DARPA Tackles Deepfakes With AI

DARPA's Information Innovation Office experiments with MediFor program platform to combat visual media manipulation.

Connor Collins

Mon, 03/11/2019 - 16:54

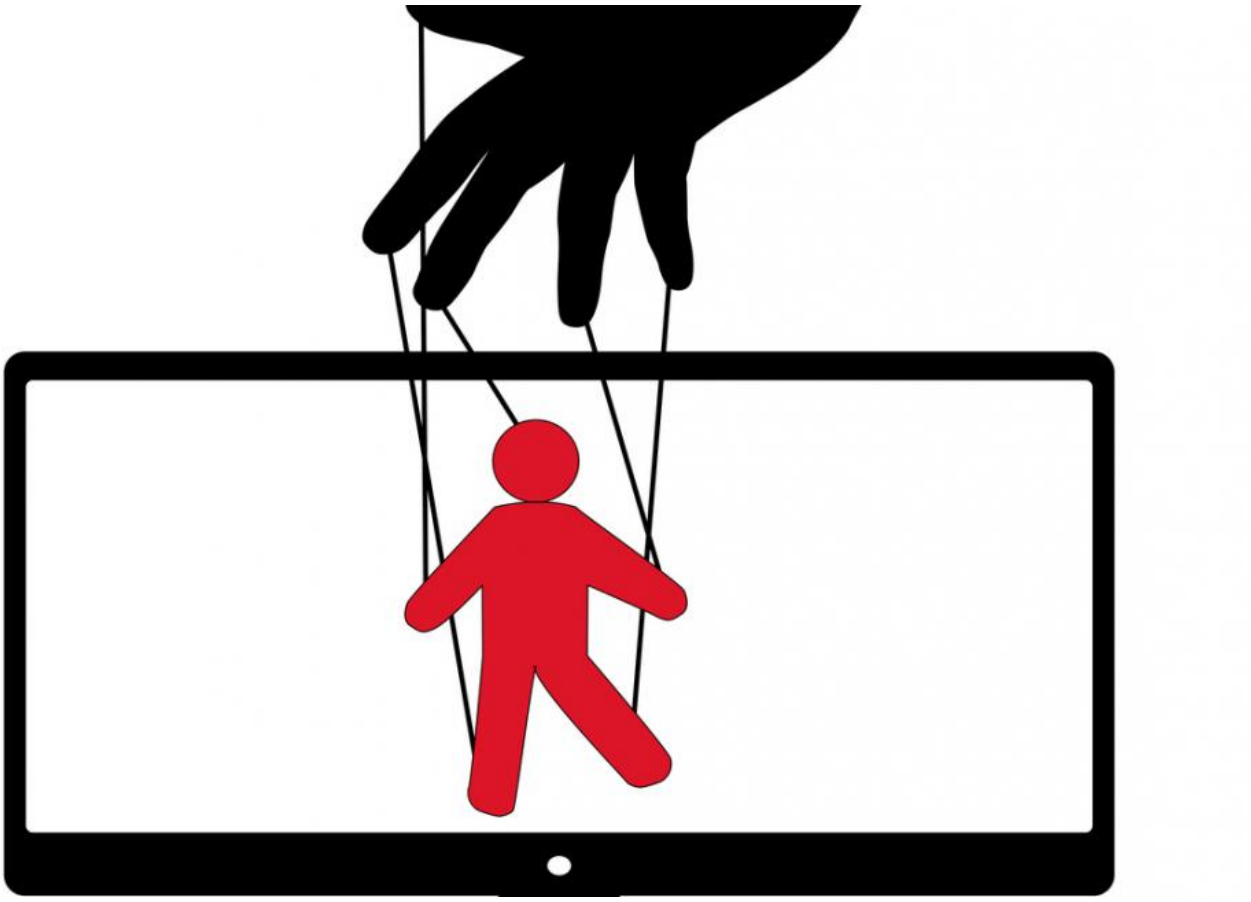

Photo Credit: Nosyrevy/iStock

The proliferation of visual digital media that can be produced and shared instantly has improved connectivity, but it comes with one major drawback. Visual media manipulation technologies have also evolved, giving amateurs and experts alike the tools to realistically manipulate media for potentially antagonistic ends like propaganda and misinformation.

The Media Forensics (MediFor) program out of DARPA’s Information Innovation Office (I2O) is trying to develop automated AI technologies to assess visual media manipulation at scale in an end-to-end platform, I2O Program Manager and MediFor

program lead Dr. Matt Turek outlined at Thursday's DARPA AI Colloquium in Alexandria, Virginia.

While manual media forensics can be time-consuming and requires expertise to identify manipulation, the MediFor platform would apply media forensics techniques automatically and at scale to compete with the online trove of visual media data, according to the MediFor program webpage.

The current visual media landscape favors manipulators because editing software is readily available and often easy to learn and use, according to the MediFor program webpage. Additionally, advanced editing techniques, such as deepfakes that use AI and machine-learning systems like generative adversarial network (GANs), can produce stunningly authentic manipulated visual media.

Advanced algorithmic manipulation is especially challenging because the pace of these technologies has accelerated rapidly over a short time. Manipulated videos and images that may be manually indistinguishable from the real thing present a series of real-world problems, including election and evidence tampering, blackmail, general propaganda and targeted social media misinformation efforts.

The MediFor program has become increasingly relevant in today’s world not only due to the explosion of visual media, but also because of the potential impact of manipulation and on such a wide scale.

Inside MediFor

To undertake automatic media manipulation detection, the MediFor program has established a framework of three tiers of information, including digital integrity, physical integrity and semantic integrity.

Digital integrity asks whether pixels in an image or video are consistent, said Turek. Physical integrity asks whether the laws of physics appear to be violated in visual media being examined. For example, with an image of a boat moving in the water, physical integrity information could be applied to see if that boat is producing a wake that corresponds to a real boat wake, Turek explained.

“The third level of information is semantic integrity, and that asks the question: is a hypothesis about a visual asset disputable?" Turek said. "This is where we can bring to bear outside information. It might be other media assets or other things that we

know are true."

For example, the MediFor program can use information on the semantic level to estimate weather properties like temperature and humidity based solely on pixels in an image. Using images with accurate metadata, the MediFor program leverages that metadata to gather weather data available in the cloud to train a deep neural network. That deep network estimates weather properties based solely on pixels. By comparing the accurate weather data from the former with the estimations of the deep network, the MediFor program can fine-tune the deep network to improve its performance.

Digital, physical and semantic integrity provide different avenues to verify the authenticity of visual media. They make it more difficult, if not impossible, for manipulators to alter visual media properly across all three levels.

The MediFor program is also combining information from these three tiers to generate one integrity score that encompasses an entire piece of visual media, be it a video or an image, according to Turek.

“In coming years, you may see technologies developed by the MediFor program that enable accurate detection of manipulated media. Such capabilities could enable filtering and flagging on internet platforms, for instance, that happens automatically and at scale and that are presented to you as the user. That gives us the ability to have much more comprehensive protection from media manipulation,” Turek said.

View printer friendly version
AI
social media
DARPA
algorithms
Standard