COMMENTARY

POWs in the Age of the Internet

Jan Kallberg, Todd Arnold, Stephen Hamilton, and Mark Visger

The emergence of deepfakes has challenged long-standing protocols concerning prisoners of war in the Geneva Conventions. The United States, its Allies, and partners must educate servicemembers about the potential exploitation of any recorded media obtained if they become prisoners of war.

A future great power conflict could potentially involve large numbers of prisoners of war (POWs)—US, Allied, and partner nations—imprisoned by regimes that could seek to utilize and exploit these captives for propaganda gain. Deepfakes and digital manipulation technology provide an advantageous opportunity for a captor aiming to mitigate international humanitarian law concerns regarding the rules for POW treatment. Such an adversary could use manipulated audio and images of POWs to forward their cause, undermine the Alliance cohesion, attack the mutual will to fight, and reduce POWs' will to resist.

The Future Conflict

The risk of becoming a POW has steadily disappeared from the minds of US military members after two decades of counterinsurgency and antiterrorism operations. The memories of the Cold War and the Soviet occupation of Eastern Europe—and the general understanding of what captivity means—are diminishing. This prospect, however, should not be forgotten as the potential for the capture of sizeable numbers of POWs in a large-scale conflict is a distinct possibility.

The general strategic direction has recently changed from counterterrorism and antiterrorism operations toward great power competition and potentially protracted conflicts involving near-peer nation-states. Since the Cold War, air mobility, standoff weaponry, capabilities for deep strikes into enemy territory, and faster decision cycles have created a new battlefield. The modern battlefield is ever changing. High-paced engagements and mobility cross multiple war-fighting domains create the potential for a fragmented, fluid fight. In this unpredictable, widespread, rapidly changing, and violent environment, the potential for large numbers of POWs is high.

Different POW Experiences

Each North Atlantic Treaty Organization (NATO) nation has a unique experience regarding prisoners of war. In most cases, events during World War II are the foundation for the general public's perception. For example, the understanding of what it means to be a POW for the United States, the United Kingdom, and France is based on the Korean and Vietnam Wars. For the British, it is the Falklands.

For today's NATO, the Balkan wars of the 1990s are the most recent conflicts with numerous POWs. Some countries obeyed the international laws, treating POWs fairly and complying with the Geneva Conventions. More common, however, are examples where regimes sought to exploit POWs as propaganda tools to forward the regime's cause and undermine home-front support by seeking to have prisoners appeal to end the war effort. This history of abuse raises the concern that utilizing deepfakes to manipulate POWs' voices and images could be a highly effective component of psychological operations.

Accelerating Information Flow

The current internet-driven information flow differs radically from the past; there was no widely embraced internet and social media until well into the global war on terrorism, decades after the Balkan wars. The internet has radically changed how information, true or not, can be delivered at unprecedented speed and scale.

Once only seen as a vehicle for positive societal impact and democracy, the world now understands the internet can also be used for less altruistic purposes. Dual-use technology is not novel or unique to the current information age; mass printing and rail commerce created social mobility and consolidated democracy in many countries.[1] Conversely, these developments were the foundation for the early Soviet totalitarian propaganda machine in the 1920s.[2] Information technology develops at an accelerating speed, and we can be informed as events happen in real time. Still, the same wide information dissemination can mislead, disinform, undermine democracy, and negatively influence entire populations in unanticipated ways.

Deepfakes

Fake imagery, audio, and videos are not novel or new concepts. Many countries' disinformation and propaganda campaigns used doctored imagery or audio in the past century. For example, when Leon Trotsky fell out of favor after Soviet Premier Vladimir Lenin's death, the government erased Trotsky's presence in all public images of early Soviet leadership.[3]

Deepfakes—detailed advanced manipulation of high-resolution digital multimedia files—can make the subject of the deepfake appear to be in a situation or location they

1. Roger Woods, "Mid-Nineteenth Century Migration from Norfolk to London: Migratory Patterns, Migrants' Social Mobility and the Impact of the Railway?" (master's diss., School of Advanced Study, University of London, 2014).

2. Adelheid Heftberger, "Propaganda in Motion: Dziga Vertovs and Aleksandr Medvedkins Film Trains and Agit Steamers of the 1920s and 1930s," *Apparatus* 1 (2015), https://www.apparatusjournal.net/.

3. J. D. Swerzenski, "Fact, Fiction or Photoshop: Building Awareness of Visual Manipulation through Image Editing Software," *Journal of Visual Literacy* 40, no. 2 (2021); Dimitris Poulopoulos, "How to Produce a DeepFake Video in 5 Minutes," Towards Data Science, April 2, 2020, https://towardsdatascience.com/; and Lisa Bode, Dominic Lees, and Dan Golding, "The Digital Face and Deepfakes on Screen," *Convergence* 27, no. 4 (2021).

never were. A highly believable deepfake can be created from as little as 15 minutes of video and audio footage.[4] In the case of POWs, the captors would have more than enough time to generate adequate audio and video of their prisoners to create robust deepfake audio and video files.

For example, the captor's interrogation team could repeatedly interrogate prisoners, a well-established tactic designed to identify an inconsistency or slip that discloses protected information. The interrogators could capture audio, video, and image files without the prisoners' knowledge or consent while routinely interrogating them about insignificant details and mundane daily events.

These repeated interviews, which for the POWs would seem a futile exercise in trying to make them cooperate, would instead generate a substantial repository of digitized material for the captor to fabricate deepfakes. Moreover, retrievals of social media, public information, and current events through open-source intelligence would provide bountiful material to inject in these deepfakes, reinforcing the perception the deepfake is genuine.

Despite POWs resisting collaboration and acting according to their defense force's code of conduct, technology could manufacture imagery and audio that makes them "artificial" collaborators.

Supporting the Captor's Regime

The adversary can use POW deepfakes for propaganda purposes in three primary ways: support the regime, undermine the will to defend and fight, and distort and influence a captive's sense of reality.

Support

If an authoritarian and totalitarian regime lacks legitimacy, it could reinforce its rationale to the population by presenting the opponent as illegitimate, promoting the message life is better living under the regime. Almost 70 years of Soviet domestic agitation and propaganda followed this construct. In the Soviet narrative, rulers of Western countries were bourgeoise that ripped the productivity and gains from the working class, leaving the Western working class abused and suppressed. Soviet leaders used their propaganda machine to convince their populace that even if the Soviet Union's conditions were less than optimal, the situation in the West was worse.[5]

Authoritarian regimes rely on propaganda units to maintain social order and suppress resistance to the regime. The Chinese government's current focus on national unity and societal stability is promoted by the political propaganda arm, which historically has presented narratives from captives as a part of its domestic propaganda.

---

4. Michail Christos Doukas et al., "Head2head++: Deep Facial Attributes Re-Targeting," *EEE Transactions on Biometrics, Behavior, and Identity Science* 3, no. 1 (2021).

5. Robert G. Kaiser, "The Soviet Pretense," *Foreign Affairs* 65, no. 2 (1986).

Undermine

The next propaganda method is the immediate release of deepfakes to the global public through social media, file sharing, social networks, and other internet platforms. Even though such a release of deepfake POW imagery is likely in violation of the Third Geneva Convention, such a deepfake could depict and push narratives of war crimes, atrocities, rejection of the Alliance war effort, pleadings to end the war, and other propaganda. Captors could distribute the videos and audio back to the POW's home nation on a broad scale to undermine the war effort and stress soldiers' families, influence politicians, fuel ethnic cleavage, and seek to break up Ally and partner cohesion to weaken support for the war.

A recent concern is deepfakes of senior military leaders and politicians in which the manipulated media presents statements these individuals never said. Deepfakes of public leaders do not serve the adversary well in the long-term because these deepfakes can be fact-checked, and the officials are accessible for media and news outlets. But such deepfakes can provide short-term benefits even after being identified as manipulated and fake. These deepfakes can also cast temporary doubt on legitimate statements by public figures. With POWs, however, there would likely be no way to communicate and verify message validity. As a result, the POW deepfakes would likely generate a more prolonged, influential, and plausible intended effect for the adversary's interest.

Distort and Influence

In the third propaganda method, the captor could show deepfakes to POWs to manipulate them. Most great power conflicts in the last century started with the actors believing they were engaging in a short-term conflict, only to realize later they were in a protracted and enduring conflict.[6] During World War II, Korea, and Vietnam, POWs were captives for several years—the sign of an enduring conflict.

In a similar situation, the captor could control the flow of information to POWs and could use deepfakes to feed false and misleading information over time. Even if a prisoner of war could ascertain each deepfake was a false portrayal of information, it is highly likely that, over time, isolation and pressure from surrounding conditions could induce a POW to accept deepfakes as legitimate. The captor could then utilize deepfakes to indoctrinate, psychologically destabilize, and manipulate the captive's mental state.

Third Geneva Convention

The protections afforded POWs have remained constant since the adoption of the Third Geneva Convention in 1949; provisions relating to the treatment of POWs were not updated in the Additional Protocols to the Geneva Conventions adopted in the

6. Stuart Hallifax, " 'Over by Christmas': British Popular Opinion and the Short War in 1914," *First World War Studies* 1, no. 2 (2010); and Karl-Heinz Frieser, *The Blitzkrieg Legend: The 1940 Campaign in the West* (Annapolis, MD: Naval Institute Press, 2013).

1970s. This fact is unfortunate, as unresolved interpretative issues remain. That said, the current treaty provisions provide some helpful guidance.

Specifically, Article 13 requires POWs be protected "against insults and public curiosity."[7] This provision builds on the baseline Article 13 requirement of "humane treatment" and Article 14 requirement of "respect for their persons and their honor."[8] Particularly egregious incidents from World War II such as the parading of POWs or publication of POW photographs in the press that resulted in a war crimes conviction and prison term, provided the impetus for these provisions.

More recently, captors have released photographs of POWs for other, arguably legitimate purposes, such as proof of life, proof of capture, or to document inappropriate treatment. While not definitively resolved, one could argue they released photos in such situations for purposes other than humiliation or public curiosity.

Despite this fact, the publication of a deepfake to public audiences either in the captor or captive's country would appear to violate the prohibition of public curiosities. The US *Department of Defense Law of War Manual* concludes "displaying POW's in a humiliating fashion on television or on the internet [is] prohibited."[9] All NATO allies would likely concur with this conclusion. It is difficult to imagine a public release of a deepfake that would not implicate the prohibition on public curiosity.

The use of deepfakes in interrogation, on the other hand, is not as clearly resolved by the Geneva Conventions. While POWs are only required to provide the standard name, rank, date of birth, and serial number, the detaining power can certainly engage in questioning to procure additional information.

Article 17 prohibits "physical or mental torture" as well as "any other form of coercion" to obtain such information in such questioning.[10] Here it is difficult to make broad or sweeping conclusions as to whether the use of deepfakes in an interrogation would constitute coercion for two reasons: (1) each instance will be dependent on the specific facts in question, and (2) states have different understandings of what constitutes coercive conduct in their domestic law.

In the United States, for example, police are not prohibited from lying to suspects during police interrogations. But such conduct may be considered when determining whether a suspect provided any resulting statement voluntarily and in accordance with the Fifth Amendment privilege against self-incrimination. One could argue the repeated use of deepfakes in an interrogation would constitute coercion, but there are no clear answers at this point.

---

7. United Nations, Geneva Convention Relative to the Treatment of Prisoners of War of 12 August 1949 (Geneva: United Nations, 1949), 97, https://www.un.org/.

8. United Nations, *Prisoners of War*, 97.

9. Department of Defense (DOD) Office of General Counsel, *Department of Defense Law of War Manual* (Washington, DC: DOD, June 2015, updated December 2016), https://dod.defense.gov/.

10. United Nations, *Prisoners of War*, 98.

Conclusion

Servicemembers must be provided education on deepfakes to prepare them for this possibility if captured. Each member nation's legal departments should develop legal positions on the use of deepfakes of POWs and be prepared to communicate those positions to work toward a common understanding under international law. Finally, additional research is needed on detecting and countering deepfakes.

The United States and its Allies and partners must transition from shared POW experiences in World War II, Korea, Vietnam, and the Balkans. Today, the tools the captors can use for propaganda and manipulation are far more advanced and have unprecedented reach. Adversaries are ready to use that reach to sow discord among the Alliance and partner nations, undermine our mutual will to fight and, in conflict, drive opinion toward a conflict resolution that favors them. ✈

Jan Kallberg, PhD
Dr. Kallberg is a research scientist at the Army Cyber Institute.

Lieutenant Colonel Todd Arnold, USA
Lieutenant Colonel/Dr. Arnold is an Army cyber officer, research scientist at the Army Cyber Institute and an assistant professor in the Department of Electrical Engineering and Computer Science at the US Military Academy, West Point.

Colonel Stephen Hamilton, USA
Colonel/Dr. Hamilton is an Academy professor, technical director at the Army Cyber Institute, and an associate professor in the Department of Electrical Engineering and Computer Science at the US Military Academy, West Point.

Lieutenant Colonel Mark Visger, USA
Lieutenant Colonel Visger, JD, is a research scientist at the Army Cyber Institute and an Academy professor at the US Military Academy, West Point.