EMERGING THREATS AND OPPORTUNITIES

WITNESS
SEE IT FILM IT CHANGE IT

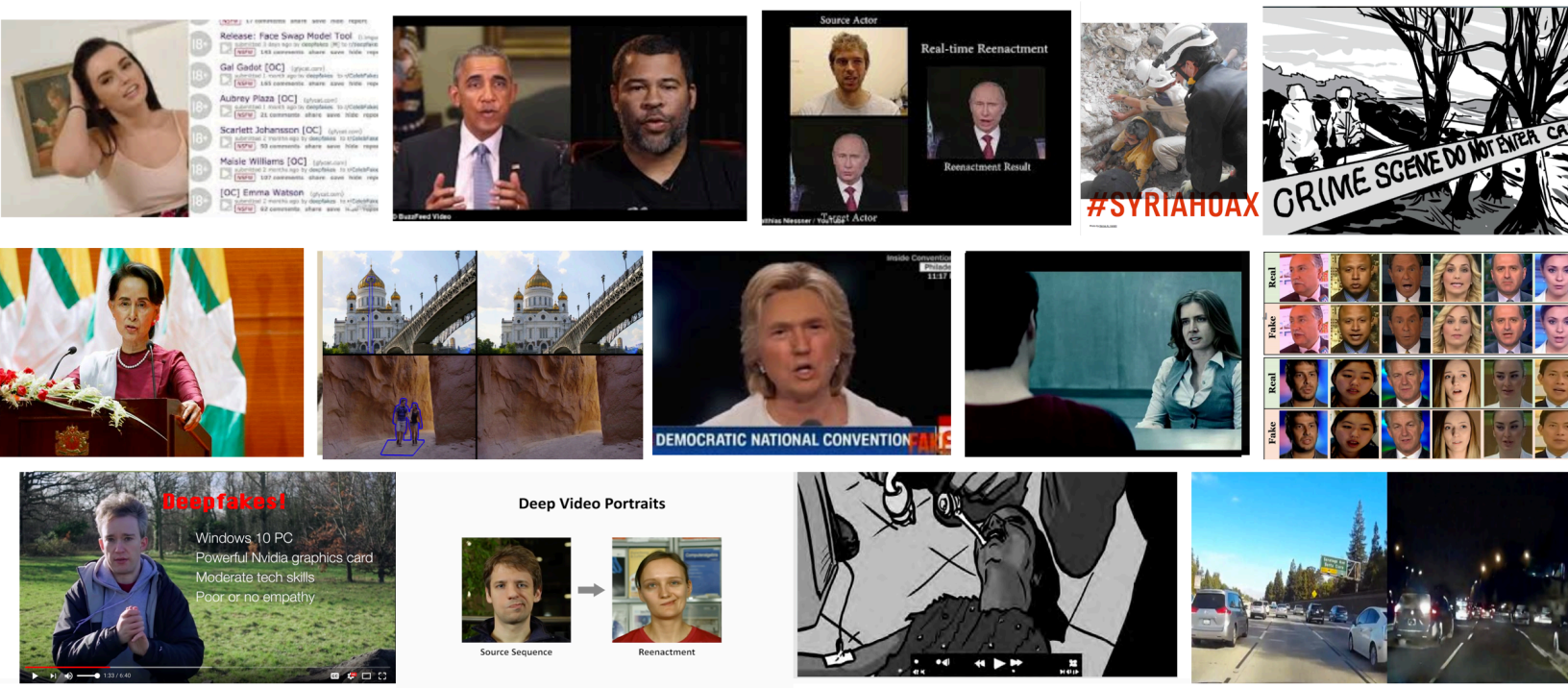
WITNESS is focused on emerging threats that accentuate risks and diminish opportunities for trustworthy media, democracy, and human rights. We bring a pragmatic perspective grounded in grassroots experiences of tech for social good; expertise articulating threats to human rights and journalism; and experience engaging with companies on their products and policies. Our current primary focus is on proactive responses to new forms of mis- and disinformation and online attacks. 'Deepfakes' and 'synthetic media' use AI to manipulate media more convincingly and make it appear people said or did things that never happened, or that events occurred that never did.

The COVID-19 pandemic as well as growing use of manipulated media in political campaigning and to attack critical civil society actors and the media, further highlight the need to support civil society and democratic institutions in efforts to counter existing and emerging forms of mis- and disinformation and support the ongoing production of trustworthy information.

To build on the work done to date, and to support our efforts at this increasingly critical moment of threat and opportunity, **WITNESS is actively raising resources to ensure our sustained leadership supporting a wider field of attention and action around media manipulation, human rights, and democracy.**

Over the coming years, our work will focus on the following activities:

1. Developing threat models and identifying solutions with at-risk communities, to ensure input from a wider range of stakeholders, and ensure quality information is available to frontline defenders of free expression, human rights, and democracy
2. Advocating to platforms for effective rights-respecting and shared approaches as they develop policy and detection and content moderation responses
3. Connecting key frontline verification experts with leading researchers to exchange methods, and discern what to adopt and how to make techniques accessible
4. Advancing a clear understanding of how to manage the trade-offs of authenticity and provenance approaches to trust via applied research and early participation in key authenticity infrastructure initiatives
5. Leading a human rights-grounded, pragmatic discussion focused on ‘prepare, don’t panic’ through media outreach, convening, collaboration, and thought leadership
6. Fostering an ethical discussion on deepfake and media manipulation norms and guiding principles



EMERGING THREATS AND OPPORTUNITIES

The explosion of video, online social networks, and technology has been accompanied by a set of opportunities and challenges for communities who work to advance justice and accountability around the world. In today's information ecosystem, digital tools have the potential to increase civic engagement and democratic participation, enabling civic witnesses, journalists, and ordinary people to document abuse, speak truth to power, and defend their rights. Unfortunately, bad actors are utilizing the same tools to spread mis- and disinformation, identify and silence dissenting voices, disrupt civil society, perpetuate hate speech, and put individual rights defenders at risk.

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. When it comes to malicious uses, 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. 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 both existing expertise and new technologies. Solutions must be global in scope, not parochial, and must be directly inter-related to solutions addressing existing forms of media manipulation such as 'shallowfakes' (forms of media that are miscontextualized, misattributed or lightly edited).

WITNESS' TRACK RECORD

WITNESS has almost 30 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, civic activists, and journalists as they stand up for human rights. We have been an early leader of this work in the journalistic, open source investigation, platform company, researcher, and human rights spheres. We are writing the guidebook for how all actors can be best prepared to respond and need to build on this momentum in a critical moment of opportunity.

Over the past two years, we have proactively addressed the emerging threat of deepfakes and synthetic media, establishing ourselves as a critical cross-sector convener, integrating essential expertise from a

non-U.S./Western perspective, advocating to platforms about what they can and should do better, and shaping public discussion through timely and visible media placements. Since organizing the first cross-disciplinary convening to identify solutions in June 2018, we have led threat modelling workshops with stakeholders in the Global South (including the first preparedness convenings in Latin America, Africa, and Asia); published surveys of solutions, and pushed the agenda with focused recommendations in meetings with technology platforms and on the Hill.

Program Director Sam Gregory – an award-winning technologist and human rights advocate – is co-chairing the Partnership on AI's (PAI) expert group on AI and Media, through which we are addressing critical challenges around disinformation, content moderation, synthetic media. As part of this, we co-hosted a convening with PAI and BBC in 2019, inviting major tech and media companies to discuss how to protect public discourse from AI-generated mis/disinformation. We engage in strategic discussions and advocacy with a range of actors (companies, 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 other forms of synthetic media has been featured in the Washington Post, MIT Technology Review, CNN, Fortune, VICE, AP, Folha de S.Paulo, and a range of other media.

WORK AREAS

Over the coming year, we will engage around the following six focal areas:

1 **Developing threat models, identifying solutions, and building playbooks for action with at-risk communities, journalists, and frontline defenders to ensure comprehensive input from a wider range of stakeholders and ensure quality information is available to frontline defenders of free expression, human rights, and democracy**

We have led a series of threat identification and modelling workshops with international and domestic participants, including the first global cross-disciplinary expert convening with First Draft in 2018, as well as an off-the-record senior journalists' convening with First Draft, Knight Foundation, and the Ethics and Governance of AI Initiative. We are now engaged with global threat modelling to ensure that solutions are driven by an understanding of real-world, non-U.S. threats and solutions, and by the input of people outside the Global North. Building on a May 2019 activist convening in Rio to prioritize threats and debate solutions, we brought together technologists, media, fact-checkers, and civic activists in Sao Paolo for an expert preparedness meeting in July. In November, we hosted a similar convening in South Africa – the first-ever deepfakes preparedness workshop on the African continent. In March 2020, the first region-wide event in Malaysia brought together journalists, civic activists and rights defenders from South and Southeast Asia. In the future, we will focus on vulnerable communities in the U.S. and additional global regions (likely to include MENA, Latin America, and Africa). These meetings have demonstrated the need to simultaneously support journalists and fact-checkers in efforts to identify and explain deepfakes; as well as to develop a robust approach to media literacy across new and older forms of media manipulation. As shareable knowledge relevant to a range of journalists and activists becomes available, we are developing practical playbooks for action and reliable backgrounders.

2 Advocating to platforms for effective rights-respecting and shared approaches as they develop policy and detection and content moderation responses

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 broadly. 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 and cons of various approaches, as well as to facilitate public discussion and serve as a neutral space for consensus building. As part of this, we are a key civil society group that provided governance on the Deepfakes Detection Challenge and identifies ways to build better shared detection systems. We provided a series of inputs on Facebook's new policy on manipulated media, resulting in a deepfakes policy we believe matches global needs. We have engaged in similar work with Twitter and other platforms. As manipulated media takes increasing center-stage on platforms, we plan to double-down on these efforts and integrate them with our work (below) on the boundary lines of satire, deception, gaslighting, and free expression.

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 and 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 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. We have been supporting continuing work in all these areas, ensuring alignment with global needs, and playing an active role in facilitating this discussion within the Partnership on AI.

4 Advancing a clear understanding of how to manage the trade-offs of authenticity and provenance approaches to trust via applied research and early participation in key authenticity infrastructure initiatives

WITNESS has been a central actor in setting new standards around how to track manipulation and edits to original images and video. In December 2019, we released Ticks Or It Didn't Happen, a comprehensive, groundbreaking report that examines optimal ways to track authenticity, integrity, provenance, and digital edits of media 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, there are significant pros and cons related to privacy, revocability, the 'ratchet effect,' and impact on vulnerable communities. We are using the report to increase the visibility of these trade-offs as these approaches begin to emerge from niche to mainstream, particularly as they apply globally for vulnerable populations. We are facilitating discussion around pros and cons with key leadership stakeholders at Adobe, Microsoft Research, Twitter, and other platforms. Our work was a key influence at the Content Authenticity Initiative launch event in January 2020, and we have been a key part of the subsequent Working Group developing an initial framework. This framework reflects key priorities around privacy, protection of vulnerable actors, technical accessibility, and centering human rights defenders and vulnerable communities as key protagonists. Looking ahead, we will engage more with company-specific efforts as well as the major media-centered Project Origin.

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

As the pervasiveness of media manipulation grows – and is further exacerbated by the spread of COVID-19, co-optation by political actors, and the impact of struggles for racial justice – it has never been more important to have a public dialogue that is profoundly focused on the potential harms to human rights, the information ecosystem, and public trust. It is equally crucial that these conversations are non-alarmist and center 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 engaged on the Hill and our viewpoint was prominently captured in a Washington Post editorial. Increasingly media, international inter-governmental, business, and other civil society organizations are also turning to us and our recommendations. We also explicitly engage with efforts by independent and academic researchers in the computer vision and media forensics community working inside and outside the DARPA Medifor/Semafor program; as part of this we have keynoted at CVPR computer vision conference workshops in 2019 and 2020, and are invited speakers at the Truth and Trust Online conference.

6 Fostering an ethical discussion on deepfake and media manipulation norms and guiding principles

Increasingly, we are witnessing the push to legislate the creation and sharing of deepfakes, spurred by broader fears of mis/disinformation. In many contexts globally, these and other 'fake news' laws capitalize on fears surrounding mis/disinformation to justify repressive measures, silence opposing voices, and stifle freedom of expression. At the same time, there is a need to define a set of ethical norms for those producing deepfake tools, as well as civil society actors and politicians deploying synthetic media in politics, advocacy, and campaigns. We are in dialogue with many actors looking at this space from within industry, start-ups, and civil society.

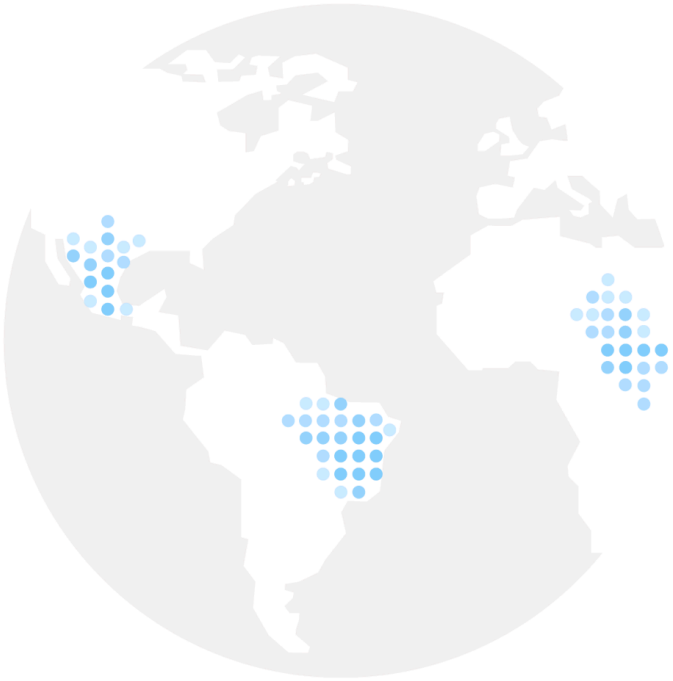
Together with MIT's Co-Creation Studio, we organized the first-ever virtual conversation series in fall 2020 to address the use and misuse of deepfakes and other forms of synthetic media in the context of satire, and how this relates to norms and expectations around mis/disinformation, journalism, political campaigning, comedy, documentary, and art. Featuring leading practitioners, artists, journalists and satirists we have raised critical questions relevant to the U.S. in 2020 and the global environment, including critical voices from Africa and Latin America. The next steps in this successful series will involve private convening work, and focus on developing norms and approaches that can be promoted and shared more broadly to key stake-holders.

TWELVE THINGS WE CAN DO NOW: 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.

ABOUT WITNESS

At WITNESS, we catalyze the human rights movement by supporting activists and human rights defenders to maximize the power of video and technology for justice. We ensure that videos can be verifiable, preserved, and curated amidst mass volume; that activists know how to stay safe and manage risk; and that technology enables greater civic participation and leads to more accountability and justice.



450
organizations partnered

11,000
people trained

130
countries represented

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; and developing ProofMode, a reference design for companies that are thinking about ways to address misinformation on their platforms.

Listen

We listen to the needs and challenges of grassroots communities and anticipate how they can use video and tech more safely, ethically, and effectively.

Collaborate

We collaborate alongside communities, addressing barriers related to the use of video and tech for human rights.

Learn and share

We extract learnings from each of our projects, and share guidance and solutions to communities facing similar issues on local, regional, and global levels.

Scale

We advocate to tech companies for changes in their policies and products, and anticipate how new developments in technology will impact human rights.