Preparing Local News for Deepfakes: Towards an ecosystems approach

— Corin Faife, Program Coordinator: Emerging Threats & Opportunities, WITNESS

Deepfakes—synthetically generated video created by artificial neural networks—have been described as "a looming challenge for privacy, democracy and national security" by leading experts on the internet and society.

Most of the recorded use of deepfakes in the wild involves "non-consensual pornography," in which the faces of celebrities (overwhelmingly female) are mapped onto the bodies of pornographic actors. Meanwhile, much of the discussion of deepfakes as a security threat centers on sophisticated disinformation operations mounted by foreign agents against high level political targets.

However, a crucial space in the fight against deepfakes will be the middle ground between these two poles: Fake videos that aim to destabilize politics at a local or regional level by targeting members of city or state legislature, business leaders, or community activists. (Moreover, internationally many community organizers are more concerned by the disinformation threat from their own governments than activities by nation-state actors.)

Over the past two years, WITNESS has become a leading voice in the push for practical, non-alarmist advocacy and legislation around deepfakes. We have developed a range of resources on synthetic media, a 12-point response framework, a review of OSINT challenges geared towards journalists, and have just released a comprehensive report on the various dilemmas presented by content authentication technology that is meant to address the problem.

In this presentation—which could be housed in either the Practice and Technology track—we will summarize our existing work on deepfakes, outline some of the technical solutions being proposed for detection of altered content and verification of trusted media, and provide guidance for incorporating these measures into local journalism workflows. We will also collectively identify gaps and opportunities to support media forensics capacity for small organizations.

In practical terms we will cover in-browser verification methods like InVid (and other forthcoming tools released between now and the conference date), and demonstrate content authenticity and provenance technology like ProofMode and Serelay—with discussion of its pitfalls and limitations, per the report.

The presentation will stress the importance of cross-disciplinary responses in adequately addressing the problem, drawing on Whitney Phillips' concept of an information ecosystem to emphasise networked rather than siloed solutions. And in discussion we will also be able to share insights from our ongoing conversations with Twitter, Facebook, Google and Microsoft over platform policy responses to synthetic media and how it intersects with digital publishing.

Rather than just presenting information, we will also be soliciting audience contributions on threat modelling, capacity gaps, and any proposed solutions that have not been covered, which we will feed back into our advocacy work towards tech platforms and on the legislative level.