AI Applications in DeepFakes: Toward A Human-Machine Trust Framework

Topic: Media Industry and Society

Abstract

This study is the phase one of a project that aims to develop a machine learning system that combats visual misinformation such as fake/altered images or videos in a news context through trust-building machine-human interactions. Specifically, the current work will investigate the factors that influence user trust in AI technology that helps detect Deepfakes. The grounding notion is that effective human-AI collaboration is likely to result in positive social, economic, and consumer outcomes, but we need to situate AI technology in human's belief, intentions, preferences, and expectation contexts (Chakraborti & Kambhampati, 2018).

Advances in visual image manipulation have enabled the recent growth of fake imagery. In fact, the use of digitally altered images is one of the most effective means of evoking emotions and inducing viral dissemination of fake news. There have been numerous incidents of violence incited by widespread manipulated political images in different parts of the world. AI technology (i.e., machine learning models) can be used to detect fake visuals at scale. While the approach is gaining steam and adopted by industry leaders such as Adobe and WhatsApp, as well as the U.S. Defense Department's DARPA Media Forensics program, it is unclear if such AI-based solutions offer the most effective means of combating fake news from the perspective of news consumers.

To transform machine learning systems from tools to partners, users must trust their machine counterpart. Trust is multifaceted, including dimensions like competence, integrity, credibility, and benevolence. The premise of this study is that technology alone is insufficient as the most advanced solutions offer no utility when they are not trusted. Based on extant literature in trust, human-machine interaction, CASA (computers are social actors), and anthropomorphism (Cui, 2019; Lankton & Shah, 2015; Nothdurft, 2014; Calhoun, 2019; Mohseni, 2020), it is proposed that trust in AI deepfakes applications is dependent on various human capacity and characteristics such as knowledge, self-efficacy, prior experience, demographics, attitudes toward AI, trust propensity, and media habits. The factor of risk and different machine characteristics as perceived by the consumers like AI ability, humanness, and intentionality would also affect trust building in this context (see Figure 1).

The research method of online survey will be adopted for the study. A pretest of 100 subjects on Mturk will be used to evaluation the validity of the variable measures, followed by the main test of 1,000 U.S. adults using Qualtrics national panels. Regression analyses will be performed to assess the relationship between the proposed independent and dependent variables. The results of this phase will inform the development of an explainable machine learning system for deepfakes diagnostics that is human centric in its design for user adoption and communication.

Keywords: artificial intelligence, deepfakes, human-machine interaction, visual misinformation

Figure 1: Proposed Analytical Framework