WITNESS

EMERGING THREATS AND OPPORTUNITIES

We are focused on proactive approaches to protecting and upholding marginal voices, civic journalism, and human rights as emerging technologies such as AI intersect with disinformation, media manipulation, and rising authoritarianism.

We have 26 years of experience using video and technology to transform lives, ensure trustworthy information, and secure fundamental rights. Bridging local communities and technology giants, we are uniquely equipped to address the critical set of challenges that threaten ordinary people, activists, and journalists as they stand up for human rights.

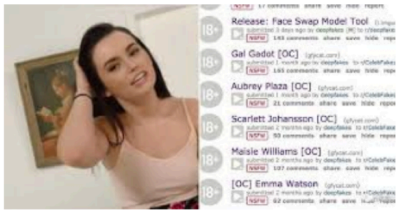
The explosion of video, online social networks, and technology has been accompanied by a set of opportunities and challenges for individuals and communities who work to advance justice and accountability around the world. In today's information ecosystem, these digital tools have the potential to increase civic engagement and participation – particularly for marginalized and vulnerable groups – enabling civic witnesses, journalists, and ordinary people to document abuse, speak truth to power, make their voices heard, and protect and defend their rights. Unfortunately, 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.

WITNESS is focused on emerging threats that accentuate these risks and diminish opportunities for human rights. In each area of focus, we bring a pragmatic perspective grounded in grassroots experiences of technologies for activism, as well as expertise in articulating threats to the human rights and journalism and engaging directly with companies on their products and policies.

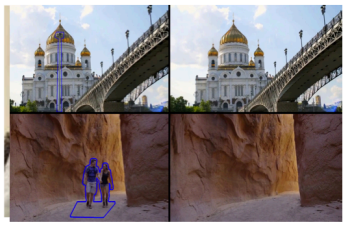
Our current primary focus is on proactive responses to how AI and deep learning systems are increasingly able to create convincing simulations of authentic media, including sophisticated audio and video manipulations called "deepfakes" and more subtle manipulations. These media forms have the potential to amplify, expand, and alter existing problems around trust in information, verification of media, and weaponization of online spaces.

Other areas of concern include:

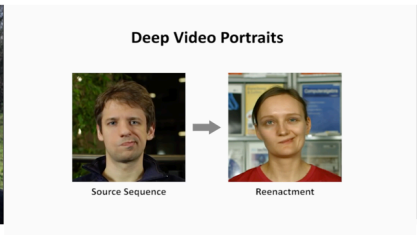
- How to create ethical but effective, tactical, technical, and strategic responses to the weaponization of online spaces against human rights: where a volume of media in "firehoses of falsehood" and increasing online/offline attacks on human rights defenders and journalists are exacerbated by synthetic media generation, bots, micro-targeting, and persuasive targeting.
- How to protect vulnerable rights defenders and civic activists as machine learning and facial recognition increases the ability to recognize and track individuals.

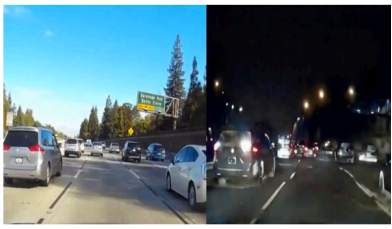










EMERGING THREATS AND OPPORTUNITIES

There is tremendous potential in the use of AI and machine learning for human rights and journalism. These technologies can aid in uncovering violations and patterns of misconduct, making sense of mass volume of media, and analyzing and presenting findings in compelling ways. However, when it comes to malicious uses, there is a critical need to bring together key actors before we are in the eye-of-the-storm, push back against apocalyptic narratives, and create proactive solutions that cut across sectors and build on existing expertise as well as new technologies. Solutions must be in global in scope, not parochial. WITNESS has been an early leader of this work in the journalistic, open source investigation, platform company, researcher, and human rights spheres, and we seek to build on this momentum in a critical moment of opportunity.

Over the past 18 months, we have been proactively addressing the emerging threat of deepfakes and synthetic media, convening the first cross-disciplinary expert convening 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 and on the Hill. We are 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 here). As of November 2018, Program Director Sam Gregory – an award-winning technologist and human rights advocate – is co-chairing the Partnership on AI's (PAI) working group on Social and Societal Impact, through which we are addressing critical challenges around disinformation, content moderation, privacy, facial recognition, and synthetic media. As part of this, we co-hosted a convening with PAI and BBC in June 2019, inviting major tech and media companies, about protecting public discourse from AI-generated mis/disinformation. We continue to engage 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.

WITNESS IN THE NEWS

Our work around deepfakes and AI-generated synthetic media has been featured in the Washington Post, MIT Technology Review, CNN, Fortune, and a range of other media.

Our work in the area of synthetic media is currently focused on five focal areas:

1 **Developing threat modelling and playbooks among at-risk communities, journalists, and frontline defenders, with a particular focus on non-US participants.**

We have led threat identification and modelling workshops with international and domestic participants at recent convenings, including a cross-disciplinary expert convening with First Draft in June 2018, workshops at MisinfoCon and StratComm, as well as an off-the-record senior journalists' convening with First Draft, the Knight Foundation, and the Ethics and Governance of AI Initiative. We are now engaged with threat modelling in Brazil in order to ensure that solutions are driven by an understanding of real-world non-U.S. threats and solutions. In May 2019, we did this work with grassroots activists and media makers and in July 2019, we will do it again with media, fact-checkers, rights activists, and technologists at a national level. Looking ahead, we seek to replicate these efforts in other regions. When there is sufficient shareable knowledge relevant to a range of size of journalists and civic activists, we will develop practical playbooks for action.

2 **Advocating to platforms for rights-respecting and shared approaches.**

We are engaging with YouTube, Google, Facebook, Adobe, Twitter and Microsoft on how platforms, social media sites, consumer tool providers, and search engines approach identifying, signaling, and moderating mal-uses of synthetic media as well as building stronger detection mechanisms. Companies producing tools for synthesis need to equally invest in making detection available. Much like the public discussions around data use and content moderation, we strongly believe there is a role for third parties in civil society to serve as a public voice on the pros/cons of various approaches, as well as to facilitate public discussion and serve as a neutral space for consensus building.

3 **Connecting key frontline verification experts with leading researchers to exchange methods, and discern what to adopt and how to make techniques accessible.**

We need better connectivity between existing practitioners and field-leaders in open-source verification and intelligence, and leading forensic analysts. Building off of the momentum from initial dialogues, we are sponsoring workshops where journalists and open-source researchers can workshop their verification workflows with researchers working on new detection techniques for synthetic media. The response to a report and series of recommendations coming out of a first workshop in 2019 confirmed the need for this collaboration from both communities. These findings were confirmed in the recent convening co-hosted by WITNESS with the PAI and BBC that focused on four key areas of preparedness for AI-generated mis/disinformation: shared detection systems, approaches to authentication, coordination between key stakeholders, and informing the wider public.

4 **Collaborating on applied research on authenticity and provenance approaches to trust in order to drive public discussion and advocacy.**

We conducted a research project on optimal ways to track authenticity, integrity, provenance, and digital edits of images, audio, and video from capture to sharing to ongoing use, addressing a critical question in the current information ecosystem. Although managing provenance and authenticity is often cited as a potential solution to synthetic media, we know there are significant pros and cons relating to privacy, revocability, the "ratchet effect," and impact on vulnerable communities. We will use the report and videos to facilitate discussion around these pros and cons with key stakeholders.

5 **Leading a rational, human rights-grounded, pragmatic discussion focused on 'prepare, don't panic' through media outreach and thought leadership.**

We need a public dialogue that is profoundly focused on the potential harms to human rights, the information ecosystem, and public trust – yet also non-alarmist and centering pragmatic, proactive solutions that are aligned with, and build upon, other approaches to AI, "fake news" and related issues. As U.S. Congress pays increasing attention to deepfakes, we have engaged on the Hill and our viewpoint was prominently captured in a Washington Post editorial. Increasingly media, inter-governmental, business, and other civil society organizations are also turning to WITNESS. We are engaged with efforts by independent and academic researchers in the computer vision and media forensics community working inside and outside the DARPA Medifor program; as well as civil society groups, such the Carnegie Endowment for International Peace's work on elections. Together with MIT's Co-Creation Lab and Mozilla, we are also exploring how the perspectives of artists and creative critics can generate ideas and provocations around synthetic media.

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. **Inclusion and human rights:** Demand responses reflect, and be shaped by, a global and inclusive approach, as well as by a shared human rights vision.

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

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

6. **Connective tissue:** Empower key frontline actors like media and civil liberties groups to better understand the threat and connect to other stakeholders/experts.

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

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

9. **Platform and tool-maker responsibility:** Determine what we want and don't want from platforms and companies commercializing tools or acting as channels for distribution, including in terms of authentication tools, manipulation detection tools, and content moderation based on what platforms find.

10. **Shared detection capacity:** 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, silenced, and empowered by the choices we make on authenticity or content moderation.

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

BACKGROUND ON WITNESS' UNIQUE CONTRIBUTION

Today, the need for WITNESS' expertise is greater than ever before. The opportunities for video and technology are coupled with significant challenges, and we know that more video and ubiquitous technology does not always equal more rights. We are fighting to change this dynamic, addressing key barriers that stand in the way of impact, and equipping people everywhere with the tools and knowledge to fight back.

Our program model is designed to ensure that activists, journalists, and ordinary people alike can use video and technology to create trusted information at the height of a disinformation crisis; that evidentiary videos can be found amidst mass volume at a time when 400+ hours of video are uploaded to YouTube every minute; and that movements and communities can stay safe and manage risk as the safety threats against human rights defenders are higher than ever before. We do this by:

Listening closely

to the needs and challenges of grassroots communities and then anticipating how activists and ordinary people can use video and technology more safely, ethically, and effectively.

Collaborating for impact

alongside vulnerable communities, filling gaps and addressing barriers and risks related to the use of video and tech for human rights.

Learning and sharing

guidance and solutions to communities facing similar issues on local, regional, and global levels.

Contributing to systems change

by advocating to tech companies for changes in their policies and products to protect the most marginalized and enable full, safe participation at scale; and by proactively bringing a human rights perspective to key emerging technology debates.

Our partners have used our resources to help secure a warlord's conviction at the International Criminal Court; to expose sectarian violence, sex trafficking, and forced evictions; and to establish legal protections for the world's most vulnerable people, from trash pickers in Delhi to elderly Americans at risk of financial, emotional, and physical abuse.

We have scaled our grassroots successes through advocacy to the tech giants, reinstating tens of thousands of videos depicting evidence of war crimes in Syria that were deleted from YouTube due to machine learning; helping YouTube integrate a blur tool functionality to help protect identities of vulnerable subjects in human rights video; and developing ProofMode, an app that makes civic media more verifiable and reference design for companies that are thinking about ways to address misinformation on their platforms.

An activist uses ProofMode to capture a scene of police disturbance in Brazil.

WITNESS' IMPACT IN ACTION

This schematic illustrates WITNESS' work in action through the thematic example of Justice and Accountability and the emergent field of citizen media as evidence, which WITNESS has been at the forefront of developing amid the explosion of grassroots documentation of conflicts in Syria, Ukraine, Yemen, and Burma. It also highlights the critical way we translate from grassroots experience of technologies to systems-level interventions.

HOW WITNESS FOSTERS IMPACT AROUND THE WORLD

LISTEN AND ANTICIPATE

We ask: how can we help human rights defenders use video and technology more safely and effectively?

COLLABORATE FOR IMPACT

We collaborate with leading activists with local influence who desire to share skills with others. Together, we fill gaps in the use of video and create human rights impact.

LOCAL CHANGE

SCALE

LEARN AND SHARE

We learn what works, then share guidance to communities facing similar issues. Our “influencers” create a multiplier effect and help us achieve scale.

SYSTEMS CHANGE

We also scale learnings on a systems level by advocating to tech companies for changes in their policies and products.