NR. 573 E
30.11.2020

LISA M. COHEN
PhD Student and Reserach 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 2)

DOES INTERNATIONAL HUMANITARIAN LAW SUFFICIENTLY REGULATE THE USE OF DEEPFAKES?

Disinformation, civilians, and the notion of "attack"

In the assessment of the legality of deepfakes, their impact on the civilian population is pivotal.

The duty of constant care (Article 57(1) API), the principles of distinction (Article 48 API) and proportionality (Article 51(5)(b) API), and the prohibition of acts whose primary aim is to spread terror among the civilian population (Article 51(2) API, Article 13(2) Additional Protocol II to the Geneva Conventions), can possibly limit the lawful usage of deepfakes in armed conflicts.

Example 3: State A produces a deepfake depicting a conversation between State B's president and military commander about an imminent nuclear attack on the capital city of State C. State A disseminates the deepfake through social media platforms to primarily cause panic across the civilian populations of States B and C.

While the duty of constant care and the principle of distinction apply in all military operations, the principle of proportionality and the prohibition of acts whose primary aim is to spread terror, only govern "attacks", acts of violence, or threats thereof. This is where specific difficulties regarding the applicability to and classification of deepfakes arise. There is virtually no case law to define what constitutes an 'attack' in the context of cyber conflicts. "Attack", as per Article 49(1) API, "means acts of violence against the adversary […]." Rule 92 Tallinn Manual 2.0 states that violence "must be considered in the sense of violent consequences and is not limited to violent acts." However, according to Rule 98, a Twitter message "sent out in order to cause panic, falsely indicating that a highly contagious and deadly disease is spreading rapidly throughout the population […] is neither an attack […] nor a threat thereof […]" and consequently does not violate Article 51(2) API. But can fake news in no circumstance be an attack, act of violence, or threat thereof?

While the exemplary tweet disseminates 'news' without claiming state involvement, the deepfake illustrated in example 3 is of an entirely different quality. As a deepfake can only be recognized as such with great difficulty and therefore will – at least initially – be perceived as authentic, the effects on State C and its civilian population will certainly not be any less grave than those of a real announcement of a nuclear attack. However, Rule 92 Tallinn Manual 2.0 deems operations causing "inconvenience or irritation" without foreseeably resulting in injury of individuals or damage of physical objects lawful. Thus, under the current framework, due to a lack of foreseeable injury, example 3 would be considered a lawful non-attack, although the resulting panic and terror could be tremendous.

Similarly, the distribution of the deepfake under example 2 via social media raises questions regarding its conformity with the principle of distinction and the prohibition of indiscriminate attacks (Article 51(4) API). The interconnectedness of cyberspace makes it practically impossible to strictly distinguish between civil and military uses of (social) media and reasonably foreseeable that any deepfake can (accidentally) fall into the hands of civilians. Regardless of whether the objectives of the deepfake are civilian or military, the notion of 'attack' is decisive. Rule 93 Tallinn Manual 2.0 (Distinction) stipulates that operations directed at civilians are only prohibited when they amount to an 'attack'; operations directed at military objectives must comply with the principle of proportionality – which only applies to attacks. According to Rule 105 Tallinn Manual 2.0, cyber weapons creating a chain of events beyond the control of the attacker are indiscriminate by nature. While social media spreads information uncontrollably, as long as the deepfake does not *foreseeably* cause injury or damage – which the deepfake of example 2 clearly does not – it is not indiscriminate. Accordingly, also example 2 would be lawful.

Conclusion

Even though IHL can somewhat grasp the concept of disinformation in warfare due to the longstanding practice, the existing legal framework is not equipped to appropriately react to the dimensions deepfakes add to the equation. Due to technological advances, it has become necessary for the law to differentiate between the available forms and contents of 'fake news'. The most pressing issues, namely the subsumption of deepfakes under the notions of 'attack', 'act of violence', or threat thereof, and the requirements of foreseeability and degree of possible harm, are in dire need for clarification, as both are determinate for the applicability of pertinent IHL norms and the protection of civilians. First steps in countering deepfakes could be the installment of safeguard-mechanisms, e.g. digital watermarks, and including the concept of deepfakes in future international manuals to facilitate the discourse between states and ascertain current *opinio iuris*.

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.**