NR. 573 E
30.11.2020

LISA M. COHEN
PhD Student and Reserch Associate
IFHV
Ruhr-Universität Bochum

QUESTIONS:
lisa.cohen@ruhr-uni-bochum.de

Völkerrechtsblog
INTERNATIONAL LAW & INTERNATIONAL LEGAL THOUGHT
IFHV

BOFAXE

THE NEW ERA OF DISINFORMATION WARS (Part 1)

DOES INTERNATIONAL HUMANITARIAN LAW SUFFICIENTLY REGULATE THE USE OF DEEPFAKES?

While the manipulation of photographs has traditionally been deemed a State intelligence privilege, today's technological evolution allows anyone to effortlessly modify digital material – deepfakes being the newest, and arguably most dangerous, trend of such practices. Deepfake algorithms use so-called 'deep learning' artificial intelligence (AI) to create new audio and video by replacing or merging one's voice and/or face with manipulated and artificial data, which automatically fits the output dimensions and conditions. A short recording of one's voice suffices for an AI to create a "voice skin" that can be processed to say virtually anything. Although deepfakes are currently mainly used for humoristic purposes (see examples here), their use for malicious and military purposes seems inevitable. Indeed, deepfakes offer the potential to deceive and misinform adversaries and gain significant military advantages, while debunking and attributing the misinformation remains highly difficult. Against this background, and with social media spreading information to a massive community within seconds, feigning an alternative reality may set off an uncontrollable chain of events with detrimental consequences for the civilian population in conflict-ridden areas.

Aiming to determine whether international humanitarian law (IHL) sufficiently regulates the use of deepfakes, this Bofax examines if and how existing IHL norms apply to deepfakes, particularly against the backdrop of the 2017 Tallinn Manual 2.0 on the International Law Applicable to Cyber Operations (Tallinn Manual 2.0).

Article 36 Additional Protocol I to the Geneva Conventions (API) facilitates the application of IHL to contemporary developments by demanding the compliance of 'new' means and methods of warfare with established IHL principles. However, as is the case with all cyber operations, the applicability of IHL rules to deepfakes proves to be no clear-cut a matter. The most comprehensive, yet non-binding, international guideline on cyber warfare is the Tallinn Manual 2.0, which unfortunately only sparsely touches on the implication of different forms of disinformation in armed conflict and makes no mentioning of deepfakes. According to Rule 80 Tallinn Manual 2.0, the *existence* of an armed conflict is a prerequisite for the applicability of IHL to cyber operations. Thus, deepfakes which are employed during an ongoing armed conflict, are governed by the same IHL rules as the 'traditional' means and methods of warfare employed in that conflict – notwithstanding that these rules might not be sufficient or appropriate in the context of information warfare.

Perfidy and ruses of war

The deception of an adversary in armed conflicts by dissemination of false information is a contemporary method of warfare. In principle, deepfakes are nothing but a more sophisticated, hyper-realistic continuation of this practice, as the following examples show:

Example 1: State A produces a deepfake of a representative of the International Committee of the Red Cross inviting both adversaries to e. g. peace talks. State A sends the deepfake to the military commander of State B with the intent to attack State B's representatives on their way to the faked meeting.

Example 2: State A produces a deepfake in which State B's military commander orders the armed forces of State B to retreat from strategically important cities. State A then spreads the deepfake video to the armed forces of State B via social media, with the intent of gaining a military advantage.

IHL offers black letter law to determine which acts of deception are permitted. According to Article 37(1) API, perfidious acts – those which invite particular confidence in the adversary and intend to betray that confidence – are prohibited. Example 1 clearly falls within the scope of Article 37(1) API and is therefore prohibited.

In contrast, under Article 37(2) API "acts which are intended to mislead an adversary or to induce him to act recklessly but which infringe no rule of international law applicable in armed conflict and which are not perfidious because they do not invite the confidence of an adversary" constitute permitted ruses of war. Article 37(2) API names "misinformation" as an example of a permissible ruse. Rule 123 Tallinn Manual 2.0 cites *inter alia* the spreading of disinformation causing an adversary to erroneously believe a false appearance of what is actually happening, and "bogus orders purporting to have been issued by the enemy commander" as prime examples of permissible ruses. Accordingly, example 2 will most likely be considered as permissible ruse, provided that no other IHL is violated.

VERANTWORTUNG Die BOFAXE werden vom Institut für Friedenssicherungsrecht und Humanitäres Völkerrecht der Ruhr-Universität Bochum herausgegeben: IFHV, Massenbergstrasse 9b, 44787 Bochum, Tel.: +49 (0)234/32-27366, Fax: +49 (0)234/32-14208, Web: http://www.ruhr-uni-bochum.de/ifhv/. Bei Interesse am Bezug der BOFAXE wenden Sie sich bitte an: ifhv-publications@rub.de. **FÜR DEN INHALT IST DER JEWEILIGE VERFASSER ALLEIN VERANTWORTLICH.**