WITNESS
SEE IT FILM IT CHANGE IT

TICKS OR IT DIDN'T HAPPEN*

CONFRONTING KEY DILEMMAS IN AUTHENTICITY INFRASTRUCTURE FOR MULTIMEDIA

December 2019

*Ticks is a British English word for checkmarks

For further information or if you are interested in participating in ongoing work in this area please contact WITNESS Program Director, Sam Gregory, sam@witness.org

Table of Contents

Why WITNESS prepared this report: A background

https://wit.to/Synthetic-Media-Deepfakes

WITNESS helps people use video and technology to protect and defend human rights.

A key element of this work is ensuring that people have the skills, tools and platforms to enable them to safely and effectively share trustworthy information that contributes to accountability and justice. Both our work on enhancing people's skills to document video as evidence and our work with the Guardian Project developing tools such as ProofMode are part of this effort. Our tech advocacy toward platforms such as Google and Facebook contributes to ensuring they serve the needs of human rights defenders and marginalized communities globally.

Within our Emerging Threats and Opportunities work, WITNESS is focused on proactive approaches to protecting and upholding marginalized voices, civic journalism, and human rights as emerging technologies such as AI intersect with disinformation, media manipulation, and rising authoritarianism.

The opportunity: In today's world, digital tools have the potential to increase civic engagement and participation – particularly for marginalized and vulnerable groups – enabling civic witnesses, journalists, and ordinary people to document abuse, speak truth to power, and protect and defend their rights.

The challenge: Bad actors are utilizing the same tools to spread misinformation, identify and silence dissenting voices, disrupt civil society and democracy, perpetuate hate speech, and put individual rights defenders and journalists at risk. AI-generated media in particular has the potential to amplify, expand, and alter existing problems around trust in information, verification of media, and weaponization of online spaces.

Our approach: WITNESS' Emerging Threats and Opportunities program advocates for strong, rights-respecting technology solutions and ensures that marginalized communities are central to critical discussions around the threats and opportunities related to emerging technologies.

New forms of AI-enabled media manipulation:

Over the past 18 months, WITNESS has led a focused initiative to to better understand what is needed to better prepare for potential threats from deepfakes and other synthetic media. WITNESS has proactively addressed the emerging threat of deepfakes and synthetic media, convening the first cross-disciplinary expert summit to identify solutions in June 2018; leading threat-modelling workshops with stakeholders; publishing analyses and surveys of potential solutions; and pushing the agenda in closed-door meetings with platforms as well as within the US Congress.

WITNESS is connecting researchers and journalists, as well as laying out the map for how human rights defenders and journalists alike can be prepared to respond to deepfakes and other forms of synthetic media manipulation (see our key recommendations below). We have particularly focused on ensuring that all approaches are grounded in existing realities of harms caused by misinformation and disinformation, particularly outside the Global North, and are responsive to what communities want.

WITNESS has emphasized learning from existing experience among journalists and activist communities that deal with verification, trust, and truth, as well as building better collaborations between stakeholders to respond to this issue. The stakeholders include key social media, video-sharing and search platforms, as well as the independent, academic and commercial technologists developing research and products in this area.

Why WITNESS prepared this report: A background

WITNESS engages in strategic discussions, planning, and advocacy with a range of actors including tech companies, core researchers, journalists, activists, and underrepresented communities, building on existing expertise to push forward timely, pragmatic solutions. As with all of WITNESS' work, we are particularly focused on including expertise and measures from a non-U.S./Western perspective, and with a focus on listening to journalists, disinformation experts, human rights defenders, and vulnerable communities in the Global South to avoid repeating mistakes that were made in earlier responses to disinformation crises. This includes a recent first national multi-disciplinary convening to discuss these issues and prioritize preferred solutions in Brazil. A comprehensive list of our recommendations and our reporting is available here, and have been covered in the Washington Post, Folha de São Paulo, CNN, and MIT Technology Review as well as many other outlets. Upcoming expert meetings will bring the prioritization of threats and solutions to Southern Africa and Asia.

WITNESS is also co-chairing the Partnership on AI's (PAI) Expert Group on Social and Societal Influence, which is focused on the challenges of AI and the media. As part of this, we co-hosted a convening with PAI and BBC in June 2019 about protecting public discourse from AI-generated mis/disinformation, and are participating in the PAI Steering Committee on Media Integrity.

TWELVE THINGS WE CAN DO NOW: WITNESS' RECOMMENDATIONS ON DEEPFAKES PRIORITIES

1 De-escalate rhetoric and recognize that this is an evolution, not a rupture of existing problems – and that our words create many of the harms we fear.

2 Recognize existing harms that manifest in gender-based violence and cyber bullying.

3 Demand responses reflect, and be shaped by, an inclusive approach, as well as by a shared human rights vision.

4 Identify threat models and desired solutions from a global perspective.

5 Promote cross-disciplinary and multiple solution approaches, building on existing expertise in misinformation, fact-checking, and OSINT.

6 Empower key frontline actors like media and civil liberties groups to better understand the threat, creating connective tissue between stakeholders and experts.

7 Identify appropriate coordination mechanisms between civil society, media, and technology platforms around the use of synthetic media.

8 Support research into how to communicate 'invisible-to-the-eye' video manipulation and simulation to the public.

9 Determine the desired responsibility for platforms and tool-makers, including in terms of authentication tools, manipulation detection tools, and content moderation based on what platforms find.

10 Prioritize shared detection systems and advocate that investment in detection matches investment in synthetic media creation approaches.

11 Shape debate on infrastructure choices and understand the pros and cons of who globally will be included, excluded, censored, and empowered by choices on authenticity or content moderation.

12 Promote ethical standards on usage in political and civil society campaigning.

TO LEARN MORE: LAB.WITNESS.ORG/PROJECTS/SYNTHETIC-MEDIA-AND-DEEP-FAKES

Executive summary

In late 2013, WITNESS and Guardian Project released a new app called InformaCam. It was designed to help verify videos and images by gathering metadata, tracking image integrity and digitally signing pieces of media taken with the app. Six years later, both the need for and awareness of apps like this have boomed, and InformaCam, now known as ProofMode, has been joined by a host of other apps and tools, often known as 'verified-at-capture' or 'controlled capture' tools. These tools are used by citizen journalists, human rights defenders and journalists to provide valuable signals that help verify the authenticity of their videos.

What we didn't predict back in 2013 is the increasing weaponization of online social media and calls of "fake news" creating an ever-increasing demand for these types of authentication tools. WITNESS has tracked these developments, and over the past eighteen months has focused on how to better prepare for new forms of misinformation and disinformation such as deepfakes. Now that synthetic media and deepfakes are becoming more common, and the debate on "solutions" is heating up, pressure is mounting for a technical fix to the deepfakes problem, and verified-at-capture technologies, like Guardian Project's 2013 authentication tools, are being heralded as one of the most viable solutions to help us regain our sense of visual trust. The idea is that if you cannot detect deepfakes, you can, instead, authenticate images, videos and audio recordings at their moment of capture.

When WITNESS hosted the first cross-disciplinary expert convening on responses to malicious deepfakes in June 2018, bringing together a range of key participants from technology, journalism, human rights, cybersecurity and AI research to identify risks and responses one of the top recommendations of those present was to do focused research to better frame trade-offs and dilemmas in this area as it started to rise to prominence. This report is a direct output of this expert recommendation..

"Baseline research … 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."

The consensus need to focus on this has been validated by the continuing growth of work in this area in the past year, both as a direct solution for deepfakes and also a response to broader 'information disorder' issues. Outside of deepfakes, a range of stakeholders are facing pressure to better validate their media, or see competitive advantage at pursuing options in this area - for example, with the recent launch of projects such as the News Provenance Project and the Content Authenticity Initiative from companies like the New York Times, Twitter and Adobe.

At WITNESS, we believe in the capacity of verified-at-capture tools and other tools for tracing authenticity and provenance over time to provide valuable validation of content in a time where challenges to trust are increasing. However, if these solutions are to be widely implemented, in the operating systems and hardware of devices, in social media platforms and within news organizations, then they have the potential to change the fabric of how people communicate, inform what media is trusted, and name who gets to decide. This report looks at the challenges, consequences, and dilemmas that might arise if this technology were to become a norm. What seems to be a quick, technical fix to a complex problem could inadvertently increase digital divides, and create a host of other difficult, complex problems.

Robert Chesney and Danielle Citron propose a similar idea for public figures who are nervous they will be harmed by hard-to-detect deepfakes They could use automatic alibi services to lifelog and record their daily activities to prove where and what they were doing at any given time. This approach to capturing additional data about the movements of public figures is not too far beyond adding additional data to each image, video, and audio recording. Both solutions have companies govern large amounts of personal data, encourage the adoption of a disbelief-by-default culture, and enable those with the most power and access to engage with these services.

Within the community of companies developing verified-at-capture tools and technologies, there is a new and growing commitment to the development of shared technical standards. As the concept of more thoroughly tracking provenance gains momentum, it is critical to understand what happens when providing clear provenance becomes an obligation, not a choice; when it becomes more than a signal of potential trust, and confirms actual trust in an information ecosystem. Any discussion of standards - technical or otherwise - must factor in consideration of these technical and socio-political contexts.

This report was written by interviewing many of the key companies working in this space, along with media forensic experts, lawyers, human rights practitioners and information scholars. After providing a brief explanation of how the technology works, this report focuses on 14 dilemmas that touch upon individual, technical and societal concerns around assessing and tracking the authenticity of multimedia. It focuses on the impact, opportunities, and challenges this technology holds for activists, human rights defenders and journalists, as well as the implications for society-at-large if verified-at-capture technology were to be introduced at a larger scale.

Dilemmas 1, 2 and 3 focus on different aspects of participation. Who can participate? How and what are the consequences of both opting in and opting out? If every piece of media is expected to have a tick signaling authenticity, what does it mean for those who cannot or do not want to generate one? Often under other forms of surveillance already, many human rights defenders and citizen journalists documenting abuses within authoritarian regimes might be further compromising their safety when they forfeit their privacy to use this technology so they can meet increased expectations to verify the content they are capturing.

Dilemma 4 looks at visual shortcuts. It could easily be imagined that color systems, tags such as “Disputed” or “Rated False,” or simply a tick (or ‘checkmark’ in American English) or a cross that indicates to the user what is “real” or “fake” could be implemented across social media platforms. However, there are various concerns with this approach. In this dilemma we explore issues such as verifying media that is “real” but used in misrepresentative contexts, and visual cues denoting verification that could be taken as signs of endorsement.

In Dilemma 5 we assess the implications that higher expectations of forensic proof might have on legal systems and access to justice in terms of resources, privacy and societal expectations. Visual material is highly-impactful when displayed in courts of law, and in most jurisdictions has a relatively low bar of admissibility in terms of questions around authenticity and verifiability. We must now ask how and if this will change with the actual and perceived increase of synthetic media and other new forms of video and audio falsification, and in what direction it will go, both in courtrooms used to photos and videos as evidence, as well as other judicial systems, where it is as yet novel.

Many of the concerns around deepfakes and synthetic media are focused on scale, in terms of creation and dissemination. It is through the lens of scale that we look at the next four dilemmas, and focus upon how to respond effectively to information disorder and mischaracterized content without amplifying, expanding or altering existing problems around trust in information and the weaponization of online spaces while preserving key values such as open internet, unrestricted access, freedom of expression and privacy.

Dilemma 6 focuses on issues that journalists and citizens who are documenting human rights violations and want to ensure the verifiability of their content might face in making this technology work for them. If those who want or are expected to verify their material face technical challenges that prevent them from using the authenticity tools assessed in this research, how will their material be treated in courts of law, by news outlets, and by social media platforms? This could create a system in which those relying on less sophisticated technology cannot produce images and videos that are accepted as "real."

The integration of verified-at-capture technology technology within social media platforms and news outlets is the focus of **Dilemmas 7 and 8.** Media and news outlets are facing pressure to authenticate media. The rising expectation is that they ensure that both the media they source from strangers as well as the media they produce and publish in-house is resilient to falsifications, and that they assess all user-generated content included in their reporting. Media outlets are concerned not only about their brand and reputation, but also about the larger societal impact this might have around trust in journalism, and the potentially disastrous consequences of reporting that is compromised by misinformation and disinformation. Related sub-challenges include liability concerns and the struggle of smaller platforms to keep up, both of which are explored within **Dilemma 7.**

Social media and messaging platforms are the key delivery mechanism for manipulated content, and provide a platform for those who want to consume and access such content. It is likely that -- due to both external pressures such as regulatory, legislative and liability concerns and changes, and internal pressures, such as maintaining and increasing levels of user engagement -- social media platforms, messaging apps and media outlets will introduce their own authenticity multimedia measures and apply more rigorous approaches to tracking provenance. These measures, if introduced, will immediately scale up the perceived need for authenticating multimedia, as well as raise awareness about the risks and harms that could accompany such changes. We discuss these measures in **Dilemma 8.**

The remainder of the dilemmas laid out in this paper focus specifically on how the technology works and, in some cases, doesn't work. **Dilemma 9** discusses one of the most critical aspects of this report: how collected data is stored. For many working on sensitive human rights issues, as well as those suspicious of platform surveillance and/or their own governments, how data is being treated, stored, deleted, and accessed, and how future changes will be accounted for, are key considerations. Alongside this, **Dilemma 9** considers a number of legal, regulatory and security threats and challenges.

The field of media forensics has only developed over the last two decades, and until recently, was still considered to be a niche field. Media forensics is not only a new field, but a disputed one. In **Dilemma 10,** we look at a number of complications with the proposed technology, and consider a number of known ways that bad actors could trick the authentication process. **Dilemma 11** then focuses on the ability of interested parties to review and, if necessary, appeal decisions and processes made by companies who have a financial interest in keeping these processes hidden. Many of the elements in media forensics are not easily readable to non-experts, and as with other processes, particularly those driven by algorithms, machine learning or AI-technologies, there is a critical need for people who are able to scrutinize them for errors and appeal poor decisions. If verified-at-capture technology is to be of use in helping interested members of the public make informed decisions on whether they can trust the media they are viewing, the data it collects must be easily comprehensible.

Both **Dilemma 12 and Dilemma 13** address the problem of devices that are not able to use verified-at-capture technologies. The basic underlying technology to create a system or set of systems for image and video authentication is still being developed, and **Dilemma 12** discusses how so far, it does not account for those with limited bandwidth and GPS, or for those using legacy devices,. **Dilemma 13** looks at those who are using jailbroken and rooted devices, and how this may hinder or even bar their ability to capture verifiable audio visual material. Many verified-at-capture tools use GPS location as an indicator that a piece of multimedia content is authentic, and need to account for the risk that jailbroken or rooted devices might have a GPS spoofer installed. The authenticity tools also rely on being able to assess the integrity of the device capturing the media, and cannot guarantee the integrity of a jailbroken or rooted device. If the expectation to use these tools in order to produce trustworthy content does scale globally, then it is essential that those who are using altered operating systems on their devices are not automatically discounted.

For the last dilemma, we look to blockchain technologies. Many companies interviewed for this report integrated blockchain technologies into their authentication tools to create a ledger of either the hash or the timestamp, or in some cases, both. People are being asked with increasing frequency to transfer their trust from human networks to technological networks, into tools built and implemented by computer scientists and mathematicians, one of which is blockchain. **Dilemma 14** explores how blockchain is being used to verify media, and whether it can be trusted.

This report is by no means inclusive of all the intricacies of verifying media at capture, and as this is a rapidly changing and growing field, it is likely that much of the technicalities discussed within this report will soon change. These technologies are offering options to better prove that a picture, video or audio recording has been taken in a particular location, at a particular time. This technology has the potential to be a tool that helps to create better quality information, better communication, greater trust and a healthier society in our shifting cultural and societal landscape. To do this, verified-at-capture technology needs to be developed in a way that it will be seen as a signal rather than the signal of trust, and that will allow people to choose to opt-in or out without prejudice, granting them the option to customize the tools based on their specific needs.

Introduction

In 1980, David Collingridge laid out the following double-bind quandary in his book, The Social Control of Technology, to describe efforts to influence or control the further development of technology:

1. An information problem: Impacts cannot be easily predicted until the technology is extensively developed and widely used.

2. A power problem: Control or change is difficult when the technology has become entrenched.

This quandary became known as the Collingridge Dilemma.

The Collingridge Dilemma goes to the heart of this report. There is a growing sense of urgency around developing technical solutions and infrastructures that can provide definitive answers to whether an image, audio recording or video is "real" or, if not, how it has been manipulated, re-purposed or edited since the moment of capture. Technologies of this type are currently being developed and used for a handful of different purposes, such as verifying insurance claims, authenticating the identity of users on dating sites, and adding additional veracity to newsworthy content captured by citizen journalists. There are indicators that these technologies are about to reach primetime, and if they are widely implemented into social media platforms and news organizations, they have the potential to change the fabric of how people can communicate, to inform what media is trusted, and even endow certain parties with the power to decide what is, or is not, authentic.

This report is designed to review some of the impacts that extensive use of this technology might have in order to potentially avoid Collingridge's second quandary. We reviewed the technologies being proposed and conducted 21 in-depth interviews with academics, entrepreneurs, companies, experts and practitioners who are researching, developing or using emerging, controlled-capture tools. This report reflects on platform, commercial and open-source activist efforts, considers the use of technologies such as blockchain, and identifies both the opportunities and challenges of different approaches, as well as the unanticipated risks of pursuing new approaches to image and video authentication and provenance. At the heart of this paper are the following questions: Is this the system that we want? And who is it designed for?

In this paper, we discuss and assess the individual, technical and societal concerns around assessing and tracking the authenticity of multimedia. This paper is designed to address a number of dilemmas associated with building and rolling out technical authenticity measures for multimedia content, and encourages a considerate, well thought-out and non-reactive approach. As the Collingridge Dilemma articulates, once these solutions are integrated within the hardware and operating systems of smartphones, social media platforms and news organizations, it becomes difficult, if not impossible, to roll back some of the decisions and the implications they have for society as a whole. To quote Collingridge, "When change is easy, the need for it cannot be foreseen; when the need for change is apparent, change has become expensive, difficult, and time-consuming."

If these infrastructures are to be prematurely implemented as a quick-fix response to a deepfakes problem that does not yet exist, then the current hype and concerns over deepfakes will have helped forge a future that would not have existed otherwise, in terms of legislative, regulatory or societal changes, introducing a whole host of other complex problems. As written by Robert Chesney and Danielle Citron, "Careful reflection is essential now, before either deepfakes or responsive services get too far ahead of us." With much of the work WITNESS carries out on deepfakes and synthetic media, the overriding message is one of "Prepare, don't panic." If this report were to build on this message, it would say **"Prepare, don't panic, but don't over prepare, either."**

Overview of current “controlled-capture” and “verified-at-capture” tools

In the wake of growing concerns over the spread of deepfakes and synthetic media, controlled-capture tools are not only being proposed by start-ups and companies as a solution to authenticating media, but are also being considered by regulatory bodies, social media platforms, and news outlets as a potential technical fix to a complicated problem. Although most of the tools discussed in this report are currently under development by start-ups and non-profits, and are based on software and apps, these technologies and related technologies of authenticity and verified provenance are also starting to be developed within news and media outlets and by social media companies themselves. Many of the players already established in this space also aspire to integrate their approach into phones and devices at the chip and sensor level, or possibly into the operating system, as well as into social media and audiovisual media-sharing platforms.

In this report we spoke to a variety of tool developers and companies at different stages of development of technologies in this area, from those with fully-fledged, market-ready tools to those that were still in their infancy stage, with tools still in the planning stage. These tools are described typically as “controlled-capture” or “verified-at-capture” tools. This next section provides an overview of the current technologies being proposed.

The general idea

Image, video and audio recordings each share similar characteristics - a moment of creation, the possibility of edits, and the capacity to be digitally reproduced and shared. In a nutshell, with controlled capture, an image, video or audio recording is cryptographically signed, geotagged, and timestamped. The idea behind verified capture is that in order to verify quickly, consistently and at scale, the applications on offer need to be present at the point of capture. Dozens of checks are performed automatically to make sure that all the data lines up and corroborates, and that whoever is recording the media isn’t attempting to fake the device location and time-stamp. The hash the media gets assigned is unique, and is based on the various elements of the pieces of data being generated. If you compare this hash with another image to see if it was an original image or not, then the test would be rerun, and if one element of the test (the time, date, location or pixelation of an image) has been changed, then the hash will not match.

While this is not a bulletproof approach, and is certainly vulnerable to sophisticated attacks, the ultimate goal for many of the tool developers is to add forensic information to this cybersecurity solution, looking at lighting, shadows, reflection, camera noise, and optical aberrations and to deploy increasing levels of computer vision and AI to detect such problems as someone taking a video of an existing video or a photo of an existing photo. This goal, according to media forensics expert Hany Farid[1], is years down the line due to the complexity of many of these techniques. In the sections below, we explore some of the elements common to many controlled-capture tools including hashing, signing, use of media forensics and access to the device camera.

Hashing and signing

Hashing and signing are cryptographic techniques.

Hashing is a cryptographic technique that involves applying a mathematical algorithm to produce a unique value that represents any set of bytes, such as a photo or video. We use different types of hashing techniques in our everyday interactions online. For example, when you enter your password into a website, this website doesn’t want to store your password on its servers so instead, it applies a calculation on it and converts your password into a unique set of characters, which it saves.

This technique can be similarly used for video, image and audio recordings. As written by James Gong “like any digital file, a video is communicated to computers in the form of character-based code, which means the source code of a video can be hashed. If I upload a video and store its hash on the blockchain, any subsequent changes to that video file will change the source code, and thus change the hash. Just as a website checks your password hash against the hash it has stored whenever you log in, a video site could check a video's upload hash against the original to see if it had been modified (if the original was known)." This means that the hash value of the original video can be checked against the value of the video being seen somewhere else, and if the video you are checking has a different hash number, then one of them has been edited. There are other approaches to hashing multimedia content, such as perceptual hashing (a method often used in copyright protection to detect near matches), which includes hashing either every frame of a video or regular intervals of frames, or hashing subsections of an image. These hashing techniques help detect manipulation, such as whether an image was cropped, and help identify and verify subsets of edited footage.[2]

Signing, in this context, uses the process of public key encryption, and uses keys that can be linked to a person, device or app, to authenticate who or which device created the file. For instance, to determine the device that a piece of media originated on, someone could compare the PGP identity and device ID. Then, to verify the data integrity, they could compare the PGP signatures with the SHA256 hash.

Media forensics ("looking for the fingerprints of a break-in")

Many of the approaches involved in verified-at-capture technologies such as hashing and signing derive from cybersecurity practices. However, there are a number of media forensic techniques, such as flat surface detection (used to detect someone taking a video or photo of an existing image), involved in some of the commercial offerings.

When discussing these tools and approaches with Dr. Matthew Stamm, he likened it to looking for fingerprints during a break-in.[3] Typically, these approaches are looking at where the multimedia signal comes from, and whether it has been processed or altered, to be able to answer questions about the source, processing history, and authenticity of the media content. This is done through signal processing, such as looking for fingerprints left by a particular type of camera, or fingerprints created by an image editor or image processing algorithm..

Due to the rapid development in machine learning and deep learning over the past five years (it used to take two to three years to develop one particular forgery detector), now, if machines are fed the right amount of training data, they can quickly learn how to detect many editing operations at once, leading to a more efficient and robust detection process. Most of the development within this timeline is in image and video, although it is increasingly trickling into audio. And as Dr. Matthew Stamm notes, many of the deep-learning techniques being developed can easily be transferred over to audio.

Cameras and microphones

Each smartphone camera has a lens, as well as a sensor that sees what the lens sees and turns it into digital data, along with software that takes the data and turns it into an image file. Additionally, smartphones might have multiple microphones.

Having access to both cameras and microphones, app developers can use an API provided by the operating system of a device to create a channel between the camera and microphone hardware and the external software. For example, WhatsApp, owned by Facebook, will use the API provided by the devices' operating system to allow WhatsApp to talk to the phone's camera and microphone.

There are two central issues with this system when it comes to authenticating media. The first is that the time lapse created when the operating system communicates with the app's API, even if it is only a nanosecond, is enough time for an adversary to insert fake content into the app. The second is that some of the apps discussed in this report only capture additional metadata when a user is taking a picture from within the app itself. This means that if you were to accidentally use your smartphone’s camera instead of the app, the image would not be authenticated.

Overview of the tools and applications reviewed

We spoke with seven companies and tool developers creating controlled-capture technology. Some of the tools have been in operation for a number of years, with others just entering the development phase. Below is a list of the tools assessed for this project, along with a short description taken from each tool’s website at the time of this report’s completion (October 2019). This list is by no means inclusive of all the tools being developed, but provides a good representative range.

Commercial offerings:

- Amber Video: “Amber Detect uses signal processing and artificial intelligence to identify maliciously-altered audio and video, such as that of deepfakes, and which is intended to sow disinformation and distrust. Detect is for customers who need to analyze the authenticity of videos, the source of which is unknown. Amber Authenticate fingerprints recordings at source and tracks their provenance using smart contracts, from capture through to playback, even when the video is cut and combined. Authenticate is for multistakeholder situations, such as with governments and parts of the private sector, and creates trustlessness so that no party has to trust each other (or Amber): parties can have unequivocal confidence with an immutable yet transparent blockchain record.”

- eWitness: “How can we protect truth in a world where creating fake media with AI techniques is child’s play? eWitness is a blockchain backed technology that creates islands of trust by establishing the origin and proving the integrity of media captured on smart-phones and cameras. With eWitness, seeing can once again be believing.”

- Serelay: “Serelay Trusted Media Capture enables any mobile device user to capture photos and videos which are inherently verifiable, and any third party that receives them to query authenticity of content and metadata quickly, conclusively and at scale.”

- Truepic: “Truepic is the leading photo and video verification platform. We aim to accelerate business, foster a healthy civil society, and push back against disinformation. We does this by bolstering the value of authentic photos and videos while leading the fight against deceptive ones.”[4]

Open-source apps:

- ProofMode: “ProofMode is a light, minimal “reboot” of our full-encrypted, verified secure camera app, CameraV. Our hope was to create a lightweight, almost invisible utility that runs all the time on your phone, and automatically embeds data in all photos and videos to serve as extra digital proof for authentication purposes. This data can be easily and widely shared through a “Share Proof” share action.”
- Tella: “Tella is a documentation app for Android. Specifically designed to protect users in repressive environments, it is used by activists, journalists, and civil society groups to document human rights violations, corruption, or electoral fraud. Tella encrypts and hides sensitive material on your device, and quickly deletes it in emergency situations; and groups and organizations can deploy it among their members to collect data for research, advocacy, or legal proceedings.

Specialized tools:

- eyeWitness to Atrocities: "eyeWitness seeks to bring to justice individuals who commit atrocities by providing human rights defenders, journalists, and ordinary citizens with a mobile app to capture much needed verifiable video and photos of these abuses. eyeWitness then becomes an ongoing advocate for the footage to promote accountability for those who commit the worst international crimes."

So, what are some of the differences between the tools?

Design

- All except TruePic are currently free tools, or offer a free version
- Most of the tools work only in English.
- Some tools are open-source, others are closed-source.
- Most are designed for smartphones and primarily work on Android operating systems. The more established companies have an iOS app.
- One of the tools focuses away from smartphones and more on integrating their technology into body-cams and cameras.

Capture

- All gather GPS location and available network signals such as WiFi, mobile, and IP addresses at point of capture.
- Most gather all available device sensor data such as altitude, the phone's set country and language preferences, and device information such as the make, model, unique device ID number, and screen size at point of capture.
- Most tools uses some kind of signing technology of media (such as PGP) at the time of capture.
- Most tools generate a SHA256 hash.
- Most use proprietary algorithms to automatically verify photos, videos and audio recordings.
- Most of the tools do not require mobile data or an internet connection to create digital signatures and gather sensor data.
- Most tools have no noticeable impact on battery life or performance.
- Some apps will still work on rooted or jailbroken devices, but others will disable the verification and the media will get written to the regular camera.
- Some tools work in the background and add extra data to media captured through the phone's camera. Other tools only work when the media is captured using the app itself.
- Some of the apps allow users to camouflage the app, picking a name and icon of their choice, such as a calculator, a weather app or a camera app icon.

Overview of current “controlled-capture” and “verified-at-capture” tools

Sharing and storage

- Some of the tools have a visual interface that offers users details such as the date, time and location of the image if the user selects to share it.
- Some companies and organizations store data only on their servers while others store data only on users’ devices.
- Some tools integrate blockchain technology to create a ledger of the hash, the timestamp, or in some cases, both.
- Two of the apps have an option that wipes all data in the app when a particular button is triggered, after which the app will uninstall itself.
- One of the tools enables users to choose how much specificity of location they want to share, such as within 10 meters, within the city, or no geolocation whatsoever.

Related dilemmas

Dilemma 5
Authenticity infrastructure will both help and hinder access to justice and trust in the legal system.

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

Dilemma 11
How to decode, understand, and appeal the information being given.

Dilemma 1:
Who might be included and excluded from participating?

“If it weren’t for the people, the god-damn people” said Finnerty, ”always getting tangled up in the machinery. If it weren’t for them, the world would be an engineer’s paradise.”
Kurt Vonnegut, Player Piano

The aspiration of many of the technical tools analyzed within this report is to become an integral part of the centralized communication infrastructure. Many of their catchy taglines suggest an aspiration to be the technical signal that can determine if an image can be trusted or not: “a truth layer for video;” “fake video solved;” “secure visual proof;” and “restoring visual trust for all.” However, through our interviews with the companies and civil society initiatives, it is clear they are not aiming to set an exclusionary standard in which a user would not be trusted unless they carry out all of the steps in authenticating a piece of media. Rather, they are working to develop a system where people can choose to add more signals to additionally verify their media. This intention, however, might not end up as reality if regulatory bodies or society at large come to expect these technical signals in order to trust visual and audio material.

If every piece of media is expected to have a tick signaling authenticity, what does it mean for those who cannot or do not want to generate one? As Zeynep Tufekci wrote in February 2018, “We need to make sure verification is a choice, not an obligation.”’

Verification technology can, as Kalev Leetaru wrote for Forbes, “offer a relatively simple and low-cost solution that could be readily integrated into existing digital capture workflows to authenticate video as being fully captured in the ‘real world’ to restore trust.” However, under the digital hood there are questions over whether this seemingly simple technology works as well as people hope it does. Those who use older smartphones, who don’t speak English, or have spotty internet or GPS might not be able to get this technology to work for them. Often the most critical videos, images and audio recordings, which are essential to authenticate, originate in places where circumstances are dangerous and stressful, connectivity is limited, or technology is older or must be hacked in order for it to work.

While these applications and tools can be both useful and necessary for those who want to authenticate their media, it is important to be cautious about implementing a technical structure that reinforces power dynamics that intertwine our ideas about who is considered truthful with who has access to technology. For instance, one of the tools in this industry, Amber Authenticate, works mainly with law enforcement in the United States to integrate their technology within the body cams of police officers. The footage captured by these officers gathers additional signals of trust and hashes the footage directly onto the blockchain.[5] However, this results in a police officer having access to technology that would authenticate their claims whereas a protester, for example, would not have access to the same technology, and would therefore be less able to authenticate the media they were collecting.
There are not just technical restraints to consider, but also environmental ones. Many of the places and instances that need these tools the most are also stressful and dangerous environments, where images and videos that push back against state-sponsored narratives will be less likely to be believed, and more likely to have doubt cast upon them. Those documenting abuses could forget to use a particular app they are “expected” to use, could be unaware that they are capturing something of significance, could use a burner phone, or might avoid using a verified-at-capture app at all because of the danger posed by being caught with such an app on their device.

The experiences of the human rights-focused app Tella illustrate this quandary. In the past, Tella allowed users to capture photos and videos using the default camera app as well as the Tella app. Both capture the additional metadata, but the default camera app stores this metadata unencrypted while the Tella app both stores and encrypts the metadata, hiding the image or video away from the user’s main camera roll. There is a tradeoff here: either the user does not capture any metadata if they accidentally use their default camera app, or they do capture it, but it is stored unencrypted, so is viewable by anyone who gains access to the device. Tella's partners expressed concern over these security risks, and in response, Tella disabled the functionality that allowed users to capture additional metadata when using their default camera app (while planning to implement a solution to address this).[6]

In order to avoid a disproportionate "ratchet effect," whereby the adoption of a new technology raises the bar both technically and practically for people who cannot afford such a risk, it is essential to consider how this technology will protect people in threatening and stressful situations, like dissidents, who may need to hide their identity or revoke information that later puts them in danger. If not, at-risk journalists and activists might not be able to fully participate in this new ecosystem. As Sam Gregory notes, these are the people who have spent decades being told they are "fake news" before that became a buzzword, and now run the risk of these "technologies of truth" being used to delegitimize their work.

QUESTIONS TO CONSIDER

- Tool developers and designers: How to design for those operating under stressful situations?
- Tool developers: How can these tools be used by those with limited access to WiFi and GPS, or those using legacy devices? How can those who are capturing media in these environments be involved and included in the design process?

Related dilemmas

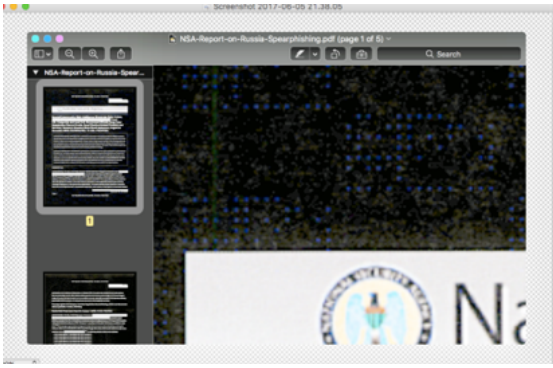
Dilemma 2
The tools being built could be used to surveil people..

Dilemma 3
Voices could be both chilled and enhanced.

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 2:
The tools being built could be used to surveil people

In 2017, Reality Winner was arrested on suspicion of leaking information concerning the potential Russian interference in the 2016 United States elections to the news outlet The Intercept. One piece of information that led to her arrest was the printer identification dots on the leaked documents. As documented by Kalev Leetaru, for decades now, "color printers have included each machine's unique signature in the form of steganographic Machine Identification Codes in every page they print. Also known as "yellow dots," most color laser printers print a hidden code atop every page that is invisible to the naked eye, but which can be readily decoded on behalf of law enforcement to identify the precise unique machine that produced the page."


Screenshot 2017-06-05 21.38.05
NSA-Report-on-Russia-Spearphishing.pdf (page 1 of 5)
Search
NSA-Report-on-Russia-Spear...
Na

A screenshot of the NSA report with the colors inverted by the blog Errata Sec

Just like these yellow dots identified the printer used to print these leaked documents, helping to identify Winner as the whistleblower, the digital signatures on images, video or audio recordings could be used by governments (repressive and otherwise), companies, law enforcement, or anyone with access to them to surveil, track and detain people of interest.

Many of the companies interviewed said that they are indifferent to identifying users, and only care about authenticating the media they are submitting. Whatever the intention of these companies, there are many actors who would want to know exactly who the photographer was and how they could use this technology to find them. Kalev Leetaru goes on to note that, "The digital signature on a secret video recording a politician accepting a bribe could be used to authenticate the video and prosecute the politician or just as easily could be used by a corrupt police official to trace its source and arrest or execute the videographer."

Verified-at-capture technology has the very real potential to compromise a user's privacy through unwarranted surveillance. As these technologies emerge and get mainstreamed, they could be susceptible to malicious use by governments, law enforcement and companies. In an article titled "The Imperfect Truth About Finding Facts in a World of Fakes," Zeynep Tufekci wrote, "Every verification method carries the threat of surveillance." Already under other forms of surveillance, human rights defenders and citizen journalists documenting abuses within authoritarian regimes might have to make a trade off of their privacy and safety when using this technology to meet increased expectations to verify the content they are capturing.

The level of surveillance can, in part, be mitigated based on the design of the technology. Some of the companies and tool developers interviewed have taken steps to reduce risks on behalf of their users. For instance, Truepic does not capture the device fingerprint as it does not add much value in terms of authenticating a piece of media, and could become a security risk as it could, in theory, be reverse engineered to identify the phone capturing the image.[7]

There are also questions as to what happens when users make mistakes by using the app imperfectly or inconsistently. For example, Truepic advises its at-risk users not to take a picture of their face, or the inside of their home. But what if this were to happen? It is then not only those capturing the media that might be at risk, but also those around them who were seen within the photo or video, but did not necessarily consent to being in a video with precise metadata recording their location and timestamping when they were there.

In a September 9, 2018 article by Kalev Leetaru entitled "Why digital signatures won't prevent deep fakes but will help repressive governments", Leetaru writes, "The problem is that not only would such signatures not actually accomplish the verification they purport to offer but they would actually become a surveillance state's dream in allowing security services to pinpoint the source of footage documenting corruption or police brutality."

Mitigation strategies deployed so far

A number of the companies interviewed for this report have been experimenting with different measures to mitigate surveillance risks and build pseudonymity into the technology's design, such as by not requiring sign-ins whatsoever, or not requiring a real email address. Recognizing that some of their users might be stopped by law enforcement or military groups and their devices searched, many developers and companies have integrated measures to cosmetically mask and/or hide their app.

As discussed above, Truepic advises that those with security concerns should not take photos inside their home, or of their immediate surroundings, in order to keep their identity anonymous. Users can also set the accuracy of the location they wish to share, from local/exact (within 65 meters), to general/city, or choose to provide no information at all. Truepic's default setting is private, meaning it is hidden and inaccessible until the user decides to share information. Be Heard, a project founded by SalamaTech, an initiative of SecDev Foundation, created a number of security-best practices for Arabic speakers using Truepic. They also advise deleting the app when crossing borders or checkpoints. Users can re-download and log into the application at anytime in the future and have their images restored.

Yemeni human rights advocates using the app Tella reported that they "were so scared and frightened that the technology and the metadata would have them identified and located that they turned that off completely."[8] As Tella focuses on modularity, they were able to adjust this feature and noted that "you are talking about contexts where people are completely fearful and we need to respect that and we need to give them the option to take that piece out."[9] Proofmode enables location capturing only if a user turns on the location capturing services. While this approach gives users a clear way to opt in and out of capturing location information, it also requires a certain level of education and clear communication of the options, and does not allow for mistakes as there is no safety net if a user forgets to turn location capturing on or off.[10]

While anyone can download the eyeWitness to Atrocities app and upload information directly to their servers, the organization behind the app often works directly with organizations and collectives with whom they have established a trusted relationship, and in most instances, with whom they have signed a written agreement. They generally partner with activist collectives, civil society organizations, NGOs and litigation groups, and in many cases travel to the location of the documentation groups that will be using their app. They not only provide the technology, but also help develop documentation plans and offer expertise in using photo and video for accountability purposes. eyeWitness to Atrocities is often introduced to users by a mutual contact, through a secure channel, and from there begins a process of building trust and forming partnerships that can take up to a year. This process may require multiple in-person meetings so eyeWitness understands what the group wants to document and why, the risks they might encounter, and their current capacity as well as the capacity they will need in order to meet their goals.[11] If needed, they may work with other partners who possess additional knowledge and litigation counsel.[12]

Amber takes a novel approach in order to enhance the privacy of both those being filmed and those using their technology. By breaking the recorded video down into segments, Amber allows the user to share only the segment or segments they wish to distribute, rather than the entire video. These segments will be assigned their own cryptographic hash as well as a cryptographic link to the parent recording. For instance, a CCTV camera is recording constantly during a 24-hour period; however, the footage of interest is only 15 minutes in the middle of the recording. Rather than sharing the entire 24 hours of footage, Amber can share just the relevant segment with interested stakeholders. As these segments are cryptographically hashed, the stakeholders will be able to confirm that, apart from a reduction in length, nothing else has been altered from the longer 24-hour parent recording.[13]

Moving forward, companies could build a data escrow function, where basic verification information could be publicly available, or provide public acknowledgement that verification data exists around a particular piece of media. Then, an interested party could request more information that the user can choose to comply with or not. The usefulness of this will depend on how much the users trust the companies offering the service and the security of their archives.

A key question to ask here is can, and how, will users be able to opt-out? Hany Farid likens it to a camera's flash function, "where you have the option, like turning a flash on and off, to securely record and not securely record."

The difference with flash is that users are able to visually understand what flash does, why it is important (in order to lighten up a dark scene), and can easily see if it is turned on or off (a bright light will shine or not); there are no great risks if you were to forget to turn the flash on or off. It is not the same for these verified media tools. It is hard to communicate what they do, easy to make mistakes when using them, and there could be serious repercussions from accidentally leaving verified capture turned off or on, resulting in media not having necessary additional data attached to it, or unintentionally adding sensitive information to circulating media.

Social media platforms and privacy

In a fairly contained environment of a smartphone app, these privacy-protecting measures can work to some success. But if these verification technologies become widely-integrated within social media platforms, they will become less successful as the amount of data associated with each piece of media expands to include social media profiles. There are real risks when disparate pieces of information combine and take on unintentional significance; this is known as the mosaic effect. If platforms add more metadata to the multimedia being posted on their platforms they will have to assess how much metadata to publish, and how much risk they are willing to take with people's privacy and safety. Like the advocates from Yemen, people in high-risk places are aware of the risks involved in sharing information and are reluctant to pick up tools that could become a safety concern.

Currently, social media platforms strip most metadata from the images and videos that are uploaded to their services. They have never been totally transparent as to why, but privacy concerns and liability risks are two decent guesses. We do know, however, that social media companies keep this metadata, as it has been requested by courts of law. If verification technology does get implemented into these platforms, then this potentially-identifying information is not only vulnerable to hacks, but also can be requested by subpoenas or court orders, leading governments or law enforcement to an activist's location. Aside from hacks and breaches, social media platforms will have to decide what information to provide to users about a "verified" image or video: too much and there will be privacy consequences; too little, and the confirmation provided will be a black box binary decision -- "trust" or "don't trust" – with no context as to the basis of the verification.

Dilemma 2:
The tools being built could be used to surveil people

QUESTIONS TO CONSIDER

- Companies: What level of identifying information should users be required to provide?
- Companies: What provisions are built into the design of the tools if someone uses them imperfectly or inconsistently?
- Companies: Can users decide to opt in or out? And is it clear how to do this?
- Companies: Are there modularity options built within the app for users who have privacy concerns and want to opt-out of particular metadata being captured?
- Companies: Who will have access to the data? Who would have third-party access to the data? Can users delete or port their data?
- Companies: What support concerning subpoena or legal threats do you offer your users?
- Companies: What level of corroborating verification data will you provide and what level of explanation as to how a verification confirmation is provided?

Related dilemmas

Dilemma 1
Who might be included and excluded from participating?

Dilemma 3
Voices could be both chilled and enhanced.

Dilemma 8
Social media platforms will introduce their own authenticity measures.

Dilemma 9
Data storage, access and ownership – who controls what?

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

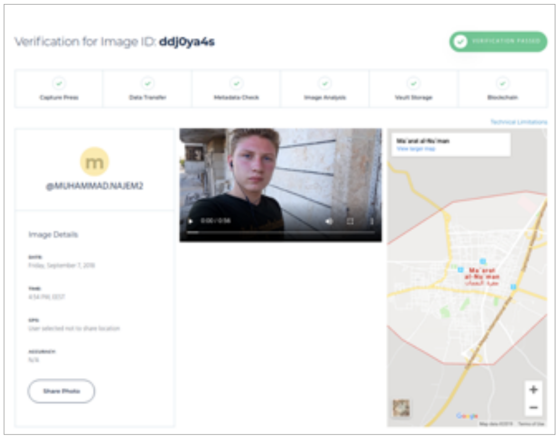
Dilemma 14
If people can no longer be trusted, can blockchain be?

Dilemma 3:
Voices could be both chilled and enhanced

Authoritarian leaders are mounting campaigns that target people's belief that they can tell true from false. "Fake news" is becoming an increasingly common reply to any piece of media or story that those in power dislike, or feel implicates them in wrong-doing, and new media-restrictive laws and regulations, as well as requirements for social media platforms and news organizations, are being proposed in democracies and authoritarian countries alike.

Verified capture is a valuable tool for critical media

In this environment, verified media capture technology can be used to improve the verifiability of important human rights or newsworthy media taken in dangerous and vulnerable situations by providing this essential and sensitive media with additional checks of authenticity. Fifteen-year-old Syrian Muhammad Najem used Truepic to capture a 56-second video in which he asks US President Trump, in English, to send international observers to protect civilians in Idlib province. In this message he states, "I know you hate fake news, that's why I recorded this authentic and verifiable video message for you." Muhammed, who has 25,000 followers on Twitter, then included a link on this tweet to the Truepic verification page for his video. Muhammad had received significant pushback on his Twitter account in the past, accusing him of lying about his location, so he began using Truepic to verify where and when he captured the videos he was posting. His video was then trusted and disseminated by mainstream media networks like Al Jazeera.



Verification for Image ID: ddj0ya4s

VERIFICATION PASSED

Capture Press ✓ | Data Transfer ✓ | Metadata Check ✓ | Image Analysis ✓ | Vault Storage ✓ | Blockchain ✓

Technical Limitations

@MUHAMMAD.NAJEM2

Image Details

DATE: Friday, September 7, 2018

TIME: 4:54 PM, EEST

GPS: User selected not to share location

ACCURACY: N/A

Share Photo

Screenshotfrom Truepic's verification page for a video recorded by Muhammad Najem

Another public example is Dimah Mahmoud, who encouraged those documenting any protest, rally or talk in Sudan in January 2019 to do so via Truepic, and included a video in Arabic on how to use the app. WITNESS, in its own work on citizen media as evidence has consistently seen how ordinary people and activists want to have options for sharing more trustworthy media (as well as options to be anonymous).

Those working to report alternative narratives against repressive governments are often looking to add more information to the media they are capturing to ensure that the evidence they are recording, often in risky situations, is trusted and admissible to courts of law. With the rise of open-source investigations and the increased amount of digital evidence being used in courts of law, those capturing the footage want to share content that can be held up to scrutiny.

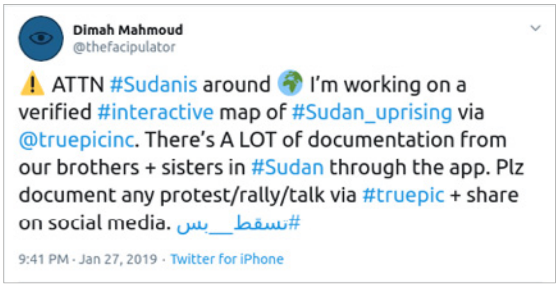
But those same activists and journalists have safety and privacy concerns

Journalists and activists have legitimate concerns in terms of these tools becoming another form of surveillance. For those who are already distrustful of their governments and of their phones, they may choose not to share the documentation they capture, or not capture important footage at all, if concerned that they might be tracked. When they use the tools inconsistently or imperfectly, then the tools themselves might be weaponized against them. Journalists and activists who do not use these apps for fear of surveillance may find their voices further dismissed.

Those who have their phones checked at military checkpoints or by law enforcement run the risk of these applications being found on their device. This invites questions by authorities concerning the videos and images that have been captured and for what purpose. Many of the human rights facing apps assessed for this report have taken steps to change the icon, create passwords and store data away from the central camera roll; however, if someone were to dig a little deeper, these apps could be relatively easily detected.

Dilemma 3: Voices could be both chilled and enhanced

Tweet by Dimah Mahmoud posted on January 27, 2019



Dimah Mahmoud
@thefacipulator

⚠️ ATTN #Sudanis around 🌍 I'm working on a verified #interactive map of #Sudan_uprising via @truepicinc. There's A LOT of documentation from our brothers + sisters in #Sudan through the app. Plz document any protest/rally/talk via #truepic + share on social media. #تسقط__بس

9:41 PM · Jan 27, 2019 · Twitter for iPhone

Tools risk being co-opted by social media and "fake news" regulation

There is an increasing global trend of decision makers regulating social media and "fake news." Under such circumstances, it is easy to find scenarios in which either governments require authenticated media, or platforms themselves default to such regulation. Similarly, increased regulation might be imposed as a requirement for media to be considered a "legitimate" news outlet. For instance, social media journalists and bloggers who reach a certain number of followers in Tanzania must register with the government and pay roughly two million Tanzanian shillings in registration and licensing fees. The Electronic and Postal Communications (Online Content) Regulations that came into effect in 2018 also forbid online content that is indecent, annoying or leads to public disorder. This new regulation has forced many content creators who cannot afford the fees offline.

Similar regulatory abuse and government compulsion around verifiable media could lead to the silencing of dissonant voices or views that could counter government narratives, especially if social media platforms or news outlets are regulated so they can only operate if they introduce this kind of technology. This could lead to companies pulling out of countries that impose such restrictions, preventing independent observers and monitors from operating in such places, and leaving the free press struggling worldwide.

Newsrooms face pressure to "not get it wrong"

It is not only those with privacy concerns that could suffer from this chilling effect, but also news organizations who may be reluctant to take risks or report rapidly on real events for, as Daniel Funke writes, "fear that the evidence of them will turn out to be fake." News organizations might report on something later proven to be faked, or become the target of a sting, or of a malicious actor trying to spread distrust or cause distress. In their paper Deep Fakes: A Looming Challenge for Privacy, Democracy and National Security, Robert Chesney and Danielle Keats Citron note that "Without a quick and reliable way to authenticate video and audio, the press may find it difficult to fulfill its ethical and moral obligation to spread truth." Conversely, to protect themselves, news organizations may choose not to use audiovisual material that does not have additional metadata attached.

Related dilemmas

Dilemma 2
The tools being built could be used to surveil people..

Dilemma 7
News outlets face pressure to authenticate media.

Dilemma 8
Social media platforms will introduce their own authenticity measures.

QUESTIONS TO CONSIDER

- Policy or decision makers: Consider how to ensure that both legal regulations and platform obligations around technical signals for authenticity do not weaponize these signals against journalists and news outlets.

Dilemma 4: Visual shortcuts: What happens if the ticks/checkmarks don't work, or if they work too well?

In the wake of concerns around the spread of fake news, a host of mobile applications and web browser extensions have been released, all designed to detect stories that contain falsehoods. In late 2017, the Reporters Lab published a report that found at least "45 fact-checking and falsehood-detecting apps and browser extensions available for download." Many of the apps analyzed in the Reporters Lab report use a color-coded system to denote the bias of each media source.

Having a simple color-coded system to indicate whether a piece of media can be trusted or not is not dissimilar to the outputs of multimedia authenticity apps. Implementing a traffic light system that indicates to the user what is "real" or "fake" could easily be imagined. In addition to color systems, tags such as "Disputed" or "Rated False," or simply a a tick/checkmark or a cross, can be used to indicate trust. While this approach could regain the general public's trust in visual and audio material, as internet users are increasingly coming into contact with such marks, various concerns with this approach arise.

Media that is "real" but misrepresented

In many misinformation and disinformation campaigns, authentic, untampered video is being used inauthentically. The most common falsified content that WITNESS encounters in its work running video verification and curation projects based on online content is genuine, but deliberately mis-contextualized video which is recycled and used in new contexts. In a report published in May 2019 by DEMOS, they found that "focusing on the distinction between true and false content misses that true facts can be presented in ways which are misleading, or in a context where they will be misinterpreted." For example, an image of a flooded McDonald's went viral after hurricane Sandy hit the US in 2012. This was a real picture, but not one taken during the hurricane; rather, it came from a 2009 art installation called "Flooded McDonald's" that was miscaptioned and misrepresented.

As Kalev Leetaru writes, "Adding digital signatures to cell phone cameras would do nothing to address this common source of false videographic narrative, since the issue is not whether the footage is real or fake, but rather whether the footage captures the entire situation and whether the description assigned to it represents what the video actually depicts." This misrepresentative media would be flagged as "real," which it is, giving media consumers a false sense of confidence that it can be trusted instead of encouraging them to investigate the media further and check if it's being recontextualised.

Authentication visual shortcuts could be seen as an endorsement

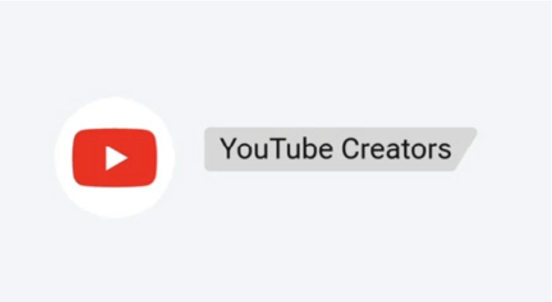
Social media platforms have well-documented issues with their verification programs. Both YouTube and Twitter face a similar issue: the ticks they assign particular accounts are seen as an endorsement rather than a confirmation of a user's identity. Twitter paused their verification program in 2017 after controversially giving a known white nationalist in the US a blue verification checkmark. In a recent blog post by YouTube, they announced that they were changing the design of their program. "Currently, verified channels have a checkmark next to their channel name. Through our research, we found that viewers often associated the checkmark with an endorsement of content, not identity. To reduce

Image of Flooded McDonalds by art collective Superflex confusion about what being verified means, we’re introducing a new look that helps distinguish the official channel of the creator, celebrity or brand it represents.” This new look is a gray banner as seen below. It seems likely that social media companies will come across the same issue with authentication checkmarks, where ticks, marks, or gray banners are seen as an endorsement of the content rather than simply a mark indicating that a piece of media has been authenticated.

Ticks could discourage a skeptical mindset

There are two known processes that occur in the brain in terms of making decisions: System 1 and System 2 thinking. These fact-checking apps take advantage of System 1’s fast, often unconscious, decision-making capabilities, as they require little attention. System 2 thinking, in comparison, is slower and controlled, requiring closer scrutiny. You can read more about how System 1 and System 2 thinking works in the context of fake news in this article by Diego Arguedas Ortiz titled “Could this be the cure for fake news?”.

Many of the proposed solutions being discussed in this report tap into System 1 thinking, whereas System 2 thinking would likely be much more effective in these scenarios. This is echoed by various scientific reports that when it comes to debunking information, it’s “useful to get the audience in a skeptical mind-set.” These visual shortcuts could become an unnecessary crutch rather than a true aid to someone’s thinking. The requirement that multimedia be captured by using these authenticity apps, or by using a particular technology, could also work towards fostering a culture of disbelief, where human networks of trust are replaced by technical understandings of trust.



YouTube’s replacement to a checkmark is a gray banner

The spillover effect

In a 2017 study by Dartmouth University, which attempted to measure the effectiveness of general warnings and fact-check tags, it was found that while false headlines were perceived as less accurate when they were accompanied by a warning tag, no matter the respondents’ political views, and that exposure to a general warning decreased belief in the accuracy of true headlines. This led the study to conclude that there is “the need for further research into how to most effectively counter fake news without distorting belief in true information.”

This study observed a spillover effect where warnings about false news primed people to distrust any articles they saw on social media. This can be interpreted as a “tainted truth” effect, where those who are warned about the influence of misinformation overcorrect for this threat and identify fewer true items than those who were not warned.

Once there is a tick on an image that states a piece of media is “fake,” then what? How would you discourage people from sharing this content? It is critical to address how people understand what it means to have a “fake” tag, and, as identified by the WITNESS/Partnership on AI and BBC convening place this in the broader context of how and why people share known false information as well as prioritize better research on how to communicate falsehood.

Dilemma 4:

Visual shortcuts: What happens if the ticks/checkmarks don't work, or if they work too well?

If these multimedia authenticity checks are indeed rolled out to be part of messaging apps and large social media platforms, these quick indicators of trust could do the exact opposite of what they were designed to do, and end up both discouraging investigation and scrutiny and bringing accurate information into question. Integrating software to decide whether something is "real" or not is just one method among a variety of approaches, and one that leaves much more to be done in terms of regaining levels of trust beyond technical indicators.

QUESTIONS TO CONSIDER

- Companies: Consider the effects of the warning labels you choose to add to the media.
- Companies and tool developers: How to manage mis-contextualized or mis-framed "real" media?

Related dilemmas

Dilemma 3
Voices could be both chilled and enhanced.

Dilemma 8
Social media platforms will introduce their own authenticity measures.

Dilemma 11
How to decode, understand, and appeal the information being given.

Dilemma 5:

Visual shortcuts: Authenticity infrastructure will both help and hinder access to justice and trust in legal systems

> **"If we become unable to discover the truth or even to define it, the practice of law, and functional society, will look much different than they do today."**
> Jonathan M. Mraunac, The Future Of Authenticating Audio And Video Evidence, July 2018

On September 21, 2018, a military tribunal in the Democratic Republic of the Congo (DRC), condemned two commanders for crimes against humanity. As part of this trial, videos were used as evidence for the first time ever in the DRC. TRIAL International and WITNESS worked together to train lawyers working on the case on the best practices of capturing and preserving video, and worked with eyeWitness to Atrocities to use their eyeWitness app to verify that the footage being captured had not been tampered with.

"During the investigatory missions, information was collected with the eyeWitness app to strengthen the evidentiary value of the footage presented in court," says Wendy Betts, Project Director at eyeWitness to Atrocities. "The app allows photos and videos to be captured with information that can firstly verify when and where the footage was taken, and secondly can confirm that the footage was not altered. The transmission protocols and secure server system set up by eyeWitness creates a chain of custody that allows this information to be presented in court." "When the footage was shown, the atmosphere in the hearing chamber switched dramatically," testified Guy Mushiata, DRC human rights coordinator for TRIAL International. "Images are a powerful tool to convey the crimes' brutality and the level of violence the victims have suffered."

This is an example of how additional layers of verification and authentication can aid in access to justice. While in some jurisdictions video has been used in trials for years, in others, like the DRC, it is new, and judicial systems are working on developing rules for admissibility.

The rules dictating procedure around the introduction of digital evidence in courts of law, in the US at least, are, as Hany Farid describes, "surprisingly lax."[14] Depending on the type of legal system in place, experts are hired by either the courts or the opposing sides, and must present their arguments, or answer questions posed by the court. In other jurisdictions, the approach differs. In Egypt, Tara Vassefi writes for a report on Video as Evidence in the Middle East and North Africa, "There is a technical committee under the umbrella of the State Television Network, also referred to as the Radio and Television Union (Eittihad El'iidhaeat w Elttlifizyun Almisri), which is responsible for authentication and verification of video evidence. Generally, if the opposing party contests the use/content/authenticity of video evidence, the judge refers the video to this committee for expert evaluation." Tara goes on to note that though Egypt has a committee that is responsible for authentication and verification of video evidence, this issue is still subject to the discretion of the judge, which means it can be used as a political tool to impact the outcome of a case. In Tunisia, video evidence is used at the discretion of the judge, who can "simply deem the video evidence inadmissible and rely on other forms of evidence."

Visual material is highly-impactful when displayed in courts of law, and in most jurisdictions, has a relatively low bar of admissibility in terms of questions around authenticity and verifiability. Tara Vassefi, in her research into video as evidence in the US, argues that "electronic or digital evidence is currently 'rarely the subject of legitimate authenticity dispute,' meaning that legitimate authenticity disputes are not coming to the fore as lawyers and judges are using digital evidence." We must now ask in what direction the actual and perceived increase of synthetic media and other forms of video and audio falsification will lead courts in determining the authenticity of media, both in courtrooms accustomed to using photos and videos as evidence, as well as in courtrooms where such media is a novelty.

Deepfakes, or synthetic media, could be introduced in courts of law. And if they are not introduced, their existence means that anyone could stand in a court of law and plausibly deny the media being presented. Other challenges to using video and images as legal evidence may be more mundane, and grounded in increasing skepticism around image integrity. In the high-profile divorce case between the actors Johnny Depp and Amber Rudd, Depp is claiming that the photos showing domestic abuse of Rudd are fake. “‘Depp denies the claims and images purported to show damage done to their property in an alleged fight contain ‘no metadata’ to confirm when they were taken,’ said his lawyer Adam Waldman.”

Depp’s claim that these images were fake, as they “contain no metadata,” could be an increasingly common argument. In this dilemma, we look at the implications of higher expectations of forensic proof in terms of resources, privacy and societal expectations.

The implications of higher expectations of forensic proof

If societies’ awareness and concerns around synthetic media grow, or if more authentication and verification is required, then there is the potential of what Associate Director of Surveillance and Cybersecurity at Stanford’s Center for Internet and Society, Riana Pfefferkorn, refers to as a new flavor of an old threat, a “reverse CSI effect.”[15]

The CSI effect refers to the popular crime drama television series, Crime Scene Investigation (CSI). Mark A. Godsey describes this effect in his article “She Blinded Me with Science: Wrongful Convictions and the ‘Reverse CSI Effect.’” “Jurors today, the theory goes, have become spoiled as a result of the proliferation of these ‘high-tech’ forensic shows, and now unrealistically expect conclusive scientific proof of guilt before they will convict.” In this context, the CSI effect might refer to juries and judges expecting advanced forensic evidence in order to trust any multimedia content being presented to them.

The reverse CSI effect often applies to cases where too much weight is placed on forensic evidence produced by the prosecution, resulting in convictions in cases where the defendant probably should have been acquitted, but were not because the juries or judges may have been “blinded by science.” Mark A. Godsey goes on to summarize these effects, “To say it another way, in cases where no forensic evidence is introduced by the prosecution, jurors give the lack of such forensic evidence too much weight to the prosecution’s unfair detriment (the “CSI Effect”), and in cases where forensic evidence IS produced by the prosecution, these same jurors give too much weight to this evidence to the defendant’s unfair detriment (the “Reverse CSI Effect”).”

Riana Pfefferkorn suggested that in the future, there might be the need to create ethical guidance for those in the legal field that takes into account their role in spreading doubt around multimedia content and the implications this might have for society at large. ”Attorneys have a special responsibility to uphold civic institutions and uphold the knowability of truth, rather than undermining big-picture interests of democracy in service of the short-sighted goal of winning a case.”[16] While acknowledging that attorneys have a duty to zealously represent their clients, Pfefferkorn cautioned that this duty to the client need not and should not blind attorneys to other considerations, including larger societal interests.[17]

This ethical guidance is essential to account for those who cannot or choose not to use the verified-at-capture technology, and who might find themselves at a disadvantage entering the courtroom, as their credibility may be questioned.

Who will take the stand?

If jurors and judges come to expect higher levels of admissibility of multimedia content, then witnesses could be asked to verify, corroborate, or authenticate multimedia evidence more frequently. As Riana Pfefferkorn notes,[18] this strategy may be a risky one for individuals who have credibility problems and are less likely to be believed to begin with, or for criminal defendants who have the right not to testify in the United States.

Raquel Vazquez Llorente, Senior Legal Advisor at eyeWitness to Atrocities notes that traditionally, authenticity tends to be proven by the videographer,

Image by TRIAL International of the South Kivu military tribunal

photographer, or a witness.[19] But with their set-up, the app has been designed not to collect any identifiable information. It also allows for the person submitting the video or image to be anonymous by choosing not to provide their name or contact details. In the majority of cases, they work with organizations who have individuals recording footage under an alias, so eyeWitness to Atrocities themselves would not even know who the videographer or photographer was. The person who pressed record is irrelevant to their system's architecture. A parallel can be made between this set up and the set up of closed-circuit television (CCTV) cameras, where the individual who set up the system or pressed play isn't relevant for authentication purposes. eyeWitness to Atrocities can provide authentication information if and when necessary, such as proving that hash values match, and submitting original files containing chains of custody to a court of law. They can also provide an affidavit or authentication certificate to court, and testify if needed.

How will people afford it?

Under Article 7 of the UDHR, every individual is entitled to equality before the law without any discrimination. When considering who has access to justice, resource limitations that bar many from accessing justice and thus obstruct equality have to be considered. Requiring additional verification for multimedia could lead to a more protracted and expensive legal process. Forensic experts with the ability to detect synthetic media are not only rare, but expensive. Media forensics is highly technical in nature. Those working in it have backgrounds in signal processing, math and engineering, and the field itself has complex pathways to entry. This not only excludes those without access to experts and resources, but is also unscalable, especially if these experts are based in particular geographic areas and are completely non-existent in others.

Truepic's initiative around the amendments made to the US Federal Rule of Evidence 902 broaches this potential imbalance. These amendments were designed to simplify the legal process and reduce the costs associated with using electronically-stored information as evidence. Truepic's machine-generated process and verification page is designed to meet new evidentiary standards by streamlining authentication for those with limited legal resources. Currently, however, as Tara Vassefi notes, "Lawyers are either unaware or not taking advantage of these amendments and only a handful of cases have drawn on the new rules."

While a forensic expert appearing in person is no doubt more persuasive than a certificate, this option is useful for smaller cases. However, as mentioned above, the forensic assessment of media is complex to truly understand and query, and if juries and judges fail to understand forensic indicators of trust, and rely on companies to authenticate media, this could be problematic, especially if these techniques and threats are not properly vetted.

Who will decide if something is authentic or not?

Quoting from Britt Paris and Joan Donovan's recent publication, Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence, "This history shows that evidence does not speak for itself. That is, it is not simply the representational fidelity of media that causes it to function as evidence. Instead, media requires social work for it to be considered as evidence. It requires people to be designated as expert interpreters. It requires explicit negotiations around who has the status to interpret media as evidence. And because evidence serves such a large role in society – it justifies incarcerations, wars, and laws – economically, politically, and socially powerful actors are invested in controlling those expert interpretations. Put another way, new media technologies do not inherently change how evidence works in society. What they do is provide new opportunities for the negotiation of expertise, and therefore power."

The field of forensic science, encompassing such processes as DNA testing, fingerprinting, pattern recognition and media forensics, has been primarily developed under the financial endorsement of law enforcement and, according to the US National Research Council, has "been primarily vetted by the legal system rather than being subjected to scientific scrutiny and empirical testing." In the US, most of the publicly-funded labs are associated with law enforcement. In a paper entitled "A call for more science in forensic science" published in 2018, the US National Research Council argued that the field is in "dire need of deep and meaningful attention from the broader scientific community. Without such guidance, forensic science and law enforcement risk withholding justice from both defendants and crime victims." In 2009, the US National Academy of Science published a report that was highly critical of many forensic practices being used to administer justice. This report led to the establishment of the National Commision for Forensic Science in 2013. The NCFS was terminated by the Department of Justice in 2017.

The amendment to the US Federal Rule of Evidence 902 discussed above accepts evidence that is "self-authenticating," meaning that evidence can be admitted without needing a witness to testify in person to its authenticity. When providing these certificates, courts of law will undoubtedly ask questions about how the technology works, how the code works, and whether it can be trusted in deciding someone's guilt or innocence. While eyeWitness to Atrocities provides an affidavit to the court detailing how their technology works, other companies may see courts of law as a stage upon which the code behind these technologies will be critically assessed. Lawyer Jonathan M. Mraunac in his July 26, 2018 article, "The Future of Authenticating Audio and Video Evidence," comments that, "In this context, the expert witness for audio and video authentication would no longer be an acoustical engineer or visual image expert but a software engineer, cryptographer and/or a representative from the hardware manufacturer."

Dilemma 5:

Visual shortcuts: Authenticity infrastructure will both help and hinder access to justice and trust in legal systems

Hiring a forensic expert is expensive, and as the need for these skills grows, the demand could lead to a cottage industry forming. Before this influx, criteria should be drawn to clarify what qualifies a media forensic expert to testify in court. For Dr. Matthew Stamm,[20] one of the worst things in terms of the development of this growing field would be for someone to go into court, misrepresent what is possible, and be more confident than they should be about the validity of an image, video or audio recording. This would undermine public and legal confidence in the field and undercut the experts.

QUESTIONS TO CONSIDER

- Lawyers: Assess your role in spreading distrust by challenging all audiovisual media as potentially falsified. Consider the impact of the spillover effect or the tainted truth effect where the distrust of multimedia in general could lead to trustworthy content being distrusted by default.

- Media forensic experts and judges: What expertise around these new tools will be required of media forensics experts? How can a stronger and broader field of experts be built around media forensics and authentication?

Related dilemmas

Dilemma 1
Who might be included and excluded from participating?

Dilemma 3
Voices could be both chilled and enhanced.

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 9
Data storage, access and ownership – who controls what?

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

Dilemma 11
How to decode, understand, and appeal the information being given.

Dilemma 14
If people can no longer be trusted, can blockchain be?

Dilemma 6:
Technical restraints might stop these tools from working in places they are needed the most.

Much of the most important human rights and civic media content being captured by smartphones is in places where connectivity both to the internet and GPS might be limited or non-existent, and is being captured by those who might not have access to the latest hardware. While the number of smartphones is growing, internet connectivity becomes an issue in certain countries for those in rural areas, with some areas having none at all. The tools assessed for this research have four technical requirements in order to work properly. They need to be newer smartphone models, and they need to have GPS, internet connectivity, and operating systems that can be verified. This means that if you are documenting an incident or sharing news from a location with no GPS, zero internet connectivity, or on a jailbroken or rooted legacy smartphone, these authentication tools will not function as effectively as possible.

Journalists and citizens who are documenting human rights violations want to ensure that the content being captured is as verifiable as possible. However, the places that need these tools the most might be the hardest places for them to work. If those who want or are expected to verify their material face technical challenges that prevent them from using the authenticity tools assessed in this research, how will their material be treated in courts of law, by news outlets, and by social media platforms? This could create a system in which those reliant on less-sophisticated technology cannot produce images and videos that qualify as "real."

Many of these technologies are being developed with little testing, by developers who have had relatively limited interactions with those who are marginalised or in positions of vulnerability, resulting in very little collaboration in terms of design processes. The basic underlying technology needed to establish a system or set of systems for image and video authentication is still being developed, and so far, does not account for those with limited bandwidth and GPS, or those using legacy and/or jailbroken devices.

There are a number of considerations and trade-offs that are often only clarified when designing collaboratively with those that experience these technical constraints. For instance, the predecessor of ProofMode, an app called Camera V, gathered significantly more sensor data, but this impacted the battery life and data of the device using it, so a trade-off was made to reduce the amount of data being collected in order to preserve both the battery life and the data usage required to use the app.[21] As this example demonstrates, it is not only internet connectivity and GPS that needs to be considered but also battery life, the download size of the app, and concerns over which devices and operating systems the apps are compatible with. In instances of capturing media in areas with no internet or GPS, Proofmode still generates metadata about the device itself, and enables the signing of the media and metadata.

Another issue that could hinder a citizen journalist or member of the public when recording a video is the length of the video. Many of the apps have limitations in terms of how long recorded videos could be, ranging from 20 seconds to 15 minutes, which may not be a realistic representation of the length of time people film across a range of settings.

Within the community of companies developing verified-at-capture tools and technologies, there is a new and growing commitment to the development of shared technical standards. It is hoped that these standards will address a number of considerations that are discussed below and within the other dilemmas in this paper that exclude participation.

The following considerations should be taken into account when designing applications that can be used by those living and working in environments with technical and societal constraints:

Device type and impact on device

- Battery life: How much battery does this app take up? How quickly does it drain battery life from a user's phone?
- Download size: How heavy is the app that people have to download? Over poor data connections, a large download could take a while, and also consume too much data while doing so. Does the app itself require large amounts of storage space on the device?

Dilemma 6:
Technical restraints might stop these tools from working in places they are needed the most

- OS: What operating systems do these apps work on?

- Device compatibility: What devices will these apps work on?

Connectivity

- Do users have to be online in order for these apps to work?

- Do users have to have access to GPS in order for these apps to work?

Related dilemmas

Dilemma 1
Who might be included and excluded from participating?

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

Dilemma 11
How to decode, understand, and appeal the information being given.

Dilemma 12
Those using older or niche hardware might be left behind.

Dilemma 13
Jailbroken devices will not be able to capture verifiable audio visual material.

Potential workarounds

- Truepic is examining the possibility of enabling offline Controlled Capture capability.[22]

- ProofMode has a functionality that enables those working and living in areas with low levels of internet connectivity to send out their content's metadata through a text message, which helps to establish that the media existed at a certain time, even if there is no internet available.

- ProofMode does not require mobile data or an internet connection to create digital signatures or to gather most of the sensor readings.

- Serelay is designed to transfer a data load less than 10kb per media item, which means that those with limited WiFi or mobile data can use the tool and their mobile data allowance won't be affected.

- One potential solution explored by the team at eWitness, a project out of the City University of New York, was crowdsourcing identifiers from other phones within short range of the phone collecting the media. This would require a sufficient number of people to be willing and able to say that they detected a particular person's phone in a particular area, and to provide some assurance that they were within a certain distance of them.

QUESTIONS TO CONSIDER

- Tool developers and designers: How do you design for those operating under stressful situations?

- Tool developers: How can these tools be used by those with limited access to WiFi, GPS and those on legacy devices? How can those who are capturing media in these environments be involved and included in the design process?

Dilemma 7
News outlets face pressure to authenticate media

Media and news outlets are facing pressure to authenticate media, both in terms of ensuring media they are sourcing from stringers, ingesting into their archives and producing and publishing in-house is resilient to falsifications, and in terms of assessing user-generated content they include in their reporting.

In addition to core questions of how authentication is done by media outlets, a number of additional potential challenges arise from the provision of increasing provenance signals in a media environmen. These include liability concerns, how paywalls around the sites that can afford authentication limit access to this information, and the struggle of smaller platforms to keep up, and are explored further below.

Authenticating media being produced in-house

Large outlets are growing increasingly concerned that the same technology being used to create fake videos of Barack Obama will be used to produce convincing falsified audio and visual material that tricks viewers into thinking the material is coming from their station, newsroom, or reporters. Media outlets are concerned not only about their brand and reputation, but also about the larger societal impact this might have around trust in journalism, and the potentially disastrous consequences fake news reporting can have worldwide. Most major media organizations have already experienced their content being edited to purposefully mislead, or their logos being appropriated on incorrect or falsified information. As Daniel Funke notes, "Without a quick and reliable way to authenticate video and audio, the press may find it difficult to fulfill its ethical and moral obligation to spread truth."

Kalev Leetaru reported on this challenge in an article entitled, "What happens when television news gets the deep fake treatment?'" He asked his readers to imagine a livestream video from a social media account that looks like an official account from a major news outlet, and the reporter announces breaking news of a financial crash or a major terrorist attack. Before the news station can issue a warning that this was not their channel but a fake version, there could already be huge repercussions. This station's reputation would be called into question, and trust in the media and journalism would be reduced. It brings into question how society will function if this threat came to fruition. As noted by the Aspen Institute, "Democracy cannot function without flourishing and trusted media, or an informed citizenry."

Some media outlets are working on systems to authenticate their own outputs using similar approaches to those outlined in this report (hashing, signing, tracking metadata, using distributed ledger technologies), including the News Provenance Project, which involves the New York Times among others, and the recently announced Content Authenticity Initiative, of which the New York Times is also a partner alongside Adobe and Twitter. These address the first challenge outlined above, of manipulation of existing content or broadcasts.

In addition, the tools assessed in this report are relevant to the verification and tracking of content produced in the field. For instance, verification tools can be used to verify the location and timeline of journalists who take media on their smartphones. While most of the tools assessed for this report work only on smartphones, Amber Authenticate technologies can be integrated into the hardware of media cameras, too. In the wake of a video filmed in the US White House that depicted what looked like a reporter dismissively pushing away a female intern's hand during a press conference, but was later deemed to be manipulated, Amber Authenticate wrote an article speculating on what might have been different if their technology had been integrated within the press camera's hardware.

Authenticating media found online or crowdsourced

Media outlets are not just concerned about issues of authenticity and provenance for material produced in-house. A growing number of news outlets are creating technical systems to verify multimedia content sent directly from citizen journalists and members of the public who may have witnessed a newsworthy event, as well as content sourced directly from social media platforms.

As part of this information gathering, news organizations themselves might become targets of a sting. The sting might involve fake videos sent in, or posted online, by malicious actors seeking to spread distrust and incite mayhem, who are hoping to be picked up by a news agency, or it may arise from less malicious intent, when manipulated content is shared online by people who do not know it is falsified.

Various international news and media outlets are preparing for an influx of fake media, and a number of news organizations are already using verified-at-capture technology. Al Jazeera has used two videos authenticated by Truepic to tell stories of civilian harm in Syria. An independent sub-Reddit has integrated Truepic technology into their "Ask Me Anything" Q&A's to verify that the questions are being answered by the person advertised. Without a doubt, others are also experimenting and designing technical systems to further authenticate media.

When large news outlets invest in this technology and infrastructure, securing additional signals of authenticity and prioritizing media that can be forensically authenticated by way of such additional signals, they hope to streamline their user-generated content verification processes, and assure their readers that what they are seeing has not been maliciously altered, and their published content can be trusted.

Regulatory and liability concerns

If policy makers begin to implement regulatory pressures around the detection of "fake" media, liability issues could arise that require more stringent verification methods to be brought in to media houses. Of course this could lead to misuse, particularly in contexts and countries where authoritarian governments might want to use these regulations to limit a free press.

News organizations could then become forced to certify output as "real" or not, unless it is for entertainment purposes, and face fines or damages if they were to "get it wrong." This has the potential to drive providers out of the market due to pricey technical systems, the cost of potential damages, and expensive insurance coverage, which could result in only the dominant news organizations being able to keep up.

Access to verified content

The current economic models of journalism are precarious, with many outlets adopting paywalls and subscription models. If only the largest international press agencies and outlets can afford to integrate multimedia authenticity technology, and these same companies increase the use of paywalls, then only those who can afford subscriptions or who are willing to subscribe will gain access to additionally-verified media, and those who cannot will not.

Capacity for forensics in news organizations

Research conducted by WITNESS assessing preparedness for deepfakes highlighted gaps in tools and expertise in media forensics. Media forensics expertise is hard to come by due to the field's relatively recent development and the complex training and academic paths it requires to become proficient in media forensics. Most of the labs working on media forensics in the US are funded in part by either the government or law enforcement agencies, and the career paths for those graduating in this field are often either academic careers funded by or in direct conjunction with law enforcement, or positions with national agencies. There is a genuine concern that by the time properly-trained media forensic analysts are needed, there will not be enough to go around. On the other hand, this could lead to more funding to improve the forensic training and capacity-building outside of a law enforcement context, and this improved funding on the side of media could also trickle down to civil society.

Right now, though, news outlets may not have the resources or capacity to attract skilled people in this area, and may be forced to rely upon and trust external companies to carry-out this kind of analysis for them.

Smaller platforms and news outlets

There are many questions concerning how smaller entities, such as underfunded media companies, platforms run by civil society and individuals, will be able to handle the increased liability risks of publishing false material as well as the technical challenges that will arise if their viewership seeks increased checks for the content they are publishing. One potential way to mitigate this is to open-source the technology being used and developed by larger media houses so that smaller news agencies can integrate it and use it on their platforms.

For example, SecureDrop, the open-source software used to share and accept documents securely, primarily from whistleblowers, is integrated within large international news outlets such as the New York Times, and, due to the fact that it is managed by Freedom of the Press Foundation in open-source form, can also be used by smaller, civil society organizations. Had this software project been developed entirely in silos, by large media outlets, it could potentially discourage whistleblowers from sending documents to their news outlets of choice. Noting the struggles that come with maintaining and sustaining open-source software, perhaps newsrooms could actively strive to provide a stable environment for further development of such technology.

Authenticity assurances are only as good as the archives that hold the material

Archives being maintained by newsrooms, which are likely to hold footage recorded by staff as well as eyewitness contributions, could be susceptible to hacks, breaches or leaks. If a newsroom's archive is penetrated and fake media is added to it, then confidence in the whole authenticity infrastructure will be lost. Thus, the public's confidence in the effectiveness of introducing authenticity technology to newsrooms is only as good as the organization's archive.

Manipulation of context is as important as manipulation of content

As noted elsewhere, most current misinformation and disinformation is not falsified content, but falsified context. This is equally true of news items, which are frequently misleadingly framed or incorrectly circulated as relating to an incorrect date or event. Content authentication approaches partially address this by providing the ability to clarify the date and verify the integrity of original media items, but they do not address the broader challenge of handling an image or a video that has been correctly signalled as "unmanipulated" in a technical sense, but has been utilized in a deceptive context.

Related dilemmas

Dilemma 2
The tools being built could be used to surveil people.

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 8
Social media platforms will introduce their own authenticity measures.

Dilemma 9
Data storage, access and ownership – who controls what?

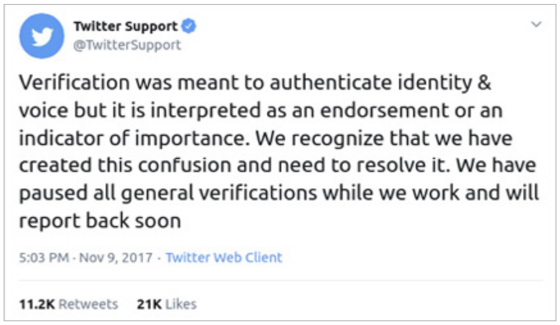
QUESTIONS TO CONSIDER

- Policy makers: What are the implications of regulations made in this area?
- Media and news outlets: Is there a way you can develop technology in the open that other smaller platforms can also use?
- Media and news outlets: What are the applicability of tools you develop for maintaining the integrity of your publishing for a broader universe of people generating media and looking for trust signals?
- Media and news outlets: How secure are the archives you are using to store this data?

Dilemma 8:
Social media platforms will introduce their own authenticity measures

Twitter accounts have always been easy to fake. A photo of the person you want to impersonate and a similar username are all you need to convincingly tweet as someone else. In 2009, Twitter began verifying accounts by adding what would soon become a status symbol: a blue tick beside a user's account. This verification process was first reserved for celebrities and public figures, then in 2016 was opened up to anyone who wanted to apply.

It became increasingly unclear how decisions were being made, who was being awarded verified checks, and in what order the company was dealing with requests. In November 2017, Twitter verified Jason Kessler, a white supremacist and organizer of the Charlottesville protest that resulted in the death of Heather Heyer. This verification compounded issues concerning how to manage the public's perception of a blue checkmark as an endorsement rather than a simple verification of a user's identity. A few days after verifying Kessler, Twitter abruptly suspended its verification program, stating that it "had been broken for a while." Despite this announcement, Karissa Bell reported in April 2019 that Twitter had continued to verify thousands of people.


Twitter Support
@TwitterSupport

Verification was meant to authenticate identity & voice but it is interpreted as an endorsement or an indicator of importance. We recognize that we have created this confusion and need to resolve it. We have paused all general verifications while we work and will report back soon

5:03 PM - Nov 9, 2017 - Twitter Web Client

11.2K Retweets 21K Likes

Screenshot from the Twitter Verified feed

YouTube themselves reported a similar issue with their verification program. On September 20, 2019, they published a blog stating "Yesterday, we announced changes to the verification badge. The idea behind this update was to protect creators from impersonation and address user confusion. Every year, we receive tens of thousands of complaints from creators about impersonation. Also, nearly a third of YouTube users told us that they misunderstood the badge's meaning, associating it with *endorsement of content*, and not an indicator of *identity*. While rolling out improvements to this program, we completely missed the mark. We're sorry for the frustration that this caused and we have a few updates to share."

Another attempt by YouTube to combat misinformation is the recent feature the company is calling "information cues," designed to add context to videos showing potential conspiracies and stop the spread of false information. This feature regularly gets it wrong. For instance, livestream videos of the Notre Dame fire were accompanied by text with information about the US 9/11 attacks. This created the false impression that the fire was linked to terrorism.

Authentication as endorsement

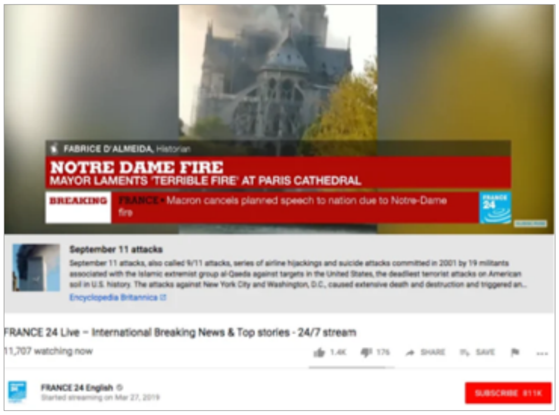
Social media platforms are the key delivery mechanism for manipulated content, and provide a platform for those who want to consume and access such content. The examples above demonstrate the complications of giving additional context on both users and content in an automated way, and how decisions made behind closed doors can lead to confusion and, ultimately, an overall decrease in trust of these authentication systems.

Giorgio Patrini, Founder, CEO and Chief Scientist at Deeptrace, argues that as we look ahead to new forms of video and audio falsification, it all boils down to scale.[23] Without the ability to scale the technology, the creation of fake content would be too expensive and time consuming; without delivery at scale, it becomes difficult to reach a wider audience, and without consumption at scale, only a fringe audience would be affected by the fake material. In this context, as the commodification of deepfake tools scales up the amount of content, the internet and social media platforms will provide the infrastructure for the scaling up of delivery and consumption.

In the past year, Facebook has been accused of fanning the flames of the genocide in Myanmar and spreading false information that led to lynchings in India. It is not only Facebook at fault: misinformation spread on WhatsApp groups helped Jair Bolsonaro win the 2018 Brazilian Presidential elections. Videos in particular are especially potent due to the convincing nature of visual media and its ability to impact non-literate groups. As Ian Bogost writes, "Video can capture narratives that people take as truths, offering evidence that feels incontrovertible."

There are trade-offs to be made as we figure out how to respond effectively to information disorder and mischaracterized content without amplifying, expanding or altering existing problems around trust in information, verification of audiovisual materials and the weaponization of online spaces. Preserving key values of the open internet, unrestricted access, freedom of expression and privacy throughout this process is crucial.

Screenshot from YouTube


FABRICE D'ALMEIDA, Historian

NOTRE DAME FIRE
MAYOR LAMENTS 'TERRIBLE FIRE' AT PARIS CATHEDRAL

BREAKING FRANCE • Macron cancels planned speech to nation due to Notre-Dame fire

September 11 attacks

September 11 attacks, also called 9/11 attacks, series of airline hijackings and suicide attacks committed in 2001 by 19 militants associated with the Islamic extremist group al-Qaeda against targets in the United States, the deadliest terrorist attacks on American soil in U.S. history. The attacks against New York City and Washington, D.C., caused extensive death and destruction and triggered an...

Encyclopedia Britannica

FRANCE 24 Live – International Breaking News & Top stories - 24/7 stream

11,707 watching now

FRANCE 24 English

Started streaming on Mar 27, 2019

Legislation and regulatory changes, as well as public pressure, could force platforms to act

It is likely that -- due to external pressures such as regulatory, legislative and liability concerns and changes, as well as internal pressures such as maintaining and increasing levels of user engagement -- social media platforms, messaging apps and media outlets will introduce their own authenticity multimedia measures.

These measures, if introduced, will immediately scale up the perceived need for authenticating multimedia, as well as the perceived risks and harms that could accompany these measures.

Many of the major social media platforms and messaging applications, such as Facebook, Twitter, YouTube and WhatsApp, are likely exploring and potentially investing in integrating technology into their platforms that tracks or assesses the authenticity or provenance of the images, videos and audio being posted. Publicly, Twitter has announced its participation in the Content Authenticity Initiative alongside Adobe and the New York Times. As reported by Karen Hao, both Truepic and Serelay are in "early talks with social-media companies to explore the possibility of a partnership, and Serelay is also part of a new Facebook accelerator program called LDN_LAB."

If these companies were to move forward with integrating this technology, then they have a number of challenges as well as opportunities to grapple with. In their paper "Deep Fakes: A looming challenge for privacy, democracy and national security," Robert Chesney and Danielle Citron discuss the technical approaches to authenticating content and conclude that they will only have limited use until ". . .one system (or a cluster of them) becomes ubiquitous and effective enough for dominant platforms to incorporate them into their content-screening systems–and, indeed, to make use of them mandatory for posting.". Similar observations have emerged in WITNESS's convening work in this area.

A common approach companies have adopted in the past in order to preempt government regulation is that of self-regulation. For instance, Hollywood studios created the Motion Picture Association of America (M.P.A.A) film rating system in order to convince Congress that a government regulatory body, which they worried would only work to censor and ban films based on moral grounds, was not required. Similarly to the technology being discussed in this report, the M.P.A.A.'s opaque film rating systems has been criticized for not revealing how or why certain decisions are made. In October 2018, to celebrate their 50 year anniversary, the M.P.AA. released a 46 page document detailing some of their criteria.

Dilemma 8:
Social media platforms will introduce their own authenticity measures

Whether or not these platforms want the role of being an arbitrator of what is a “real” image or video could be irrelevant. They could be forced into the role, not only by their users, but also by regulators, due to concerns over liability and defamation risks, and legislation that bans promoting and publishing maliciously altered multimedia. A recent bill passed in Australia includes penalties of up to $525,000 for corporations found guilty of spreading deepfakes, while recent legislation proposed in the US proposes obligatory watermarking for synthetic media and deepfakes. As noted by Robert Chesney and Danielle Citron, while Section 230 of the US Communications Decency Act “immunizes from (most) liability the entities best situated to minimize damage efficiently: the platforms,” pressure is already mounting on platforms and messaging apps with concerns over visual material contributing to lynchings, terrorism and encouraging genocidal acts. Facebook is currently integrating technology to spot deepfakes and, like Google, Microsoft, Adobe and others, investing in research on how to detect deepfakes.

Confirming authenticity favors larger incumbents and silences some voices

In the future, governments could silence dissonant voices through regulatory abuse and government compulsion that requires social media platforms and news outlets to use this kind of technology in order to operate. This could lead social media companies to pull out of countries that impose such regulations, bar independent observers and monitors from operating and cause the free press to struggle.

When the burden of confirming authenticity and determining whether a piece of multimedia can, ultimately, be trusted is placed on a company, it will create barriers for lower-resourced, smaller companies and decentralized platforms. They might not have the capacity, either technical or human, to take on this role, leading to voices being silenced and increasing the chance that the content they do post on smaller, decentralised platforms will be open to liability risks. People using these sites could may find their content is less trusted. To avoid these issues, , regulatory bodies can explore providing exceptions for smaller and decentralized platforms.

However, if large tech companies running social media platforms and messaging apps are able to offer their users advanced forensic information on not only their multimedia but also the multimedia of others, this would further lock users into using their services or risking their content being perceived as false. This dependency would impact vulnerable creators, and many may feel forced to upload their content to companies that will retain the rights to the material as well as grant third-party access to it.

Integration within the native cameras of social media apps would partially address the problem of providing access to those using legacy technology, since they will be able to authenticate their media taken through these apps. However, this could result in sensitive material being posted on social media accounts in order to verify it, leading to a larger mass of content available to law enforcement through court orders, subpoenas and/or hacking and breaching attacks. Many indicators of authenticity in multimedia files provide identifying information, like time and location. This is hugely valuable identifying information to both law enforcement, authoritarian governments, and malicious actors.

If, in the future, search results and news feeds are curated based on whether a piece of media has been authenticated or not, this would leave those who have opt-ed out of the process for whatever reason down-ranked or excluded. Such funneling of information would create a situation where individuals are exposed only to what is deemed the most “authentic” media, thus narrowing the range and diversity of information and ideas available to the public.

How would platforms roll these approaches out, and what challenges would they face?

There are particular forensic challenges when it comes to social media platforms, such as how to deal with the varied transformations, like filters or text over images, and the emojis and stickers that can be added to images and videos. And then, in turn, comes the question of how to communicate the nature of these transformations to users. Will this amount of alteration, common on a platform like Instagram, mean that images might be flagged as "fake?". Although deployed for malicious purposes, generally these alterations are not only harmless, but also done for entertainment and to increase the communication value of social media.

How will the distinction be made between a "fake" image or video and an image or video with few enough transformations to be considered "real," and can the technology keep up with this distinction-making? Most of the "false" content being circulated has not been altered, and it is often "real" content that has been recaptioned, mis-contextualized and repurposed. Then the question might be one of tracking provenance rather than focusing on the authenticity of the image itself.

How platforms determine which technology to use is another question. They will likely bring on academics and experts, and acquire and partner with companies already working on this technology, as they decide what to use in their closed, proprietary systems. Based on the relatively small number of media forensic experts that work in the field and have the necessary skills to assess the authenticity and provenance of multimedia, this will probably lead to the commercialization of media forensic expertise, leading to these experts being in short supply and possibly only available to the highest bidders.

Due to the proprietary nature of the companies and the technology they build, when the technology is actually rolled out, it will be essential for companies to allow for independent audits of the technology. Metcalfe's Law, which states that the value of a network provides competitive advantage, is at play here because although these social media platforms might not have the best authenticity technology, they have the most users, which is difficult to replicate, so the power of the network will drive out other competition. If this technology is used to tell 2.2 billion monthly users which images and videos they can trust or distrust, then this technology has to work, and it has to be auditable.

This underlying presence of a human network could also be an advantage here. For instance, a user can see who posted the video in question, determine if it is someone they already know, and quickly assess if they have provided trustworthy content in the past. And in cases where this human trust is not present, then the user can decide to take extra cautionary measures.

Dilemma 8:
Social media platforms will introduce their own authenticity measures

QUESTIONS TO CONSIDER

- Social media companies: How to create a system that allows users to opt-out without having their content automatically distrusted or down-ranked?
- Social media companies: How can users appeal or challenge decisions being made?
- Policy makers: What are the implications of regulations imposed in this area? What are the implications of regulations in the US on platforms used globally?

Related dilemmas

Dilemma 2
The tools being built could be used to surveil people.

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 7
News outlets face pressure to authenticate media.

Dilemma 9
Data storage, access and ownership - who controls what?

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

Dilemma 11
How to decode, understand, and appeal the information being given.

Dilemma 9:
Data storage, access and ownership – who controls what?

For many working on sensitive human rights issues, how data is being treated stored, deleted, accessed, and how future changes will be accounted for, are all key considerations when using authenticity technology. Additionally, there are a number of legal, regulatory and security threats and challenges that must be taken into account. As with other dilemmas, we look at this question through the lens of Collingridge's Dilemma, considering the crucial implications of these technologies when at a global scale in addition to reviewing the current commercial offerings of verified-at-capture technologies.

Who controls the data? Portability, deletion and third party access

In March 2019, it was made public that the once-leading social network MySpace had lost millions of photos, videos and songs that had been uploaded to the site before 2015. While some people may feel relieved that embarrassing photos, uploaded onto MySpace accounts to which they could no longer remember their log-ins, were gone, for many others, this data loss meant that precious audio-visual material could never be recovered. MySpace released a press statement that stated the loss was due to an error that occurred when the company was migrating the data between servers. Some commentators have asked whether MySpace did, in fact, lose this data through a server migration mishap, or if this was used as an excuse to avoid spending the resources necessary to transfer and host millions of files.

ome | Browse | Search | Invite | Film | Mail | Blog | Favorites | Forum | Groups | Events | Videos | Music | Comedy
View My: Pics | Videos
Contacting Tom
Send Message
Forward to Friend
Add to Friends
Add to Favorites
Tom's Videos

Tom

":-)"

Male
31 years old
Santa Monica,
CALIFORNIA
United States

Last Login:
5/25/2007

Tom's Latest Blog Entry [Subscribe to this Blog]

sherwood record coming out tuesday! (view more)

profile trackers and other bogus websites (view more)

phished accounts .. unblock! (view more)

get the quicktime update (view more)

bulletins that you didn't post? (view more)

Screenshot from Tom from Myspace's profile

With any system there is a risk of data loss, and in situations where users are not storing data on their devices and instead rely on companies' cloud storage, or have deleted audio-visual content from their devices due to security concerns and are relying on an app to host their content, it is important to ask what measures and policies are in place to mitigate data loss as well as what procedures have been developed if data is deleted or lost.

Perhaps more importantly than the risk of a company deleting the content is the question of how individuals control the life of their data. For those recording and uploading verified-at-capture content that potentially contains a rich range of personal and contextual data, as a single file or in aggregate via the mosaic effect, the ability to delete media with confidence is essential, especially if their personal safety situation changes.

Article 20 of the General Data Protection Regulation in the EU has enshrined within it the right to data portability. Data portability was introduced to allow users to transfer personal data from one service into another service. In Article 20, they make the distinction that the data's owner should be able to transmit their personal data themselves when technically feasible. In reality, many of the companies being interviewed will be working with data protection experts in order to be GDPR-compliant, and it is unknown how such regulations will affect data portability.

The companies we spoke to store data differently; some store it on their servers while others store it only on the user's device and never retain a copy. The companies we interviewed who do store data on their servers do not currently track user data, behavior data or location data, and do not monetize this data as part of their business models. However, this is not to say that these, or other, companies might not take a different approach in the future. Depending on the private agreements between companies, data shared with one company can be accessed, harvested, mined, and processed by another, or multiple, companies.

Social media companies, for example, would not use external technology that requires granting public access to users' private data, so instead, they will likely create their own in-house technology. With the very public and damaging Cambridge Analytic scandal in 2018, social media companies are extremely aware of the risks of granting third-party companies access to their data, especially the risks and reputational damage that arise if such access is misused or data mishandled.

Changes over time must be managed with care

On August 10, 2016 the United States Defense Advanced Research Projects Agency (DARPA) announced they have a contract for the source code, data and non-exclusive license of Izitru, a patented image authentication technology. Izitru was developed by Fourandsix Technologies in 2011 and used commercially for two years before it was licensed by DARPAs, and codenamed project MediFor, for $500,000. Those creating it were ahead of their time, and at the time of development, there wasn't the sense of urgency that there is now.[24]

As commercial companies like Izitru shutting down their business and selling their technology to a government program is a good example of the type of changes that can occur over time.. In this case, as the number of Izitru users was declining, and the company gave them a two-month warning about the change of service so they could delete all their data.[25]

With any software or hardware technology, there are future-proofing challenges for both for-profit and non-profit tool developers. These challenges center around commercial and economic sustainability, the addition and deprecation of services offered, the service being discontinued because it was bought off or folded, or because a non-profit's funding model changed, the company or non-profit lacks the ability to adapt to changing threat models in time, or the company needs to change their business models regarding the treatment of data.

Garbage in, garbage out

This expression, accredited to computer programmer Wilf Hey, is typically used to describe a bad output as a result of a bad input. In the case of verified-at-capture technology, if the archives that house the data being generated are insecure and untrustworthy (garbage), then the multimedia being verified will also be untrustworthy (garbage). Garbage in, garbage out. Below we look at a number of issues that may create a "garbage" archive.

The possibility of remote tampering may draw into doubt the reliability of media in third-party archives[26]

In a 2018 security penetration test presented at the security conference DefCon, researcher Josh Mitchell analyzed five body camera models from five different companies who all market their devices to law enforcement. In this research, he found that many of them were vulnerable to remote digital attacks, some of which could "result in the manipulation of footage," leaving no indication of any change. The possibility of security flaws leading to remote tampering of third-party databases and storage could place the reliability of the videos, images and audio recordings being stored into question.

"I haven't seen a single video file that's digitally signed," noted Mitchell. Amber Authenticate, whose clients include law-enforcement agencies in the US, is already integrated within body-cam technology, and is currently working on this very issue. Amber brought on Josh Mitchell in December 2018 as an adviser on cybersecurity threats facing police body cams.

We can conclude, then, that the authenticity proofs being generated by the companies we interviewed are only as good as their archives. Meaning that, if the archives storing this data are, or become, vulnerable, then the verification of media they provide will essentially become meaningless. Many of the companies interviewed hire external penetration testers to check their systems for vulnerabilities.

Dilemma 9:
Data storage, access and ownership – who controls what?

Court orders and subpoenas could threaten the privacy of sensitive content

Companies may receive legal threats, in the form of court orders and subpoenas, demanding they grant law enforcement access to the material they house in their databases. Ceentralizing essential media content in a database with a set of corresponding rich data on source and context makes this material susceptible to legal threats. For example, an authoritarian government could seek access to information a human rights defender or dissident uploaded to a company's servers. As countries and populations get more tech savvy, those documenting human rights abuses are becoming more aware of the risks of sharing information.

With the storage of data comes the threat of hacks, breaches and leaks, alongside legal threats such as subpoenas and court orders. Threats also arise from employees being able to access certain content, and from questions concerning the legal duty to report illegal acts detected within data. If a vulnerability is detected, or the data being stored becomes compromised, then as laid out in The Universal Declaration of Human RIghts at 70. Putting Human Rights at the Heart of the Design, Development and Deployment of Artificial Intelligence, there should be policies and practices in place to inform those affected and address the incident so it does not recur.

Here are a number of general recommendations from the Responsible Data community:

- Carrying out regular data audits and mapping what data is held, where and why.
- Collecting and storing only the minimum data necessary to avoid leakage or subpoena of data by third parties.
- Limiting data retention: only holding data collected for as long as needed.
- Storing data securely and informing users about what data is stored locally on users' devices and what is stored in places they cannot directly access.
- Setting appropriate permissions and access mechanisms: only people who need to see sensitive data should have access to it, and these permissions should be reviewed regularly.
- Carrying out regular risk assessments, Privacy Impact Assessments, or threat modelling to assess potential risks and harms.
- Designing a well-considered consent process which transparently identifies risks to users, clearly states the purpose for which the data will be used, and ensures that not giving consent for a particular use of digital data does not prevent access to support.

Dilemma 9: Data storage, access and ownership – who controls what?

QUESTIONS TO CONSIDER

- Users: What measures have been built in to account for changes over time?
- Companies and organizations: What steps have you taken to be able to respond to increasing and changing threats?
- Companies and organizations: What steps have you taken to secure viability over time?
- Tool makers and companies: Is it clear to users that their media is being uploaded and they have the option to delete it?
- Companies: Are you working with lawyers to assess the legal threats in the countries you are operating in? The knowledge gained from these discussions should be shared with your users so they can understand the risks associated with uploading content to your servers.
- Tool makers and companies: Can the media being uploaded be deleted by the user, and if so, can it be recovered from the servers you are using?
- Tool makers and companies: Who retains the rights to the data; which, if any entity, is granted third-party access, and will the companies grant data-mining opportunities to any companies, academic institutions or researchers? And if the data is open to these third parties, then what form of consent is being asked of the content creators, if any?

Related dilemmas

Dilemma 5
Authenticity infrastructure will both help and hinder access to justice and trust in legal systems.

Dilemma 8
Social media platforms will introduce their own authenticity measures.

Dilemma 14
If people can no longer be trusted, can blockchain be?

Dilemma 10: The technology and science is complex, emerging and, at times, misleading

Cropped version of Paul Hansen's photograph (cropped by the author to remove distressing images of two children killed by an airstrike).

In 2012, computer scientist Neal Krawetz claimed that the winning photograph of the World Press Photo was a fake. Hany Farid, a media forensic expert, disagreed, and the two had one of the first public disputes between media forensic experts over an image's authenticity. Krawetz alleged that the image in question, as seen below, was a composite image of three files, while Farid, an independent expert for the World Press Photo competition, said it was not. The two experts argued about a range of factors including metadata, lighting, whether different files had been combined, and the degree of certainty that is possible in these situations. The discussion highlights the complexity of forensic analysis and the challenges of establishing certainty.

The field of media forensics has only developed over the last two decades, and until recently, was still considered to be a niche field. Now, due to the rapid development in both machine learning and deep learning to create and detect altered content, alongside the increased public awareness of the impact this can have on society, the field is set to boom. While it used to take a number of years to develop one particular detector for a forgery, now machines being fed the right amount of training data can learn how to detect many editing operations at once, leading to a more efficient and robust detection process.[27]

Media forensics is not only a new field, but a disputed one. In a paper entitled "A call for more science in forensic science," published in 2018, the authors argue that the entire field of forensic science is in "dire need of deep and meaningful attention from the broader scientific community." However, with few formal career pathways, and many of the math and engineering-heavy courses available only to those within or entering into law enforcement, there will be undoubtedly be a shortfall in terms of skilled individuals able to work in media forensics for the public good in years to come.

Verified-at-capture technology should aim to provide signals, but avoid claiming to be a definitive signal. As this section details, it is more complicated than that. On each individual Truepic image page there is a link stating "technical limitations," which leads any interested user to a list, written in English, of a number of technical limitations within the technologies that Truepic "feels they should disclose publicly"-- namely what to do about older devices and software, and the issue of re-broadcasting, discussed below. This transparency is a positive approach to discussing the technical limitations that come with this software, and is better than implying that a technology is able to provide a definitive signal.

This dilemma looks at the various challenges that exist within the emerging field of media forensics, a field that could come to decide what audio-visual material is considered trustworthy or not. For recent, more detailed surveys, consider:

- Digital Image Integrity – A Survey of Protection and Verification Techniques by Pawel Korus (2017)

- Video content authentication techniques: a comprehensive survey by Raahat Devender Singh (2018)

- Fake Photos by Hany Farid (2019)

A cat and mouse arms race

When fingerprinting techniques were introduced into courts of law, and accepted as evidence for arrest and imprisonment, in response, criminals began to wear gloves. Techniques and malicious actors adopt to one another. The question is, can defensive media forensics keep up with deepfakes and adversarial attacks? Media forensics expert Hany Farid said in an 2018 interview, “We’re decades away from having forensic technology that you can unleash on a Pornhub or a Reddit and conclusively tell a real from a fake.” There are a number of media forensic researchers working on producing convincing forgeries and tricking detention systems in order to improve the detection of maliciously-altered media. Their goal is to make it difficult and expensive to maliciously alter media, and to dissuade bad actors by setting the bar high and making it a costly endeavor.

Bad actors will try to trick the authentication process

There are a number of ways that an image or video can be recorded to trick a device into authenticating it.

- **Attack the gap between sensor firmware and the software:** This gap between the firmware of a device and the authentication software becomes a place where content can be inserted or altered.

- **Stage an event:** Bad actors might stage an event that never took place, or recreate an event that happened using lookalike actors and real locations, and upload it to an authentication tool that will then add additional metadata in attempt to convince whoever is looking that the image can be trusted.

- **Use the analog hole:** Imagine that you have created a very convincing fake image, you print it on high-quality paper and take a photo of it through one of the authenticity tools mentioned in this report. Or you make a convincing synthetic video of a scene that never took place and take a video of it with your smartphone. Your smartphone digitally signs this image and it passes as an authenticated piece of media. This is known as the “analog hole” or “rebroadcast effect,” where someone takes a picture of a picture, or a video of a video, and it passes as authentic. While various companies interviewed in this research are now developing techniques to detect this type of attack, it is still considered an active threat.

- **Spoof the GPS:** GPS location data can be faked. Conclusively stating the location of an image is harder than it sounds, and many of the apps have to cross-reference GPS data with other devices and sensor checks. Media forensic expert Nasir Memon from eWITNESS, says that the reliance on GPS greatly concerns them.[28] GPS location data can be easily spoofed, and a synthetic GPS signal can be generated by those with resources and expertise. Some apps have the ability to detect whether the device taking an image or video is using a mock location application; however, this will be a race to stay up-to-date as new mock location applications and spoofing tools are released.

- **Counter-forensics challenges:** Media forensic researchers have developed fake media that is able to falsify camera model traces to show that the image, video or audio recording came from a different device (for example a iPhone6 rather than a Google Pixel).[29] This is an example of a counter-forensics approach. If, and once, malicious actors find a way to navigate the signing process to allow signing of any media from any computer or smartphone, or to simulate signing, then this malicious content will have a sign of approval that it should be trusted, leading to less trust in the authentication system.

- **Achieving 100% certainty:** Experts might not find evidence of something fake or altered in an audio recording or video or photo, but that doesn't mean the media can be 100% trusted; it might just mean the forgery hasn't been detected yet.

Complications with the technology being proposed

- How many changes are considered an acceptable number of changes? There are particular forensic challenges when it comes to imagery generated on smartphones, via apps and shared on social media platforms. If you add a filter to an image, does it stop being real? Some perceptual qualities of the image have changed, but the content remains the same. There are a whole host of innocuous manipulations that do not affect content. For instance, applying a bokeh filter on a picture of a cup of coffee to obscure the background and create the effect of a shallow depth of field is a manipulation that, in most cases, we would not consider a fundamental alteration of the integrity of the image.

- These considerations of degrees of manipulation include, for example, how to deal with the varied amount of transformations such as filters or text over images, or emojis and stickers that can be added to images and videos. Will this amount of alteration, common on a platform like Instagram, cause images to be flagged as fake content? How will the distinction be made between a fake image or video and one that has an acceptable number or type of transformations, so is still considered "real," and can the technology keep up with this?

"Bokeh Coffee" by EthanDiamond, CC BY-NC-SA 2.0

- The destructive nature of compression and metadata removal: A further complication that comes with social media platforms is that currently, most platforms perform two operations on images and videos that can be forensically destructive: they shrink the image down, compressing it in order to make the file smaller, and they delete the metadata associated with it.[30] Compression is forensically destructive, so platforms and companies will face authentication challenges and might have to retain copies of media in the most pristine format achievable to confidently authenticate it.

- PRNU might be used to identify individual devices: Photo Response Non Uniformity (PRNU) is a term used to express errors in the output from sensors caused by the physical properties of the sensor. This non-uniformity is caused by the camera sensor itself, and is considered to be a normal characteristic of the sensor. If you think of every sensor on every smartphone having their own PRNU, like a snowflake, then if someone were to have enough photos or videos from the same device, this could be used to identify which individual camera captured the images or videos. So one image taken with a verified-at-capture approach that contains location, date and time data can be correlated to other images taken with the same camera but not using the controlled capture tool (perhaps in a different circumstance where privacy and anonymity was prioritized by the users over adding authenticity signals). So PRNU signatures on a controlled-capture image could be used by governments or companies to trace back which media was coming from which device (and by extension, to potentially identify the person behind this device). Those in the media forensic field are working on techniques to reduce this risk. [31]

Dilemma 10:
The technology and science is complex, emerging and, at times, misleading

- Low false alarm rate: With any automated system using machine learning, there will be errors. The question is how to keep the overall false alarm or false negative rate acceptably low, and how to provide a system where people can ask questions and appeal decisions. Even if companies are able to ensure a high level of accuracy in the detection of maliciously-altered content on a sample dataset in a controlled lab environment, this will complicated when replicated in complex and messy real-life situations. An open discussion on whether false positives or false negatives are more damaging, particularly to vulnerable or high public-interest media creators, is critical.

- The length of video that can be captured: Many of the companies interviewed for this report had limitations in terms of how long the videos being recorded could be, ranging from 20 seconds to 15 minutes. This may not be a realistic representation of the length of time people film across a range of settings.

- Authenticity proofs are only as good as the archives that store them: Being able to provide authenticity certificates to courts of law, or to be able to state with a high amount of confidence that an image or video was captured on a certain device, at a certain time, in a certain location, is only as reliable as the company's or organization's archive or backend. If this archive or storage system is not reliable, or vulnerable to security breaches, this assurance of authenticity is meaningless.

Related dilemmas

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 8
Social media platforms will introduce their own authenticity measures.

Dilemma 9
Data storage, access and ownership - who controls what?

Dilemma 11
How to decode, understand, and appeal the information being given.

QUESTIONS TO CONSIDER

- Media forensics experts: How can more people be brought into this emerging field, and how can they have the option to work in other fields aside from law enforcement?

- Companies and tool developers: How are you keeping your archive safe?

- Companies and tool developers: How are you communicating these technical challenges to those making decisions about a piece of media's authenticity?

- Companies and tool developers: How are you fostering public conversation around acceptable false positive and negative rates?

Dilemma 11:
How to decode, understand, and appeal the information being given

The technology being proposed puts the onus and burden of proof on content creators, media consumers, and media distributors. It requires content creators to record media using particular apps and devices, and puts responsibility on media consumers to know what indicators of trust or signals of fake content they are looking for, and how to investigate further if they find them. Lastly, it requires those that distribute media from news outlets and social media companies to either integrate the technology or provide methods for content creators to authenticate the media they are producing and publishing, in a format which consumers are able to assess.

A comparison can be made to spam filters. It is estimated that 78% of emails sent globally are spam. If left unchecked, this would completely disrupt the way that people use email, obscuring who they can trust and impacting how much content they are receiving. The term "spam" refers to a comedy sketch by Monty Python where a group of singers would sing "Spam" louder and louder, drowning out other conversations. This was reused to refer to unwanted emails being sent to a large amount of email addresses, drowning out authentic communication on the internet.

To combat this, spam filtering was introduced. Spam filters look automatically at the source of the email, the reputation of the sender, the content of the email, and previous subscriber engagement to make a decision as to whether to add the email to a user's inbox or not. The user is then able to go into their spam folders to check which emails are being captured, adjust the filters and add email senders to their "safe" list. Interestingly, similar to the slogan being used by WITNESS when discussing deepfakes, a typical slogan for those working on spam filters for the public, is "Don't Panic."

In 2019, spam is under control. Users are more aware of the emails they get, the risks of phishing emails, and how to manage their spam filters. When they receive content from people they know, or who are in their network, they are more likely to trust it, and are able to make a judgment based on this. Some say that spam has moved from emails over to social media platforms, and spam could now be another word for misinformation, disinformation and malinformation.

Spam messages risked making email useless as email users found it increasingly hard to trust that the messages they were receiving were genuine. Through spam filtering and large-scale education campaigns on detecting spam and phishing emails, people began to regain trust in the system once more.

However, there are real implications with who has power in that space, who creates the white and black lists in a space dominated by large players, and who decides which email addresses and content can be trusted or not.

In terms of verified-at-capture technology, for these tools to be of use, interested members of the public have to be able to make sense of the data being collected in order to make informed decisions about whether or not they can trust the media they are viewing. For many companies interviewed for this report, the data accompanying the media being verified is technical, only useful to a trained expert. Some of the companies are taking measures to produce understandable information that goes alongside the media. This points to the need for the metadata to be simplified; people should see it and understand what data is being collected and how to read it. Tella, who have designed their app for human rights activists and journalists, has received feedback that it is essential that the metadata is readable.[32] The latest version of Tella allows users to export their metadata as a CSV file. Likewise, ProofMode users can download the data captured via a CSV file.

How to understand and influence the decisions being made behind the scenes

On March 7, 2017, an article was published on the technology blog Lieberbiber entitled "The Guardian Project's 'Proof Mode' app for activists doesn't work." This article looked specifically at the technicalities of ProofMode, how it worked, the different attack vectors, and the publicly-available code behind it.

The detailed nature of this article was only possible due to the code being available online for anyone interested to see and audit. Having code published online means that others can take this code, build on it, add to it, remix it, fork it, and check it. Due to the open-source nature of ProofMode, the Tella app has been able to integrate ProofMode's library into their own tool and is now working on adapting it for their needs, something that has saved considerable development time. They are also operating as an open-source tool, in the hopes of adding to the project.

Open-source projects usually require a community willing to volunteer their time to preserve and build on the tool. Much has been written on how difficult it is to sustain this kind of infrastructure, and some argue that having for-profit companies develop these tools is a more sustainable approach in the current climate.

How are decisions being made? And how can they be challenged?

No system is error-free, and many of the elements in media forensics are not easily readable to non-experts. As with other processes, particularly those that are driven by algorithms, machine learning or AI-technologies, there is a critical need for people to be able to review, scrutinize and appeal decisions and processes made by these systems.

With any automated system using machine learning and neural networks, there will undoubtedly be errors, and these systems will get it wrong. Even if, in the controlled environment of a lab, there is close to 100% accuracy in the detection of maliciously-altered content on a sample dataset, the system in question will not work as well in complex and messy real-life situations. This raises necessary questions about how to keep the overall false alarm rate acceptably low, and how to provide a system where people can ask questions and appeal decisions. These questions are similar to many other current queries around algorithmic transparency and decision-making, such as for example, their usage in content moderation.

The question that this dilemma explores, then, is how can interested people review, scrutinize and appeal decisions and processes made by companies who have a financial interest in keeping these processes hidden? How would people be able to query complex, patented formulas that are informing decisions on what images, videos and audio recordings are altered or not? And how would these systems become more transparent without enabling malicious actors to see potential holes and exploit them?

The approach that Dr. Matthew Stamm, a researcher who creates maliciously-altered content to try to break these forensic systems, takes is to publish full technical details of what they are working on. In his words, "If I can figure this out, someone else can figure it out."[33] For his team, it is important that people know the vulnerability exists, and understand both how it works and how it can be detected. He shared an example of this. A few years ago, he published a paper that showed it was possible to wipe away evidence that you had compressed an image more than once. This seemed like a devastating blow, as this was once one of the ways you could detect if an image had been altered and re-saved. However, within a year, several people had figured out how to detect evidence of the attack and published their findings.

Dilemma 11:
How to decode, understand, and appeal the information being given

QUESTIONS TO CONSIDER

- Companies and tool developers: How can you make the metadata as useful and readable as possible so the tool's users can understand all the information captured?
- Companies and tool developers: How would people be able to query complex, patented, formulas that are informing decisions on what images, videos and audio recordings are altered or not?
- Companies and tool developers: How would these systems become transparent without enabling malicious actors to see potential holes and exploit them?
- Companies and tool developers: Did the change of the media materially change the image? Or did the image just have minor changes?
- Companies and tool developers: How reliable is the test?
- For tool developers and platforms: What is the right to appeal? How do people who think that the system has made a mistake appeal and query the results of the test?
- Platforms: Would platforms have a new content moderation issue in terms of distinguishing "fake" from "real" images?

Related dilemmas

Dilemma 8
Social media platforms will introduce their own authenticity measures.

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

Dilemma 13
Jailbroken devices will not be able to capture verifiable audio visual material.

Dilemma 14
If people can no longer be trusted, can blockchain be?

Dilemma 12:
Those using older or niche hardware might be left behind

On November 7, 2018, Sarah Sanders, the White House Press Secretary, posted a video on Twitter that purports to show CNN reporter Jim Acosta dismissively pushing away a female intern's hand during a press conference. Questions over whether this video was authentic or maliciously-altered soon followed, and in the wake of this discussion, Amber Authenticate, a tool assessed for this research, published an article discussing how this video's reception might have been different if their technology had been integrated within the press camera's hardware.

Much of this report has focused on discussing software solutions, but many of the companies interviewed have already begun exploring integrating their technology at the hardware level. That additional time between a device capturing a video and the software authenticating is a vulnerability in and of itself, and all that malicious actors might need to exploit and insert altered content.[34] It is recognized that in order to verify-at-capture confidently, this process needs to take place at the hardware level so it is as close to the moment of capture as possible.

There is a huge diversity of phones, and with older phones producing lower-quality images, those who might need to additionally verify their media the most might not be able to do so. As Kalev Leetaru writes, "The sheer volume of camera devices in circulation and the simple fact that not everyone can afford to upgrade their device annually means that even if every phone manufacturer rolled out digital signatures tomorrow, unauthenticated footage would be with us for years to come." Truepic, in a statement on their website, explains that they "can not fully ascertain whether older devices conform to high enough security standards to ensure the data originating from them is trustworthy. We flag device integrity issues, if detected, on our verification pages as a yellow warning flag." In a recent article on the tool Roy Azaelay, the CEO of Serelay, wrote, "I think it's fair to say most startups, at least in the US and Europe, tend to start development with a bias towards high-end devices. However a solution with poor device proliferation support or an outsized SDK size is unlikely to succeed outside some very narrow industry verticals. At Serelay we currently support Android all the way back to Lollipop 5.0 (API 21) which means we can support a 6 years old Nexus 4 or a new phone you can buy for as little as £10 in the UK."

Companies will no doubt be working to persuade hardware manufacturers to integrate their controlled-

Image via AP Photo/ Evan Vucci)

capture technology. Whether these manufacturers will be persuaded is another question, and could depend on whether it is affecting customers' purchasing decisions. Leading media forensic expert Hany Farid is relatively positive, and thinks all it would take is two to four of the major manufacturers to agree on a standard, and then incorporate this standard into their pipeline.[35] There are also questions and concerns over how companies match manufacturers with authenticity solutions. In the future, choosing an Android device over an Apple device could also mean making a choice over which verification standards they have built into the device.

This can take time. Making changes to hardware is a slower process than developing software. The longer timeframe could provide some necessary breathing room for hardware manufacturers to carefully consider the consequences of verified-at-capture technology that have been raised in this paper. Once a particular company's technology is integrated within a hardware manufacturer's workflow, it becomes a timely and expensive task for them to switch to another.

Authenticated or watermarked recording devices and outputs are not necessarily going to be equally available to all socio-economic communities, so there is a possibility that people who cannot afford top-line technology will be doubted to even greater degrees, and less able to receive equal access to justice. Those with older hardware, or for those unwilling or unable to upgrade their devices regularly, could be left behind in terms of authenticating their media. Depending on which hardware manufacturers adopt which technology, there could be a chance that the less mainstream manufacturers might not be able to, or might be unwilling to, add this technology into their production cycles, leaving those with more niche providers of smartphones unable to generate the authenticity proofs that could become expected of them.

Related dilemmas

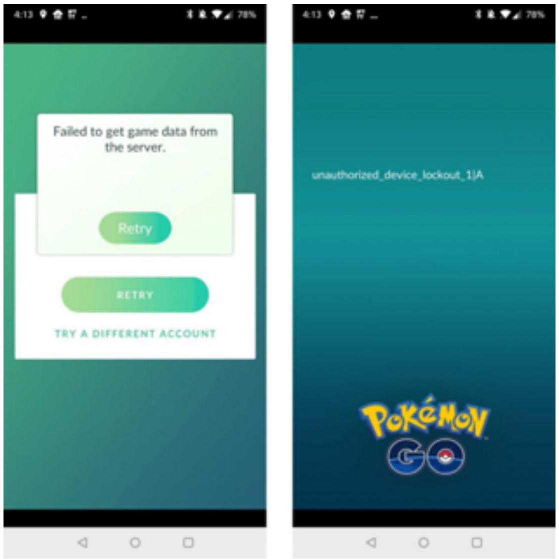
Dilemma 1
Who might be included and excluded from participating?

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 9
Data storage, access and ownership - who controls what?

Dilemma 13
Jailbroken devices will not be able to capture verifiable audio visual material.

Dilemma 13:
Jailbroken or rooted devices will not be able to capture verifiable audiovisual material


Failed to get game data from the server.

unauthorized_device_lockout_1|A

The lockout message, Ryan Whitwam, 2018

In July 2016, a new game was released for mobile devices that quickly became one of the most downloaded apps that year. Pokémon GO uses augmented reality and GPS to allow players to find, battle, capture, and train hundreds of creatures called Pokémon. Three years later, the app has been downloaded a billion times worldwide, and has been both praised for promoting physical activity and criticized for the poor taste of placing Pokémons in hazardous areas, cemeteries, and memorial sites.

The app itself requires GPS in order to work, and checks if the user is playing on a jailbroken or rooted mobile device. If the app detects that a jailbroken or rooted device is being used, then Pokémon GO will crash, ensuring that the potential Pokémon catcher can no longer use the app. This was implemented to crack down on cheaters, as those who have jailbroken or rooted devices can install GPS spoofers in order to fake their location, making it easy to level up without any of the hard work. However, others who were using jailbroken or rooted devices for personal or technical reasons found themselves "being lumped in with cheaters and effectively losing access to the game."

What does it mean to jailbreak or root a device?

Jailbreaking literally refers to a phone having been broken out of its operating-system prison. It is not to be confused with unlocking your device so it can work with a different phone carrier's network. People jailbreak their phones in order to access certain apps that have been banned from the App Store in certain countries. As detailed by Lorenzo Franceschi-Bicchierai and Brian Merchant, "To jailbreak an iPhone means exploiting one or more bug to disable a security mechanism called code-signing enforcement. This allows the hacker to run code that's not signed and approved by Apple. By doing that, the hacker opens up the possibility to install apps not approved by Apple and make changes and tweaks to the operating system."

Jailbreaking an iPhone's iOS is on the decline in recent years, as Apple has implemented many of the changes that people wanted to make to Apple's iOS. However, many people globally are using older versions of the operating system that require the flexibility a jailbroken phone provides.

For an Android device, rooting is the nearest comparison, "Everything that iOS users hope to accomplish with jailbreaking their device is already included as basic functionality within Android." Rooting allows a user to completely remove and replace the device's operating system, and allows them to grant themselves "superuser" permissions, for example.

Dilemma 13:
Jailbroken or rooted devices will not be able to capture verifiable audiovisual material

The verified-at-capture apps reviewed for this report face a similar dilemma, with considerably higher stakes. Similar to Pokémon GO, verified-at-capture tools rely on GPS location, which they use as an indicator that a piece of multimedia content is authentic, and run the risk that jailbroken or rooted devices might have a GPS spoofer installed. The authenticity tools also rely on being able to assess the integrity of the device capturing the media, and cannot guarantee the integrity of a jailbroken or rooted device.

Pokémon GO uses Google Notary as one technique to inspect the devices using their applications. One of the tools reviewed for this report, ProofMode, also uses this system to determine if a device has been jailbroken, rooted, or modified in some way. If such modifications are detected, ProofMode will still work, but will note in the metadata that the device did not match a “certified device” profile.[37]

Many of the app developers interviewed for this report recognized the trade-offs associated with the serious concerns and technical challenges around assessing the veracity of media captured through rooted or jailbroken devices, and the problem of the app being incompatible for such users, or restricting their use of the tools.

Truepic and Serelay both allow their app to work on jailbroken and rooted devices, but flag the media and communicate to those assessing it that the media is coming from an unreliable operating system, which might be a concern for those evaluating its authenticity. In an interview with Truepic’s Vice President for Strategic Initiatives Mounir Ibrahim, he notes that they are not working to be the arbiters of what is true or false, but to deliver holistic information to decision makers. Part of this includes educating decision makers about media coming from an unsound operating system, and alerting them that this might indicate the media, or the metadata associated with it, is not accurate.[37]

Building on this, these tools could allow the viewer, or decision maker, to be able to look at the content, understand the context and why the media creator might need or want to use a jailbroken or rooted device, and if the person is known, assess whether they have provided trustworthy content in the past and established a level of trust beyond technical indicators. Those who have chosen to jailbreak or root their phones for legitimate reasons, such as navigating censorship and gaining access to apps that have been blocked in countries such as China and Myanmar, will have their media flagged when using these tools.
There are many aspects to this problem that could change at any time. For instance, fascist political climates, increasing surveillance and censorship laws could impact the demand or need for jailbroken or rooted devices, or methods of detecting forged media could change so that they no longer depend on whether an operating system is stable or not.

If the expectation to use these tools in order to produce trustworthy content does scale globally, then it is essential that those who are using altered operating systems on their devices are not automatically discounted. These technologies should not exclude those who want to add additional identifiable metadata to their multimedia content, but are unable to due to their device set-up. These users are not capturing Pokémon creatures, but may be capturing important recordings of human rights abuses and civic journalism.

Dilemma 13:
Jailbroken or rooted devices will not be able to capture verifiable audiovisual material

QUESTIONS TO CONSIDER

- For tool developers: Are those who want to use the tool able to use it on jailbroken or rooted devices?
- For tool developers: How are you communicating to your users the trade-offs and reasons as to why people might be using jailbroken devices in order to assist in their decision making?

Related dilemmas

Dilemma 1
Who might be included and excluded from participating?

Dilemma 3
Voices could be both chilled and enhanced.

Dilemma 6
Technical restraints might stop these tools from working in places they are needed the most.

Dilemma 11
How to decode, understand, and appeal the information being given.

Dilemma 14: If people can no longer be trusted, can blockchain be?

In the wake of the 2008 financial crisis, the Bitcoin blockchain was introduced as post-trust technology that would allow people to automate transactions. It was supposedly removing, for the first time ever, the need for the societally-appointed and mandated record-keeper.

The distributed consensus that powers the Bitcoin blockchain, and every blockchain that has been invented since, allows multiple untrusting parties to cryptographically reach a point of agreement about the information being processed. In other words, as creative programmer Darius Kazemi describes it, the concept behind blockchain is about putting your trust in mathematics instead of into people. The types of distributed consensus used in the bitcoin blockchain were not new, nor were any of its other foundational technologies, namely, p2p networks and PGP key signing. However, what was new was the combination of these technologies into a blockchain.

Blockchains were applied first to the financial world, but are now being used in various other sectors. Many of the companies interviewed for this report are using blockchain technologies as a way of tracking the chain of custody of digital evidence, and providing a decentralized and public way to track the provenance and authenticity of an image, video or audio recording.

A blockchain, pending a variety of caveats about its actual performance, can provide verification of data. For most of the companies that we interviewed, this was a major selling point. Image or video hashes can be written to the chain in a transaction's data field, timestamps can be recorded, and, barring a variety of hacks/breaches, the verified data will persist and can be revisited for verification in the future.

One of these companies, Amber Authenticate, describes their use of the blockchain in the following statement as abstracting trust away from any particular power-holder: "Amber doesn't abstract trust away to itself; Amber creates trustlessness. People don't need to trust any stakeholder, not even Amber, if they have doubts. They just need to trust mathematics and can run the fingerprinting process themselves and compare it to the original ones: if they don't match, the user knows the video has been altered since capture. The user can also look at the provenance of the video as it and the fingerprints are stored in a transparent, yet immutable, database."[38]

These tools are not designed to increase trustlessness in users; rather, people are being asked to transfer their trust from any given human networks to technological networks, into tools built and implemented by computer scientists and mathematicians, one of which is blockchain. This section looks to describe blockchain, how it is being used to verify media, and whether it can be trusted.

What is the blockchain and how does it work?

Blockchain is a technology that allows you to transact with anyone, including people you don't know or trust, without a central party (like a bank or Paypal). The internet, as we know it today, is reliant on central parties to manage interactions between users and to store data. However, blockchain transactions are managed and stored via a cryptographically-secure process that allows them to function in a completely decentralized manner.

At its core, blockchain is a distributed ledger that is usually powered by a network of servers and computers that are operated by independent parties. The "chain" is a ledger of transactions that are stored by all of the parties in the network, known as "nodes." The cryptographic security of blockchain means that, in a well-functioning network, the chain is immutable; that is, it cannot be altered. Any outside observer is able to verify that a certain transaction happened by examining the chain of any one node in the network, and can correlate that proof across all other chains. At its most simplistic, we can think of a blockchain as a distributed Google Sheet, where old entries cannot be edited.

A properly functioning blockchain will have mathematical proof that any given transaction is valid. This proof is generated via a verification process known as "distributed consensus." Transactions that enter the network are organized into lists, or "blocks," that are added to the chain following this distributed consensus.

Dilemma 14:
If people can no longer be trusted, can blockchain be?

Anything that is written to a public chain can be retrieved by someone who either has the transaction hash, or the time to sort through hundreds of thousands of transactions. Though this does not require being particularly technically skilled (there are many online interfaces, like Etherscan, which you can use to view blockchain transactions), it does require having a basic understanding of what you're looking at (and therefore, of blockchain).

At most, from the companies that we spoke with, image or media hashes and time stamps can be retrieved from the blockchain, along with the public key of the party responsible for publishing them (which, in all cases, is either the tool creator or OpenTimeStamps).

How can blockchain technology be used for verifying media?

Given that blockchain is a mathematically-sound method of storing information about transactions that operates without a central party, it is one solution that can be considered in situations where "trust" is lacking. Blockchain removes the authority from one party or another and gives it to seemingly objective technology.

Amber, a company using blockchain in their work, asks the following question:

"Which would you trust more:

a) a file sitting on the cloud server of an organization with a vested interest in the outcome of a case, and within a system where numerous people have access privileges and whose security practices are opaque; or

b) a file stored in a decentralized system, the permissions of which are transparent, and whose fingerprints–which you can independently confirm yourself–and audit history are transparently viewable yet immutable?"[39]

A blockchain transaction can contain two things that are useful in terms of media provenance and authenticity tools, used in roughly similar ways by Amber, Truepic and ProofMode.

First, it has a timestamp, making it possible to know when, exactly, the transaction was written to the chain. Blockchain can be used as a time-stamping method to prove that a certain image was captured at, or around, a certain time. This time-stamping mechanism is one of the most valuable uses of a correctly-functioning blockchain. For example, Proofmode provides timestamps through both Google and OpenTimeStamps, which is a volunteer-run time-stamping network using the Bitcoin blockchain; in doing so, they achieve timestamps from two different parties. OpenTimeStamps batches their transactions in order to reduce the costs of writing to the chain; this means that the timestamps provided by OpenTimeStamps will not be as precise as those provided by Google. However, they will at least (in most cases) correctly reflect the date the image was taken.

Second, blockchain transactions have an encoded data field, which allows for the information in the blockchain transaction to be stored on the chain. This data can be a reference to an event or a hash of an image. Some chains, such as Bitcoin, limit the amount of data that can be sent with a transaction, making it impossible to embed media in the chain. Other chains, such as Ethereum, have limits that vary from block to block, or unlimited block size. In this instance, the media is hashed with sha-256 or sha-512 hashing algorithms. This hash is then submitted to a permissionless (public) blockchain (such as Ethereum), which either includes the hash in the data field of the blockchain transaction, or submits the hash to a smart contract that is capable of storing the data. In either case, the hash will remain on the chain indefinitely, allowing anyone with the associated transaction hash to look it up, find the image hash, and check this hash against the media file in question.

Signing to blockchain at point of creation

So how does this work for an image or video? An image is captured. It is hashed with a SHA-256 or SHA-512 hashing algorithm. Blockchain is then used as a place to publicly store the unique SHA-256 code that serves as an immutable reference to the piece of media being verified. This happens either on the user device or on the company's servers. The image hash is submitted to a blockchain in the data field of a transaction sent from one wallet to another, or to a smart contract, which is designed to process and store the hash. A smart contract could attach other data to the hash, either relevant to the image or arbitrary (like the weather at the time that the hash was written). Regardless of how the transaction is processed by the chain, there will be a timestamp attached to it.

Signing when media has been edited

One of the organizations that tracks edits to media on the blockchain is Amber. Every timeAmber's system processes a video, they deploy a unique smart contract that connects it to the Ethereum blockchain or, as Amber is blockchain agnostic, any blockchain that allows smart contracts. This blockchain is then used to store and track the hashes of the video in question. The video is broken down into segments, which are hashed and sent to the smart contract via blockchain transactions. In subsequent edits of the video, the altered sections are rehashed and sent to the smart contract, so it becomes an active record of the video's edits. As each new hash is submitted, it is timestamped by its block, a chain of custody emerges within the smart contract.

Amber takes this novel approach in order to enhance the privacy of both those being filmed and those using their technology. By breaking the video down into segments, Amber allows the user to share only the segment or segments they wish to share, rather than the entire video. These segments will be assigned their own cryptographic hash as well as a cryptographic link to the parent recording. For instance, a CCTV camera is recording constantly during a 24 hour period; however, the footage of interest is only 15 minutes in the middle of the recording. Rather than sharing the entire 24 hours of footage, Amber can share just the relevant segment with interested stakeholders. As these segments are cryptographically hashed, the stakeholders will be able to confirm that, apart from a reduction in length, nothing else has been altered from the longer, 24-hour parent recording.[40]

The main consequence of relying on the blockchain for time-stamping and image chain of custody comes in the form of costs. The cost of gas (money paid to run a transaction) through one of the more reputable blockchains is often difficult to predict and fluctuates regularly, such that if you pay a standard fee and the price rises above that, your transaction can be de-privileged and run through the chain more slowly, causing a delay in the timestamping. If every piece of media captured needs to be logged in a blockchain transaction, as is the case with Truepic, ProofMode, and Amber, costs can become extravagant. When asked, all companies pointed to using a permissioned consortium blockchain, or asking customers or donors to shoulder the transaction costs on a permissionless chain. Others anticipate that these costs will decrease significantly as these technologies evolve. While developments in blockchain technology (such as side-chains, concurrent processing, lightning networks, and so on) may allow for faster transaction speed and different storage capabilities in the future.

Will using the blockchain expose people's identities?

This is one of the most important questions concerning verified-at-capture technology's use of blockchain. Is it possible to take a string of incoherent digits, reverse the algorithm that created it, and access the data behind it, such as the time, location and even the image itself?

The media authenticity and provenance tools included in this report never write anything that would expose user identity to the blockchain, so this is not an issue for these tools. Analyzing the blockchain transactions written by any of the companies we interviewed would not allow for correlation of user photos or identification of users in any foreseeable manner. For companies like Truepic, you would need to either need to crack their algorithmic system or hack the company to access this information.

The data written to blockchain by all of these tools consists of, at most, an encoded hash of the media file in question. This hash can be seen (along with the date of the transaction and the public key of whatever tool/party is writing the data) by anyone who has the transaction hash, or by anyone who is browsing through the chain and examining the individual transactions.

The only data written to the blockchain are image or media file hashes, which cannot, without quantum computing, be decoded to reveal the media. The hashes are either stored via inclusion in the transaction, in the input data field, or on Ethereum via a smart contract that is capable of storing hashes.

The image hashing algorithms are one-way functions. For example, if I have the image, I can generate the same hash, but I cannot, without the help of a quantum computer, generate the image from the hash. Thus, there is a reasonable amount of privacy inherent in this technique, and no direct identification of attack vectors. The media files themselves, and all other metadata related to them, are stored by the companies in centralized storage (other than ProofMode, where media remains on the user's phone).

Each company that we spoke with takes the threat of user identity exposure on blockchain seriously, and, as such, never writes anything beyond image or video hashes to the chain. Without quantum computing, these SHA-256 or SHA-512 hashes are unable to be "cracked," so the image cannot currently be "backsolved" from the hash. Because the companies are simply writing the hash to the chain, without any other correlating user information, there is no possibility to link a user identity to a blockchain transaction.

Privacy concerns associated with the blockchain

There are several relevant privacy concerns to address, though most of them are in the infrastructure that companies use to accompany their blockchain deployments, and not in the blockchain deployments themselves.

The companies we spoke to have solid blockchain deployments in that they are only publishing timestamps and/or SHA-256 or SHA-512 hashes to the chain. The hashes can be of a video still or of an image, often combined with information about the media creation time and location. Though edge-cases like quantum computing will weaken SHA-256 hashes, they will still not be easily breakable (and RSA encryption would break before them, compromising so much information that the SHA hashes written to the blockchain would not be a concern, as all blockchains will be broken when RSA breaks), and there will be much easier attack vectors on any individual who might be a target.

Thus, the writing of hashes to the blockchain is not a big privacy concern in and of itself. However, in spite of the fact that the hashes may not pose a direct threat to individual privacy, many of the companies that we spoke with, including Truepic via Chainpoint and ProofMode via OpenTimestamps, automatically write data to a public chain, such as hashes and timestamps, and there's no opt-out method. While this decision was made to make the process of media capture as smooth as possible for the end-user, it does open up some questions about whether or not every photograph captured by a user with the app (or with an app that integrates the technology) justifies publication to the chain. The bigger concerns for user privacy come in the architecture that surrounds the blockchain-writing process, as detailed in Dilemma 9's discussion of data storage.

Can blockchain as a technology be trusted?

Blockchain technology is changing every day, both with the evolution of existing chains and the creation of new ones. Most chains are still stabilizing, and those that are already well-established are facing fundamental issues when trying to scale. Blockchains are still brittle and prone to failure, though the hype around them has largely downplayed the fact that most technology being built to incorporate blockchain is still very much in vaporware stage.

The issue with using blockchain to engender trust is that it's placing a lot of reliance upon a technology that is in its fledgling stages and still has a long way to go to be fully functional, robust, and reliable. Blockchain is promising as an "objective" store of value, but, as large existential issues such as the battle over governance of public chains and their scalability show, the technologies are still being invented. To ask the public and the legal system to trust these nascent technologies is a leap. Most people don't understand blockchain as a technology, so asking people to place trust in it could be asking them to make an arbitrary and disempowering decision.

How decentralized are the blockchains being used?

Though all blockchains are theoretically secure and immutable, whether they are in practice depends on the degree to which they are decentralized.

Blockchains are either permissioned or permissionless. Permissionless blockchains, or public blockchains, such as the Ethereum or Bitcoin blockchains, are the most well-known. In these networks, anyone who is technically capable of running a computer server could join the network as a "node" in order to participate in the process of creating, storing, and validating blockchain transactions. Permissioned blockchains, also known as "enterprise" or private blockchains, such as Quorum and Hyperledger, are controlled by a single party or consortium, so not everyone can join. Any open-source blockchain architecture can be cloned and run as an enterprise chain. The issue with permissioned blockchains is that due to the generally low number of nodes, as well as the shared ideological grounding indicated by consortium subscription, they miss the point of the blockchain. On a private chain, the parties can do whatever they choose in terms of consensus, for example reduce the difficulty of the puzzles to solve. This freedom to "do whatever you like in terms of consensus" goes against the very idea of blockchain. If the majority of nodes are compromised (as in hacked or taken over by a malicious party), the data in the chain can be altered. Further, an ideologically homogeneous group of actors goes against the spirit of blockchain. If you don't trust the group politically, why would you trust their verification method?

Any blockchain is capable of being used for timestamping. However, more decentralized chains are more "trustworthy" as timestamp machines because they're less susceptible to hacks. Any chain that is controlled by only a few parities is susceptible to a 51% attack, in which a malicious actor takes over 51% or more of the network and is therefore able to alter transactions (including timestamps). When such a small number of actors control a blockchain, its claim to immutability becomes deeply suspect. It's much simpler for 51% of a private blockchain to be compromised than for 51% of the Ethereum blockchain, so hacking becomes a concern. Also, in scenarios in which it's one party's word against another's, it's difficult to foresee how much a private blockchain (run by one party) would sway legal or societal opinion. A permissionless blockchain, in contrast, is run by an ever-growing network of strangers with completely different incentivisation than a private blockchain. Not only is it more difficult to hack, but it is also more "objective."

In general, any hack on a public or private chain that can disrupt time-stamps would destroy chain of custody claims. Such a hack could alter not only timestamps, but any other data written to the chain. Therefore, a hash of an image could be changed (or in the unlikely event that an entire image were stored on chain, that image could also be altered).

For this reason, most of the companies that we spoke with are prototyping on public blockchains. Proving authenticity of media relies on having a chain that can be independently verified via the presence of a truly distributed network of nodes. A private blockchain cannot be objectively verified since the single party, or consortium, running it could easily compromise the chain.

Dilemma 14:
If people can no longer be trusted, can blockchain be?

Can you change data once it has been added to the blockchain?

Properly-functioning blockchains are immutable. That means that data written to them cannot be changed. It also means that if you send your cryptocurrency to the wrong wallet, you can't get it back. If you forget your private key, you can't access the coins. There are no safeguards, and nothing can be changed. This means you have to be very careful about what you write to the chain, and should be especially careful when you're writing anything related to personal data or sensitive information.

All of the companies interviewed for this report write data to the blockchain for their users, and so users of these media authenticity tools will never grapple with blockchain key management or decisions on what data to include or not include on the chain.

Can you trust the content being posted on the blockchain?

Blockchain works to verify transactions; it provides a record that attests that a certain transaction happened at a certain time. It doesn't verify any of the data added along with the transaction, and so it's completely feasible that the blockchain component of a media authentication tool could be exploited to add the hash of a fake image to the chain. Once a hash of a fake image exists on a chain, it would be impossible to remove. In this scenario, the media would have to be moderated by whatever platform or user was "verifying" it to indicate that the blockchain hash links to a valid image.

The idea behind blockchain is that it is a decentralized, immutable ledger synonymous with trust, a source of truth and, by extension, that the data sent to it is also true. This may work to validate fake video that has been written to a blockchain and thus become automatically trusted. As those at Amber write, "The only perception should be that the data is unaltered since being written to a blockchain." It is similar to a sealed wax envelope, where you can see whether the contents of the envelope have been altered after being sealed by checking to see if the wax seal is broken; however, this does not provide a mechanism to check how genuine the contents of the letter are in the first place.

Can you delete something once it has been added to the blockchain?

Following immutability, nothing written to the chain can be revoked; it is there forever. In the case of the technology we are assessing, the data being written to the blockchain is a hash. If, for instance, someone wants to delete a video they had authenticated through one of the tools using blockchain, they would be able to delete the video, but its corresponding hash would remain publicly viewable. However, as discussed above, this hash is essentially meaningless and cannot be reverse-engineered to reveal the video, or any of its details.

In other uses of blockchain technologies, applications that want to assert that data has expired, or is no longer valid, generally do so not by erasing the data, which is impossible, but by issuing a revocation transaction certifying that the data is no longer valid. In most cases, this is a transaction from the original transaction-issuing authority that cites the original transaction's hash and declares it no longer valid. While the none of the companies interviewed here can cite any instances of this happening, revocation transactions could be used in the future if, for example, a piece of media that has been authenticated later proves to be altered, or "fake" media.

Can you trust the governance structures of blockchains?

The most well-known and widely-used blockchains, Bitcoin and Ethereum, are likely controlled by fewer parties than decentralization proponents would acknowledge given the massing of computing power in China and its ability to dominate the network due to the proof-of-work protocol. Even beyond concerns around how decentralized the systems really are, Ethereum has been the subject of numerous attacks, most notably the DAO hack of 2016, in which a single hacker drained an enormous investment fund by exploiting a single bug. The fallout from this hack resulted in a splitting of the Ethereum chain, with the Ethereum Organization "rolling back" the hack to return the hacked funds to their owners, a major violation of the supposed immutability of the blockchain. A disaffected group formed a new chain, Ethereum Classic, that allowed the hacker to keep the Ether he had stolen, as a way of maintaining the concept of decentralization.

In January 2019, Ethereum Classic was hit by a 51% hack, in which the majority of the nodes in the network were compromised and the chain was rewritten in order to allow double-spend operations, where the same coins are used multiple times. A sufficiently decentralized network would generally not be at risk for this kind of attack, but the low number of nodes on the Ethereum Classic network allowed it. Given how difficult it is to determine the extent to which the larger blockchains, like Ethereum and Bitcoin, are actually decentralized (given the amassing of computing power in the hands of a small pool of miners), it's possible that they could be susceptible to 51% attacks as well.

These are only two of the major vulnerabilities that face the largest blockchains, which are generally understood to be the most secure due to their purported decentralization. But there are other, less technical problems that present themselves. As the Ethereum's foundation to the DAO hack showed, trusting blockchains is actually trusting the governance structure around them. Most agreed that the Ethereum Foundation made the right decision in reverting the chain, but it was a decision fundamentally at odds with the idea that the blockchain is an automated, trustless, truth-making technology.

When media authenticity and provenance companies decide to rely on any given blockchain technology to verify the data that is at the core of their products, they are also, ultimately, relying on the governance of that blockchain.

QUESTIONS FOR COMPANIES AND TOOL DEVELOPERS WHO ARE CONSIDERING USING BLOCKCHAIN IN THEIR AUTHENTICITY TECHNOLOGY:

- Do you really need a blockchain? Could timestamps be delivered at a fraction of the cost by a company like Google? Could you use PGP signing of images in order to generate proofs that live off the blockchain?

- Are you considering a permissioned chain? What benefits are you gaining by running a chain instead of merely having a consortium-run, distributed database with time-stamping provided by Google or another non-chain service and, potentially, PGP signing of data?

- Do you have a strategy for how your tool will be able to evolve with the blockchain that you're working with? How will you accommodate the subsequent infrastructure changes?

- Are you planning on writing anything to the chain that could reveal user identity, or allow for correlation of user activity to be embedded in the chain? If so, reconsider this decision, as this data can never be removed.

- Do you have a plan in place for cases in which you digitally sign and authenticate maliciously-altered content and put the corresponding hash on a blockchain? Consider the use of revocation transactions to communicate that a piece of media has been found to be inauthentic.

QUESTIONS FOR INDIVIDUALS OR ORGANIZATIONS CONSIDERING USING A VERIFIED-AT-CAPTURE TOOL THAT USES BLOCKCHAIN TECHNOLOGIES:

- Consider what you actually gain from using blockchain versus another tool.
- Who is the audience of your verified media? Will using blockchain cause them to trust your media more, or less?
- What does the tool do with your media? What are the default settings for media privacy? Does the media remain on the company's servers? For how long? Can you choose to delete the media?
- All of the companies that we spoke to only write image or media hashes to the blockchain. When you use a tool that does anything with the blockchain, you should know exactly what gets written to the chain, paying special attention to anything that could be related to your identity (and don't use the tool if it writes anything related to your identity to the chain).

Further reading

- MIT Technology Review, Explainer: What is Blockchain?
- Tow Center for Digital Journalism, Blockchain in Journalism
- World Economic Forum, Blockchain Beyond the Hype
- Bruce Schneier, There's no good reason to trust blockchain technology

Related dilemmas

Dilemma 9
Data storage, access and ownership - who controls what?

Dilemma 10
The technology and science is complex, emerging and, at times, misleading.

Dilemma 11
How to decode, understand, and appeal the information being given.

Conclusion

This report is by no means inclusive of all the intricacies of verifying media at capture, and with this rapidly changing and growing field, it is likely that much of the technicalities discussed within this report will soon change. These technologies are offering an option to be better able to prove that a picture, video or audio recording has been taken in a particular location, at a particular time.

Adopted from a number of questions asked by Bruce Schneier around the involvement of blockchain in systems of trust, we asked a number of questions as we carried out this research:

- Do the tools change the system of trust in any meaningful way, or just shift it around?
- Do they try to replace trust with verification?
- Does it strengthen existing trust relationships, or work against them?
- How can trust be abused in the new system? And is this better or worse than the potential abuses in the old system?

When thinking about the consequences of any technological development, it is crucial to not only focus on what these consequences might be for the next few years, but also for future generations. It is also critical that we focus on the implications for technologies when they are implemented at a global scale, in countries with limited rule of law, and in contexts where marginalized communities have already been harmed by both misinformation and disinformation, and by technology built without their input. A set of decisions made now, out of fear or desire for profit, might lead to further harm today, and not lead to us being good ancestors to those in the future.

Throughout this paper we have introduced a number of precedents: verified user badges, spam detection; Pokémon Go. We have done this in order to highlight worries (and in some cases, solutions) that have been raised around previous technologies. In many of these cases, technology did help, but it worked in conjunction with human judgment, networks, context and education.

As the threat of more sophisticated, and more personalized audio and video manipulation emerges alongside existing problems of misinformation, disinformation and "fake news," we see a critical need to bring together key actors before we find ourselves in the eye of this new storm. It is essential that we prepare in a more coordinated way and challenge "technopocalyptic" narratives that in and of themselves damage public trust in video, images and audio. We are encouraged by the fact that many of the companies developing this technology are addressing the need to develop a set of shared technical standards in order to avoid a number of the constraints and harms raised in this report.

Ideally, verified-at-capture technology will be developed in a way that it will be seen as a signal rather than the signal of trust, and people will be able to opt-in or out without prejudice, and have the option to customize the tools based on their specific needs. If managed wisely and justly, this technology has the potential to become an important tool that contributes to more quality information, better communication, greater trust and a healthier society.

Acknowledgements

Thank you to all those who agreed to be interviewed for this research. We are grateful to those we interviewed who shared their time, expertise and experiences with us.

- Arul Prakkash, Program Manager for Asia and the Pacific, WITNESS
- Aviv Ovadya, Founder of Thoughtful Technology Project
- Brendan Dolan-Gavitt, Assistant Professor in the Computer Science and Engineering Department at NYU-Poly
- Britt Paris, Assistant Professor in the Library and Information Science Department at the School of Communication and Information at Rutgers University
- Cristian Canton, Research Manager, Facebook AI
- Hany Farid, Professor at the University of California, Berkeley
- Maneesh Agrawala, Professor of Computer Science at Stanford University.
- Matthew C. Stamm, Associate Professor in the Department of Electrical and Computer Engineering at Drexel University
- Mounir Ibrahim, Vice President Strategic Initiatives at Truepic
- Nasir Memon, Vice Dean for Academics and Student Affairs and a Professor of Computer Science and Engineering at the New York University Tandon School of Engineering
- Nathan Freitas, founder and director of The Guardian Project
- Raphael Mimoun, founder of Horizontal and maker of Tella
- Raquel Vazquez Llorente, Senior Legal Advisor, eyeWitness to Atrocities
- Riana Pfefferkorn, Associate Director of Surveillance and Cybersecurity at the Stanford Center for Internet and Society.
- Roy Azoulay, Founder and CEO of Serelay
- Shamir Allibhai, Founder and CEO of Amber Video
- Shweta Jain, Associate Professor in the Department of Mathematics and Computer Science at John Jay College of Criminal Justice, Graduate Faculty in the Digital Forensics and Cyber Security program and Doctoral faculty of Computer Science at the Graduate Center of CUNY.
- Tara Vassefi, former Washington Director of Strategic Initiatives at Truepic

A special thanks to all those who reviewed this report: Ambika Samarthya-Howard; Darius Kazemi; and Giorgio Patrini.

Credits

Editing: Jackie Stowers
Blockchain research and writing: Nina Fernclaw
Editorial support: Ambika Samarthya-Howard and Sam Gregory
Design: Adam Cohen
Research and writing: Gabi Ivens and Sam Gregory

This report was made possible through the Mozilla Foundation and the Ford Foundation, who fund and manage the Ford-Mozilla Open Web Fellowship program. As a part of this fellowship, Gabi Ivens was hosted at WITNESS and was able to dedicate ten months to thinking through some of these challenges and opportunities along with WITNESS's Program Director Sam Gregory.

About WITNESS

Founded in 1992, WITNESS helps people use video and technology to protect and defend human rights. The work of WITNESS, and the work of their partners, demonstrates the value of images to promote 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 with which videos and audio, often crudely edited or even simply recycled and recontextualized, can perpetuate and renew cycles of violence.

WITNESS's Tech Advocacy program engages with key social media and video sharing platforms to develop innovative policy and product responses to challenges that high-risk users and high-public interest content face, and to ensure that those most likely to be harmed by technologies are centered in the discussion. WITNESS's Emerging Threats and Opportunities initiative focuses on proactive approaches to protecting and upholding marginalized voices, civic journalism, and human rights as emerging technologies, such as AI, intersect with disinformation, media manipulation, and rising authoritarianism.

References

1. Hany Farid, Professor at the University of California, Berkeley, interview with Gabi Ivens, December 18, 2018. Dr. Hany Farid is on the board of advisors for Truepic, one of the verified-at-capture companies interviewed for this report.

2. Nathan Freitas, founder and director of The Guardian Project, in email correspondence with Gabi Ivens, September 18, 2019.

3. Matthew C. Stamm, Associate Professor in the Department of Electrical and Computer Engineering at Drexel University, interview with Gabi Ivens, January 8, 2019.

4. Note: As of November 2019, in final design of this report, TruePic announced it would no longer support a free version of its tool.

5. Shamir Allibhai, Founder and CEO of Amber Video, interview with Gabi Ivens, January 23, 2019.

6. Raphael Mimoun, Founder of Horizontal and maker of Tella, interview with Gabi Ivens, March 6, 2019.

7. Mounir Ibrahim, Vice President Strategic Initiatives at Truepic, interview with Gabi Ivens, December 21, 2018.

8. Raphael Mimoun, Founder of Horizontal and maker of Tella, interview with Gabi Ivens, March 6, 2019.

9. Raphael Mimoun, Founder of Horizontal and maker of Tella, interview with Gabi Ivens, March 6, 2019.

10. Nathan Freitas, Founder and Director of The Guardian Project, interview with Gabi Ivens, December 18, 2018.

11. Raquel Vazquez Llorente, Senior Legal Advisor, eyeWitness to Atrocities, interview with Gabi Ivens, January 3 2019.

12. Raquel Vazquez Llorente, Senior Legal Advisor, eyeWitness to Atrocities, interview with Gabi Ivens, January 3 2019.

13. Shamir Allibhai, Founder and CEO of Amber Video, in email correspondence with Gabi Ivens, September 11, 2019.

14. Hany Farid, Professor at the University of California, Berkeley, interview with Gabi Ivens, December 18, 2018.

15. Riana Pfefferkorn, Associate Director of Surveillance and Cybersecurity at the Stanford Center for Internet and Society, interview with Gabi Ivens, January 21, 2019.

16. Riana Pfeffrrerkorn, Associate Director of Surveillance and Cybersecurity at the Stanford Center for Internet and Society, interview with Gabi Ivens, January 21, 2019.

17. Riana Pfefferkorn, Associate Director of Surveillance and Cybersecurity at the Stanford Center for Internet and Society, in conversation with Gabi Ivens, September 4, 2019.

18. Riana Pfefferkorn, Associate Director of Surveillance and Cybersecurity at the Stanford Center for Internet and Society, interview with Gabi Ivens, January 21, 2019.

19. Raquel Vazquez Llorente, Senior Legal Advisor, eyeWitness to Atrocities, interview with Gabi Ivens, January 3 2019.

20. Matthew C. Stamm, Assistant Professor in the Department of Electrical and Computer Engineering at Drexel University, interview with Gabi Ivens, January 8, 2019.

21. Nathan Freitas, founder and director of The Guardian Project, interview with Gabi Ivens, December 18, 2018.

22. Mounir Ibrahim, Vice President Strategic Initiatives at Truepic, interview with Gabi Ivens, December 21, 2018.

23. Giorgio Patrini, Founder, CEO and Chief Scientist at Deeptrace, in email correspondence with Gabi Ivens, September 14, 2019.

24. Hany Farid, Professor at the University of California, Berkeley, interview with Gabi Ivens, December 18, 2018.

25. Hany Farid, Professor at the University of California, Berkeley, interview with Gabi Ivens, December 18, 2018.

26. This header is from Riana Pfefferkorn's article "Too good to be true? "Deepfakes" pose a new challenge for trial courts."

27. Matthew C. Stamm, Associate Professor in the Department of Electrical and Computer Engineering at Drexel University, interview with Gabi Ivens, January 8, 2019.

28. Nasir Memon, Vice Dean for Academics and Student Affairs and a Professor of Computer Science and Engineering at the New York University Tandon School of Engineering, interview with Gabi Ivens, January 7, 2019.

29. Matthew C. Stamm, Associate Professor in the Department of Electrical and Computer Engineering at Drexel University, interview with Gabi Ivens, January 8, 2019.

30. Matthew C. Stamm, Associate Professor in the Department of Electrical and Computer Engineering at Drexel University, interview with Gabi Ivens, January 8, 2019.

31. Nasir Memon, Vice Dean for Academics and Student Affairs and a Professor of Computer Science and Engineering at the New York University Tandon School of Engineering, interview with Gabi Ivens, January 7, 2019

32. Raphael Mimoun, Founder of Horizontal and maker of Tella, interview with Gabi Ivens, March 6, 2019.

33. Matthew C. Stamm, Associate Professor in the Department of Electrical and Computer Engineering at Drexel University, interview with Gabi Ivens, January 8, 2019.

34. Shamir Allibhai, Founder and CEO of Amber Video, interview with Gabi Ivens, January 23, 2019.

35. Hany Farid, Professor at the University of California, Berkeley, interview with Gabi Ivens, December 18, 2018.

36. Nathan Freitas, founder and director of The Guardian Project, interview with Gabi Ivens, December 18, 2018.

37. Mounir Ibrahim, Vice President Strategic Initiatives at Truepic, interview with Gabi Ivens, December 21, 2018.

38. Shamir Allibhai, Founder and CEO of Amber Video, in email correspondence with Gabi Ivens, September 11, 2019.

39. Shamir Allibhai, Founder and CEO of Amber Video, in email correspondence with Gabi Ivens, September 11, 2019.

40. Shamir Allibhai, Founder and CEO of Amber Video

WITNESS
80 HANSON PL
BROOKLYN
NY 11217