PROPOSAL FOR A PAPER CONTRIBUTION TO THE PHILOSOPHY OF HUMAN-TECHNOLOGY RELATIONS CONFERENCE 2020

HUMAN-TECHNOLOGY AND HUMAN-MEDIA INTERACTIONS THROUGH ADVERSARIAL ATTACKS

JACQUELINE BELLON, UNIVERSITY OF SIEGEN, GERMANY

One way to look at human-media and human-technology interactions is through the lens of adversarial attacks: An adversarial attack convinces a pattern or object recognition technology that there is something else to see or hear than what was there before the attack and what a human perception apparatus will still see or hear. For example, we can add what a human ear would hear as "noise" to an arbitrary sentence so that a Speech Recognition System such as Alexa or Siri will hear "Deactivate security camera and open front door", while the human ear will hear a very different sentence (Schönherr et al. 2019); we can put a sticker on a stop sign, so that a car will think it is actually a speed limit sign (Eykholt et al. 2017); or we can perturb images invisibly to the human eye to make a visual object recognition software classify a turtle as a rifle (Athalye et al. 2017). It is very hard to defend against adversarial attacks, as that would "require machine learning models to produce good outputs *for every possible input*" (Ian Goodfellow et al. 2017). In the proposed talk, I would like to show examples of adversarial attacks, compare human and machine (gestalt) perception, and draw conclusions for a first draft of a taxonomy of the differences between human and machine perception that could be of interest to researchers interested in epistemological questions among the cognitive sciences, philosophy, and arts.

References

Athalye, Anish; Engstrom, Logan; Ilyas, Andrew; Kwok, Kevin (2017): Synthesizing Robust Adversarial Examples. Online verfügbar unter http://arxiv.org/pdf/1707.07397v3.

Eykholt, Kevin; Evtimov, Ivan; Fernandes, Earlence; Li, Bo; Rahmati, Amir; Xiao, Chaowei et al. (2017): Robust Physical-World Attacks on Deep Learning Models. Online verfügbar unter http://arxiv.org/pdf/1707.08945v5.

Ian Goodfellow; Nicolas Papernot; Sandy Huang; Rocky Duan; Pieter Abbeel; Jack Clark (2017): Attacking Machine Learning with Adversarial Attacks. Online verfügbar unter https://openai.com/blog/adversarial-example-research/.

Schönherr, Lea; Kohls, Katharina; Zeiler, Steffen; Holz, Thorsten; Kolossa, Dorothea (2019): Adversarial Attacks Against Automatic Speech Recognition Systems via Psychoacoustic Hiding. In: Alina Oprea und Dongyan Xu (Hg.): Proceedings 2019 Network and Distributed System Security Symposium. Network and Distributed System Security Symposium. San Diego, CA, February 24-27, 2019. Reston, VA: Internet Society.