Summary of Discussions and Next Step Recommendations from

"Mal-uses of AI-generated Synthetic Media and Deepfakes: Pragmatic Solutions Discovery Convening"

June 11, 2018

WITNESS

SEE IT
FILM IT
CHANGE IT

with

FIRSTDRAFT

CONTACT INFORMATION

For media inquiries, please contact Matisse Bustos-Hawkes, WITNESS Associate Director of Communications and Engagement, matisse@witness.org.

For further information on the convening and next steps, please contact Sam Gregory, WITNESS Program Director, sam@witness.org.

PUBLISHED JULY 2018

TABLE OF CONTENTS

INTRODUCTION: ABOUT THE "MAL-USES OF AI-GENERATED SYNTHETIC MEDIA + DEEPFAKES: PRAGMATIC SOLUTIONS DISCOVERY CONVENING"

On June 11, 2018, WITNESS in collaboration with First Draft, a project of the Shorenstein Center on Media, Politics and Public Policy at Harvard Kennedy School, brought together thirty leading independent and company-based technologists, machine learning specialists, academic researchers in synthetic media, human rights researchers, and journalists. Our goal was to have an open discussion under the Chatham House Rule about pragmatic proactive ways to mitigate the threats that widespread use and commercialization of new tools for AI-generated synthetic media such as "deepfakes" and facial reenactment potentially pose to public trust, reliable journalism and trustworthy human rights documentation. This document is a summary of that discussion and recommendations.

For 25 years, WITNESS has enabled human rights defenders, and now increasingly anyone, anywhere to use video and technology to protect and defend human rights. Our work and the work of our partners demonstrates the value of images to drive a more diverse personal storytelling and civic journalism, to drive movements around pervasive human rights violations like police violence, and to be critical evidence in war crimes trials. We have also seen the ease in which videos and audio, often crudely edited or even simply recycled and re-contextualized can perpetuate and renew cycles of violence.

WITNESS' Tech + Advocacy program frequently includes engaging with key social media and video sharing platforms to develop innovative policy and product responses to challenges facing high-risk users and high-public interest content. As the threat of more sophisticated, more personalized audio and video manipulation emerges, we see a critical need to bring together key actors *before* we are in the eye-of-the-storm, to ensure we prepare in a more coordinated way and to challenge "technopocalyptic" narratives that in and of themselves damage public trust in video and audio. The convening goals included:

- **Broaden journalists, technologists, and human rights researchers' understanding** of these new technologies.

- While recognizing positive potential usages, begin building **a common understanding of the threats created by—and potential responses to—mal-uses of AI-generated imagery, video and audio** to public discourse and reliable news and human rights documentation, and map landscape of innovation in this area.

- Build **shared understanding of existing approaches in human rights, journalism, and technology** to deal with mal-uses of faked, simulated and recycled images, audio and video, and their relationship to other forms of mis/dis/mal-information.

- Based on case studies (real and hypothetical) **brainstorm potential pragmatic tactical, normative and technical responses to risk models of fabricated audio and video** by companies, independent activists, journalists, academic researchers, open-source technologists, and commercial platforms.

- Identify **priorities for ongoing discussion between stakeholders.**

SUMMARY OF RECOMMENDED NEXT STEPS

1. **Baseline research** and a **focused sprint** on the **optimal ways to track authenticity, integrity, provenance and digital edits of images, audio and video from capture to sharing to ongoing use**. Research should focus on a rights-protecting approach that a) maximizes how many people can access these tools, b) minimizes barriers to entry and potential suppression of free speech without compromising right to privacy and freedom of surveillance c) minimizes risk to vulnerable creators and custody-holders and balances these with d) potential feasibility of integrating these approaches in a broader context of platforms, social media and in search engines. This research needs to reflect platform, independent commercial and open-source activist efforts, consider use of blockchain and similar technologies, review precedents (e.g. spam and current anti-disinformation efforts) and identify pros and cons to different approaches as well as the unanticipated risks. WITNESS will lead on supporting this research and sprint.

2. **Detailed threat modelling around synthetic media mal-uses for particular key stakeholders (journalists, human rights defenders, others).** Create models based on actors, motivations and attack vectors, resulting in identification of tailored approaches relevant to specific stakeholders or issues/values at stake.

3. **Public and private dialogue on how platforms, social media sites and search engines design a shared approach and better coordinate around mal-uses of synthetic media**. Much like the public discussions around data use and content moderation, there is a role for third parties in civil society to serve as a public voice on pros/cons of various approaches, as well as to facilitate public discussion and serve as a neutral space for consensus-building. WITNESS will support this type of outcomes-oriented discussion.

4. **Platforms, search and social media companies should prioritize development of key tools** already identified in the OSINT human rights and journalism community as critical; particularly **reverse video search.** This is because many of the problems of synthetic media relate to existing challenges around verification and trust in visual media.

5. More shared learning **on how to detect synthetic media that brings together existing practices from manual and automatic forensics analysis with human rights, Open Source Intelligence (OSINT) and journalistic practitioners** - potentially via a **workshop where they test/learn each other's methods** and work out what to adopt and how to make techniques accessible. WITNESS and First Draft will engage on this.

6. **Prepare for the emergence of synthetic media in real-world situations** by working with journalists and human rights defenders to build **playbooks for**

upcoming risk scenarios so that no-one can claim “we didn't see this coming” and so as to facilitate more understanding of technologies at stake. WITNESS and First Draft will collaborate on this.

7. Include **additional stakeholders** who were under-represented in the June 11, 2018 convening and are critical voices either in an **additional meeting or in upcoming activities including:**

   - Global South voices as well as marginalized communities in U.S. and Europe;
   - Policy and legal voices and national and international level;
   - Artists and provocateurs.

8. **Additional understanding of relevant research questions and lead research** to inform other strategies. First Draft will lead on additional research.

MAL-USES OF SYNTHETIC MEDIA: WHAT ARE WE TALKING ABOUT?

It was clear from the convening discussions that the terms to describe advances in video and audio manipulation are not yet well-defined. The current conversation is dominated by the term “deepfakes” which refers to the result of software that “swaps” a face between one person and another.

In the convening we focused on a wider range of mal-uses of video and audio synthetic media. These included uses that deliberately try and falsify images to deceive an audience as well as techniques where many viewers will be aware the material is not “real,” as is true with many malicious usages of deepfakes. We also focused on the intersections of these developments with other “information disorder” problems of mal/mis/dis-information and computational propaganda.

Potential tools susceptible to mal-uses included:

- **Individualized simulated audio**: The enhanced ability to simulate individuals’ voices as developed and available commercially via providers such as Lyrebird or Baidu DeepVoice.

- **Emerging consumer tools that make it easier to selectively edit, delete or change foreground and background elements in video.** Concepts such as Adobe Cloak are advancing image editing currently available in tools like Photoshop or Premiere and competitors such as Pixelmator to allow better potential seamless editing of elements within video.

- **Facial reenactment:** This refers to using images of real people as “puppets” and manipulating their faces, expressions and upper body movements. Tools such as Face2Face and Deep Video Portraits allow the transfer of the facial and upper body movements of one person onto the realistic appearance of another real person’s face and upper body.

- **Realistic facial reconstruction and lip sync created around existing audio** tracks of a person as seen for example with the LipSync Obama project.

- **Real people with exchange of one region, typically a face**: Most commonly seen via deepfakes created using tools like FakeApp or FaceSwap these approaches also relate to technologies utilized in consumer tools like Snapchat, in which a simulation of the face of one person is imposed over the face of another person or in which a hybrid face is produced.

- **Combinations such as a deepfake matched with audio (simulated or real) and additional retouching**, e.g. the Obama-Jordan Peele video in which the actor-director Jordan Peele made a realistic Obama say words that Peele himself was saying.

CORE CONCEPTS AND RECENT RESEARCH INNOVATIONS

It is outside the scope of this report back to provide a detailed technical coverage of advances in machine learning, deep learning and graphics.

However, one relevant approach that has enabled the development of new forms of image synthesis is the growth of the subfield of machine learning known as Deep Learning, which uses architectures for artificial intelligence similar to neural networks. Generative Adversarial Networks (GANs) are the technology used in deepfakes. Two neural networks compete to produce and discern high quality faked images. One is the "generator" (which creates images that look like an original image) and the other is the "discriminator" (which tries to figure out if an image is real or simulated). They compete in a cat and mouse game to make better and better images.

The cost of producing new forms of synthetic media has decreased significantly in the last few years given increasing amounts of training data, computing power and effective publicly shared approaches and code. However, there are still significant limits—for example the cost of computational power, the need for a large number of good quality images as training data, artefacts in images, and that GAN models can frequently be brittle and fail to generate effective fakes. These factors are significant when we look at questions of detecting synthetic media and deepfakes—for example via artefacts—as well as the importance of training data within these models and within automatic forensics models for detecting fakes (such as the FaceForensics approach noted below).

For further detail on production of deepfakes see for example, Alan Zucconi's guide.

THE ARMS RACE BETWEEN SYNTHESIS AND FORENSICS

There is an ongoing arms race between manual and automatic synthesis of media, and manual and automatic forensic approaches.

Manual synthesis is characterized by the *explicit modeling* of geometry, lighting and physics that we see in Hollywood effects. CGI has been a part of movie industry for 30 years, but it is time-consuming, expensive, and has required domain expertise. On the other hand, automatic synthesis involves use of *implicit synthesis* of texture, lighting or head motion as we have seen for example in LipSync Obama, Deep Video Portraits or of course, deepfakes. Techniques here often involve a combination of computer vision and computer graphics, and in some cases uses of neural networks. Tools such as LipSync Obama build on a 20-year research trajectory of exploring how to create 3D face models from existing images. There are a range of positive applications of enhanced synthetic media including video and virtual telepresence, VR and AR and content creation, animation and dubbing. There will also be uses in autonomous systems and in human computer/human-robot interaction.

Editing software and manual and automatic synthesis can increasingly create perceptually realistic images that are not visible as manipulated to the naked eye and visual analysis.

Manual forensics does explicit checks of perspectival geometry, lighting, shadows and the "physics" of images, as well as detecting for example copying and splicing between images and evidence of the camera model for a photo. A recent notable example of manual forensics specific to deepfakes is the idea of using a technique known as Eulerian Video Magnification, to see the visible pulse rate of real people that would be absent in a deepfake.

An emerging field is automatic forensics. Approaches explored in this include looking at larger datasets and using machine learning to do forensic analysis. Recent experimentation includes:

- Detection of copy and splicing or use of two different camera models on origin images;
- Detection of "heat map" of fake pixels in facial images created using FaceSwap;
- Identification of where elements of a fake image originate;
- Use of GANs themselves to detect fake images based on training data of synthetic video images created using existing tools (the FaceForensics database).

Most systems are trained on specific databases, and might detect mainly the inconsistencies of specific synthesis techniques, although there is work in progress that addresses these shortcomings. There are also new approaches that use GANs to fight back against forensic analysis—for example, by wiping the forensic traces of multiple cameras and creating an image that appears to have the uniform camera signature of another camera.

Researchers disagree on whether the "arms race" is likely to be won by the forgers or the detectors. Humans are not good at detecting the difference between a real and a fake video (see data in FaceForensics) but machines are. Detection is currently easier than forgery and for every forgery AI there is a powerful detection model. Provided there is sufficient training data showing new types of faked images, audio and video, the use of GANs should be able to keep up in enabling AI-assisted identification of nonvisible faking. There might be a time lag but detection should keep improving.

THE CONTEXT OF 'INFORMATION DISORDER' AND F*** NEWS

Synthetic media must be considered in the context of how we understand mal/dis/mis-information spreading and being received by audiences. We should consider how to view the use of synthetic media within the broader context of "information disorder" as well as within patterns of how information is released and shared in the current information environment (e.g. via coordinated timely leaks, or in DMs to a journalist, or via state lead multipronged campaigns). Of particular concern is how poorly we as humans are equipped to discern around video and audio, how it can create false memories, and the limitations of current approaches to 'fact checking' or platform-based solutions when it comes to video and audio.

We should be concerned about the impact of visual information. It's a much bigger threat than text: it's more prevalent in low literacy cultures; users don't have to leave social networking sites to watch so they are more likely to glance at visual information, and our brains are wired to be more trusting of visuals. We also don't have much context for fabricated realistic audio (except '[The War of the Worlds'](#)) but our ears are trusting and we don't have the framework to stop and think twice.

An information disorder approach to understanding mal-uses of synthetic media will consider their integration into a space characterized by axes of "falseness" and "intent to harm" that are captured in Figure 1.

Mal-uses of synthetic media will primarily be disinformation and misinformation.

Likely usages of synthetic media will include: as part of imposter content that uses news outlet logos and journalist names shared to build credibility, satire shared on without users realizing it is satire or non-satirical content labelled as satire to avoid fact checkers, as well as recycled, manipulated and wholly fabricated content.

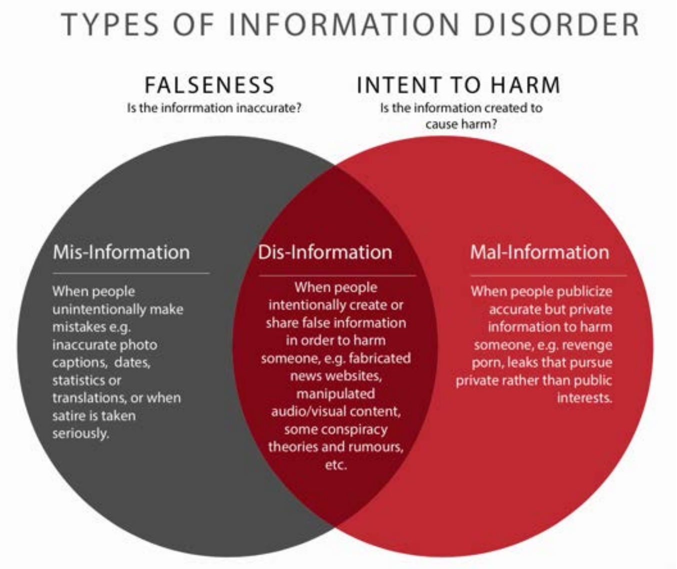
TYPES OF INFORMATION DISORDER
FALSENESS
Is the information inaccurate?
INTENT TO HARM
Is the information created to cause harm?
Mis-Information
When people unintentionally make mistakes e.g. inaccurate photo captions, dates, statistics or translations, or when satire is taken seriously.
Dis-Information
When people intentionally create or share false information in order to harm someone, e.g. fabricated news websites, manipulated audio/visual content, some conspiracy theories and rumours, etc.
Mal-Information
When people publicize accurate but private information to harm someone, e.g. revenge porn, leaks that pursue private rather than public interests.

Figure 1

Any conversation about synthetic media must focus on global implications (not just U.S.) especially in situations with low levels of trust and high levels of cultural, social, ethnic or religious division given the growing mal-information, misinformation and disinformation problems in these contexts.

We need to learn from existing practices in the verification community for both fast and slow debunks, and fast and slow verification. These include the 30-second checks of seeing whether content is indexed on a search engine with a Google Reverse Image Search, as well as the harder challenges of verifying an image from Ghouta, Syria that takes three days or finding content that is appearing in real time and doesn't have a presence on search engines.

Approaches to identifying synthetic media might also focus on real time tools that get to the sourcing (as with other mis/mal-information problems) rather than tools to help fact-check the content.

While is beneficial for the broader community that most verification tools are free and open-source, they are often poorly resourced and mainly built by hobbyists. A better tool set is needed.

MAL-USES OF SYNTHETIC MEDIA UNDERSTOOD IN RELATIONSHIP TO OTHER MIS/DIS-INFORMATION TRENDS

Of critical importance in understanding potential mal-uses synthetic media we noted:

- Overlap with the rest of the mis/dis/mal-information sphere;
- Overlap with personalization and microtargeting in platforms and hacks on the "attention economy" on which tools and platforms are built;
- Relationship to bots and computational propaganda;
- Relationship to fundamental differences of frame and understanding experienced in polarized societies;
- Relationship to established patterns of information warfare by state actors and distribution planning in spaces like 4Chan and 8Chan.

DISCOVERY, VERIFICATION AND PRESENTATION OF AUDIO AND VIDEO IN JOURNALISM AND HUMAN RIGHTS

Contemporary verification of open-source photos, audio and video by human rights and journalism incorporates a range of open-source and proprietary tools (for one example of a verification toolkit of a prominent open-source intelligence-based OSINT journalist see here). There is no one single magic tool—most practitioners will use 5-10 tools on a daily basis. Currently Reverse Image search is the most powerful tool used in day-to-day work, as it can take 30 seconds to debunk a picture/video. However, there is not yet a good reverse video search—practitioners can only use image thumbnails to search. For more complex stress testing on images an OSINT practitioner will conduct a series of tests using different technologies and doing visual inspection, feature extraction, spatial analysis, satellite imagery and multi-source corroboration with other footage sources and with individuals who were on the ground, or with uploaded content to social media. A range of these methods are explained by groups like First Draft or made available in tools such as Citizen Evidence Lab and InVid.

More complex models for event reconstruction and for confirming the integrity of particular videos via spatial analysis and comparison to other videos, audio and sources and via spatial analysis can be found in the rigorous multi-source documentation work of groups like Situ Research and Forensic Architecture.

However, in many cases the tools to debunk are there, and they are easy to use—in fact a debunk can take 30 seconds. Yet the same photo or video might still circulate years later, recycled in multiple crisis contexts.

WHAT MAKES THIS MOMENT DIFFERENT FROM BEFORE?

The potential mal-uses of synthetic media have precedents in existing and historical patterns of recycling, re-editing and manipulating photos, video and audio for malicious purposes. So, what makes these emerging uses different from a fake public record, a manipulated image, a faked tweet purporting to be from a public figure, a “cheapfake” or “shallowfake” swapping the audio on an existing video to incite violence or presenting one video as being from another context, or a virulent hate-meme?

One approach to understanding this looks at how synthetic media (combined with other forms of AI, computational propaganda and disinformation trends):

- **Expands** existing threats. *For example, by expanding the range of people who are able to create threats, or the range of potential targets.*

- **Introduces** new threats. *For example, the previously unfeasible use of widespread malicious use of personalized faked voice recordings.*

- **Alters** existing threats. *For example, by making it easier or faster to carry out particular types of mal-uses or to more finely target those attacks?*

- Is **reinforced by** other threats. *For example, by the interaction of synthetic media with computational propaganda.*

Other key characteristics appear to mark a potential difference from previous precedents, in some cases in lasting ways, in other ways that are addressable including:

- Visual and audio manipulation at scale as opposed to in limited, labor-intensive ways

- Vastly increased possibility to simulate real people including public figures

- Combined with AI-enhanced micro-targeting, the increased possibility to target individuals and small groups within messaging apps

- Significant reduction in cost and time to produce and share

- Increasing audio and video production, created and shared on mobile devices

- A context where we are poorly equipped cognitively and in our media literacy skills to discern falsified audio and video

- A journalistic and platform response to fake news issues that has largely focused on text-based mis/dis/mal-information and text-based response

WHAT ARE THE THREAT SCENARIOS AROUND MAL-USES OF SYNTHETIC MEDIA?

Threat scenarios range from more ephemeral, localized and fast-moving, to more structured attacks on institutions. AI-generated synthetic media also has the potential to create or contribute to long-term negative effects on the information ecosystem and levels of public trust. Examples of how these threats could emerge are not hard to find if we look at existing practices and history. For example, micro-targeted, personalized, weaponized attacks and incitement could become more effective—imagine if the hate radio of Radio Mille Collines in 1994 Rwanda were now individualized and personalized. Gender-based attacks on female journalists, politicians and activists occur all too frequently and have already been fomented with faked images — most recently to Ranya Ayyub in India. Low-res images circulate in volatile communities and function to incite a “digital wildfire” of malicious rumors rapidly shared via social networking sites and messaging apps: this has most recently happened with edited image on WhatsApp.

THREAT STORIES AND SCENARIOS

(prepared by the conveners and added to during the discussion)

- Digital wildfire (aka “shouting fire in a crowded theater”) shared rapidly and locally

- Digital wildfire shared rapidly and locally, primarily in closed messaging apps

- Induced public panic via “fake” emergency

- Manipulated “sting” videos in a short news cycle

- “Exclusive” leaks to journalists within rapid news cycle

- Swamping newsroom operations with unverifiable media

- Political speech in public is manipulated

- Local politicians manipulated and leveraged for national impact

- Credible doppelganger of a leading political figure calling for an action
- Fakes shared during the critical pre-election media lockdown period common to many democracies
- Credibility-based attacks on public figures, human rights defenders and journalists— corruption, adultery, racism
- Gender-based attacks on credibility of human rights defenders and journalists, including in repeated DDOS approach
- Increasingly sophisticated spoofing of opponent identities
- Targeting of dissidents in authoritarian contexts with ubiquitous, unconstrained surveillance
- “Poisoning the well” in a leak with a few well-faked videos
- Poisoning the well of an investigation
- Altered documentation of war crimes violations compromises credibility of investigators and journalists
- Attacks on social movement narratives and credibility
- Autogenerated, microtargeted, individualized visual content aka “laser phishing” (Ovidya)
- Simulated, individualized hate speech audio
- New forms of “astroturfing” for fake public opinion at scale with convincing content
- “Floods of falsehood” involving volume of faked images shared as part of a computational propaganda attacks and using individualized microtargeting, that contribute to disrupting the remaining public sphere
- Integration of faked audio/video into ongoing state/parastatal disinformation campaigns
- Under-resourced courts and legal processes increasingly reject video and image evidence due to the uncertainty created by prevalence of synthetic media
- Integration by a range of disinformation actors

- Widespread use to destroy trust in institutions in authoritarian societies

- Plausible deniability for the powerful on any image as captured by Hannah Arendt: *"… 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."*

In summary, what does this look like in terms of potential human rights, journalism and public trust scenarios?

- **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.

- **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.

- **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.

As a next step, participants recommended using approaches based on incentives for different actors (e.g. the incentives that companies have to understand non-organic, non-promoted content circulation on commercial platforms) as well as use of a formal threat modeling framework that considers actors, motivations and vectors and identifies which actors to invest in supporting. We also considered identifying more explicitly what we are protecting: Is it elections, people's understanding of their health, diplomacy, governance, trust?

We noted the variance here between discreet threats that have existing localized, fast-moving analogues (for example, digital wildfire of rumors on WhatsApp in India), or existing information paradigms they plug into (for example, dumps of leaked documents shared anonymously) as well as broader systems effect that will place increased cost on institutional gatekeepers like newsrooms and human rights organizations or create an overall climate on "plausible deniability" on any news or information item or increased trust in rumor and conspiracy theory. Another risk is in how a generalized response may reduce freedom of expression broadly whether implemented at a governmental level or a platform level.

A key consideration is how these different levels of threats are positively or negatively correlated. Does the reduction of one contribute to the reduction of the other? Can we accept increase in one area for reduction in another? Similarly, how do solutions trade off?

This becomes most apparent when we look at the opportunity and challenges of increasing authentication of media at source (discussed further in solutions below). Increased authentication will have tradeoffs on security and privacy in the moment and over time and these are most likely to create risks for those who are already vulnerable. It will also exclude people who don't have the economic means to participate in new authentication systems and at least until globally accepted, will credit an arbitrary "in" or "out" system of authentication.

Detection mechanisms need to be thought of as an extra signal and research is needed on whether this would be best communicated to the public or provided to more expert fact-checkers. There are significant dangers of false positives on detection at scale. On a large platform even a .001 false positive rate would be an extremely serious problem, particularly if it affected a particularly vulnerable group. It is also noteworthy that some studies have shown that marking content as false can increase the velocity of its spread; and alongside this, people consciously share untrue information.

A key question is how multiple actors, including consumers, civil society, journalists, technologists, and companies will contribute to addressing the increased risk of synthetic media maliciously used to simulate reality.

Technological approaches may exist in centralized contexts—for example there are analogs of companies coming together to create tools and identify classifiers to combat child exploitation imagery. However, the analogs here are less strong given the political or satirical nature of potential synthetic media content as well as dual usage of synthetic media tools for creative purposes. There is also not yet the legal imperative that drives the removal of child sexual content from platforms. Other potential historical and contemporary examples include counter-violent extremism, content moderation more broadly and spam. Current challenges around hate speech and controversial content moderation, especially in locations such as Myanmar, illustrate the challenges of both keeping up and taking down content. A centralized solution would also raise questions about the integrity and authority of blackbox systems that are not vetted by a broader technical community, as well as about the value of a diversity of detectors and systems to detect rather than one shared system. Tools for sourcing and for seeing patterns (e.g. of a bot-driven attack, or coordination in 4Chan) might be preferable to tools for detection and might not require being only a machine learning (ML) technique.

Many of these approaches will be less relevant for content that is not hosted on centralized, controllable platforms—but rather on other non-centralized media sharing where it's impossible to take down. Other challenges include the growing importance of end-to-end encrypted messaging apps and small groups as vectors.

Regulation (self or government) will become key because it will drive decisions about company investments in particular solutions, as has been happening with the current wave of 'fake news' legislation worldwide.

HOW AFRAID SHOULD WE BE ABOUT AI-GENERATED SYNTHETIC MEDIA?

Participants discussed their level of concern around AI-generated synthetic media right now. From a technical perspective some researchers were less concerned as they believed detection would keep up and that much of the current capacity of AI-generated synthetic media is overstated and overhyped. Others working more directly in platforms and vulnerable communities were concerned that the stakes are too high and worry about existing patterns of Behavior where faked content is causing communal riots, and how a particularly prominent singular usage (e.g. a faked politician) could have huge implications. Others were more concerned about localized harms that would come from usage of these tools at a local level, including to disparage other businesses and communities. Others were concerned about how the threat of deepfakes will be used as another excuse to reduce freedom of expression and to throttle access to the Internet.

Looking ahead five years, participants hoped that as deepfakes become more mainstream people will have enough visual media literacy to ask the right questions, and that platforms will have built out robust infrastructure to detect them. However, others worried that in five years' time we will not have developed sufficient literacies and resiliency. Equally risky is the broader effect of the presence of realistic simulated or manipulated personalized audio and video in destabilizing the nature of truth. There will be tradeoffs between pursuing detection on individual media items (and creating risk of 'implied truth' on fakes that get through) versus investing in broader resiliency on how to ascertain truth in an environment with increased and more personalized audiovisual fakery.

This may lead to a world where consumers opt-into more controlled platforms, essentially trading off against the problem and choosing to turn the 'fake news dial up or down'.

Whether it will still be likely that the detectors of synthetic media are better than the creators will depend on whether the detector models are publicly available and also on the availability of public data on synthetic media creation tools, including examples that can be added to training data to enable generalization from existing models. For synthetic media created on consumer-oriented tools built by western companies (e.g. on an Adobe product) these companies do not have the motivation to make manipulations non-detectable at a forensics level.

It seems more likely that in the long run there will be ways to maintain the integrity of court cases (at least in well-resourced jurisdictions) and long-term investigative journalism. The damage will be mostly on media that moves fast, such as digital wildfire, and the broader effects on bad actors in society to claim plausible deniability.

OBSERVATION: There are two separate classes of threats: one is “the inability to show that something real is real, and then the ability to fake something as if it was real”. We should be more worried about the inability to show anything is real.

KEY QUESTION: When will the cost of synthetic media fakery meet the needs and expectations of people to use them broadly? When will the cost-benefits of deepfakes and other forms of synthetic media bring them into as widespread usage as cheapfakes?

WHAT ARE POSSIBLE SOLUTION AREAS?

There are already a range of potential solution areas that have been suggested to confront the mal-uses of synthetic media.

In the convening we discussed possible frameworks for thinking about solutions including Lessig’s framework of options emerging from Norms, Architecture (Code), Law and Markets, as well as models focused on different actors such as individuals, broader individual and societal resilience, institutional actors such as journalists and human rights workers or election monitors, as well as technologists and platform companies.

We reviewed a range of potential analogies that could shape thinking at this early stage in the discussion and to avoid the conversation being bound by one problem framing.

Among the analogies we noted that could shape responses:

- Banknotes and anti-forgery? *In this case, an institution makes it hard to create the notes and provides an easily accessible way for individuals and businesses to verify them.*

- Legal documents and chain of custody? *Established systems track content within a legal process.*

- Spam detection and harm reduction? *Including consideration of acceptable levels of false positives.*

- Image detection and hashing? *As used in policies and technologies deployed against violent extremism, copyright violations, and child exploitation imagery.*
- Drug resistance? *If we don't come up with a common approach, it allows a threat actor to build skills and ability in a particular under-resourced context.*

- Cyber-security? *Suggesting that this will be a constant offense-defense, with use of blue and red-team testing on risk areas.*

- Asymmetrical warfare?

- Journalism and experience of verification of user-generated context? *Including that open-source tools, widely shared can help build a strong community of practice around better verification.*

- Human rights and particularly existing patterns of attacks on the basis of gender?

POTENTIAL APPROACHES AND PRAGMATIC OR PARTIAL (AND NOT-SO) SOLUTIONS

As part of the convening we reviewed information gathered by WITNESS on the range of external solutions proposed to-date around existing and potential mal-use of synthetic media. These included *(with expanded detail provided for readers unfamiliar with these areas):*

Invest in media literacy and resilience for news consumers and invest in resiliency and discernment against disinformation

There is already an increase in funding and support for efforts to promote media literacy around disinformation both among educators, foundations and media outlets. These initiatives could further integrate commonsense approaches to spotting both individual items of synthetic media (e.g. via visible anomalies such as mouth distortion that are often present in current deepfakes) as well as developing approaches to assessing credibility more broadly and to supporting people on how to engage with this content. This article by Craig Silverman for Buzzfeed is an example noting some simple steps that could currently be taken to identify a deepfake at this point—some such as checking the source are common to other media literacy and verification heuristics, while advice to "inspect the mouth" and "slow it down" are specific to the current moment in deepfakes detection.

Other work will need to build on the growing body of research on social media and disinformation. A recent Harvard Business Review article, "Big Idea on 'Truth, Disrupted'," provides a summary of recent research and approaches. More depths is provided in resources such as the Council of Europe's "Information Disorder" report, the

Social Science Research Council report on the state of the field in Social Media & Democracy, the recent scientific literature on "Social Media, Political Polarization, and Political Disinformation" and The Science of Fake News as well as the ongoing work of Data & Society on and First Draft. Research and responses lag significantly in how to deal with visual and audio information and in non-U.S. contexts.

Build on existing efforts in civic and journalistic education and tools for practitioners

There is a range of existing efforts in journalistic, human rights and OSINT discovery and verification efforts that support practitioners in those areas to better find, verify and present open-source information, including video, audio, and social media content. New approaches to recognizing and debunking deepfakes and synthetic media can be built into the toolkits, browser extensions, and industry training provided by First Draft, Google News Initiative and similar non-governmental and industry peers, as well as the efforts of groups like Bellingcat and WITNESS' Media Lab and Video As Evidence efforts. They will undoubtedly also be integrated by industry leaders working in social media verification such as Storyful and the BBC.

Reinforce journalistic knowledge and enhanced collaboration around key events by supporting better collaboration on understanding and rapid identification and response in the journalism community

In response to misinformation threats, competing journalistic organizations have worked together around elections and other potential crises via initiatives such as Crosscheck and Verificado. In preparation for potential deepfake deployment that will try to target "weak links" in the information chain in upcoming elections in the U.S. and elsewhere, coalitions of news organizations working on shared verification can integrate an understanding of deepfakes and synthetic media, threat models and response approaches into their collaboration and planning as well as coordinate with researchers and forensic investigators.

Explore tools and approaches for validating, self-authenticating and questioning individual media items and how these might be mainstreamed into commercial capture/sharing tools

An increasing range of apps and tools seek to provide a more metadata-rich, cryptographically signed, hashed or otherwise verifiable image from point of capture. These include apps for journalists and human rights defenders such as ProofMode and commercial tools such as TruePic. A potential additive element here that a range of companies are exploring is the use of the blockchain as a distributed media ledger to track content and edits. The use and validation of 'confirmed' live video, recorded from a live broadcast, might also play a role. More whimsically certain creators could pursue a re-emphasis on analog media for trustworthiness.

Invest in rigorous approaches to cross-validating multiple visual sources

Approaches pioneered by groups like Situ Research with its Euro-Maidan Ukraine killings reconstruction, Forensic Architecture, Bellingcat and the New York Times Video Investigations Team utilize combinations of multiple cameras documenting an event as well as spatial analysis to create robust accounts for the public record or evidence. These approaches could overlap with improved tools for authenticating and ground-truthing eyewitness video, allowing for one authenticated video to anchor a range of other audiovisual content.

Invest in new forms of manual video forensics

As synthetic media advances, new forms of manual and automatic forensics could be refined and integrated into existing verification tools utilized by journalists and factfinders as well as potentially into platform-based approaches. These will include approaches that build on existing understanding of how to detect image manipulation and copy-paste-splice, as well as approaches customized to deepfakes such as the idea of making blood flow more visible via Eulerian video magnification with the assumption that natural pulse will be less visible in deepfakes (*note: some initial research suggests this may not be the case).*

The US government via its DARPA MediFor Program (as well as via media forensics challenges from NIST) continues to invest in a range of manual and automatic forensics approaches that include refinements on existing approaches for identifying paste and splice into images and tracking camera identities and fingerprints. Other approaches look for physical integrity ('does it break the laws of physics') issues such as ensuring there is not inconsistency in lighting, reflection and audio as well reviewing the semantic integrity of scenes and identifying image provenance and origins (pdf). Many are looking for additional new forms of neural network-based approaches described below.

Invest in new forms of deep learning-based detection approaches

New automatic GAN-based forensics tools such as FaceForensics generate fakes using tools like FakeApp and then utilize these large volumes of fake images as training data for neural nets that do fake-detection.

These and similar tools developed in programs like MediFor such as the use of neural networks to spot the absence of blinking in deepfakes (pdf) could be incorporated into key browser extensions or dedicated tools like InVid. They could also form part of a platform, social media networks and search engines' approaches to identifying signs of manipulation. Platforms have access to significant collections of images (including increasingly, the new forms of synthetic media) and could collaborate on maintaining updated training data sets of new forms of manipulation and synthesis to best facilitate use of these tools. Platforms, as well as independent repositories such as the Internet

Archive, also have significant databases of existing images that can form part of detection approaches based on image phylogeny and provenance that detect the use of elements of existing images via both neural networks and other approaches.

Track and identify malicious deepfakes and synthetic media activity via other signals of activity in the info ecosystem

As identified elsewhere, the best way to track deepfakes or other synthetic media may be to focus on real-time tools for sourcing enhanced bot activity, detecting darknet organizing or creating an early warning on coordinated state or para-statal action. Recent reporting from the Digital Disinformation Lab at the Institute for the Future and the Oxford Internet Institute, among others, explores the growing pervasiveness of these tactics but also signals of this activity that can be observed.

Identify, incentivize and reward high-quality information, and rooting out mis/mal/disinformation

A distributed approach could include analogies and lessons learned from cyber-security, for example, the use of an equivalent to a "bug bounty."

Support platform-based approaches (social networks, video-sharing, search, and news) including many of the above elements

Platform collaboration could include detection and signaling of detection at upload, at sharing, or at search. They could include opportunities for cross-industry collaboration and a shared approach as well as a range of individual platform solutions from bans, to de-indexing or down-ranking, to UI signaling to users, to changes to terms-of-service (as for example with bans on deepfakes by sites such as PornHub or Gyfycat). Critical policy and technical elements here include how to distinguish malicious deepfakes from other usages for satire, entertainment and creativity, how to distinguish levels of computational manipulation that range from a photo taken with "portrait mode" to a fully engineered face transplant, and how to reduce false positives; and then how to communicate this to regular users as well as journalists and fact-finders. As Nick Diakopolous suggests, related to solutions around supporting journalism, if they 'were to make media verification algorithms freely available via APIs, computational journalists could integrate verification signals into their larger workflows'.

Human rights and journalists' experience with recent platform approaches to content moderation in the context of current pressures around 'fake news' and countering violent extremism—with Facebook in Myanmar/Burma and with YouTube's handling of evidentiary content from Syria— highlights the need for extreme caution around approaches focused on takedowns of content. WITNESS' recent submission to the United Nations Special Rapporteur on Freedom of Opinion and Expression highlights many of the issues we have encountered, and his report highlights steps companies should take to protect and promote human rights in this area.

In addition, there remain gaps in the tools available on platforms to enable solutions to other existing verification, trust and provenance problems around recycled, faked and other open-source images. NOTE: One key recommendation out of the expert convening was that platforms, search and social media companies should prioritize development of key tools already identified in the OSINT human rights and journalism community as critical; particularly ***reverse video search***.

Ensure commercial tools provide clear forensic information or watermarking to indicate manipulation

Companies such as Adobe producing consumer-oriented video, image, and audio manipulation tools have limited incentives to build counter-forensics measures into the outputs of their products since they are designed to be convincing to the human eye but not machines. There should be a unified consensus that consumer video and image manipulation should be machine forensics readable to the maximum extent possible, even if the manipulation is not visible to the naked human eye. Another approach would look at how to include new forms of watermarking - for example, as suggested by Hany Farid, to include an invisible signature to images created using Google's TensorFlow technology, an open-source library used in much machine learning and deep learning work. Such approaches will not resolve the analog hole where a copy is created of a digital media item but might provide traces that could be useful to signal forensically many synthetic media items.

Protect individuals vulnerable to malicious deepfakes by investing in new forms of adversarial attacks

Adversarial attacks include invisible-to-the-human-eye pixel shifts or visible scrambler-patch objects in images that disrupt computer vision and result in classification failures. Hypothetically these could be used as a user or platform-lead approach to "pollution" of training data around specific individuals in order to prevent bulk re-use of images available on an image search platform (e.g. Google Images) as training data that could be mobilized to create a synthetic image. Others such as the EqualAIs initiative are exploring how similar tools could be used to impede increasingly pervasive facial recognition and preserve some forms of visual anonymity for vulnerable individuals (another concern of WITNESS' within our Tech + Advocacy program).

Consider the pros and cons of an immutable authentication trail, particularly for high profile individuals

As suggested by Bobby Chesney and Danielle Citron, this concept of using lifelogging to voluntarily track movements and action to provide the potential of a rebuttal to a deepfake via "a certified alibi credibly proving he or she did not do or say the thing depicted" might have applications for particular niche or high-profile communities, e.g. celebrities and other public figures although not without significant collateral damage to privacy and the possibility of facilitating government surveillance.

Ensure communication between key affected communities and the AI industry

The most-affected by mal-uses of synthetic media will be vulnerable societies where misinformation and disinformation are already rife, where levels of trust are low and there are few institutions for verification and fact-checking. Many of the incidents of 'digital wildfire' where recycled or lightly edited images have spread violence have recently taken place in the context of closed messaging apps such as WhatsApp in India.

Most recently in the human rights space, there has been mobilization in the Global South Facebook Coalition to push Facebook to listen more closely and resource and act on to real-world harms in societies such as Myanmar/Burma and Sri Lanka. These groups, and the likely risks and particular threat paradigms in these societies need to be at the center of solutions.

Confront shared root causes with other dis/mal/misinformation problems

There are shared root causes with other information disorder problems around how audiences understand and share mis and disinformation. There are also overlaps with broader societal conversation around micro-targeting of advertising and personalize content and how "attention economy" focused technologies reward fast-moving content and that are oriented towards an attention economy approach.

Develop industry and AI self-regulation and ethics training/codes, as well as 3rd party review boards.

As part of the broader discussion of AI and ethics, there could be a stronger emphasis on training on human rights and dual-use implications of synthetic media tools (for example, drawing on operationalization of the Toronto Principles on AI); this could include discussion of these in research papers and of use of independent, empowered 3rd party review boards.

Pursue existing and novel legal, regulatory and policy approaches

The convening did not discuss this area in depth. Since the convening date, Professors Bobby Chesney and Danielle Citron have published an advance draft of "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," a paper outlining an extensive range of primarily US-centric legal, regulatory and policy options that could be considered. Legal options include new narrowly targeted prohibition on certain intentionally harmful deepfakes, the use of defamation or fraud law, civil liability including the possibility of suing creators or platforms for content (including via potential amendments to CDA Section 230), the utilization of copyright law or right to publicity, as well as criminal liability. Within the US there might be potential limited roles for the Federal Trade Commission, the Federal Communications Commission and the Federal Elections Commission.

Other areas to consider that have been raised elsewhere include re-thinking image-based sexual abuse legislation as well as in certain circumstances and jurisdictions expanding post-mortem publicity rights or utilizing the right to be forgotten around circulated images. The options that would be available globally and in other jurisdictions than the US remain under-explored.

HOW SHOULD WE PRIORITIZE SOLUTIONS?

Participants were asked to prioritize for “low-hanging fruit” solution areas; the most important solution areas; and the ones requiring most collaboration. This was not a prolonged process but an initial read from the participants.

Top low-hanging fruit solution areas: Mainstream authentication tools, invest in tools and approaches for authentication and provenance, support better collaboration with journalism and human rights, invest in new forms of forensics specific to synthetic media, platform-based approaches

Most important solution areas: Platform-based approaches, civic education, invest in new forms of video forensics e.g. anomaly detection

Solutions areas requiring most collaboration: Industry and AI-self regulation and ethics codes, building on existing efforts in open source journalism and human rights

Moderate prioritization: Adversarial attacks, existing and novel legal strategies, public policy responses

Little prioritization: Blockchain, communication between industry and AI-affected communities, integrate new forms of video forensics into tools, rigorous documentation, immutable authentication trails.

SOLUTION AREA: WHAT ARE THE OPTIONS FOR SCALING SOLUTIONS VIA PLATFORMS?

A group discussed potential options for how platforms could address synthetic media.

Key topline questions that were raised included the role of platforms in indicating truth vs. indicating technical manipulation of media and the “implied truth” problem with both of these approaches; definitions of synthetic media needing clarity to include both legitimate freedom of expression (e.g. satire) as well as the boundaries of computational photography as being considered synthetic media; and the desirability of shared vs platform-specific solutions. Another key element was the need to do this in end-to-end encrypted messaging apps and options for doing this.

As with the threat scenarios, the clarity on who/what we’re protecting and from what will help define product choices. For example, are fast debunks the priority and if so, how do we support this?

Potential solution areas discussed included:

- **Better collaboration between platforms** on identifying and labelling manipulated content for example by running a range of updated screening algorithms on incoming video and audio and sharing this information between platforms. A critical question here will be whether to make truth judgements or simply provide information on manipulation/synthesis identified. With the question of whether one shared approach or a diversity of approaches is most productive, it’s also important to recognize user and regulator concerns about what is included or excluded, and the existing problems around algorithmic content moderation. The legitimacy model for any intervention needs to be clear as well as communication to the public on the how/why of tools.

- **Coordination on sharing found-fakes, training data and detection methods** since the scale problem gets harder with ongoing iteration and motivated adversaries. Also, many automatic forensics and detection approaches will rely on adding examples to training data to generalize from existing GAN models.

- Providing **clear indices of what has been manipulated or altered**, particularly where this is not visible to the ”naked eye,” would “fool a human” or is outside the normal bounds of image manipulation. We recognized that by definition computational photography includes image manipulation, and most images are altered. The question here could be providing access to the manipulation data that companies have available (Facebook from Instagram, Apple with its camera, Snapchat with its filters) either to consumers or in a legally-restricted context.

- A **“slow news” approach to content** spread including the option to use approaches to detect event spikes or surges as a signal to prioritize content for review, and in some cases throttle acceleration. However, this will play into concerns re platform editorial roles.

- Incorporation of **capture-based authentication and provenance tracking information** – for example confirmation of a digital signature to author or of a signed hash showing underlying data not altered -- into platform as signals or visible information

- Use of **DRM-based approaches to protect content** within a stack and confirm no edits

- Use of **Content ID type approaches (as used for copyright owners to see use of their content on YouTube)** to track malicious synthetic images once they are in circulation across multiple platforms, and to **utilize shared hashes** of mal-used synthetic content

- **“Verified” image information** connected to authorship (noting concerns around anonymity and known authors)

- Incorporation of more **automated ways to identify splices from videos/images**, and use of an existing background in a synthetic video

- **Bounties that incentivize users** to find and report mal-uses

- Option within closed messaging systems to have an **installed “helper” button to “check this out”** to test signature of image/video you’re seeing against server for an exact or fuzzy match or to test whether synthetic media and to support community-based leaders to know how to use

- Identify how to **counter-message against digital wildfire** in closed messaging apps

- A **watermarking or adversarial perturbation system** on original videos from high profile sources that could prevent their use in synthetic media

- Rather than focus on arms race of detection focus on **getting users to be more on alert** including target particularly gullible users via a “gullibility detector” or provide options to choose level of notification on manipulation that a given user wants

- Increase **literacy in synthetic audio and video among users**, particularly around teachable moments

Productive next steps initially suggested:

- Identify what is **necessary to develop a sharing approach on research and new forms of attacks and detection**

- Prepare **research into right set of common standards** based on previous precedents

- Do a **small-scale research test on a platform** since as yet may not be economically viable to do a full-scale approach

- Identify **what it takes to get platforms to move on this** outside of scaring them with a real-world crisis? Who needs to be convinced? What prevalence is required? What level of bad consequences needs to be articulated? What is the potential economic impact?

- Support **building out training data sets** of fake images created by new synthetic media techniques

SOLUTION AREA: PROVENANCE, CHAIN OF CUSTODY, FILE INTEGRITY AND AUTHENTICITY ACROSS THE LIFECYCLE OF VIDEO, IMAGES AND AUDIO

A significant amount of the public discussion on solution approaches to synthetic media focuses on more authenticable original media and how this can be facilitated particularly by the blockchain.

A focus of discussion was on how to track provenance, chain of custody, file integrity and authenticity across the lifecycle of video, images and audio. Here the emphasis is not on detecting synthetic media but emphasizing the integrity of non-synthesized media. Participants noted that proving something is fake is probably easier than tracking chain of custody all way from source through distribution: proving fake has been easier than proving real in recent OSINT practices.

A group discussion concluded that this was more challenging than public discussion suggests but that there is potential to identify how authenticity infrastructure could complement other approaches.

Discussion on solution areas:

1. **Track an image from capture to display, ideally across platforms**
   - Provenance data could include rich metadata, camera signature, user signature; with original image and changes signed to the blockchain or another form of database.

- Tracking provenance has a range of hard challenges: e.g. the “analog hole” where someone takes a picture of a picture or a video of a video, to problems of managing revocation of keys used for identification within PKI, as well as security risks to individuals around either irrevocable notarization or requirements to provide identities linked to media.

- However, tracking provenance is one of the only ways you can prove that an image is real, so there is a value tradeoff.

- Strong source authentication is not part of the “soul” or architecture of Internet though that may be changing.

- Provenance and metadata are a problem of scale rather than technical feasibility. It’s definitely solvable in a closed ecosystem (e.g. crime scene photos) but for every phone in the world, from every manufacturer in the world it is a lot harder.

- What are unintended consequences of this approach? As tracking technologies emerge, then there will be susceptible to malicious use by less scrupulous government?

- Conversely, how do you handle content that emerges outside the system without making it subject to the “ratchet effect” in a way that will penalize content made by marginalized groups or by people who can’t risk providing authorship information or irrevocable ledgering of their content.

- Mechanisms would need to reflect users’ privacy requirements. For example, could you start with a zero-knowledge proof to show whether content meets certain authenticity standards?

2. **Track content as it enters “public sphere” of a platform** with an identifier associated with the content using similar approaches to ContentID or child exploitation images

    - This is more feasible but it doesn’t solve for the wild fire scenario where content does not reach public platforms. Could fuzzy matching of images work here?
    - Content could be tracked on basis of image integrity but also other measures, e.g. virality, social graph.

3. **Provide options for users to query questionable images within a closed messaging app**. Fact-checkers in Africa, Colombia and Mexico have provided one model where users can query a human on suspect stories. Another option would be to provide an installed “helper” button to “check this out” to test signature of image/video you’re seeing against server for an exact or fuzzy match or to test whether synthetic media. There are questions about scalability here

and reaching users who don't want to know if true or not. But with scaled virality at least some people will want to know.

4. **Shareable “playbooks for action” as well as support to critical community curators within social sites**, so that people who have trusted contact with communities can do verification, and then push it out as hub/spokes to other people. This could include more “neighborhood watch” oriented approaches to supporting individuals in communities to watch in their corner of the Internet for fake content outside of journalists and platforms. Technical solutions such as authentication techniques will be part of this fingerprinting on top of that.

All of these approaches must also address the same UI and UX problem of how you engage with users to encourage them not to share ‘once proved fake’ content further. Provenance can be more strongly enforced from devices and via platforms but it’s also even more critical to address how people understand what it means and in the bigger context of media flows to understand why people share this content.

Productive next steps initially suggested:

- Produce a **comprehensive analysis of the opportunities and problems with authentication-based approaches** to identify if the problems are solvable or feasible from technical point-of-view or business model. What is doable and what is not in creating authenticity infrastructure?

SOLUTION AREA: THE NEED FOR INCLUSION AND GREATER STAKEHOLDER ACCOUNTABILITY

The participants in this initial convening were primarily from Silicon Valley and the USA. Among suggestions on **follow-up engagement were to focus strongly on participation from the Global South, from high-risk communities in the U.S., Europe and East Asia, from policy and legal communities, as well as artists and provocateurs** with a different voice and solution approach on this.

SOLUTION AREA: HOW COULD TECHNOLOGISTS, PLATFORMS AND KEY INSTITUTIONAL ACTORS LIKE HUMAN RIGHTS DEFENDERS AND JOURNALIST COLLABORATE. WHAT WOULD BE USEFUL ACROSS SECTORS?

Building better collaboration between existing communities of practice in this area, those communities who will be most adversely affected, and key technologists and platforms is essential.

Among suggestions in this area for discussion was the need to find the optimal way to support intersection of tools and practices – i.e. what’s the optimal mechanism for connecting the community of practice around open-source verification with the ongoing research, experimentation and earning in the academic and company ML space?

Some suggestions included:

- **Improve access for journalists/verification specialists to the technical innovation/research**: For example, via short presentations to smaller groups of journalists who already lead in this space and understand online verification and discovery work.

- **Pro-active platform and technologist conversations with key journalists and civil society leaders** who have been good at pointing out problems on information flow in platform, and advocating for people impacted by vulnerabilities.

- Build shared understanding through a **collaborative workshop of ML and verification specialists testing out different proof approaches**. e.g. a simulated proof "war game" with ML and journalists, taking a set of videos, and figuring out which are real or not and why.

- **Prepare playbooks for 'prepping for synthetic media'** for key upcoming national and international events so that we are not left with "unexpected scenarios": upcoming elections in the U.S. and around the world, the U.S. Census etc. as well as other relevant scenarios on a global level including in: China, Russia, India.

SOLUTION AREA: HOW DO WE ENGAGE THE PUBLIC ON THESE ISSUES?

There is currently a dangerous hype cycle around the threat of deepfakes and other synthetic media. Public education and engagement is key - alongside engaging key journalists and human rights defenders who participate in verifying, debunking and utilizing information.

This conversation included:

- **Facilitate better communication on timelines and feasibility of tools** via more direct conversation between researchers and journalists so news coverage can engage in more informed speculation.

- **Engage public on this issue via deliberately created synthetic videos and public education around them**, noting the success and reach of the recent Jordan Peele simulation of President Obama in mainstreaming this discussion

- Identify ways to **make more visible the information disorder framework of mal/mis/dis-**information as it proved very helpful to convening participants.

- **Support trusted community figures at 1 or 2 hops from the source problem**, particularly to address digital wildfire in closed messaging contexts. In this context research what are the most useful tools to put in the hands of these and other users?

SOLUTION AREA: ADDITIONAL RESEARCH

A non-comprehensive survey of potential research questions surfaced in this convening, and an earlier discussion at the Information Disorder conference (June 2018) surfaced the need for further discussion on research questions and these initial incomplete research areas.

- What % of people believe images produced with new synthetic media approaches are real? (both ordinary people and famous people images)

- What is the motivation behind sharing various types of synthetic media content (from deepfakes to facial reenactments)?

- What does an academic literature review tell us about to better understand how people view, understand and analysis realistic faked visual media?

- What cues make people doubt a synthetically-produced image?

- What metadata when surfaced for viewers is most useful/persuasive to people?

- Within the journalistic community how much do they think they are currently being targeted by this media? And how do they perceive/gauge the threat?

RECAP OF RECOMMENDED NEXT STEPS

1. **Baseline research** and a **focused sprint** on the **optimal ways to track authenticity, integrity, provenance and digital edits of images, audio and video from capture to sharing to ongoing use**. Research should focus on a rights-protecting approach that a) maximizes how many people can access these tools, b) minimizes barriers to entry and potential suppression of free speech without compromising right to privacy and freedom of surveillance c) minimizes risk to vulnerable creators and custody-holders and balances these with d) potential feasibility of integrating these approaches in a broader context of platforms, social media and in search engines. This research needs to reflect platform, independent commercial and open-source activist efforts, consider use of blockchain and similar technologies, review precedents (e.g. spam and current anti-disinformation efforts) and identify pros and cons to different approaches as well as the unanticipated risks. WITNESS will lead on supporting this research and sprint.

2. **Detailed threat modelling around synthetic media mal-uses for particular key stakeholders (journalists, human rights defenders, others).** Create models based on actors, motivations and attack vectors, resulting in identification of tailored approaches relevant to specific stakeholders or issues/values at stake.

3. **Public and private dialogue on how platforms, social media sites and search engines design a shared approach and better coordinate around mal-uses of synthetic media**. Much like the public discussions around data use and content moderation, there is a role for third parties in civil society to serve as a public voice on pros/cons of various approaches, as well as to facilitate public discussion and serve as a neutral space for consensus-building. WITNESS will support this type of outcomes-oriented discussion.

4. **Platforms, search and social media companies should prioritize development of key tools** already identified in the OSINT human rights and journalism community as critical: particularly **reverse video search.** This is because many of the problems of synthetic media relate to existing challenges around verification and trust in visual media.

5. More shared learning **on how to detect synthetic media that brings together existing practices from manual and automatic forensics analysis with human rights, Open Source Intelligence (OSINT) and journalistic practitioners** - potentially via a **workshop where they test/learn each other's methods** and work out what to adopt and how to make techniques accessible. WITNESS and First Draft will engage on this.

6. **Prepare for the emergence of synthetic media in real-world situations** by working with journalists and human rights defenders to build **playbooks for**

upcoming risk scenarios so that no one can claim “we didn't see this coming” and so as to facilitate more understanding of technologies at stake. WITNESS and First Draft will collaborate on this.

7. Include **additional stakeholders** who were under-represented in the 6/11 convening and are critical voices either in an **additional meeting or in upcoming activities:**

   - Global South voices as well as marginalized communities in US and Europe
   - Policy and legal voices and national and international level
   - Artists and provocateurs

8. **Additional understanding of relevant research questions and lead research** to inform other strategies. First Draft will lead on additional research.

ACKNOWLEDGEMENTS AND THANKS

WITNESS particularly thanks the John D. and Catharine T. MacArthur Foundation for their support of its Tech + Advocacy Program and this convening.