patterns already established during clustering. Whether identifying anomalies (i.e., groups of data which depart from what is generally regarded as common[102]) or outliers (i.e., an *individual* object which presents an uncommon characteristic[103]), the result is that the network is able to actively filter.




[*Fig. 4.* The above graphic, devised by Dr. Deepak Padmanabhan of Queen's University Belfast, depicts a hypothetical geographic region split into grids with each grid colored in accordance with its average temperature. As the largest pattern in this set is that most grid areas possess average temperatures, the cluster of dark squares to the top-right are an abnormality because they represent a sub-region that experiences higher-than-normal average temperatures. The dark square at the bottom left represents an outlier.[104] A network tasked with finding a place for someone to spend a weekend in a comfortable climate could, using unsupervised learning-devised anomaly detection, search in only the white grids in order to improve the chances of finding the most-desired vacation spot.]

This filter training, combined with immense computing power, is what makes the GANs discussed above work and by extension can make today's deepfake technology threat so potent. Because unsupervised learning principles help inject discrimination into machine learning, GANs receive the discriminator needed to enable the tasked program to

[102] P. Deepak, *supra* note 100 at 1-2.
[103] *Id.* at 2.
[104] *Id.*

constantly improve results. Researchers have seized on this strength by pairing GANs in unsupervised learning contexts with other neural networks such as VAEs to develop even more sophisticated content production,[105] resulting in more robust deepfake capabilities. However, sometimes requirements necessitate hybrid machine learning which is where semi-supervised learning gains purchase, and sometimes where deepfake content is best made.

### 3. Semi-Supervised Learning – Cooperating with the Machine

There is no chicken-egg, which-came-first conundrum about deepfake images. The person desiring to obtain a deepfake image comes first. This person supplies sometimes basic, sometimes sophisticated, parameters into a GAN in the hopes of getting a desired result. By doing so, the person has weighted and labeled at least some data sets. However, the GAN works to produce an image that is not only equivalent to the labels provided by the person but, for those features which do not carry an express label, is also consistent with patterns identified during the GAN's generate-and-reject volleys.

Deepfake images, therefore, can often be the product of semi-supervised machine learning. Consider the work built upon the pioneering GANs which have enabled today's deepfake technology. Within a year after Mr. Goodfellow published his work on GANs, advocates for a semi-supervised learning approach to GANs advanced the concept of a third network–known as a classifier–to deepen and improve the performance of the discriminator network, and thereby the generator network.[106]

A few years thereafter, researchers in China expounded upon the semi-supervised GAN approach with the development of the Margin

---

[105] *See e.g.,* Ming-Yu Liu et al., Unsupervised Image-to-Image Translation Networks 2 fig. 1 (Jul. 23, 2018, 3:39 AM), https://arxiv.org/pdf/1703.00848.pdf (proposing UNIT Networks as a combination of VAEs and GANs to leverage each network structure's strengths in order to better refine image generation accuracy and quality).

[106] *See e.g.,* Jost Tobias Springenberg, Unsupervised and Semi-Supervised Learning with Categorical Generative Adversarial Networks 2-4 (Nov. 19, 2015, 21:26 UTC) (presented at the 2016 International Conference on Learning Representations), https://arxiv.org/abs/1511.06390; Aug.us Odena, *Semi-Supervised Learning with Generative Adversarial Networks* 1 (Jun. 5, 2016, 11:42 PM), https://arxiv.org/pdf/1606.01583.pdf (citing Springenberg). This structure is more commonly known today as a "Triple-GAN." *See* Chongxuan Li et al., *Triple Generative Adversarial Nets* 2 (Dec. 20, 2019, 12:17 PM), https://arxiv.org/abs/1912.09784 (describing the growing utility of classifier or classifier-like networks in teaching GANs to produce more precise results).

Generative Adversarial Network (MarginGAN), a GAN which has a classifier network designed not only to help the discriminator sort data in order to identify fake images but also, by influence of "pseudo labels" provided to the generator, to increase margins of real images and decrease margins of fake images.[107] With additional proliferations of similar off-shoots such as CatGANs,[108] Triangle GANs,[109] and SGANs,[110] and the realization through semi-supervised learning that even greater network precision can occur by introducing further adversity between not only the generator and discriminator but also the generator and the classifier[111], semi-supervised learning has helped foster tremendous progress in synthetic content development. Combine these advances with developments in "reinforcement learning," described as a "crowning achievement of deep learning,"[112] in which the AI improves its output through a trial-and-error/reward-punishment system imposed by a programmer,[113] and it becomes easier to see how deepfake technology has arrived at its current sophisticated state.

## III. IDENTIFYING VIOLATIONS AND VIOLATORS: CLASSIFICATION, ATTRIBUTION, AND AGENCY.

In order to know how to enforce the laws on deception in combat, a State must understand what a violation of those laws looks like and how to identify perpetrators. The first challenge in combating deepfake content is knowing when content is in fact fake. After swiftly notifying partners about the fake content, the second challenge is identifying the responsible actors as quickly as possible. The third and potentially most

---

[107] Tong Lin & Jinhao Dong, *MarginGAN: Adversarial Training in Semi-Supervised Learning*, *in* ADVANCES IN NEURAL INFO. PROCESSING SYS. (H. Wallach et al. eds., 32nd ed., 2019), https://papers.nips.cc/paper/2019.

[108] *Id.* at 2 (citing Jost Tobias Springenberg, Address at International Conference on Learning Representations: Unsupervised and Semi-Supervised Learning with Categorical Generative Adversarial Networks (Nov. 19, 2015)).

[109] *Id.* (citing Zhe Gan et al., Triangle Generative Adversarial Networks, (2017 Neural Information Processing Systems conference paper, 2017), https://arxiv.org/abs/1709.06548).

[110] *Id.* (citing Zhijie Deng et al., Structured Generative Adversarial Networks, (2017 Neural Information Processing Systems conference paper, 2017), https://arxiv.org/abs/1711.00889).

[111] *See* Wenyuan Li et al., *Semi-Supervised Learning Using Adversarial Training with Good and Bad Samples*, 31 MACH. VISION AND APPLICATIONS 49 (2020) (also available at https://doi.org/10.1007/s00138-020-01096-z).

[112] GOODFELLOW, *supra* note 70, at 25, 103.

[113] *Id.*; *see also* Surbhi Arora, *Supervised vs Unsupervised vs Reinforcement*, AITUDE (Jan. 29, 2020), https://www.aitude.com/supervised-vs-unsupervised-vs-reinforcement/.

sensitive challenge is determining whether the actions of any actors can be attributed to a State or an organization. Deepfake technology poses unique difficulties in all three efforts.

### *A. Chasing and Catching Mirages*

The persistent problem among those who wish to regulate digital and cyber activities is that human ingenuity often has the appearance of staying one step ahead. The same is true for those currently hoping to find ways to quickly identify deepfake content. As Dr. Alexa Koenig has observed, detecting deepfakes presents several challenges including the "increasing sophistication and decreasing costs of deep learning technologies," an "information ecosystem" degraded by a continuous influx of misinformation, and a lack of legal professionals trained to verify fakes—a skill Dr. Koenig describes as "a first line of defense against being duped."[114]

Although these challenges exist, several projects are nonetheless underway to combat AI-enhanced deception—and some of these projects employ just as much ingenuity as their adversaries. The most common intuition is to design automated deepfake detection systems—i.e., combat AI with AI—in order to maximize detection timing, sophistication, and capacity while reducing the potential for human error.[115] To this end, hosts of computer scientists and engineers have researched various methods that can algorithmically detect deepfake-enabled content.[116] Diverse research has competed to develop machine learning algorithms that detect deepfakes by the various subtle errors that today's technology still exhibits, such as co-motion patterns,[117] the

---

[114] Alexa Koenig, *"Half the Truth is Often a Great Lie": Deepfakes, Open Source Information, and International Criminal Law*, 113 AJIL UNBOUND 250, 252 (2019). Dr. Koenig is the Executive Director of the Human Rights Center at the University of California Berkley School of Law.

[115] *See* Alex Engler, *Fighting Deepfakes When Detection Fails*, BROOKINGS INSTITUTE (Nov. 14, 2019), https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/.

[116] *See e.g.*, *id.* (citing, *inter alia*, Yuezun Li & Siwei Lyi, Exposing DeepFake Videos by Detecting Face Warping Artifacts, (Nov. 2018), https://arxiv.org/abs/1811.00656; David Guera & Edward J. Delp, DeepFake Video Detection Using Recurrent Neural Networks (Nov. 2018), https://engineering.purdue.edu/~dgueraco/content/deepfake.pdf).

[117] Gengxing Wang, Jiahuan Zhou, & Ying Wu, Exposing Deep-Fake Videos by Anomalous Co-Motion Pattern Detection (Aug. 11, 2020), https://arxiv.org/abs/2008.04848. "Co-Motion Patterns," as described by the authors here, occur when a person's face in a deep-fake video exhibits slight movements (a.k.a. landmarks) or lacks slight movements in a way that is atypical in genuine facial movements. These are more identifiable in deep-fake videos that have

lack of miniscule changes of skin color in a face that an actual normal heartbeat would presumably cause,[118] and atypical eye blinking.[119] Observers also acknowledge the early efforts of Gfycat to combat deepfake pornography through its Project Angora and Project Maru initiatives which scour the internet and find images of the depicted individual in order to compare facial features and make an analytical assessment about whether the concerned content is synthetic.[120] Recently, Facebook has also invested in deepfake detection technology through its 2020 Deepfake Detection Challenge (DFDC) which incentivized over 2,000 competitors to devise a program which would have the highest detection rate among a selection of video images.[121]

The United States government has also been a vigorous player in the effort to develop deepfake-combating AI. Through its Guaranteeing AI Robustness against Deception (GARD) Program, the Defense Advanced Research Projects Agency (DARPA) has a robust portfolio of approaches for identifying deepfake and other similarly faked content with the specific aim of creating "deception-resistant [machine learning] technologies"[122] which can competently defeat both current levels of deepfake technology and expected future evolutions.[123] Finding biological inspiration in the immune system, GARD looks to develop a defense system that "identifies attacks, wins and remembers the attack to create a more effective response during future engagements."[124]

---

both real and deep-fake content (for example, where a genuine video of a President giving a real speech is altered to make the President say only a few things that he or she did not actually say). A similar focus influenced some of the first work on counter-deepfake AI. *See e.g.* Darius Afchar, et al., MesoNet: A Compact Facial Video Forgery Detection Network (Sep. 4, 2018), https://arxiv.org/abs/1809.00888.

[118] Hua Qi et al., DeepRhythm: Exposing DeepFakes with Attentional Visual Heartbeat Rhythms (Aug. 26, 2020), https://arxiv.org/pdf/2006.07634.pdf.

[119] Yuezun Li et al., In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking (Jun. 11, 2018), https://arxiv.org/abs/1806.02877.

[120] *See* Louise Matsakis, *Artificial Intelligence is Now Fighting Fake Porn*, WIRED (Feb. 14, 2018), https://www.wired.com/story/gfycat-artificial-intelligence-deepfakes/; Koenig, *supra* note 114, at 254; Citron, *supra* note 41, at 1787n.145.

[121] *Deepfake Detection Challenge Results: An Open Initiative to Advance AI*, FACEBOOK AI, https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/ (last visited Feb. 25, 2021).

[122] *Defending Against Adversarial Artificial Intelligence*, DEF. ADVANCED RSCH. PROJECTS AGENCY (Feb. 6, 2019), https://www.darpa.mil/news-events/2019-02-06.

[123] Anticipated future evolutions in deep-fake technology include "multi-sensor and multi-modality variations" as well as generative AI capable of making predictions, decisions, and adaptations. *Id.*; *see also* Bruce Draper, *Guaranteeing AI Robustness Against Deception (GARD)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception (last visited Feb. 24, 2021).

[124] *Defending Against Adversarial Artificial Intelligence*, *supra* note 122.

Many of the programs intended to bring GARD's immune system-inspired deception-defeat capability to life show high creative and practical potential, and equally high ambition. The Reverse Engineering of Deceptions (RED) program seeks to employ AI capable of reverse engineering media content's algorithmic toolchains (i.e. the sequential series of steps in a machine's operation from start to finish) not only to determine if content is fake but also to determine the content's point of origin–enabling the U.S. to actually identify the adversary sending the deepfake.[125] The Media Forensics (MediFor) program builds on work already done in the fields of digital and other media forensics by developing an "end-to-end" platform which can employ techniques relevant across the media spectrum to detect expected manipulations, explain how the programmers made the manipulations, and quantify the likelihood that target content is actually fake.[126] Finally, the Semantic Forensics (SemaFor) program would train AI to latch on to semantic errors such as problems with facial structure, coloration, or eye-blinking discussed above to develop a catalogue of errors which would impose a burden on creators to "get every semantic detail correct, while defenders only need to find one, or a very few, inconsistencies."[127] Additionally, the SemaFor program would also train AI, like the MediFor program, to determine not only that content is fake but also where the content originated in order to aid in attribution.[128]

While the combined results of these efforts, both within DARPA and within the larger computer sciences communities, show tremendous progress in combating deepfake, many of these approaches still have inherent weaknesses. Gfycat's Projects Maru and Angora, for instance, would appear useless when faced with videos that do not have any source content from the internet. The DeepRhythm methodology, which would look for semantic errors if an image's facial coloration did not correlate to a normal heartbeat,[129] does not appear immediately able to account for

---

[125] Matthew Turek, *Reverse Engineering of Deceptions (RED)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/reverse-engineering-of-deceptions (last visited Feb. 24, 2021) [hereinafter *RED]*.
[126] Matthew Turek, *Media Forensics (MediFor)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/media-forensics (last visited Feb. 24, 2021) [hereinafter *MediFor*].
[127] Matthew Turek, *Semantic Forensics (SemaFor)*, DEF. ADVANCED RSCH. PROJECTS AGENCY, https://www.darpa.mil/program/semantic-forensics (last visited Feb. 24, 2021) [hereinafter *SemaFor*].
[128] *Id.*; *see also Uncovering the Who, Why, and How Behind Manipulated Media,* DEF. ADVANCED RSCH. PROJECTS AGENCY, *https://www.darpa.mil/news-events/2019-09-03a* (last visited Jun. 25, 2022).
[129] Qi, *supra* note 118 at 1.

biological variables such as might present in a person with a heart condition or blood pressure issues. And the winner of the Facebook DFDC achieved an 82.56 percent accuracy[130]—certainly impressive but still allowing for an error rate that could permit significant harm in a national security or armed conflict scenario.

Potentially most problematic, even for all of the work and resources expended in deepfake detection efforts, is the "detection dilemma."[131] Simply put, this is the notion that the more work that goes in to detecting deepfake, the more deepfake creators learn how to avoid detection. As discussed, and cited to above, much of the research done into detection strategies is open source. For every publication that describes how a new set of algorithms can detect unnatural blinking patterns, deepfake developers learn to improve blinking. Even a mass-effort style approach by entities like DARPA can seem from afar like a Sisyphean task. A recent article on the subject by members of three highly-influential AI advancement enterprises called for an all-hands "multistakeholder" coalition effort among academia, media, technology, and civil society organizations in order to effectively counter the coalition of adversarial interests that can cause deepfake proliferation.[132] This kind of broad-based cooperability between government and non-governmental entities has also been proposed in seeking ways to confirm and counter GAN-enabled manipulation of satellite imagery.[133] A bulletproof, long-term solution may ultimately not be likely. The real best defense, and thereby best ability to detect deepfakes, may at least for now be the fact that we know they exist, that we continue to talk about them, and that major social players maintain dialogue to determine methods of cooperation as deepfake threats grow.

---

[130] *Deepfake Detection Challenge Results: An Open Initiative to Advance AI*, *supra* note 121.

[131] Claire Leibowicz et al., The Deepfake Detection Dilemma: A Multistakeholder Exploration of Adversarial Dynamics in Synthetic Media (Feb. 11, 2021), https://arxiv.org/abs/2102.06109.

[132] *Id.* The three enterprises are The Partnership for AI, the XPRIZE Foundation, and the Thoughtful Technology Project. It is worth noting that this article did not list government explicitly as a "stakeholder" in this effort.

[133] *See* Patrick Tucker, *The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth*, DEFENSE ONE (Mar. 31, 2019), https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/.

### *B. Agency and Attribution: Technical Analysis*

Once a deception has been revealed, and authorities have spread the word to assure that the deception or similar variants of it do not continue to succeed, the next task is to identify the source of the deceptive action. In the context of armed conflict, this determination must occur immediately in order to stop the bleeding, both figuratively and possibly even literally, as well as to determine follow-on responses.

Specifically, the targeted force must be able to identify persons and belligerents. Like with other forms of cyber warfare, this task can be difficult, as a cyberattack does not often leave a literal trail of smoke. Furthermore, the asset which deploys the attack does not have to even be in the same hemisphere as the target.

Many of the deception identification efforts discussed above seek not only to confirm that content is fake but also to begin to detect agency i.e., the confirmation that human actors are involved, their identities, and their level of responsibility. Once agency is established, attribution of the concerned people or entities to States or non-State actors can begin. DARPA's RED program, for example, acknowledges that "identifying an adversary" is one of many desired outcomes from its automated toolchain reverse engineering approach.[134] Their SemaFor program specifically seeks to employ "attribution algorithms" in order to help determine if the content originated from an individual or an organization.[135]

Significant academic research over the past three years has produced a steady stream of analyses helpful for finding actors and entities employing deepfake. One such study, financed in part by DARPA's MediFor program, has developed attribution algorithms which train on and identify "GAN fingerprints" in images in order to increase a classification network's ability to specifically identify GANs and conduct image and model attribution.[136] Their classifiers, even when tested against attribution defenses, often demonstrated accuracy rates well in excess of 90%.[137]

Currently, however, it is unclear whether any of these efforts are effective enough to solve the Gordian Knot that has become cyberspace attribution. First, many attribution methods still suffer from exploitable

---

[134] *RED*, *supra* note 125.
[135] *SemaFor*, *supra* note 127.
[136] Ning Yu et al., *Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints*, 7555, 7556 (Feb. 27, 2020), https://ieeexplore.ieee.org/document/9010964.
[137] *Id.* at 7562.

vulnerabilities. The GAN fingerprint-detecting attribution algorithm discussed above, for example, is only trained for scouring images of still faces.[138] It does not analyze videos, audio data, or any other data other than still human faces. Also, it can only "attribute" the image to being a GAN construct or not being a GAN construct—it is not yet sufficiently sophisticated to attribute a GAN to a server.[139]

Second, several means currently exist for mal-intended actors to hide their involvement. Tor is a popular anonymity platform which prevents IP address tracking.[140] However, other utilities such as Mixmaster, Onion Routing, and AN.ON further complicate the picture because they use anonymity networks via proxy servers to code, re-code, and re-order (scramble) data in order to make a content's route or source strenuously difficult to track.[141] While robust work and resources have been invested in developing anonymity network hacks and have seen some success,[142] to the delight of privacy advocates none so far have proven effective enough to lift the veil.

Third, none of the known automated attribution systems are anywhere near the level of sophistication necessary to identify fakes and find actors to the level needed to be truly effective real-time. If a skilled developer constructs a deepfake video appearing to show U.S. soldiers mocking the Quran and cursing the Prophet Muhammed and then manages to release it anonymously on the internet claiming it came from an area of active operations in a Muslim-majority region, forensics work

---

[138] *Id.*

[139] *Id.*; however, *cf.* Tianyun Yang et al., *Deepfake Network Architecture Attribution* 1 (Mar. 14, 2022), https://arxiv.org/abs/2202.13843 (providing an architecture-based approach to deepfake "fingerprint" detection as opposed to model-based detection).

[140] *See* Citron, *supra* note 41, at 1792. Tor pre-dates deepfake technology, having been used famously in 2009 by Iranians trying to protest the elections there and the subsequent crushing of popular unrest by then-President Mahmoud Ahmadinejad. *See* Cyrus Farivar, *Geeks Around the Globe Rally to Help Iranians Online*, FRONTLINE (Jul. 8, 2009, 3:56 pm), https://www.pbs.org/wgbh/pages/frontline/tehranbureau/2009/07/geeks-around-the-globe-rally-to-help-iranians-online.html.

[141] *See* Simone Fischer-Hbner & Stefan Berthold, *Privacy-Enhancing Technologies*, in COMPUTER AND INFORMATION SECURITY HANDBOOK 759-78 (John R. Vacca ed., 3d ed. 2017) (discussing the mix net concept).

[142] *See e.g.*, Zhongxiang Wei et al., Fundamentals of Physical Layer Anonymous Communications: Sender Detection and Anonymous Precoding (Oct. 18, 2020), https://arxiv.org/abs/2010.09122 (discussing the ability of signaling patterns and channel characteristics to provide inferences which can help identify a sender, while also acknowledging that precoding can help defeat sender identification efforts); Wenlin Han & Yang Xiao, *Privacy Preservation for V2G Networks in Smart Grid: A Survey*, 91 COMPUT. COMMC'N 17, 17-28 (2016) (concluding that adversarial algorithms can detect individuals otherwise clouded in an anonymity network by compiling various data outside the anonymity network which provides inferences about the individual item's presence in the anonymized group).

may reveal that the image is fake and even begin to point towards a particular country or even individuals inside a country. However, this will take a few days to confirm. By the time this task is complete, the realistic-looking video will have already done its damage.

Additionally, combatants are in even more trouble if the deepfake content is not a photo or video image of a person. If, for example, an enemy unit devises a deepfake-enabled voice recording or voice masker and manages to call their adversary unit's commander directly to make the unit commander think his superior is instructing him to surrender to the enemy (a capability made progressively more real today by such developers as WellSaid Labs[143] and Google),[144] nothing in the arsenal of computer science research can currently combat this tactic.

Certainly, the computer sciences would not be alone in any of these scenarios to help reveal a fraud. Sophisticated deepfake content still requires extremely skilled developers. So, the synthetic content in both of these scenarios may demonstrate enough imperfections to trigger quick scrutiny and provide signs, along with various degrees of intelligence collection, that can point to a responsible office or even person.[145] Also, the use of deepfake in several armed conflict scenarios, such as in an international armed conflict between two states or a non-international armed conflict between long-time familiar enemies, will logically facilitate finger-pointing before digital forensics can even tie its proverbial shoes. However, while progress has proceeded quickly, we still remain quite a long way from having automated networks which can detect deepfakes across the media spectrum and quickly attribute them to human actors.

---

[143] WELLSAID LABS, https://wellsaidlabs.com/ (last visited Feb. 25, 2021).

[144] Google's artificial intelligence development group DeepMind, for example, works expressly to "solve intelligence" by replicating the brain's physiological and mathematical progressions in order to imitate and train human-like thought processes in artificial intelligence. Part of DeepMind's various programs is one called "WaveNet" which seeks to train artificial neural networks to develop realistic-sounding text-to-speech audio capabilities in order to assist the disabled. DEEPMIND, https://deepmind.com/ (last visited Jan. 5, 2021); *see also* Yutian Chen et al., *Using WaveNet Technology to Reunite Speech-Impaired Users with Their Original Voices*, DEEPMIND (Dec. 18, 2019), https://deepmind.com/blog/article/Using-WaveNet-technology-to-reunite-speech-impaired-users-with-their-original-voices.

[145] This is essentially how the Zelenskyy deepfake was so quickly debunked. Its production value was relatively low likely owing to the hasty nature of its creation. Viewers were able make out lighting inconsistencies, odd head-to-body proportionality, image blurriness, and could perhaps most easily tell that the video was fake due to the poor quality of Mr. Zelenskyy's depicted voice. *Supra* note 13.

### *C. Agency and Attribution: Legal Analysis*

Legally attributing an act of deepfake-deception to an individual, an organization, or a country is also complicated. Because deepfake is an act of artificial intelligence that utilizes node-based neural networks within cyberspace often to achieve objectives through cyberspace, deepfake invokes legal equities related to cyberspace operations.[146] Deepfake technology can also pair with other classic examples of cyber activities, such as ransomware, to conduct a cyberattack.[147] However, deepfake does not have the same purpose as typical cyberspace operations such as distributed denial of service attacks on servers supporting an adversary's headquarters. Deepfake is a means of deception and hence also bears legal equities related to information operations.[148]

The current best source for modern perspectives on attribution for cyber activities conducted prior to or during an armed conflict are not in a law, but in a manual. Published in 2017, the Tallinn Manual 2.0[149]

---

[146] *See* U.S. DEP'T OF DEF., LAW OF WAR MANUAL ¶ 16.1.2 (May 2016) [hereinafter DOD LAW OF WAR MANUAL] (citing JOINT PUBLICATION 3-0, *Joint Operations* (Aug. 11, 2011)); JOINT PUBLICATION 3-12, *Cyberspace Operations*, GL-4 (Feb. 5, 2013) (defining cyberspace as a "global domain within the information environment consisting of interdependent networks of information technology infrastructures and resident data, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers.").

[147] *See e.g.*, Jovi Umawing, *The Face of Tomorrow's Cybercrime: Deepfake Ransomware Explained*, MALWAREBYTES LABS (Jun. 26, 2020), https://blog.malwarebytes.com/ransomware/2020/06/the-face-of-tomorrows-cybercrime-deepfake-ransomware-explained/.

[148] *See* JOINT PUBLICATION 3-13, *Information Operations*, GL-3 (Nov. 27, 2012 (incorporating Change 1, Nov. 20, 2014)) (defining information operations as the "integrated employment . . . of information-related capabilities . . . to influence, disrupt, corrupt, or usurp the decision-making of adversaries and potential adversaries while protecting our own."). While U.S. Department of Defense doctrine, which is currently silent on deepfake, would not organize deep-fake operations into cyberspace operations, this perspective does not seem to be universal. *Compare* DOD LAW OF WAR MANUAL, *supra* note 146 at ¶¶ 16.1.2.1 and 16.1.2.2 (stating cyber operations "use computers to disrupt, deny, degrade, or destroy information resident in computers and computer networks" but that "operations to distribute information broadly using computers would generally not be considered cyber operations.") *with* Citron, *supra* 41, at 1801 (highlighting domestic liability for deepfake-based crimes in federal cyberstalking laws under 18 U.S.C. § 2261A); DANIELLE CITRON, HATE CRIMES IN CYBERSPACE (2014); Mike Faden, *Malicious Deepfake Technology: A Growing Cyber Threat*, MIMECAST (Jul. 13, 2020), https://www.mimecast.com/blog/malicious-deepfake-technology-a-growing-cyber-threat/; *see also* Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224, 227 (2011) (discussing impacts from "cyberspace harassment").

[149] TALLINN MANUAL ON INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS (Michael N. Schmitt ed., 2d ed. 2017) [hereinafter TALLINN MANUAL 2.0].

absorbed the first iteration of the Tallinn Manual issued in 2013[150] and, despite being only a statement by international legal experts and not a law itself, it is nonetheless an influential source in an armed conflict-related field that has virtually no subject-specific multilateral treaties[151] and is otherwise light on expressions of customary international law.[152]

The Tallinn Manual 2.0 articulates several elements of international law that are vital for ascertaining attribution. It recognizes for instance that international law provides States with sovereignty in cyberspace.[153] States also have a duty to exercise due diligence to ensure they do not allow their territory or cyber infrastructure to be used to produce "serious adverse consequences" for other States.[154] States may exercise territorial and extraterritorial jurisdiction over cyber activities or persons involved in cyber activities that cause substantial effects in the State.[155] Taking inspiration from the Draft Articles on Responsibility of States for Internationally Wrongful Acts,[156] the Tallinn Manual 2.0 also

---

[150] TALLINN MANUAL ON INTERNATIONAL LAW APPLICABLE TO CYBER WARFARE (Michael N. Schmitt ed., 2013) [hereinafter TALLINN MANUAL 1.0].

[151] Some multinational treaties that might have application in a non-armed conflict context include the 2000 Palermo Convention and the 2001 Budapest Convention. United Nations Convention Against Transnational Organized Crime, Nov. 15, 2000, 2225 U.N.T.S. 209 [hereinafter Palermo Convention]; Convention on Cybercrime, Nov. 8, 2001, E.T.S. 185 [hereinafter Budapest Convention]. The United States has ratified and is party to both treaties.

[152] For a discussion on how the Tallinn Manual iterations accompany expressions of international law, *see* Eric Talbot Jensen, *The Tallinn Manual 2.0: Highlights and Insights*, 48 GEO. J. INT'L L. 735, 738 (2017). Mr. Jensen was a member of the International Group of Experts who met at the NATO Cooperative Cyber Defense Center of Excellence in Tallinn, Estonia to develop the Tallinn Manuals.

[153] TALLINN MANUAL 2.0, *supra* note 149, at 11 r. 1., 16 r. 3, 17 r. 4 (providing that "[t]he Principle of Sovereignty applies to cyberspace" and that "it is a violation of territorial sovereignty for an organ of a State, or others whose conduct may be attributed to the State, to conduct cyber operations while physically present on another State's territory against that State or entities or persons located there."). The Manual acknowledges that it is not settled international law as to whether violation of sovereignty in cyberspace by itself constitutes an internationally wrongful act or whether sovereignty just acts as a mere rule. *See* Jensen, *supra* note 152, at 741-42 (citing Gary Corn, *Tallinn Manual 2.0 – Advancing the Conversation*, JUST SECURITY (Feb. 15, 2017, 8:41 am), https://www.justsecurity.org/37812/tallinn-manual-2-0-advancing-conversation/#more-37812).

[154] TALLINN MANUAL 2.0, *supra* note 149, at 30 r. 6. As Mr. Jensen points out, this rule does not prohibit all harm – just that harm which results in serious adverse consequences. Jensen, *supra* note 152, at 744.

[155] TALLINN MANUAL 2.0, *supra* note 149, at 51 r. 8. Particularly, Rule 9 provides that States can exercise jurisdiction over "cyber infrastructure and persons engaged in cyber activities on its territory," cyber activities "originating in, or completed on, its territory," or cyber activities causing "substantial effect" in its territory. *Id.* at 55 r. 9.

[156] For related discussion *see* Jensen, *supra* note 152, at 750.

provides that States "bear international responsibility for a cyber-related act that is attributable to the State . . ."[157]

But who is "the State"? Rule 15 seeks to clarify that "[c]yber operations conducted by organs of a State, or by persons or entities empowered by domestic law to exercise elements of governmental authority, are attributable to the State."[158] This clarification of course only raises more questions about what qualifies as an "organ," what domestic law would need to do to show empowerment, where does the divide lay between element and non-element, and so forth.

The Manual explains that the term "State organ" has "broad meaning to ensure that States do not escape responsibility by asserting an entity's non-status as its organ in domestic law."[159] It provides that the "clearest case" occurs when State military or intelligence agencies commit the acts, listing U.S. Cyber Command and Israel's Unit 8200 as examples.[160] In order to cast a wide net, however, the Manual adopts the perspective of the Draft Articles on Responsibility as well as the International Court of Justice, stating:

> "[P]ersons, groups of persons or entities may, for the purposes of international responsibility, be equated with State organs even if that status does not follow from internal law, provided that in fact the persons, groups or entities act in 'complete dependence' on the State, of which they are ultimately merely the instrument."[161]

But despite this language, quickly the net begins to narrow. The burden to show that a person or entity must act in "complete dependence" of the State is not low. There must be a showing that a "particularly great degree of State control" exists over the person or entities concerned[162] and that when determining this, the key factors are "the function of the entity" and the "State's intention" concerning the person or entities because even State ownership of an entity is not

---

[157] *Id.* (citing TALLINN MANUAL 2.0, *supra* note 149, at 84 r. 14).
[158] TALLINN MANUAL 2.0, *supra* note 149, at 87 r. 15.
[159] *Id.* at 87-88 ¶ 3 (referencing Int'l Law Comm'n, Draft Articles on Responsibility of States for Internationally Wrongful Acts, Rep. of the Int'l Law Comm'n on the Work of Its Fifty-Third Session, U.N. Doc. A/56/10, at art. 4(2) (2001) [hereinafter Draft Articles on Responsibility]).
[160] *Id.* at 87 ¶ 1.
[161] *Id.* at 88 ¶ 4 (quoting Application of Convention on Prevention and Punishment of Crime of Genocide (Bosn. & Herz. v. Serb. & Montenegro), Judgment, 2007 I.C.J. 47 ¶ 392 (Feb. 2007) [hereinafter I.C.J. Genocide Case]).
[162] *Id.* (citing I.C.J. Genocide Case, *supra* note 161, at ¶ 393).

enough to demonstrate requisite State control.[163] While responsibility may still attach for *ultra vires* acts that exceed State grants of authority, the actor must nonetheless still appear "under colour of authority."[164]

The purpose of this narrowing is that the Tallinn Manual 2.0—like international law when it comes to activities in general in cyberspace—only sees international legal support for holding *states* responsible for internationally wrongful acts. Individuals or entities in most contexts would only be subject to the jurisdiction of domestic law or certain specific treaties. This is why Rule 17 provides that cyber operations by non-state actors are only attributable to states if the non-state actor acts "pursuant to [the state's] instructions or under its direction or control" or if the state "acknowledges and adopts" the non-state actor's activities as their own.[165] Thus even if a state gave malware to a terrorist organization and the terrorist organization then decided on its own to independently plan and execute an offensive cyber operation with that malware, the Manual would not legally attribute the cyber operation to the state unless the state later adopted the cyber operation as its own.[166]

The result is that in the context of cyberspace operations, *lex generalis* provides that acts are legally attributable to only one entity—states—and therefore in such a case only states would need to be concerned about countermeasures. By this view, terrorists, insurgents, stateless militias, hacktivists, non-governmental organizations, Silicon Valley titans, Silicon Valley start-ups, protestors, risk-inclined college students, and bored teenagers would not face jeopardy under international law for deploying deepfake deception which, in times of peace, is not a *per se* international crime. However, as the doctrine of *lex specialis derogat legi generali* explains, specified international laws override general law.[167] The laws of armed conflict are precisely the *lex specialis* which might bridge the gap in legal attribution for deepfake-derived deception—in both cyber and information operation contexts—when it may not seem to otherwise exist.

---

[163] *Id.* at 88 ¶ 5.
[164] *Id.* at 89 ¶ 7.
[165] *Id.* at 94 r. 17.
[166] *Id.* at 97 ¶ 8.
[167] *See id.* at 80 ¶ 5 (citing commentary to Draft Articles on Responsibility at art. 55); *see also* Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. Rep. 226, at 25 (Jul. 8); Anja Lindroos, *Addressing Norm Conflicts in a Fragmented Legal System: The Doctrine of Lex Specialis*, 74 NORDIC J. INT'L L. 27, 35-39 (2005) (tracing the history of the doctrine back to Roman law and its later development from Hugo Grotius to the International Court of Justice).

## IV. DISINFORMATION AND THE LAWS OF ARMED CONFLICT (LOAC).

### *A. Ruse*

The law generally categorizes any lawful deception as a ruse.[168] The U.S. Department of Defense (DoD) broadly defines a ruse as a "trick of war designed to deceive the adversary, usually involving the deliberate exposure of false information to the adversary's intelligence collection system."[169] Joint Publication 3-13.4, which provides baseline DoD policy on military deception activities, characterizes a ruse as a "cunning trick designed to deceive the adversary to obtain friendly advantage."[170]

Both of these definitions derive primarily from two sources of international law—the Hague Conventions and the 1977 Additional Protocol I to the Geneva Conventions. The regulations featured in the 1907 Hague Convention concerning the Laws and Customs of War on Land ("Hague Convention IV") provide at Article 24 that generally

---

[168] *See e.g.*, U.S. DEP'T OF ARMY, FIELD MANUAL 6-27, THE COMMANDER'S HANDBOOK ON THE LAW OF LAND WARFARE, ¶ 2-171 (7 Aug. 2019) [hereinafter FM 6-27]. This is also true in a domestic law sense though the use of deception, particularly in law enforcement circumstances, can encounter significantly more skepticism than in a combat scenario. *Compare e.g.* Nadia B. Soree, *Thank You All the Same, but I'd Rather not be Seized Today: The Constitutionality of Ruse Checkpoints Under the Fourth Amendment*, 66 BUFFALO L. REV. 385, 433-34 (2018) (arguing that the use of "ruse checkpoints" violates the Fourth Amendment) *with* Daniel R. Dinger & John S. Dinger, *Deceptive Drug Checkpoints and Individualized Suspicion: Can Law Enforcement Really Deceive its Way into a Drug Trafficking Conviction?*, 39 IDAHO L. REV. 1, 29-55 (2002) (arguing that deceptive checkpoints can be just as lawful a manner of ruse as the use of undercover techniques and are not per se violative of the Fourth Amendment). This paper, however, does not seek to explore domestic impacts of deep-fake technology outside of the context of armed conflict.

[169] JOINT CHIEFS OF STAFF, JOINT PUB. 1-02, DEPARTMENT OF DEFENSE DICTIONARY OF MILITARY AND ASSOCIATED TERMS 207 (8 Nov. 2010, as amended through 15 Feb. 2016); *see also* NAT'L SEC. LAW DEP'T, THE JUDGE ADVOCATE GEN.'S LEGAL CTR. & SCH., OPERATIONAL LAW HANDBOOK (2018) (this publication was amended in 2020 at which time the publication opted instead to lean more on the definition of ruse provided in the Hague Regulations discussed *infra*).

[170] JOINT CHIEFS OF STAFF, JOINT PUB. 3-13.4, MILITARY DECEPTION, ¶ 11(c)(3) (26 Jan. 2012). JP 3-13.4 takes the additional step of distinguishing ruses from other similar acts of deception such as feints ("an offensive action involving contact with the adversary conducted for the purpose of deceiving the adversary as to the location and/or time of the actual main offensive action") and displays ("the simulation, disguising, and/or portrayal of friendly objects, units, or capabilities in the projection of the MILDEC story"). *Id.* at ¶ 11(c)(1),(4). Notably this regulation, promulgated at the same echelon and near in time to JP 1-02, avoids the JP 1-02 narrowing of the definition of ruse to those acts which interact with "[an] adversary's intelligence collection system," focusing instead on the objective of employing a ruse, namely, to obtain "friendly advantage." With respect to the Army, however, FM 6-27 goes to significant effort to encompass both dynamics so that the definition of ruse is not limited in either respect. FM 6-27, *supra* note 168, at ¶¶ 2-172, 2-173.

speaking "ruses of war . . . are considered permissible."[171] However, this article does not attempt to redefine ruse—instead, it implies that a ruse is anything that is not expressly forbidden by the regulations.[172]

Additional Protocol I to the Geneva Conventions ("API"), however, provides clarity to the concept of ruse that still controls today.[173] Bearing in mind that API applies to Common Article 2 international armed conflicts only,[174] Article 37 of API provides at Section 2 that "[r]uses of war are not prohibited."[175] Article 37 defines a ruse as those acts "intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict."[176] Furthermore, it distinguishes a ruse from an act of perfidy by explaining that a ruse "[does] not invite the confidence of an

---

[171] Convention (IV) Respecting the Laws and Customs of War on Land and its Annex: Regulations Concerning the Laws and Customs of War on Land art. 24, Oct. 18, 1907, 36 Stat. 2277 [hereinafter 1907 Hague Convention IV]. The Hague Conferences, both the 1907 meeting and the earlier 1899 meeting, were themselves inspired by preceding benchmark regulations on armed conflicts, most notably the famous Lieber Code promulgated by the Lincoln Administration during the American Civil War as well as the 1874 Brussels Declaration and the 1880 "Oxford Manual" on the laws of war on land by the Institute of International Law. *See* Sean Watts, *Law-of-War Perfidy*, 219 MIL. L. REV. 106, 125-37 (2014).

[172] 1907 Hague Convention IV, *supra* note 171, at art. 24. This is most likely due to the fact that this language came from the 1899 Hague Conventions which themselves borrowed enormously from the 1874 Brussels Declaration. Convention (II) Respecting the Laws and Customs of War on Land and its Annex: Regulations Concerning the Laws and Customs of War on Land art. 24, Jul. 29, 1899, 32 Stat. 1803 ("1899 Hague Convention II"); *see also* Watts, *supra* note 171, at 137 n.103.

[173] While the United States is not a party to Additional Protocol I to the Geneva Conventions, the United States acknowledges that several portions of the Protocol are customary international law and therefore seeks to abide by those portions. As for Article 37, the United States recognizes it in its entirety to be customary international law. Michael J. Matheson, *Remarks in Session One: The United States Position on the Relation of Customary International Law to the 1977 Protocols Additional to the 1949 Geneva Convention*, 2 AM. U. J. INT'L L. & POL'Y 419, 425 (1987).

[174] Protocol Additional to the Geneva Conventions of 12 Aug. 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 1, Jun. 8, 1977, 1125 U.N.T.S. 3 [hereinafter Additional Protocol I]. Article 1(4) seeks to apply API beyond the scope of international armed conflicts to conflicts involving efforts to throw off colonial domination, alien occupation, or fight racist regimes (a.k.a. conflicts of "national liberation"), circumstances which often involve non-international armed conflict. Many countries including the United States expressly reject this expansion. *See* Matheson, *supra* note 173. Hence, as is discussed *infra*, the discussion of perfidy to follow would not apply in Common Article 3 conflicts. *Compare also* Additional Protocol I at arts. 37-39 *with* Protocol Additional to the Geneva Conventions of 12 Aug. 1949, and Relating to the Protection of Victims of Non-International Armed Conflicts, Jun. 8, 1977, 1125 U.N.T.S. 609 [hereinafter Additional Protocol II] (pertaining solely to Common Article 3 non-international armed conflicts yet omitting any focused discussion on perfidy, misuse of recognized emblems, or misuse of emblems of nationality).

[175] Additional Protocol I, *supra* note 174, at art. 37(2).

[176] *Id.*

adversary with respect to protections under that law."[177] Article 37 then provides a short list of acts which would qualify as a ruse to include "the use of camouflage, decoys, mock operations, and misinformation."[178]

The Article's explicit reference to "misinformation" as a lawful example of deception accepts the reality that trickery has and to some degree should be a part of war. The Diplomatic Conference which promulgated API expressly recognized this when it considered how to draft Article 37. As the International Committee of the Red Cross (ICRC) Reading Commission wrote in its Commentary on the Additional Protocols of 8 June 1977 when discussing the Conference's perspective on ruse, "[t]he art of warfare is a matter, not only of force and of courage, but also of judgment and perspicacity. In addition, it is no stranger to cunning, skill, ingenuity, stratagems and artifices, in other words, to ruses of war, or the use of deception."[179] The Commentary even goes so far as to concede that, while it can cause significant problems, deception is "a just and necessary means of hostility."[180]

At the same time, however, Article 37's drafters acknowledged that setting parameters on lawful deception was harder to do than setting parameters on unlawful deception. Its list of examples of ruse is purposefully broad and non-exclusive because the Conference understood it would be a fool's errand to try to predict the limits of human creativity.[181] This appears to have everything to do with why the Article defines perfidy, discussed more below, first[182] and then defines ruse in contradistinction of perfidy.

The Commentary does offer an affirmative definition of ruse by explaining that a ruse "consists either of inducing an adversary to make a mistake by deliberately deceiving him, or of inducing him to commit an imprudent act, though without necessarily deceiving him to this end."[183]

---

[177] *Id.*
[178] *Id.*
[179] INT'L COMM. RED CROSS, Commentary, ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, at 439-440 (Yves Sandoz et al. eds, 1987) [hereinafter API Commentary] (referencing the use of the term 'ruses of war' by Carl Von Clausewitz).
[180] *Id.* at 440 n.49 (citing Adjutant Gen.'s Office, U.S. Dep't of War, Gen. Orders No. 100, *Instructions for the Government of Armies of the United States in the Field*, Art. 101 (Apr. 24, 1863) [hereinafter *Lieber Code*]).
[181] *Id.* at 443 (explaining "It was impossible to enumerate in the Protocol all the operations described under this heading . . ."). The Commentary also noted that the examples proposed, and ultimately included, did not provoke any debate at the Conference. *Id.*
[182] Additional Protocol I, *supra* note 174, at Art. 37(1).
[183] API Commentary, *supra* note 179, at 441.

However, its discussion quickly branches again into contrasting this characterization from acts that would not qualify as a ruse.[184]

More helpful insight on what could qualify as a ruse under Article 37 comes from the Commentary's list of examples. This includes such "commonly described" ruses of war as ambushes, simulated operations on land, air, or sea, camouflaging troops or positions "in the natural or artificial environment," and even laying dummy mines.[185] Most notably for the purposes of this article, the Commentary also listed acts which remarkably seem to foretell the evolution of 20th century electronic and cyber warfare. These ruses included:

> ". . . [T]ransmitting misleading messages by radio or in the press; knowingly permitting the enemy to intercept false documents, plans of operations, despatches [sic] or news items which actually bear no relation to reality, using the enemy wavelengths, passwords and wireless codes to transmit false instructions; pretending to communicate with reinforcements which do not exist . . . using signals for the sole purpose of deceiving an adversary; resorting to psychological warfare methods by inciting the enemy soldiers to rebel, to mutiny or desert, possibly taking weapons and transportation; inciting the enemy population to revolt against its government etc."[186]

In fact, in attempting to delineate the multiple ways that a ruse could lawfully occur, the Conference ultimately had to toss up its hands and declare "the imagination of man is too inventive for one to think that everything it could come up with can be covered in a list."[187] The drafters wisely conceded that the evolution of combat is "unforeseeable" and presciently that its nature "will always give rise to new ideas."[188]

### *B. Perfidy*

For the above reasons, then, much more effort has gone into defining what a ruse is not rather than what it is, and if a lawful deception

---

[184] *Id.* The Commentary particularly here contrasts a ruse from a "prohibited ruse" discussed further *infra* which itself contrasts against perfidy.
[185] *Id.* at 443.
[186] *Id.* at 443-44.
[187] *Id.* at 444.
[188] *Id.*

is a ruse, its opposite is perfidy and treachery. International law strictly defines p—more so than is often realized. However, whether a use of deepfake-generated content would amount to perfidy is not correspondingly easy to define.

Article 37, Section 1 of API observes that perfidy is those actions "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."[189] As the Commentary explains about this definition, perfidy "consists of the deliberate use of international law protection[s] . . . to deceive the adversary."[190]

Just like in its discussion on ruse, Article 37 also provides a list of acts which would constitute perfidy. These acts include: feigning an intent to negotiate under a flag of truce or surrender, feigning incapacitation by wounds or sickness, feigning civilian, non-combatant status, and feigning protected status by the use of signs, emblems, or uniforms of the United Nations or of states not party to the conflict.[191]

Based on the language in Article 37, perfidy has four elements: (1) inviting the confidence of an adversary, (2) with the intent of betraying that confidence, (3) betraying that confidence through the claim or demonstration of an unmerited protection afforded by international law applicable in armed conflict, and (4) killing, injuring, or capturing an adversary as a result.[192] The first two elements tend to be somewhat straightforward, though cyber dynamics can muddy what it would take to prove an "inviting."[193] The third element can be very broad in practice. It does not confine perfidy to acts which invoke Article 37 protections, or API protections, or even protections under just the Geneva Conventions. The Commentary points out that Article 2 of API provides that the laws and rules applicable to armed conflict extend outside of the Geneva Conventions to also encompass not only bilateral agreements between parties to a conflict but also "generally recognized principles and rules of

---

[189] Additional Protocol I, *supra* note 174, at Art. 37(1).
[190] API Commentary, *supra* note 179, at 444.
[191] Additional Protocol I, *supra* note 174, at art. 37(1)(a)-(d). Acts which seek to use the flag, emblem, uniform, or other insignia of a State which is not a party to the conflict, or which seek to use such emblems of an adversary, are also prohibited independently by Article 39. *Id.* at art. 39(1)-(2).
[192] *Id.* at art. 37(1). *See also* API Commentary, *supra* note 179, at 435 (only enumerating the first three elements but recognizing the fourth in later discussions).
[193] For example, a line of malicious code which does not activate until a target clicks a link or downloads a file may or may not "invite" a particular confidence yet could theoretically still be perfidious in nature.

international law which are applicable to armed conflict."[194] In other words, the protection claimed can possibly have no direct correlations to the Geneva Conventions or the Hague Conventions but still constitute perfidy.[195]

The fourth element, however, imposes a quizzical limitation. By narrowing perfidy to only those situations that produce specific consequences i.e., "to kill, injure, or capture an adversary,"[196] Article 37 dramatically narrows the purpose of enumerating perfidy at all and threatens the ability to regulate other actions that would undercut trust in the laws of armed conflict. To be fair, this possibility was not lost on the API Conference. Questions arose immediately as to why the use of deception that disabuses protections under the laws of armed conflict to achieve any other objective other than killing, injuring, or capturing prisoners—such as seizing an enemy fighting position or delaying an attack—would not also constitute perfidy.[197] Additionally, delegates wondered openly whether an *attempt* to kill, injure, or capture an adversary through perfidy would also constitute perfidy since the Article and the rest of API were silent on inchoate offenses.[198]

Unfortunately, the Conference does not appear to have provided Jean Pictet and his peers drafting the Commentary much to answer these questions. The Commentary instead notes that the drafters of the Article considered that there remained "a sort of gray area of perfidy which is not explicitly sanctioned as such, in between perfidy and ruses of war," leading to a "permanent controversy in practice as well as in theory."[199]

The Commentary did manage, however, to construct an analysis on how to read Article 37 in relation to the rest of API which helps to broaden the application of perfidy to consequences beyond killing, injuring, or capturing prisoners. First, the Commentary observes that attempts or unsuccessful acts of perfidy also fall within the definition of

---

[194] API Commentary, *supra* note 179, at 435 (citing Additional Protocol I, *supra* note 174, at Art. 2(b)).
[195] The Commentary observes "the definition of perfidy extends beyond the prohibition formulated . . ." *Id.* It notes, for example, that there are protections at sea provided by the laws of armed conflict that API does not entertain.
[196] Additional Protocol I, *supra* note 174, at Art. 37(1). The Commentary observes that this finite, exclusive list was a direct adoption from the 1907 Hague Convention IV which sought to make it illegal "to kill or wound treacherously." API Commentary, *supra* note 179 at 432 (citing 1907 Hague Convention IV, *supra* note 171, at Art. 23(b)). The Conference decided to add "capturing" as an additional nod to the combat nature of perfidy but seem to have limited it there because agreements on expanding the definition to other acts had become too difficult.
[197] API Commentary, *supra* note 174, at 432-33.
[198] *Id.*
[199] *Id.* at 433.

perfidy, although without providing any explanation why other than "it seems evident."[200] Second, the Commentary reminds readers that the Vienna Convention on the Law of Treaties demurs against interpreting treaties to conflict with a peremptory norm of general international law, and that any related peremptory norms should be read into perfidy as a result.[201] Third, Articles 38 and 39 reinforce Article 37 by prohibiting the misuse of recognized emblems under the Geneva Conventions as well as the misuse of emblems of either non-party or adversary states, respectively,[202] thus capturing a large bulk of related concerning behavior.

This latter interpretation may not sit on firm ground. It implicitly relies on Articles 37-39 to become customary international law, if not universally ratified and adopted. It does not contend with the fact that powers such as the United States would not and still have not ratified API and, unlike Articles 37 and 38, does not consider Article 39 to be customary international law.[203] It is even less ready to resolve applicability in the face of inter-government disagreements about applicability, such as how the United States has accepted that Article 37 reflects customary international law[204] (and is therefore binding on the United States) but its own Department of Defense has declared that Article 37's "capture" is actually not a part of customary international law.[205] Nonetheless, these observations remain helpful for discerning a wider landscape in which to declare acts beyond those resulting in killing, injuring, or capturing individuals to be perfidious.

---

[200] *Id.*

[201] *Id.* (citing Vienna Convention on the Law of Treaties art. 53, May 23, 1969, 1155 U.N.T.S. 331; 8 I.L.M. 679 [hereinafter Vienna Convention]).

[202] *Id.*

[203] *See* Matheson, *supra* note 173, at 425.

[204] *Id.*

[205] DOD LAW OF WAR MANUAL, *supra* note 146, at ¶ 5.22.2.1. The DoD Law of War Manual does not provide any authority on which it bases its interpretation. FM 6-27 is careful to only define perfidy as "wounding or killing" the enemy while possessing a protected status. FM 6-27, *supra* note 168, at ¶¶ 1-82, 2-91, 2-109, 2-151, 2-152 (citing the DoD Law of War Manual; the latter paragraph only invites the reader to "consider" API, Art. 37). The only discussion of capture and perfidy in FM 6-27 instead relates that "any combatant who feigns death in the hope of evading capture has not engaged in perfidy," a notion which would not offend Articles 37-39. *Id.* at ¶ 2-152.

### *C. Treachery a.k.a. Violations of Honor*

Scholars do not usually discuss this third category explicitly as "violations of honor" but instead in terms of "treachery" to touch on the concept's historical roots, usage, and proximity to perfidy.[206]

#### 1. Chivalry and Honor

Battlefield concepts of "chivalry" referenced in international humanitarian law today are well-documented as originating in Europe during the Middle Ages[207] and have still managed to persist, albeit to varying degree, into the 21st century.[208] The API Commentary acknowledged the role of chivalry in fostering concepts of honor also found in contemporary laws of armed conflict, noting that "[t]his sense of honour, which was nourished during the Middle Ages of Europe by chivalry, particularly in tournaments and in jousting, has contributed to the establishment of the rules which finally became assimilated into the customs and practices of war . . ."[209] The Commentary characterizes the battlefields of Medieval Europe, or at least the Christian warriors at that time, as steeped in "rules for attack and rules for defence," and that undergirding conduct in battle was the notion that "the knight always trusted the word of another knight, even if he were an enemy."[210] This notion was so strong, the Commentary posited, that "[p]erfidy was

---

[206] *See generally* Watts, *supra* note 171. Whether treachery is actually a distinct concept from perfidy is still debatable. Significant historical evidence does support the position that they are substantively different with the former more concerned about violations of ethical or chivalrous expectations of good-faith behavior and the latter concerned about abuses of international law's protections in ways which could neutralize the law itself. *Id*. at 109, 113-14, 125-29, 134-37, 140-41 (discussing distinctions between the two concepts as found in, among other things, the Lieber Code, the 1907 Hague Regulations, the prosecution of defendants at the Tokyo International Military Tribunal, and the 2009 Military Commissions Act).

[207] *See* Watts, *supra* note 171 at 106, 157-58 (citing Geoffrey Parker, *Early Modern Europe*, and Robert C. Stacey, *The Age of Chivalry*, *in* THE LAWS OF WAR 29-31, 54 (Michael Howard, George J. Andreopoulos, & Mark Shulman eds., 1994)).

[208] *See e.g.*, JUDGE ADVOC. GEN., CANADIAN ARMED FORCED, LAW OF ARMED CONFLICT AT THE OPERATIONAL AND TACTICAL LEVELS, B-GJ-005-104/FP, 2-1 (2001) (declaring chivalry to be a core principle of the laws of armed conflict). Indeed for 63 years before it was updated in 2019, the U.S. Army's primary field manual on the laws of armed conflict expressly required that U.S. Soldiers abide by chivalry as a core principle of armed conflict. U.S. DEP'T OF ARMY, FIELD MANUAL 27-10, THE LAW OF LAND WARFARE, para. 3(a) (Jul. 1956) [hereinafter FM 27-10] (superseded by FM 6-27, *supra* note 168).

[209] API Commentary, *supra* note 179, at 434.

[210] *Id.*

considered a dishonour which could not be redeemed by any act, no matter how heroic."[211]

However, the Commentary had to acknowledge that notions of chivalry, honor, and good faith had no single origin, and that no particular culture could claim sole ownership over any of the concepts.[212] While chivalry *per se* may have originated in Christian Europe, it could not reliably inform modern notions of treachery. Medieval Christian warriors, of course, did not always abide by their own oaths of chivalry. Often, they were prone to abandon their chivalric code—with no legal or immediate political consequence—when facing non-Christian adversaries, such as during the First Crusade when Crusader armies brutally stormed Jerusalem in 1099, slaughtered many of the city's inhabitants, and desecrated several holy sites.[213]

However, in contrast to the limited application of chivalry, honor and good faith have been facets of armed conflict across the world. The Islamic warriors who opposed Crusaders at Jerusalem, for example, and in later Crusades had their own ethics of honor, founded in their own religious beliefs and world views rather than European or Christian culture.[214] The famous samurai warriors of Japan were required to prize

---

[211] *Id.* The view that acts of treachery by knights could incur a lifelong bounty has some support. *See, e.g.*, Watts, *supra* note 171, at 106 (citing Parker, *supra* note 207) (stating "medieval notions of honor and chivalry sanctioned unending blood feuds to avenge knights killed by treachery.").

[212] API Commentary, *supra* note 179, at 434 (observing "Perfidy is injurious to the social order which it betrays, regardless of the values on which this social order is founded.").

[213] *See, e.g.*, JAY RUBENSTEIN, ARMIES OF HEAVEN: THE FIRST CRUSADE AND THE QUEST FOR APOCALYPSE 290 (2011) (quoting Raymond of Aguiler, a French participant in the sack of Jerusalem, who boasted that in the city, "[s]ome had their heads cut from their bodies (which was fairly merciful) or were hit with arrows and forced to jump from towers. Others suffered for a long, long time, and were consumed and burned up in flames. Horses and men on public roads were walking over bodies. But these things I say are trifling. Let us go to the Temple of Solomon.").

[214] The distinguished 12th century Muslim warrior and writer Usama Ibn Munqidh, a native of modern-day Syria and witness to the Second Crusade, wrote extensively about the multi-cultural world of the contemporaneous Near East. His writings often contrasted his views on honor and good behavior, informed by his own Islamic beliefs, with the behavior and lack of honor he perceived of "Franks" (as he called all Europeans, even if they did not come from France) who he characterized as unintelligent and "[possessing] nothing in the way of regard for honour or propriety." USAMA IBN MUNQIDH, THE BOOK OF CONTEMPLATION: ISLAM AND THE CRUSADES 144, 148 (Paul M. Cobb trans., Penguin Group 2008) (1183). In fact, in one anecdote he conveys that the invitation from a close Christian friend for Usama's son to come to Europe to "acquire reason and chivalry" was kind but laughable and was carefully refused. *Id.* at 144. Historian Will Durant has observed that during the Crusades the Islamic forces, while themselves not strangers to inflicting suffering or division, on the whole "seem to have been better gentlemen than their Christian peers; they kept their word more frequently, showed more mercy to the defeated, and were seldom guilty of

honor as "more important than life itself."[215] Today, the famous *pukhtunwali* code in Afghanistan, revered most fervently by the country's Pashtun population who are also called to abide by it even in battle, maintains honor as one of its three pillars (the other two being hospitality and revenge/justice).[216]

While chivalry may not be as fashionable a concept today as it once was in international humanitarian law, honor and good faith are still relevant. The United States Army continues, as it has for several decades, to declare "honor" to be one of its seven Army Values along with related concepts such as "respect, duty, loyalty, selfless service, integrity, and personal courage, in everything Soldiers and Marines do."[217] Additionally, the Army very recently confirmed the continued legal relevance of honor in FM 6-27 which effectively sidelines "chivalry" in favor of "honor." Characterizing honor as a "core Army and Marine Corps value,"[218] FM 6-27 declares honor as a "basic LOAC principle" in line with the other four historically-accepted principles of distinction, proportionality, military necessity, and humanity.[219] FM 6-27 defines the concept as "[t]he LOAC principle [sic] that demands a certain amount of fairness in offense and defense and a certain mutual respect between opposing forces."[220] Honor "gives rise to rules that help enforce and give effect to LOAC"[221] and "provides legitimacy to the entire endeavor."[222] While the concept does not define its limits, FM 6-27 observes that the

---

such brutality as marked the Christian capture of Jerusalem in 1099." WILL DURANT, THE STORY OF CIVILIZATION: PART IV, THE AGE OF FAITH 341 (1950).

[215] *See* Nicholas W. Mull, *The Honor of War: Core Value of the Warrior Ethos and Principle of the Law of War*, 18 CHI.-KENT J. INTL'L & COMP. L. 1, 23 (2018) (citing YAMAMOTO TSUNETOMO, HAGAKURE: THE BOOK OF THE SAMURAI 30 (William Scott Wilson trans., Kodansha Int'l 1979) (1716)).

[216] *See, e.g.*, Ken Guest, *Dynamic Interplay Between Religion and Armed Conflict in Afghanistan*, 92 INT'L REV. RED CROSS 877, 886 (2010). Mr. Guest observes in his article that *pukhtunwali* "represents an ideal rather than an absolute—not dissimilar to Western concepts of chivalry." However, Mr. Guest also remarks that such similarity also leaves *pukhtunwali* susceptible to issues similar to chivalry, namely "it is subject both to personal interpretation (which can be very creative) and to common abuse." *Id. See also* ANDREA CHIOVENDA, CRAFTING MASCULINE SELVES: CULTURE, WAR, AND PSYCHODYNAMICS IN AFGHANISTAN 41-44, 46, 190 (2020)(discussing two separate concepts of honor in Afghan Pashtun male culture, particularly, *izzat* (masculine honor requiring revenge for insults) and *namus* (honor which demands modesty)).

[217] *See* FM 6-27, *supra* note 168, at ¶ 1-31.

[218] *Id.*

[219] *Id.* at ¶¶ 1-18 to 1-21.

[220] *Id.* at Glossary-3. The most treatment that chivalry gets from FM 6-27 is an equation to "honor" (in fact, the next sentence in the definition is "[a]lso called chivalry."). However, FM 6-27 does not treat chivalry as a separate concept either in definition or in consequence.

[221] *Id.* at para. 1-32.

[222] *Id.* at para. 1-21.

principle "require [sic] that parties accept . . . that certain legal limits exist."[223]

2. Good Faith

As for "good faith" in relation to the laws of armed conflict, sources today apply requirements and expectations of good faith almost as broadly as sources of yesteryear. As early as the sixteenth century, scholars on the laws of war such as the Dutch military jurist Balthazar Ayala observed that throughout history "there was no grander or more sacred matter in human life than good faith."[224] The 1863 Lieber Code instructed federal armies in the American Civil War that deception was permissible so long as it "does not involve the breaking of good faith either positively pledged . . . or supposed by the modern law of war to exist."[225]

In the twentieth century, good faith gained new *ius ad bellum* purchase as the post-World War II global order ardently embraced international law. The United Nations Charter now demands that Member States "shall fulfil in good faith the obligations assumed by them."[226] The Vienna Convention on the Law of Treaties provides at Article 26 that "[e]very treaty in force is binding upon the parties to it and must be performed by them in good faith."[227] In a *ius in bello* context, while the Hague and Geneva Conventions do not expressly use the term, the API Commentary shows that good faith often featured in the Additional Protocol's underlying philosophies, making pronouncements about the "rules on good faith"[228] and that these same rules "prohibit killing or wounding the enemy treacherously, as well as deceiving him by the improper use of the flag of truce, of national emblems or of enemy uniforms, and also by the improper use of the red cross emblem."[229] The Commentary even makes sure to stress that the obligation to think with good faith when engaged in armed conflict does not just sit with the lawyers "but is also imposed on those who enjoy a

---

[223] *Id.* at para. 1-32.
[224] Watts, *supra* note 171 at 174-75 (citing BALTHAZAR AYALA, 2 THREE BOOKS ON THE LAW OF WAR AND ON THE DUTIES CONNECTED WITH WAR AND ON MILITARY DISCIPLINE 55 (John P. Bate trans., 1912) (1582)).
[225] *Lieber Code*, *supra* note 180, at art. 15.
[226] U.N. Charter art. 2, ¶ 2.
[227] *See* Vienna Convention, *supra* note 201, at art. 26.
[228] API Commentary, *supra* note 179, at 382.
[229] *Id.* This is of course notwithstanding the objections made by the United States to the related provisions in Article 39 discussed *supra*.

certain degree of freedom of action in the field, even though the heat of battle does not favour an objective view of things."[230]

Today, both the DoD Law of War Manual and FM 6-27 make good faith a distinct concept in the United States Armed Forces, with FM 6-27 particularly declaring that "absolute good faith" is an essential component of armed conflict and its violation garners separate consequences.[231] Furthermore, these regulations impose expectations of good faith in several aspects of combat from assessing whether a person or object is a lawful target[232] to making agreements for the removal of vulnerable populations during a siege,[233] implying that good faith is a concept essential not only to actions done while interacting with the external enemy but also to decision-making situations requiring internal honesty.

## V. ENFORCING THE LAWS ON DECEPTION IN ARMED CONFLICT

Consequences can vary widely for an actor's violation of the laws and principles related to deception in armed conflict. Distinctions arise not just in what kind of deception is used but how it is used, for what purpose, where, and by whom, the final being the hardest question to answer due to attribution challenges.

### *A. Perfidy – Grave, Prohibited, and Simple*

Perfidy is the act with the most immediate severity and consequences. As United States Military Academy Professor Sean Watts explains, "[p]erfidy and treachery are among the gravest law-of-war violations . . . perfidy and treachery provoke draconian and irreversible reactions."[234] Amassing an impressive survey of perfidy from its treatment over the centuries, Professor Watts correspondingly articulates three kinds of perfidy in existence today which have different roots and different enforceability. The first is simple perfidy, described as "all acts" that falsely invite an enemy to provide law-of-war protections and then betray that confidence.[235] The second is prohibited perfidy,

---

[230] *Id.*
[231] FM 6-27, *supra* note 168, at ¶¶ 2-146, 2-147, 2-148, and 2-149. See discussion on consequences *infra*.
[232] DoD Law of War Manual, *supra* note 146 at ¶ 11.18.2.1; FM 6-27, *supra* note 168, at ¶ 2-17.
[233] FM 6-27, *supra* note 168, at ¶ 2-102.
[234] Watts, *supra* note 171, at 106.
[235] *Id.* at 154.

described as perfidious acts that "result in death, injury, or capture of the betrayed enemy."[236] The third is grave perfidy, described as acts of prohibited perfidy that willfully use the recognized emblems under the Geneva Conventions such as the Red Cross or Red Crescent against persons protected under the Geneva Conventions.[237]

The source of simple perfidy in this interpretation appears to be customary international law (described here as "broad custom")[238] that informed notions of honor and prohibited corresponding acts of treachery, and which still remains the only source of international law to prohibit such acts when not covered by written laws. The source of prohibited perfidy, by contrast, is Article 37 of Additional Protocol I with its distinct consequence limitations. The source of grave perfidy is equally concrete, this time originating from Article 85(3)(f) of Additional Protocol I which declares the "perfidious use, in violation of Article 37" of recognized emblems or other protective signs as "grave breaches."[239] The enforcement mechanisms for prohibited perfidy and grave perfidy, however, are not equally concrete, and enforcement mechanisms for simple perfidy are difficult to define.

1. Grave Perfidy

Grave perfidy enjoys the largest degree of certainty. By declaring this very specific vein of Article 37 perfidy a "grave breach," states parties must automatically promulgate domestic legislation in accordance with the grave-breaches provisions in all four 1949 Geneva Conventions to enact "effective penal sanctions" to repress grave perfidy.[240]

---

[236] *Id.*
[237] *Id.*
[238] *Id.*
[239] Additional Protocol I, *supra* note 174, at art. 85(3)(f). While Article 85 declares the acts listed under section 3 to be grave breaches when "causing death or serious injury to body or heath," Professor Watts posits that because subsection (3)(f) explicitly nests into Article 37, even here only acts which misuse recognized emblems in order to cause "killing, injury, or capture" would constitute grave perfidy. Watts, *supra* note 171, at 153.
[240] Geneva Convention for the Amelioration of the Condition of the Wounded and Sick in Armed Forces in the Field art. 49, Aug. 12, 1949, 6 U.S.T. 3114, 75 U.N.T.S. 31 [hereinafter Geneva Convention I]; Geneva Convention for the Amelioration of the Condition of the Wounded, Sick, and Shipwrecked Members of Armed Forces at Sea, art. 50, Aug. 12, 1949, 6 U.S.T. 3217, 75 U.N.T.S. 85 [hereinafter Geneva Convention II]; Geneva Convention Relative to the Treatment of Prisoners of War art. 129, Aug. 12, 1949, 6 U.S.T. 3316, 75 U.N.T.S. 135 [hereinafter Geneva Convention III]; Geneva Convention Relative to the Protection of Civilian Persons in Time of War art. 146, Aug. 12, 1949, 6 U.S.T. 3516, 75 U.N.T.S. 287 [hereinafter Geneva Convention IV].

The "grave breach" designation also triggers several other effects in API. Article 85 declares that grave breaches "shall be regarded as war crimes,"[241] Article 86 requires states parties to "repress" grave breaches[242], and Article 87 requires state parties both to direct military commanders under their control "to prevent and, where necessary, to suppress and to report . . . breaches of the Conventions and of this Protocol"[243] and to require any commander who is aware that a breach has or will occur "where appropriate, to initiate disciplinary or penal action against violators thereof."[244]

Furthermore, Article 88 encourages maximum cooperation in criminal proceedings between aggrieved state parties.[245] Article 90 International Fact-Finding Commissions can investigate grave breaches independent of state party efforts to investigate.[246] Finally, Article 91 provides a minimum requirement for a "[p]arty to the conflict which violates the provisions of the Conventions or of this Protocol" to "pay compensation" (although API does not discuss the method, amount, currency determination, or means for deciding compensation disputes).[247]

Additionally, grave breaches of the Geneva Conventions as well as API are subject to universal jurisdiction.[248] While observers and jurists have debated the actual extent of this jurisdiction,[249] the fact remains

---

[241] Additional Protocol I, *supra* note 174, at art. 85(5). This provision is notable because the Convention for the 1949 Geneva Conventions purposefully did not equate grave breaches (a novel term at the time) to war crimes. The Conference felt the term "crimes" had too many different meanings and so sought to avoid it. *See* Gary D. Solis, INTRODUCTION TO GENEVA CONVENTIONS 25 (Kaplan Publishing, 2010).
[242] Additional Protocol I, *supra* note 174, at art. 86(1).
[243] *Id.* at art. 87(1).
[244] *Id.* at art. 87(3).
[245] *Id.* at art. 88(3). This subsection provides that "the law of the High Contracting Party requested shall apply in all cases." *Id.* How this choice-of-law decision would resolve would likely revolve around political considerations, although legal considerations could certainly be determinative if, for example, a State could not muster the resources to conduct prosecutions because armed conflict had crippled its law enforcement infrastructure.
[246] *Id.* at art 90(2)(c)(i).
[247] *Id.* at art. 91.
[248] Universal jurisdiction comes from the 1949 Geneva Conventions' demand that the States Party are "under the obligation to search" for people accused of committing grave breaches and "shall bring such persons, regardless of their nationality, before its own courts." They may also exercise, through the principle of *aut dedere aut judicare*, the option to extradite the accused to the custody of another state party for trial. Geneva Convention I, *supra* note 240, at art. 49; Geneva Convention II, *supra* note 240, at art. 50; Geneva Convention III, *supra* note 240, at art. 129; Geneva Convention IV, *supra* note 240, at art. 146.
[249] *Compare, e.g.*, Roger O'Keefe, *The Grave Breaches Regime and Universal Jurisdiction*, 7 J. INT'L CRIM. J. 811 (2009) (arguing that universal jurisdiction only

that this jurisdictional component is unique to grave perfidy compared to the other two varieties discussed here.

2. Prohibited Perfidy

Prohibited perfidy, by comparison, enjoys only a portion of the codified support necessary for a state to embark on a prosecution or a related legal sanction. The biggest substantive distinction is that while prohibited perfidy captures the *actus reus* qualifications under Article 37 of API, it does not concern the misuse of recognized emblems which when paired with a perfidious act would elevate the crimes to the grave-breaches threshold. Because prohibited perfidy here does not rise to the level of a grave breach, prohibited perfidy does not automatically qualify as a "war crime" under API.[250] None of the States Party have to criminalize it distinctively as "grave breaches" in their domestic criminal codes. They are also not under an affirmative obligation to "repress" prohibited perfidy and none of the States Party are required to bring anyone to trial for committing prohibited perfidy.[251] Finally, universal jurisdiction does not apply to prohibited perfidy, meaning that a state not party to the conflict would have no unilateral ability[252] to prosecute an actor who the state felt committed perfidy (albeit not constituting a grave breach) under Article 37—an omission that can have real-world impacts on efforts to prosecute perfidious uses of deepfake technology.

On the other hand, states party to the API still have an affirmative obligation to assure under Article 87 that their military commanders understand and execute their duties to prevent, suppress, report, and,

---

applies in those circumstances when no other country has made a proper claim to jurisdiction), *with* Arrest Warrant of 11 Apr. 2000 (Democratic Republic of the Congo v. Belgium), Judgment, 2002 I.C.J. 1, 24-25, § 59 (Feb. 14) (finding that universal jurisdiction and the "customary international law" which informs it permits one country's court to have jurisdiction to issue arrest warrants and another country's court to have jurisdiction to afford immunity).

[250] *See supra* note 239 and discussion.

[251] *See* API Commentary, *supra* note 179, at 159 (providing "Although the Parties to the conflict are under the obligation to take measures necessary for the suppression of all acts contrary to the provisions of the Conventions and Protocol I, they are only bound to bring to court persons having committed grave breaches of these treaties . . .").

[252] This presumes no other treaties—bilateral or multilateral—exist at the time which would provide said third-party State with jurisdiction. Additionally, Jean Pictet reasoned in the API Commentary that customary law supports the application of universal jurisdiction to "serious violations of the laws of war" regardless of whether they qualify as grave breaches. API Commentary, *supra* note 179, at 1011. This position, however, may not reflect actual customary international law or even a consensus among States Party. *See, e.g.,* Matheson, *supra* note 173.

"where appropriate, to initiate disciplinary or penal action" against acts which violate the Geneva Conventions, to include prohibited perfidy.[253] Additionally, the Geneva Conventions impose on states party the particular obligation to take domestic legal measures to prevent and repress "at all times" acts by any entity, public or private, that make unlawful use of recognized emblems.[254] The call for investigatory cooperation in Article 88 also applies[255] as may also the requirement to cooperate with the United Nations during investigations of "serious violations" under Article 89.[256] While the Article 90 International Fact-Finding Commission does not have unilateral authority to investigate non-grave breaches, it may still conduct inquiries into "other situations" so long as both parties to the conflict consent to the investigation.[257] Finally, the minimum penalty under Article 91 still applies as well.[258] While the United States does not believe that Articles 90 and 91 reflect customary international law and is unlikely to enforce them, the United

---

[253] Additional Protocol I, *supra* note 174, at art. 87(1), (3).
[254] *See, e.g.*, Geneva Convention I, *supra* note 240, at arts. 53-54.
[255] *See supra* note 245 and discussion.
[256] Additional Protocol I, *supra* note 174, at art. 89. API does not define the term "serious violations" which is purposefully distinct from "grave breach" terminology. The Commentary provides that the Conference initially intended this section to address reprisals in order to avoid breaches being answered by more breaches. However, the revision process neutered that intent and resulted in a broad article simply requiring cooperation with the United Nations. The Commentary says that "[t]he terms 'violation' and 'breach' may be considered to be synonymous." The Commentary, though, does not equate "serious violations" with "grave breaches" which it acknowledges the Conference purposefully made distinct from all other violations. The Commentary states flippantly "[w]e do not need to have in mind exactly what conduct could fall under this definition" in order to avoid proposing a definition. Instead, it posits three categories of acts which would equate to a "serious violation": (1) non-grave isolated acts "of a serious nature," (2) non-grave acts which because of frequency or other circumstances "takes on a serious nature," and (3) "'global' violations" described as "acts whereby a particular situation, territory or a whole category of persons or objects is withdrawn from the application of the Conventions of the Protocol." API Commentary, *supra* note 179, at 1032-33. Research does not show any application of this three-tier definition of "serious violations." Instead, practice appears to show that declaring an act to be a serious violation can be more by feel and circumstance than adherence to a rigid definition. Furthermore, the finding of an act to constitute a serious violation does not garner any more resources or heightened sanctions under the Geneva Conventions or Additional Protocol I than any other non-grave violations. *See, e.g.*, Tadić, *supra* note 25, at ¶¶ 90-91 (finding "It is therefore appropriate to take the expression 'violations of the laws or customs of war' [found in Article 3 of the ICTY Statute] to cover serious violations of international humanitarian law" and that the intent to hitch "serious violations" to the broader concept of violations of laws or customs of war in the ICTY Statute (itself drafted very closely to the Geneva Conventions and Additional Protocol I) was to make the Tribunal's jurisdiction "watertight and inescapable.").
[257] Additional Protocol I, *supra* note 174, at art. 90(2)(d).
[258] *Id.* at art. 91.

States has expressed the intent to require "[a]ppropriate authorities" to take "all reasonable measures to prevent acts contrary to the applicable rules of humanitarian law," "to bring to justice all persons who have willfully committed such acts," and "to cooperate" with other States Party in related proceedings.[259]

### 3. Simple Perfidy

By comparison, the enforcement mechanism for prohibiting so-called simple perfidy would be unpredictable. Some acts may fall within prohibitions on the misuse of recognized emblems, such as in Article 53 of Geneva Convention I,[260] but not involve any killing, injury, or capture—acts described by Jean Pictet as "prohibited ruse."[261]

For example, some acts may deliberately make an adversary falsely believe they have law-of-war protections but not involve either the misuse of a recognized emblem or a killing, injury, or capture. This is aplausible scenario should deepfake technology proliferate in combat, for example, to convince a belligerent to send supplies to an adversary or to waste instead of to the intended recipient. Other acts may engage in the seemingly perfidious behavior but have no other intended and actual effect than to sow confusion and distrust—also equally plausible as a utility for deepfake. So long as an act of deception can invoke some portion of the Geneva Conventions or the Additional Protocols, these instruments can provide some mechanism to repress and punish those actions similar to the above discussion on prohibited perfidy. Where they do not invoke either document, however, the alleged simple perfidy is likely to blend into a correspondingly simple notion of violating honor—and encounter corresponding repression challenges.

---

[259] *See* Matheson, *supra* note 173, at 428.
[260] Geneva Convention I, *supra* note 240, at art. 53.
[261] API Commentary, *supra* note 179, at 441, 443. Mr. Pictet argues here that "a distinction should be made between a ruse, a prohibited ruse, and an act of perfidy," with a prohibited ruse constituting primarily those acts of deception which unlawfully employ recognized emblems but do not meet the requirements of Article 37 to constitute perfidy. Mr. Pictet goes on to surmise that "prohibited ruse" could also theoretically apply to acts involving delayed-action weapons such as mines and certain booby-traps. However, the extent to which international humanitarian law has adopted this suggestion is not clear. Notably, the 1999 Convention on the Prohibition of the Use, Stockpiling, Production and Transfer of Anti-Personnel Mines and on Their Destruction ("Ottawa Treaty") contains no reference to concepts of ruse, perfidy, or deception in general.

### *B. Violations of Honor and the Problem with Treachery*

Violations of honor as described above are difficult to legally articulate, especially under international humanitarian law, particularly because they are not, by definition, perfidious. They can be offensive, cause tremendous confusion, inflict irreparable damage to property, and themselves also make peace harder to establish. But they do not enjoy express positive prohibition in the modern laws of armed conflict. As Professor Watts has observed, "claims to a complementary, broad-based perfidy prohibition derived from notions and principles of chivalry and honor are overstated. Such claims seem grounded in little more than nostalgia, hardly worthy of legal recognition."[262] However, these claims are more easily addressed in domestic laws and codes.

Today chivalry is, from an international humanitarian law perspective, dead letter.[263] Some modern efforts have attempted to re-define the legal notion of chivalry,[264] but the concept is instead often wrapped into discussions on honor and good faith.

Violations of honor and good faith, by comparison, enjoy more robust treatment in international law. The API Commentary, for example, notes particularly that "[Articles 37-39 of API] appeal to the good faith of the combatant which is a fundamental condition for the existence of law."[265] However, neither the Hague nor the Geneva Conventions, nor their Protocols define deceptive actions that constitute explicit violations of "honor" or "good faith." Instead, the laws offer general expressions that States Party must abide by their obligations honorably and/or in good faith[266] and that states remain bound to "principles of the law of nations" (presumably including principles of honor and good faith) as derived from "the laws of humanity and the

---

[262] Watts, *supra* note 171, at 174.
[263] *Id.* at 160 (observing "Chivalry as a principle . . . would be unlikely to actually regulate the conduct of hostilities or form a reliable basis for law-of-war enforcement efforts such as criminal prosecution.").
[264] *See, e.g.*, Evan J. Wallach, *Pray Fire First Gentlemen of France: Has 21st Century Chivalry Been Subsumed by Humanitarian Law?*, 3 HARV. NAT'L SEC. J. 431, 443-60 (2012) (seeking to define modern chivalry by the concepts of courage, trustworthiness, mercy, loyalty, and courtesy; also argues that "[c]hivalry mandates actions and punishes inaction that IHL can only recommend.").
[265] API Commentary, *supra* note 179, at 473.
[266] 1907 Hague Convention IV, *supra* note 171, at pmbl.(commending the instrument to, *inter alia*, the "dictates" of the public conscience); Additional Protocol I, *supra* note 174, at art. I(1) (requiring that States Party "respect" the Protocol "in all circumstances.").

dictates of public conscience" even if they try to withdraw from or denounce the Conventions.[267]

This is not to say that these aspirations are not important or do not present jeopardy for a potential violator. The latter aspirations particularly, reflective of the famous "Martens Clause" found in the preamble of the 1899 Hague Convention[268] and further codified in the Geneva traditions,[269] make it plain that states remain bound to customary international law even if they try to withdraw from international multilateral treaties and will remain stubbornly so especially when these treaties reflect customary international law. However, the fact that armies of scholars and international jurists have proclaimed that honor and good faith are central components of the laws of armed conflict does not guarantee that a perpetrator who employs deepfake technology in odious but not perfidious ways during armed conflict can easily face trial.

This does not mean, however, that an actor hoping to employ deepfake technology in such a manner does so free of any consequences. Uses of deepfake technology could make the actor a lawful target for non-lethal and even potentially lethal force. Consider, for example, a scenario in which a non-state actor in a Common Article 3 non-international armed conflict[270] has crafted a successful deepfake campaign which has contributed significantly to losses of vital war-fighting materiel for an opposing state force. The opposing state force has through various

---

[267] *See, e.g.*, Geneva Convention I, *supra* note 240, at art. 63; Geneva Convention II, *supra* note 240, at art. 62; Geneva Convention III, *supra* note 240, at art. 142; Geneva Convention IV, *supra* note 240, at art. 158; Additional Protocol I, *supra* note 174, at art. 1(2); 1907 Hague Convention IV, *supra* note 171, at pmbl.

[268] 1899 Hague Convention II, *supra* note 172, at pmbl. (declaring "[u]ntil a more complete code of the laws of war is issued, the High Contracting Parties think it right to declare that in cases not included in the Regulations adopted by them, populations and belligerents remain under the protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of the public conscience"). This provision, advocated for by Russian delegate Friedrich Martens at the Conferences to the 1899 Hague Conventions, was a compromise intended to keep disagreements about the Conventions' applicability and enforceability from scuttling the Conventions' creation. Today the Clause itself has received recognition and enforcement in the highest international forum. *See, e.g.*, Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, 257 (Jul. 8) (describing the Martens Clause as "an effective means of addressing the rapid evolution of military technology" and proclaiming that the Clause's "existence and applicability is not to be doubted".)

[269] Geneva Convention I *supra* note 240, at art. 63.

[270] Recall that Additional Protocol I and its prohibited perfidy could not apply here because Additional Protocol I only governs international armed conflicts. *See* discussion *supra* note 174; *see also* discussion about Tadić factors *supra* note 25.

means been able to positively identify the actor at a location which is in an area of active hostilities. The opposing state force believes that the actor's skill in deploying deepfake derived deceptions poses a threat to their lawful military objectives and to the personal safety of their own forces. If the opposing state force can correctly conclude that the actor is directly participating in hostilities, the opposing state force would be within its right under the laws of armed conflict to take lethal action against the actor.[271]

The next level of potential consequence would be prosecution in the opposing state force's domestic criminal system. If the actor, seized in a raid, were detained by the opposing state force, the opposing state force could prosecute the actor in a regularly constituted court[272] under domestic laws which assert personal jurisdiction over the actor. For example, if the opposing state force were the United States—which has a well-established (albeit it controversial) practice of employing military tribunals to try violations of the laws of war[273]— the actor could face trial by a United States military tribunal or even a general court-martial by way of Rule for Court-Martial 201 which applies personal jurisdiction over "any person" who "is subject to trial by military tribunal for any crime or offense against the law of war . . ."[274] In such a case, the actor faces severe legal jeopardy as the Uniform Code of Military Justice authorizes various punishments including the death penalty in cases where violations of the laws of war result in death.[275]

---

[271] This presumes that other non-lethal means, such as conducting a reciprocal malicious cyberattack against the actor's computer or servers or even a raid to arrest the actor, are not feasible. At any rate, if the actor is directly participating in hostilities at the time the actor is observed for targeting, the opposing state force would have no legal obligation to pursue non-lethal means first.

[272] *See* Geneva Convention I, *supra* note 240, at art. 3(2).

[273] *See, e.g*,. Ex parte Quirin, 317 U.S. 1, 45-46 (1942) (upholding the trial of German saboteurs by a U.S. military commission for violations of the laws of war); Hamdi v. Rumsfeld, 542 U.S. 507, 537 (2004) (plurality opinion) (observing that an enemy combatant detainee could be prosecuted by and have habeas corpus petitions entertained by a "properly constituted military tribunal"); for a perspective skeptical of the notion of using U.S. military tribunals to prosecute enemy combatant detainees, *see* Michael R. Belknap, *Alarm Bells from the Past: The Troubling History of American Military Commissions*, 28 J. SUP. CT. HIST. 300 (2003).

[274] MANUAL FOR COURTS-MARTIAL, UNITED STATES, R.C.M. 201(f)(1)(B)(i)(a) (2019) [hereinafter MCM]. This same subsection of R.C.M. 201 also declares that a general court-martial in such a case "may adjudge any punishment permitted by the law of war." *Id.* at R.C.M. 201(f)(1)(B)(ii); *see also id.* at R.C.M. 1003(d)(10)(explaining a general court-martial may, in cases tried under the law of war, adjudge any punishment "not prohibited by the law of war."); UCMJ art. 18.

[275] *See, e.g*., UCMJ art. 81(a)(b) (discussing the potential application of the death penalty in the case of a conspiracy to violate the laws of war that results in death).

Ultimately, whether and to what extent an actor may face prosecution for violations of honor in relation to a use of deepfake technology would be heavily fact dependent. Whether the forum is a U.S.-style military commission or court-martial, a domestic court, or even in an *ad hoc* international tribunal, if the governing code or tribunal charter does not carefully account for the distinctions discussed here then it will set its prosecutors up to fail.[276]

### *C. An Argument in Favor of Deepfakes: Lawful Ruse*

Although much of this article has discussed circumstances in which deepfake manipulation would violate the laws of armed conflict, it is equally important to acknowledge that the employment of deepfake-based deception is not, by itself, illegal. Just like any other medium of deception, deepfake technology is not *per se* banned from war.

Deepfake deception can be as perfectly lawful a utility during the conduct of military operations as many acts of deception have been throughout history. For example, a belligerent could use deepfake-derived content to make an enemy think an attack was occurring on one outpost in order to create a distraction allowing the belligerent to attack a different outpost. In order to thwart an attack, a besieged belligerent could fake its numbers by broadcasting deepfake videos seeming to show hundreds of defenders at a base when in reality the base may only have a couple dozen defenders. As discussed below, even the deepfake

---

[276] A classic example of the folly inherent in trying to prosecute violations of honor without a concrete understanding of the offense occurred during the proceedings of the 1946 International Military Tribunal for the Far East (a.k.a. the Tokyo War Crimes Tribunal). There, prosecutors charged the Japanese defendants with, *inter alia*, violating Article 23(b) of the 1907 Hague Convention by committing Article 23(b) treachery which allegedly occurred when Japan attacked the United States at Pearl Harbor. The prosecutors and the Tribunal both failed to understand the fundamental divide in international law between *ius ad bellum* and *ius in bello*. The prosecutors confused *ius in bello* treachery under Article 23(b)—which would occur when the deception works to affect a hostile act while engaged in combat—with the *ius ad bellum* facts charged i.e., that Japan had engaged in diplomatic deception to affect a hostile act in furtherance of securing an advantage in a war that had not yet come, which the Hague Regulations are powerless to regulate. While the Tribunal did not rule against the prosecutors because of their erroneous charge, the result was the same—the Tribunal did not convict the defendants, applying a *ius in bello*-style rationale that the United States was in possession of too much information about Japan's intentions before the attack for the bombing of Pearl Harbor to constitute a violation of Article 23(b). *See* Watts, *supra* note 171, at 141 (citing NEIL BOISTER & ROBERT CRYER, THE TOKYO INTERNATIONAL MILITARY TRIBUNAL: A REAPPRAISAL 171 (2008)). Had the prosecutors and the Tribunal understood that they needed to apply *ius ad bellum* law to the *ius ad bellum* facts before them, the Tribunal's ruling on the Pearl Harbor attack may have been different.

manipulation of an enemy's satellite-based geo-spatial imagery could be done lawfully as part of a ruse. Synthetic content created and delivered by AI could support a "feint"[277] to deceive an adversary as to the time or place of a knockout assault, thereby winning a war or causing a belligerent to lose one.

To a military theorist, perhaps deepfake's most effective use would be to infiltrate an opponent's OODA Loop. The OODA Loop is the Observe-Orient-Decide-Act cyclic chain pioneered by the late U.S. Air Force Colonel (Ret.) John Boyd to describe the means to out-think, out-maneuver, and overwhelm an opponent's mental processing abilities and defeat them by getting "inside [their] decision cycle."[278] This occurs by using speed and unpredictability to create confusion in the enemy so severely that the enemy loses the mental ability to take in information and react in time to avoid losing. As Colonel Boyd's biographer described the effect, "the losing side rarely understands what happened."[279] A deepfake information and cyber campaign powered by algorithms designed precisely to hijack an opponent's OODA Loop could do just that with historic efficiency—and without legal ramification.

So long as these acts do not take advantage of or cause distrust in the protections under international law in order to achieve their objectives, and do not cause the enemy to unknowingly harm protected people or places, international law does not prohibit them. Such uses of deepfake technology more likely require political or military options, not legal recourse.

## VI. CHALLENGES OF DEEPFAKE TECHNOLOGY ON PRESENT AND FUTURE CONFLICTS

### *A. Democratization*

Although deepfake technology is still young, it has evolved quickly. The learning curve, which at first appeared too steep for most to

---

[277] *See* JP 3-13.4, *supra* note 170, at para. 11(c)(1).

[278] Colonel (Ret.) Boyd did not write a book or an article when creating the OODA Loop or its underlying concepts but instead featured them in a slide deck entitled "Patterns of Conflict" which he briefed to military leadership for decades. *See* John Boyd, Patterns of Conflict (Dec. 1986) (available at http://www.ausairpower.net/JRB/poc.pdf). The quote here, while often stated by Col. (Ret.) Boyd as a goal of the OODA Loop concept, actually comes from U.S. Army General Colin Powell as he described how coalition forces were able to secure a sweeping victory during Operation Desert Storm. ROBERT CORAM, BOYD: THE FIGHTER PILOT WHO CHANGED THE ART OF WAR 425 (2004).

[279] Coram, *supra* note 278, at 334.

handle, is barely visible today. Several programs now exist for creating deepfake content that anyone can buy. Applications such as Reface,[280] DeepFaceLab,[281] Descript,[282] and ZAO[283] which can produce high-quality deepfake content in under an hour, are widely accessible. Additionally, where previously a person would need some degree of training or programming experience to use these applications, YouTube now has several videos which seek to train people to create deepfakes using these applications, sometimes in under an often-claimed "10 minutes."[284]

The fruit of the feverish labor to democratize deepfake technology is, like the nature of the internet itself today, both entertaining and hazardous. Certainly, YouTube abounds with deepfake images composed for benign purposes such as depicting a *Star Wars* movie recast with a different actor or for satirical purposes.[285] However, the hazards of deepfake technology, which asserted themselves from the beginning, have evolved beyond the scatological.

Even before the Zelenskyy deepfake, actors had already used artificial intelligence to create synthetic content of world leaders for political and social purposes. For example, a January 2020 video by Alethea Group purports to show U.S. President Donald Trump and

---

[280] REFACE, https://play.google.com/store/apps/details?id=video.reface.app&hl=en_US&gl=US (last visited Feb. 28, 2021).

[281] *See* Ivan Perov, et al., *DeepFaceLab: A Simple, Flexible, and Extensible Face Swapping Framework* (May 12, 2020), https://arxiv.org/abs/2005.05535 (boasting that DeepFaceLab, an open-source deep-fake system, allows users to modify content "to achieve their customization purpose . . . with high fidelity and indeed indiscernible by mainstream forgery detection approaches . . .").

[282] DESCRIPT, https://www.descript.com/overdub (last visited Feb. 28, 2021).

[283] When ZAO became available on China's iOS App Store, it became China's most downloaded app overnight. *See, e.g.*, Zak Doffman, *Chinese Deepfake App ZAO Goes Viral, Privacy of Millions 'At Risk'*, FORBES (Sep. 2, 2019, 4:27 AM), https://www.forbes.com/sites/zakdoffman/2019/09/02/chinese-best-ever-deepfake-app-zao-sparks-huge-faceapp-like-privacy-storm/?sh=3a391bf84700. Considerable controversy ensued when ZAO's privacy policy turned out to allow the Chinese government to retrieve data input through ZAO. *Id.*; *see also* Laura He, Jack Guy, & Serenitie Wang, *New Chinese "Deepfake" Face App Backpedals After Privacy Backlash*, CNN (Sep. 3, 2019, 6:33 AM), https://www.cnn.com/2019/09/03/tech/zao-app-deepfake-scli-intl/index.html.

[284] *See, e.g.,* Tom Baranowicz, *How to Make DeepFake in 10 Mins – Tutorial* (Aug. 12, 2020), https://www.youtube.com/watch?v=eq55Qy4RPiA; Amrit Aryal, *Create Deepfakes with Just One Picture in Under 10 Minutes* (Oct. 31, 2020), https://www.youtube.com/watch?v=TY2DEP-C-O4.

[285] *See, e.g.,* Shamook, *Harrison Ford in Solo: A Star Wars Story [DeepFake]*, YOUTUBE (Aug. 16, 2020), https://www.youtube.com/watch?v=bC3uH4Xw4Xo.

British Prime Minister Boris Johnson, among others, admitting they were wrong about denying climate change.[286]

Some videos like these can be discredited almost instantaneously because they depict globally-known figures. The January 2020 climate change deepfake completely contradicted the politicians' long-held and well-known positions as well as their own personalities. As a result, the content had no likelihood of convincing anyone that the 'speakers' had suddenly changed their views just minutes after delivering remarks to the contrary. They were easily and naturally identifiable as fake. In fact, the high-profile nature of political life is often the best utility for combating deepfake content depicting high-profile politicians, as the resolution of the 2022 Zelenskyy video incident also proved.[287]

The challenge grows, however, when confronting content that depicts relatively low-profile people, such as tactical-level military commanders, or people who otherwise do not have a large public profile and so the content cannot as quickly be disproven. Furthermore, content that is purposefully incomplete such as voice-only deepfake can make detection difficult and aggravate confusion, especially if transmitted in high-intensity situations.

This is no academic concern. If anyone thinks this technology could not reasonably fool someone into thinking that they are interacting with someone they personally know, much less effect any significant outcome, they should think again. It's already happened.

In 2019, a criminal enterprise used deepfake technology to make a U.K.-based CEO believe he was talking to the Germany-based CEO of his parent company.[288] The AI managed to perfectly mimic the German CEO's voice. As an insurance investigator for the company reported to the Wall Street Journal, the AI replicated the German CEO's "slight German accent" and even the "melody" of the German CEO's cadence.[289] It only took one phone call. The criminals used the AI to make the British CEO believe an emergency was occurring and that the British CEO

---

[286] Alethea Group created and posted the videos shortly after President Trump made comments at the 2020 World Economic Forum in Davos, Switzerland where he denied that the environment was an economic concern. While the quality of the images and audio produced in the faked videos was raw, the timing and swiftness of the videos were still remarkable. CBS News, *President's Words Used to Create "Deepfakes" at Davos*, YOUTUBE (Jan. 24, 2020), https://www.youtube.com/watch?v=4A9LAxhi68I.

[287] *Supra* note 13.

[288] Catherine Stupp, *Fraudsters Used AI to Mimic CEO's Voice in Unusual Cybercrime Case*, WALL ST. J. (Aug. 30, 2019, 12:52 PM), https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

[289] *Id.*

needed to transfer $243,000.00 to a Hungarian supplier's bank account in one hour. The British CEO, skeptical but nonetheless convinced he was talking with his boss, transferred the money. The money, however, went to a bank account in Mexico where it disappeared. A few moments afterwards, so did the criminals—needing mere minutes to succeed.[290]

The victims involved were not traditionally vulnerable. They were not under-resourced, under-educated, or over-leveraged. To the contrary, the AI fooled a European businessman, someone of presumably significant acumen entrusted with co-leading a multinational corporation.[291] Only hubris could argue that a military commander could not be fooled as well and be convinced in part by an AI-derived manipulation to surrender forces or even unknowingly commit a war crime themselves. Because of the democratization of deepfake technology, near-perfect media manipulation capabilities—and the resulting complications they can cause—are within reach of any actor, state or non-state, with a motivation, an internet connection, and some free time.

### *B. Satellite Imagery Manipulation*

Another advent in deepfake proliferation that is growing quickly does not involve depicting people at all—but it is a serious threat to the multi-domain battlespace. GAN-powered manipulation has begun hitting satellite imagery.

The concept is both elegant and nefarious. An actor infiltrates an enemy's satellite link. The actor identifies the geographic area of an enemy's expected operations. The actor then uploads a deepfake-generating program that doesn't make major manipulations, such as wiping out mountains on a digital map, but makes subtle manipulations such as thinning a forest to make an area seem passable or depicting a small bridge over a stream where a bridge in reality does not exist. The satellite link transmits these manipulations throughout the enemy's formations who believe their convoy has a clear route to a waypoint on the other side of the stream. Only when the convoy reaches the stream and sees no bridge does the convoy realize the deception. Then the ambush begins.

---

[290] *Id.*

[291] While media has so far not published the business's name, as a possible sign of the business's robustness, the entire loss was swiftly covered by Euler Hermes Group, a multi-billion-dollar global insurance firm. *Id.*

This is the exact scenario which leaders in defense artificial intelligence development already acknowledge is here.[292] At a Genius Machines summit in 2019, Mr. Todd Myers, automation lead for the CIO-Technology Directorate at the U.S. National Geospatial-Intelligence Agency, publicly acknowledged this capacity and beyond even that, Mr. Myers conceded that "[t]he Chinese are well ahead of [the United States]."[293] Mr. Andrew Hallman, director of the C.I.A.'s Digital Directorate, speaking at the same summit, observed that "[w]e are in an existential battle for truth in the digital domain" and when asked if he felt that the C.I.A. was up to the task of defeating satellite imagery manipulation, responded "I think we are starting to. We are just starting to understand the magnitude of the problem."[294]

This vulnerability presents several problems beyond the one detailed above. The GANs which would manipulate geo-spatial imagery may also adversely influence the machine learning that other neural networks within the satellite are constantly conducting. If those neural networks lack effective classifiers to identify that a tree or a road is fake, they will exacerbate the manipulation by classifying the manipulation as authentic—thus causing allied neural networks to learn errantly and make the problem harder to detect. Also, defenses against infiltration and manipulation would be very expensive, requiring redundancies of all

---

[292] *See* Patrick Tucker, *The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth*, DEF. ONE (Mar. 31, 2019), https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/; *see also* Chunxue Xu & Bo Zhao, *Satellite Image Spoofing: Creating Remote Sensing Dataset with Generative Adversarial Networks*, Article No. 67, p. 1-6 (Jun. 10, 2018) (10th International Conference on Geographic Information Science paper), https://drops.dagstuhl.de/opus/volltexte/2018/9395/pdf/LIPIcs-GISCIENCE-2018-67.pdf (examining how satellite image manipulation works through the use of CycleGAN and Pix2Pix networks and advocating their use for urban planning purposes).

[293] *Id.* Russia may also be experienced in using AI to spoof navigation technology. In late June 2021 Russian naval and air forces stationed in the contested Crimean peninsula scrambled to confront UK and Dutch warships transiting the Black Sea on the basis of data it claimed showed the ships threatening Russian-claimed economic exclusion zones near the Crimean peninsula. The UK and Dutch ships and crews denied being close to the peninsula, arguing that they had been almost 200 nautical miles and 70 nautical miles away, respectively, from the peninsula – not the 12 nautical miles that Russia claimed. Fact-finding later appeared to show that Russia likely had spoofed the warships' radio transponders to give off an incorrect position to justify the subsequent show of Russian force. David Axe, *Harassing Ships and Spoofing Radios, Russia is Telling a Story – That Occupied Crimea is Russian*, FORBES (Jul. 1, 2021 at 8:00 AM), https://www.forbes.com/sites/davidaxe/2021/07/01/harassing-ships-and-spoofing-radios-russia-is-telling-a-story-that-occupied-crimea-is-russian/?sh=782ec8ba414b.

[294] *Id.*

imagery so that if one set is compromised, the compromise can be found by comparing the concerned set to a second set.[295] Furthermore, if a state's armed forces succeed in defending their satellite networks from deepfake manipulation, they would remain defenseless against infiltrations of privately-owned satellite mapping services unless they could enlist the help of the private entities who own the services.[296]

There is no question that this capability could impose mission- and possibly life-threatening costs. However, combating these costs is not just a matter of resource allocation or plan execution. These developments present moral, philosophical, and legal complications that stand to deeply challenge how and even whether actors observe certain facets of the laws of armed conflict.

### *C. The Liar's Dividend Weaponized, and the Competency Paradox*

A unique challenge that AI-driven manipulation creates is the so-called "Liar's Dividend"[297] where someone actually does something or says something but then denies doing so, by falsely claiming that the content depicting the speech or action was a deepfake. All of these complications can impact the battlefield.

Cases of the Liar's Dividend have already impacted Gabon, as well as Sino-Australian relations.[298] Combatants could blatantly attack a civilian population, abuse protected emblems for military gain, execute prisoners of war, or commit a number of other offenses against the laws of armed conflict but take advantage of the Liar's Dividend to argue that even the clearest evidence of these crimes are just deepfake concoctions.

To be sure, bad faith actors can and often do argue regardless of basis that legitimate evidence against them is fraudulent or made-up and have done so long before the invention of deepfake technology. What makes the Liar's Dividend particularly nefarious is that it would arise not in a manicured court of law, where it could be strangled, but in a court of public opinion, where it could thrive and then deflect the trial that might strangle it.

In a court of law, a painstaking digital forensics evidentiary audit and related expert witness testimony, various degrees of corroborating evidence, the unique intensity of focus that trials muster, and procedural

---

[295] *Id.* (quoting Mr. Myers).
[296] *Id.* (detailing concerns about Google Maps or Tesla being infiltrated).
[297] Citron, *supra* note 41 at, 1785-86.
[298] *Supra* notes 20, 44, respectively.

rules that guide evidence presentation, credibility, challenge, ultimate acceptance, and fact-finder consideration could deliver a stout haymaker to a Liar's Dividend-style defense. That haymaker, however, requires an enormous wind-up. This delivery would only come after months if not years pass while the trial comes together. The court of public opinion never provides that time. It arraigns, holds trial, considers evidence, and delivers a verdict all before breakfast.

Furthermore, the Liar's Dividend takes advantage of a competency paradox. The most credible circumstance for a Liar's Dividend defense would be where the falsely accused party is in fact adept at engaging in deception themselves. In other words, the better a belligerent is at using deepfake technology or deception in general, the stronger the Liar's Dividend defense. In turn, as a state becomes more vulnerable to the Liar's Dividend, the actual perpetrator's platform becomes more powerful. The perpetrator can use that platform to make a trial appear unjust or evidence appear untrue.

Russia appears to have recently attempted both of these approaches. For example, in the early phase of its invasion of Ukraine, when its forces attacked from every point of the compass except west and attempted to seize Kyiv, it occupied the town of Bucha a short distance outside of the Ukrainian capital.[299] Ultimately Russian forces failed to take Kyiv and withdrew to focus on an offensive in the east. Almost immediately after Russian soldiers left Bucha, dozens of videos and photographs emerged showing that hundreds of Ukrainian citizens had been executed, many of them bound and tortured before the killing shot.[300]

Instead of launching an investigation or seeking to bring the perpetrators to justice, the Russian government launched a campaign declaring that the videos and photographs were fake.[301] Employing the state-run Russian Telegram (RT) network, Russia aired a piece to its viewers entitled "War on Fakes" which claimed that the images were "staged" by Ukrainian and Western media outlets, attempted to point out inconsistencies in the videos, and portrayed timelines involving the Russian occupation of Bucha to argue that the content was fake.[302] They

---

[299] Cara Anna, *War Crimes Watch: A Devastating Walk Through Bucha's Horror,* ASSOCIATED PRESS (Apr. 10, 2022), https://apnews.com/article/russia-ukraine-europe-war-crimes-7791e247ce7087dddf64a2bbdcc5b888.
[300] *Id.*
[301] Yevgeny Kuklychev, *Fact Check: Russia Claims Massacre in Bucha 'Staged' by Ukraine*, NEWSWEEK (Apr. 4, 2022 at 11:41 AM), https://www.newsweek.com/fact-check-russia-claims-massacre-bucha-staged-ukraine-1694804.
[302] *Id.*

even employed some of the same techniques used in Western media to discredit the Zelenskyy deepfake, labeling images of bodies as "fake" and holding "antifake" panel discussions purporting to inform viewers that they should not believe what they see.[303]

Various Western and Ukrainian media outlets have worked to debunk Russia's campaign, pointing to witness testimonies, drone footages, satellite images, and other means.[304] And while Ukrainian prosecutors have already begun war crimes trials to seek justice for the killings,[305] the victims and their families may have to agonizingly witness justice delayed and possibly denied for the very real crimes the perpetrators commited,[306] especially as prosecutors may need to exert significantly more time and resources to lay the evidentiary foundation for video or photographic evidence than would have been required in another age.

## VII. RECOMMENDATIONS FOR IMPROVED GOVERNANCE OF DEEPFAKE

While the invasion of Ukraine has ushered deepfake technology into the records of war, as of the writing of this article, purely by the numbers, the vast majority of problems with deepfake media manipulation remains relegated to the domestic realm. However, like with other inventions such as barbed wire or the airplane that were not born for war but were nonetheless enlisted, there is no reason to believe that AI-derived media manipulation will not be further weaponized. It is important, therefore, to figure out now how to better handle its impact.

First, international agreements seeking to govern artificial intelligence or cyberspace operations in armed conflict must expressly

---

[303] Robert Mackey, *Russian TV is Filled with Images of Bucha's Dead, Stamped with the Word "Fake"*, THE INTERCEPT (Apr. 12, 2022 at 7:51 AM), https://theintercept.com/2022/04/12/bucha-massacre-russia-tv-fake-ukraine-war/.

[304] *Id.*; *see also* Aude Dejaifve, *Fresh Round of Fake Videos Claim the Bucha Massacre was Staged*, FRANCE24 (Jun. 4, 2022 at 6:40 PM), https://observers.france24.com/en/europe/20220408-fresh-round-of-fake-videos-claim-the-bucha-massacre-was-staged; Malachy Browne, *Satellite Images Show Bodies Lay in Bucha for Weeks, Despite Russian Claims*, THE NEW YORK TIMES (Apr. 4, 2022), https://www.nytimes.com/2022/04/04/world/europe/bucha-ukraine-bodies.html.

[305] Victor Jack, *Ukraine Files First War Crimes Charges Against Russia Over Bucha Killings*, Politico (Apr. 28, 2022 at 6:17 PM), https://www.politico.eu/article/ukraine-first-war-crimes-charges-against-russia-over-bucha-killings/.

[306] *See e.g.* Erika Kinetz, *War Crimes Watch: Hard Path to Justice in Bucha, Ukraine, Atrocities*, Frontline (Apr. 4, 2022), https://www.pbs.org/wgbh/frontline/article/bucha-ukraine-civilian-deaths-justice-tribunal-international-criminal-court/ (detailing the myriad difficulties in prosecuting Russian soldiers for the alleged killings).

address the use of artificial intelligence in deception or misinformation activities. These agreements, in whatever form they may take, should acknowledge the reality that AI can create synthetic content that seems to change reality. They should expressly govern such deployment of AI under a regime that criminalizes its use to engage in perfidy, whatever the style.

Second, such instruments should articulate perfidy definitions that not only align with Article 37 of Additional Protocol I but also build upon it. Article 37 has long been derided for being too narrow with its "kill, injure, or capture" limitation.[307] This list should remove consequence requirements all together, and replace them instead with a general intent *mens rea* of intending to secure a military advantage. If the wrongfulness of perfidy is that the abuse of protections afforded under international law will cause a destruction of trust necessary to secure peace, it should not matter whether that sin serves the purpose of killing or the purpose of confusing.

Third, and in assistance with the first two recommendations, U.S. Department of Defense doctrine on perfidy should align with the representations the U.S. government otherwise has made as expressed in the Matheson Memorandum.[308] If the Department of Defense believes it necessary not to acknowledge "capture," because the Department believes customary international law allows a combatant to fake a protected status in order to avoid capture, Article 37 does not conflict with this view. Article 37 only prohibits claiming a protected status in order to commit a capture. Updating this posture will be a net positive for the U.S. as it will foster intra-governmental unity of vision, intra-governmental unity of expectation, communicate to the rest of the world that the U.S. is of the same mind about Article 37, and better assure that its forces do not become subject to behavior that it would likely want to object to if such behavior occurred to U.S. forces.

Fourth, U.S. Department of Defense information operations and artificial intelligence doctrines should expressly address deepfake capabilities, threats, and counters, with corresponding training inserted into Information Operations and LOAC training to signal and military intelligence occupations and to senior leaders regardless of branch or occupation specialty on how to recognize and react to a deepfake ruse. Furthermore, deepfake technology implications should also be trained in

---

[307] *Cf.* U.S. Department of Defense refusal to recognize "capture" as part of customary international law discussed *supra* note 205.
[308] Matheson, *supra* note 173.

concert with the recently released DoD AI Ethics Principles.[309] Training in either context would not need to be overly-detailed – simply enough to apprise commanders and impacted subject matter experts of the issues and what they should and should not do in response.

Fifth and finally, given that deepfake manipulation is unlikely to lose its attractiveness anytime soon, and until counter-deepfake methods reach the same level of productivity as their opponents, the international community, spearheaded by the United States, should embark on a concerted public awareness and education campaign about deepfake technology problems. The best way to combat such sophisticated deception before it can do serious harm may be to just make sure everyone knows it exists and what it can really do. This approached proved itself when media outlets and the Ukrainian government identified and discredited the Zelenskyy deepfake almost as quickly as it was broadcast, with no reported surrenders or slackening in the Ukrainian war effort.[310] The Ukrainian government had even launched a deepfake public awareness campaign two weeks before the Zelenskyy deepfake broadcast, further aiding in the later content's quick debunking.[311] Without an awareness campaign, the resulting skepticism, while not without its own negative social impacts, may present a targeted entity with enough time to uncover the deception before anyone acts in a way that could achieve the deception's objectives.

## CONCLUSION

Deepfake technology only promises to gain more traction in the affairs of armed conflict. Experts in artificial intelligence and armed conflict suggest that the technology has already in the short space of a couple years advanced from a first generation to a second-generation capability and that combating it now will require a "whole of society approach."[312]

However, despite its penchant for victimization and its clear potential to cause irreparable harm to notions of trust from the ballot box to the bunker, its growing uses in popular media have already endeared deepfake technology to an entire generation of consumers. These

---

[309] DEF. INNOVATION BD., AI PRINCIPLES: RECOMMENDATIONS ON THE ETHICAL USE OF ARTIFICIAL INTELLIGENCE BY THE DEPARTMENT OF DEFENSE (Oct. 2019).
[310] *Supra* note 13.
[311] Simonite, *supra* note 13.
[312] *See, e.g.,* Brigadier General R. Patrick Huston & Lieutenant Colonel M. Eric Bahm, *Deepfakes 2.0: The New Era of "Truth Decay,"* JUST SEC. (Apr. 14, 2020), https://www.justsecurity.org/69677/deepfakes-2-0-the-new-era-of-truth-decay/.

consumers may understandably cheer the sight of a circa-1980s Luke Skywalker appearing in 2021 *Star Wars* content, gush at the thought of swapping in themselves as the lead in their favorite movie, or adore the technology's capacity to engineer biting political satire. However, the legal community must remain vigilant to help the greater global community continue to always bear in mind that while deepfake technology may have harmless entertainment value in some contexts or even net positive effects in others,[313] as examples from Gabon and Ukraine show, it still bears a capacity for great harm and significant legal instability.

[313] *See, e.g.*, Jessica Silbey & Woodrow Hartzog, *The Upside of Deepfakes*, 78 MD. L. REV. 960, 962-64 (2019) (observing positive effects of deepfake technology such as creating new teaching utilities in education or strengthening journalistic integrity standards).