

# Deepfakes, human rights, and archives

Association of Moving Image Archivists conference, Portland  
1 December 2018

Yvonne Ng, WITNESS

(Adapted from the work of my colleague Sam Gregory)

# Presentation overview

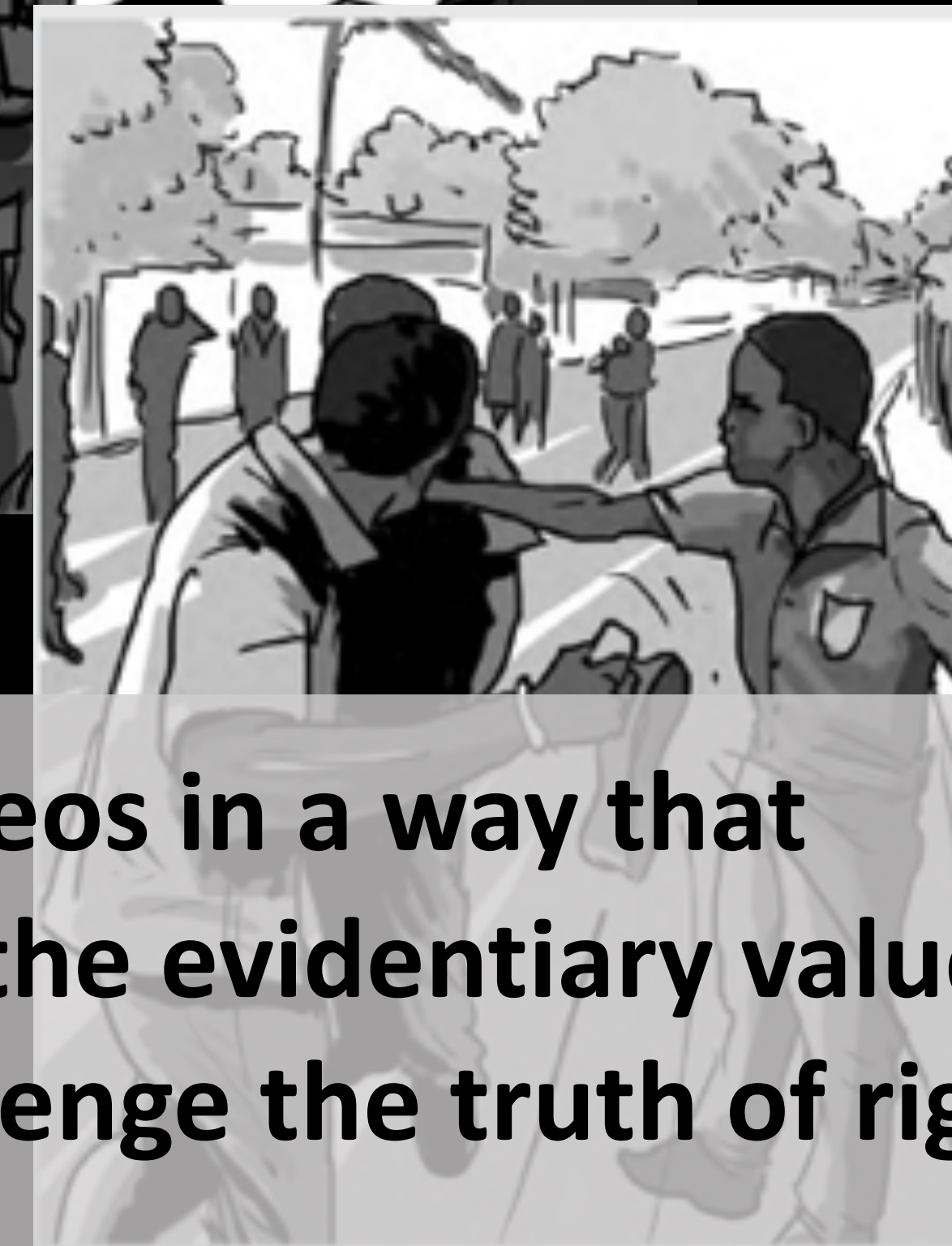
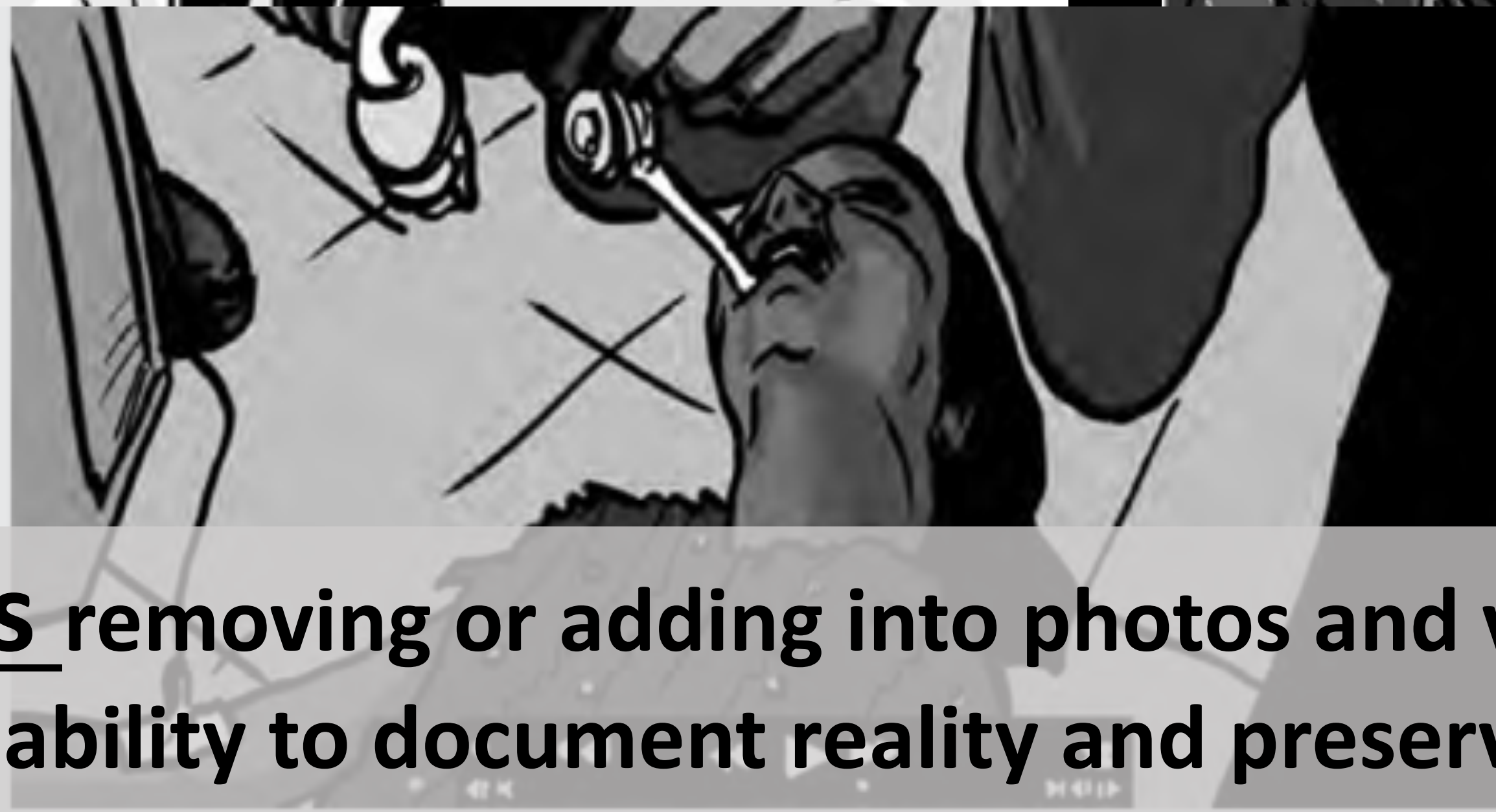
- Threats to human rights.
- Potential approaches.
- Role of archivists/archives.

**WITNESS** makes it possible for anyone, anywhere to use video and technology to protect and defend human rights.

# Deepfakes

- In this presentation, “deepfake” as shorthand for all kinds of “synthetic media” that might have more subtle effects:
  - Face swapping, hybrid faces
  - Simulated audio (e.g. Lyrebird)
  - Deleting foreground or background elements (e.g. Photoshop, Cloak)
  - Facial re-enactment (e.g. Face2Face)
  - Facial reconstruction with lip-sync of audio

# Deepfake Threats



**Reality edits** removing or adding into photos and videos in a way that challenges our ability to document reality and preserve the evidentiary value of images, and enhances the ability of perpetrators to challenge the truth of rights abuses.



**Credible doppelgangers of real people that enhance the ability to manipulate public or individuals to commit rights abuses or to incite violence or conflict.**

## #ElectionWatch: Doctored Video Goes Viral Ahead of Elections in Moldova

A doctored video viewed half a million times ahead of the local elections in Chisinau



On the left, the doctored video to suggest the news segment was about Chisinau, Moldova. On the right, the original news segment showing footage from Yemen. (Source: @DFRLab)

**News remixing that exploits peripheral cues of credibility and the rapid news cycle to disrupt and change public narratives.**





**Plausible deniability for perpetrators to reflexively claim “That’s a deepfake” around incriminating footage or taken further, to dismiss any contested information as another form of “fake news”.**

“... A people that no longer can believe anything cannot make up its own mind. It is deprived not only of its capacity to act but also of its capacity to think and to judge. And with such a people you can then do what you please.”

Hannah Arendt, 1974

**Floods of falsehood created via computational propaganda and individualized microtargeting, contributing to disrupting the remaining public sphere and to overwhelming fact-finding and verification approaches.**

# Potential outcomes

- Loss of trust in institutions, rising authoritarianism.
- “Digital wildfire” of misinformation leading to widespread mob violence.
- Demonization and targeting of individuals (human rights defenders, journalists, public figures) and social movement narratives/credibility.
- More reliance on AI in content moderation / content censorship and amplification.
- “Zero trust” becomes the baseline.

# Window of opportunity

- Broader malicious uses to disrupt political debate, undermine national security, confuse human rights investigations and attack businesses and civil society groups are not yet widespread.
- Neck-and-neck race between deep fake synthesizers and deep fake detectors / forensic techniques.


WORLD ECONOMIC FORUM

Agenda Initiatives Reports Events About

English TopLink

Digital Economy and Society | Internet Governance

## Heard about deepfakes? Don't panic. Prepare



The first wave of maliciously synthesized media is likely to be here in 2019. Image: Witness

23 Nov 2018

**Samuel Gregory**  
Programme Director, WITNESS

You may have noticed the panic around the [threat of 'deepfakes'](#). They are one form of so-called 'synthetic media', which draw on advances in AI and machine learning to change elements of a photo, video or audio track, or to recreate a person's voice or face with life-like subtlety. The insidious idea of anyone's face

# Potential approaches

- Build media literacy about these technologies, threats of mal-use, and discernment of deepfakes.
- Learn from existing practices and collaborate on solutions (human rights investigators, journalists, archivists, digital forensics, cybersecurity, etc)
- Develop pragmatic responses that can be taken by different actors, at different scales: from tech companies to independent activists, journalists, researchers.
- Address issue through variety of frameworks: legal, market, norms, technology.

# Role of archivists?

- **Media literacy:** How can archivists promote public literacy around deepfakes, and general methods of assessing authenticity or trustworthiness?
- **Learning from existing practices and collaborating:** How can archivists contribute their deep knowledge around issues of authenticity, provenance, description to this conversation?
- **Pragmatic responses:** How can archivists respond to specific misinformation threats or specific key events?
- **Frameworks:** How else can archivists influence laws, norms, technologies, market /commercial solutions that can address deepfakes?

# Rights-based approach

- We want to:
  - Maximize how many people can access / use these tools.
  - Minimize suppression of free speech, while protecting rights to privacy and freedom from surveillance.
  - Involves and minimizes risk to most vulnerable societies, creators and custody-holders.
  - Potentially integrate approaches into platforms so they are by default available to all users.

# Resources

- Summary of Discussions and Next Step Recommendations from “Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening.” [http://witness.mediafire.com/file/q5juw7dc3a2w8p7/Deepfakes\\_Final.pdf/file](http://witness.mediafire.com/file/q5juw7dc3a2w8p7/Deepfakes_Final.pdf/file)
- Deepfakes and Synthetic Media: What should we fear? What can we do? <https://blog.witness.org/2018/07/deepfakes/>
- Deepfakes and Synthetic Media: Survey of Solutions against Malicious Usages. <https://blog.witness.org/2018/07/deepfakes-and-solutions/>
- How Archivists Could Stop Deepfakes From Rewriting History. <https://gizmodo.com/how-archivists-could-stop-deepfakes-from-rewriting-hist-1829666009>



**Thanks!**

*[yvonne@witness.org](mailto:yvonne@witness.org) | [ng\\_yvonne](#)*

*My colleague Sam:  
[sam@witness.org](mailto:sam@witness.org)*