INSIGHTS FROM THE VISUAL GENERATIVE AI AND PROPAGANDA CONVENING: KEY THEMES, REFLECTIONS, AND OUTCOMES

Mona Kasra, PhD
Digital Technology and Democracy Lab
Karsh Institute of Democracy
University of Virginia

September 2024

TABLE OF CONTENT:

1. EXECUTIVE SUMMARY

This report provides a summary of Visual Generative AI and Propaganda Convening held in September 2024 at the Digital Technology and Democracy Lab, University of Virginia. The event brought together experts from industry, academia, and policy to examine the challenges posed by increasingly sophisticated AI-generated visual content, particularly deepfakes and manipulated media, on public trust and democratic processes.

Participants emphasized the risks of computational propaganda and adversarial techniques, the limitations of current detection methods, and the need for transparency, accountability, and trust in information. Key discussions focused on the role of policy innovation in addressing the evolving landscape of AI-driven threats, the development of ethical and regulatory frameworks to govern the use of generative AI on a global scale, and the necessity of advancing media literacy and cross-sector, interdisciplinary collaboration to combat the spread of visual misinformation.

2. INTRODUCTION AND BACKGROUND

Rapid advancements in artificial intelligence (AI) and deep learning techniques have led to the proliferation of increasingly sophisticated visual misinformation, posing a multifaceted challenge for individuals, organizations, and democratic societies. The potential consequences of AI-generated visual content, such as the erosion of public trust in information and institutions, demand urgent and collaborative action from industry, policy, and academia experts.

This document reflects on the Visual Generative AI and Propaganda Convening, a two-day event hosted by the Digital Technology and Democracy Lab at the University of Virginia on September 26–27, 2024. Organized by UVA faculty Mona Kasra, the event brought together a diverse group of industry and academia experts from various fields and backgrounds to explore the opportunities and challenges of AI image generators.

Participants included 11 invited guests from outside the University, as well as faculty, post-doctoral researchers from the Digital Technology and Democracy Lab, and several members of the University.

Over the two days, participants engaged in short presentations and discussions on a range of topics, including the technical, ethical, and societal challenges of synthetic media, the threats to photojournalism, the role of AI in art and media authenticity, and the risks of computational misinformation spreading through private channels. The convening also examined the use of political deepfakes in the 2024 U.S. election and the need for transparent and interoperable tools to ensure information integrity.

Participants shared research on the difficulties individuals face in identifying AI-generated content, the challenges in detecting and countering manipulated media, and learned about emerging initiatives such

as the Content Authenticity Initiative (CAI) and the Coalition for Content Provenance and Authenticity (C2PA), an open-source technical standard for digital media provenance designed to counteract manipulated and AI-generated content.

The event also featured smaller group discussions and opportunities for shared insights.

This document aims to capture the key insights and recommendations from the event, encourage continued engagement and action in addressing these challenges, and inspire further collaboration and research.

For further reference, a list of participants is included in Appendix A.

In addition, the participants shared readings and research, which are included in Appendix B, a folder containing research papers, case studies, and other relevant readings compiled by the participants.

3. PUBLIC PANEL SESSION

A key activity of the convening was a public panel discussion open to both the university community and the larger Charlottesville community. The panel featured Renée DiResta, Social Media Researcher and Author; Santiago Lyon, Head of Advocacy and Education, Content Authenticity Initiative; and Samuel Woolley, William S. Dietrich II Endowed Chair in Disinformation Studies, University of Pittsburgh. The discussion was moderated by Mona Kasra.

The discussion centered on AI-Generated visual misinformation and its impact on democracy. Panelists shared research, experiences, and recommendations.

Woolley argued that while popular discourse often warns of AI's threat to democracy, some academic and institutional studies, particularly in the West, downplay its measurable influence on global political processes and elections. He challenged these findings, asserting they are based on limited, controlled data and overlook real-world effects in the "majority world" context, where AI-driven disinformation has had severe, sometimes deadly, consequences. He stressed that lack of measurable impact is often used by tech platforms to justify inaction on content moderation, and advocated for nuanced, context-aware research and policy responses to the manipulation of digital platforms, which are far from the free, open marketplaces of ideas they claim to be.

DiResta shifted focus to the adversarial abuse of generative AI by spammers and scammers on social media, and its implications not only for information integrity but for trust and safety. She emphasized that adversarial actors quickly adopt new technologies to manipulate users at scale, noting that understanding the mechanisms of spam/scams can help in identifying and countering propaganda techniques. Traditional methods of identification and moderation are becoming less effective, requiring innovative and adaptive approaches. Regulatory frameworks should combine technological solutions,

industry collaboration, and legal enforcement where appropriate. Policies must be agile enough to adapt to the evolving landscape of AI-driven threats, with an emphasis on international cooperation to address these issues on a global scale.

Lyon presented efforts by the Adobe-led Content Authenticity Initiative in establishing protocols for digital content provenance, including its source, creation, editing, and any AI modifications. Highlighting the importance of transparency in digital media, he introduced the Coalition for Content Provenance and Authenticity (C2PA), an open technical standard for publishers, creators, and consumers to establish and verify the origin and edits of digital content. Lyon concluded by emphasizing three key pillars for combating increasingly sophisticated AI-generated visual content and ensuring transparency and accountability: 1) provenance technologies (secure metadata, invisible watermarking, fingerprinting); 2) education (media, societal, and consumer); and 3) policy (formation, implementation, and enforcement).

__Panelists' Biographies

The biographies of the panelists are provided in Appendix C.

__Video Recording of the Public Panel Discussion

A full copy of the video recording and a downloadable transcript of the panel discussion is available on YouTube and is also included in Appendix D.

4. PARTICIPANT ESSAYS AND REFLECTIONS

To ensure that the knowledge gained from the event is shared with a wider audience, participants were invited to reflect on their perspectives, insights, and research related to the topic of visual generative AI and misinformation. These essays represent participants' positions, perspective, insights, current or past work, and potential future research directions, offering valuable insights into the broader implications of the event's themes.

__SUMMARY OF ESSAYS

Bilva Chandra argues that content-based approaches cannot be framed as comprehensive solutions for ensuring information integrity in the era of generative AI. She emphasized the need for Interoperable tools and collective commitment from AI developers, platform providers, and governments are necessary to establish a reliable, trustworthy, open, and secure information environment. Without such measures, challenges to human agency in the digital space will persist, posing a threat to the future of knowledge and understanding.

Claire Leibowicz highlights the broad societal implications of synthetic media, spanning child safety, politics, and artistic expression. She calls for a human-centered approach and meaningful transparency to help audiences critically engage with media in the AI era. The challenge lies not merely in identifying the presence of AI in media creation and manipulation, but in ensuring that transparency solutions reveal the type, degree, and impact of manipulation or synthesis. By integrating interdisciplinary insights from

fields such as journalism and museums—where media authenticity and trust have long been central concerns, synthetic media regulation can better align with user needs.

Sophie Nightingale Her extensive research shows that even with training, individuals struggle to reliably identify AI-generated content, particularly AI-generated faces, with accuracy often around 50%. GAN-generated faces are frequently perceived as more trustworthy than real ones, raising concerns about their potential misuse in creating fake online profiles. Nightingale also highlights the risk of racial bias in generative AI models, where White faces often appear more realistic. Further research into both human and computational methods is needed to improve detection capabilities and build trust in online content.

Daniel and Kaylyn Schiff Their collaborative research reveals a complex picture of the evolving landscape of political deepfakes during the 2024 U.S. election. They found that many deepfakes circulated as satire or positive content rather than direct threats, with no major "Deepfake coup" occurring. However, political deepfakes remain a significant area of concern. The normalization of deepfakes—even when used as art, satire, or entertainment—poses a risk of fostering cynicism, distrust in reality, and weakening democratic discourse. There is a critical need for continued vigilance, enhanced fact-checking, and greater digital literacy to address the growing threat posed by deepfakes in shaping public perception and eroding trust in institutions.

Sandra Stevenson points out that while traditional methods like Photoshop and darkroom editing techniques are still used for image manipulation in photojournalism, generative AI now presents new, more sophisticated threats. In the interconnected nature of the news media ecosystem, a single manipulated image can have widespread consequences. The 2024 British Royal image serves as a stark example of how even non-AI-generated images can be misinterpreted and spread as factual. To combat misinformation and maintain public trust, clear AI policies, ethical guidelines, media literacy in journalism, as well as education and transparency in how AI is used in news media production are essential.

Victoria Szabo explores how generative AI is reshaping artistic creation, authorship, and media authenticity, raising critical questions about truth, control, and the evolving role of art in an AI-driven world. She draws on examples from art and film, along with the Ars Electronica 2024 exhibition in Linz, highlighting shifts in artistic roles and intellectual property. Key prescriptions include: complementing deepfake and appropriation discussions with creative examples; critically examining media indexicality; exploring parallels between past and present media misappropriation; emphasizing source attribution and trust; acknowledging the democratization of fakery while recognizing AI's potential for equity; considering provenance, authority, and process in all media, including AI; investigating intersections between media arts and archives/documentary; embracing AI glitch aesthetics; and reflecting on the emotional impacts of generative media, regardless of its origin, and evaluating the ethics of its production and effects.

Sam Wooley discusses the spread of computationally enhanced propaganda, including disinformation and deepfakes, not only on public social media platforms but also through private, encrypted messaging apps (EMAs). His research highlights the growing use of platforms like WhatsApp and Telegram by partisan groups to spread disinformation through close, trusted networks. Computational misinformation can spread both upstream to mainstream platforms and downstream into private conversations, distorting and decontextualizing messages—a phenomenon known as "cascade logic." There is a pressing need for further research into how propaganda operates in these private, intimate spaces, as well as the ethical and societal implications of generative AI-driven influence campaigns. The goal is to better understand and counteract the harmful effects of these technologies before they become more pervasive.

__FULL ESSAYS AND REFLECTIONS

Bilva Chandra–

The impact of generative AI systems on platforms, the infrastructure they operate on, and how users and populations communicate digitally reveals an acceleration of pre-existing issues and challenges in the information ecosystem. Specifically: who is a real person online, what makes information "trustworthy" and how can the internet and platforms be optimized for their intended purpose (e.g. connection), rather than enabling toxic filter bubbles, misleading or harmful information, targeting women and minors with non-consensual intimate imagery (NCII), and low-quality spam?

Mitigations to combat harms from generative AI in the information ecosystem have largely focused on content-based methods, mainly provenance efforts to track the origin and history of digital content (both synthetic and nonsynthetic) such as authenticating the source and metadata of content (content authentication) and detection methods to determine whether content is AI-generated or not. Adoption for these efforts is slowly increasing, though many challenges remain including but not limited to: the interoperability of authentication protocols across platforms, security of techniques and infrastructure to support the application of techniques, impact and consequences for users on a global scale, and simply further academic research needed to assess the effectiveness of techniques and how transparency is helping or hurting trustworthiness in content and media. Advancing the adoption of these methods to increase the transparency of content will require multi stakeholder alignment with AI developers, hardware and software providers, and social platforms to start, iterative R&D that learns from implementation issues, as well as the formal standardization of technical methods such as watermarks and secure metadata. Even so, transparency approaches for content are just one small part of the puzzle, and should not be communicated as holistic solutions for information integrity.

Content-based approaches will fall short as the "synthetic" and "nonsynthetic" content-binary fails as a meaningful descriptor as humans rely more on AI systems and tools and more content becomes hybrid and difficult to classify or define (both for the end-user and for platforms). Lastly, mitigations will not be evenly effective across various harms and use cases. For example, provenance techniques fall short in

reducing harm to victims of NCII as the damage is already done with the creation of content, though authenticated provenance signals could be useful by law enforcement for victim identification and to support evidentiary claims. Implementing mitigations with an understanding of these nuances will shape their effectiveness.

Efforts to protect the integrity of the information ecosystem, especially as AI systems become more advanced, should not stop at current digital content transparency approaches. Moving forward, efforts to empower and delineate human agency and our control over our own information consumption online, particularly that of a prosocial nature, from AI systems, agents, and more advanced algorithmic unknowns, will be an effective long-term strategy. To empower human agency is to provide individuals with more control and informed consent over their own data and how it is used, the ability to shape their own user experience, to fund collective processes that improve information ecosystems, and build and implement new digital infrastructures to allow for authenticated online activity while maintaining privacy. Some examples of efforts to increase human agency include Twitter's "Birdwatch" pilot, which shaped its current "Community Notes" feature, deliberative democracy platforms such as Polis built to harness the power of plurality for real governance decisions in Taiwan, and new ideas about what privacy-preserving digital credentials could look like to better distinguish human behavior on the internet. Though the operationalization of these efforts has presented its own challenges and requires further research and development, such as Community Notes being too slow of an approach to reduce the virality of misinformation, the purpose of these approaches is still valuable. What underlies the fear around the degradation of our current and future information ecosystem is a question of human agency and empowerment over our own information consumption and participation in the public sphere, including digital spaces. Without the implementation of interoperable tools, techniques, and the prioritization of a trustworthy, transparent, and safe information ecosystem by AI developers, platforms, software providers, and governments, issues of human agency online will continue to exacerbate and put our epistemic future at risk.

Claire Leibowicz–

For centuries, visual technologies have enabled new ways of depicting and understanding the world. As Kelsey and Roberts suggest, visual technologies "extend beyond [their] typical wires-and-switches connotation to [include] any innovation that has extended, structured, or transformed visual perception and communication." Synthetic media, also known as generative media, is one such example of a modern visual technology that does just this, implicating vast societal dynamics ranging from child safety to politics to artistic expression. While the capacity to misrepresent through media has existed for centuries, the realism, ease, and accessibility of synthetic media, as well as the increased primacy of visuals for human communication and information consumption, require express attention from policymakers, practitioners, and academics.

Through my work at Partnership on AI, a global multistakeholder nonprofit devoted to responsible artificial intelligence, since 2018 and my scholarship at the Oxford Internet Institute, I have developed and informed multistakeholder synthetic media governance for supporting audience understanding of visuals in the AI age. This has involved working alongside over 120 partners from civil society, industry, academia, and media to develop solutions that support trust in information, civic discourse, and free expression – including our Synthetic Media Framework providing guidance on responsible creation, development, and distribution of synthetic media with support from 18 organizations like OpenAI, BBC, and Adobe.

Through this work, it has become clear that synthetic media must be understood not simply as a challenge of technology, but as one of misrepresentation and impersonation that warrants humanistic attention on how to empower audiences to understand media in the AI age. As NIST recently wrote in their synthetic content report, there is a need "to provide awareness about whether and how AI was involved in creating the content, or about other relevant provenance information" to "humans." However, AI is only one detail that may be relevant to supporting audience understanding of misrepresentation or impersonation made possible via synthetic media.

The 18 organizations supporting the Synthetic Media Framework contributed long-form case reports about how they implemented key principles in the voluntary technology policy in practice, and many focused on how to label and describe media manipulations to audiences. OpenAI described how they bake disclosure into their DALL-E image generator; WITNESS described a need for disclosure even when synthetic media is used for storytelling purposes; Meta highlighted how they updated their label from "Made With AI" to "AI Info"; and Microsoft described how very subtle changes in label language (e.g., "certified" vs. "verified" by) can significantly impact consumer understanding of content on LinkedIn.

Taken together, and alongside forthcoming qualitative research on the same topic, several themes on meaningful transparency emerge. First, "AI or not" is not the right way to frame transparency since it forces people to think of a false binary – that AI content is inherently problematic or misrepresentative. As Meta wrote, "AI development is moving fast. Soon, AI will be meaningfully embedded in much of, if not most of, the content people see online." Therefore, simply disclosing the presence of AI is an imprecise form of transparency to help support audience understanding of media.Instead, transparency solutions should shed light on the type, degree, and impact of manipulation or synthesis more than those that solely emphasize AI's presence.

When asked for analogies from other fields to illustrate media transparency best practices, one stakeholder suggested a "recipe" that reveals the complexity and layers involved in visual misrepresentation. This approach would help audiences understand a "process rather than a product" and understand all the elements that shape media. Additional features of meaningful transparency that emerged included standardizing visual signals, ensuring audiences trust the authority implementing the disclosure, moving beyond a label that relies upon binaries to show a process of how content was made,

and simplicity and ease of understanding for audiences who might not care about media transparency in the first place.

Ultimately, research suggests the limits, but importance, of transparency solutions for evaluating and contextualizing media, with specific focus on identifying and disclosing more than just AI's presence. Turning to how fields beyond AI that have dealt with media transparency and trust in information for ages – like the journalistic or museum context– can tie emerging technology governance questions to the actual animating questions on visual governance more broadly. Only then can synthetic media rulemaking be responsive to the needs of actual people interpreting online content.

Sophie Nightingale–

Over the past decade, we've seen a rapid rise in the sophistication and realism of synthetic media. The advent of Generative Adversarial Network- (GAN) and diffusion- based synthesis models has led to an ability to create increasingly realistic images. What's more, the democratization of these technologies means that the ability to create synthetic images is available to almost anyone. Although technically impressive and beneficial in certain contexts, inevitably generative image models have been used to synthesize images for nefarious purposes, for example in fraud, non-consensual intimate imagery, and disinformation campaigns. Accordingly, there is growing attention on how to detect manipulated or AI-generated content. At the Karsh Institute of Democracy's Digital Technology & Democracy Convening in September 2024, I presented research examining people's ability to distinguish a real face from an AI-generated face. With the prevalent and problematic use of synthesised faces when creating inauthentic online accounts, understanding this ability is of particular importance.

Recent research has examined how effective these types of fake profile pictures are in convincing the average user that a real human is behind the online account. In Nightingale and Farid (2022), we conducted three experiments to test the realism of GAN-generated faces. We started with 400 GAN-generated and 400 matched (in terms of age, race, gender, and overall appearance) real faces. We asked 315 participants to judge if a face – shown one at a time – was real or AI-generated. Participants' mean accuracy on this task was 48.2% close to chance performance of 50%.

In a second experiment, a new set of 219 participants received a short training session to raise their awareness of common synthesis artifacts, and received image-by-image feedback. Participants' mean accuracy improved only slightly to 59.0%. In a third experiment, participants were asked to assess the trustworthiness of faces. Much to our surprise, GAN-generated faces were rated as more trustworthy than real faces.

These findings already raise significant concern about the potential for misuse—a problem that may be amplified by the recent development of diffusion-based synthesis models that combine the generative

image engines with linguistic prompts to allow users to generate novel images based on descriptive text prompts.

To examine the realism of these diffusion-generated faces, McGuire et al. (2024) supplemented Nightingale and Farid's (2022) stimulus set with an additional 400 diffusion faces (matched in terms of age, race, gender, and overall appearance). A total of 169 participants judged if the faces were real or AI-synthesized. Mean accuracy was 58.4%, only slightly better than chance (50%), with a slightly better accuracy for real (65.1%) and diffusion (62.1%), than GAN (48.0%) faces.

One widespread concern about generative image models is the issue of bias, specifically that the models render more realistic White faces than of other racial groups. We found some evidence to support this concern, namely that GAN-generated White faces were hyperrealistic (Miller et al., 2024), i.e., they were similarly likely to be judged real as faces that were in fact real—a finding that applied to a much lesser extent for East Asian or South Asian faces and did not generalize at all to Black, faces (McGuire et al., 2024). Although our results did not find this racial bias for diffusion-generated faces, other work has demonstrated racial and cultural stereotyping within these models (example).

Attempts have been made to improve human detection of AI-generated images. Research has, for example, examined the effect of encouraging people to look for the visual artifacts left behind by any editing or synthesis process. Although the results are mixed, we found that informing participants about the common artifacts associated with synthesizing faces enhanced accuracy slightly, compared to participants who were given no information (Nightingale & Farid 2022). Yet as technology improves, artifacts will likely become less apparent and in turn further limit the usefulness of detection techniques that rely on informing people to check for such artifacts. Nonetheless, these and other techniques might still prove useful, even if only for the poorer attempts at manipulation. It is, however, overly optimistic to expect people to reliably detect the most sophisticated fakes that are now so easily created and shared. As such, it is important to encourage more research in this area, including the development of improved computational methods of detection and technical standards to bolster people's ability to trust content online.

Daniel S. Schiff & Kaylyn Jackson Schiff– "Is the Deepfake Threat Dead or Dormant? Reflections on the U.S. Election"

During the recent U.S. election, the long-feared scenario of a decisive deepfake turning the tides never quite materialized. At least, this is how many analysts and commentators interpret the outcomes. That is, despite dire warnings, we did not witness a last-minute, hyper-realistic deepfake that upended voter confidence or single-handedly tilted the balance at the ballot box. (More than a few instances were clearly in this category, however, like deepfakes of burned ballet boxes, an audio deepfake of Joe Biden peddling misinformation about electoral processes, and more.)

But overall, what can we take away from what happened and failed to happen? Was the threat of deepfakes overhyped? Or perhaps our informational environment, our political institutions, our news media, and our social media platforms were more resilient than we expected?

The reality is quite a bit more complicated. Rather than a decisive deepfake shutting down the polls or leading to immediate violence, the deepfake landscape was instead marked by a mix of satirical, artistic content, and isolated examples, many of which reinforced existing partisan narratives. And the full story may take years to present itself as researchers, deepfake detection experts, and communication specialists attempt to make sense of the role of misinformation in the election. (Indeed, counterfactual scenarios prevent us from understanding what might have occurred under other circumstances!)

Yet, a few notes are worth considering as an initial reflection:

Before the election, many experts, including ourselves, worried that deepfakes—a technology that once seemed confined to Hollywood and YouTube actor swaps—would be weaponized to spread misinformation en masse and destabilize our political discourse. In practice, however, what emerged was more subtle. Many or most deepfakes that circulated were seemingly crafted for satire or entertainment. Think of meme-like videos and images ranging from humorous to absurd, poking and prodding, sometimes in good humor and sometimes not.

Another trend that surprised us: many of the most prominent deepfakes were also 'positive' in nature rather than 'negative,' aiming to bolster the reputation and image of a politician, not detract from it. Yet, even this information can be problematic for a variety of reasons, and can certainly constitute subtle to egregious misinformation. If anything was clear, these positive deepfakes were very likely to tap into pre-existing political biases, including ugly ones.

While somewhat unexpected, these outcomes ask us to reconsider our concerns about political deepfakes. On one hand, the absence of an overt 'deepfake coup' may suggest that our political system proved resilient in the face of this emerging threat. We are personally skeptical that our information environment, voter literacy, norms of our political institutions, and more would satisfactorily testify to this rosy picture. In fact, it may be very reasonable to make the case that misinformation remains one of the (if not the) decisive factors in our political system.

Yet, it is important to recognize the substantial work of advocates, victims, social and technical experts, and more. Legislators passed dozens of bills instituting bands and penalties for abuses of political deepfakes; civil society groups held hearings; technical experts created standards and startups; and social media companies did implement some relevant changes, all of which could have made real differences in terms of limiting the penetration of harmful deepfakes.

Thus, while the election did not serve as a perfect testing ground for the worst-case applications of negative, photo-realistic deepfakes, it did reveal a more insidious trend. Political deepfakes, even in their 'lighter' forms, are now increasingly and gradually normalizing the idea that reality can be bent, twisted, or reinterpreted by an average citizen or a PAC, and shared by a political or social media mogul without consequence (and indeed, with legal protection as expression or satire). This normalization, represented by the rapid dissemination and popularization of more entertainment or artistically-oriented deepfakes, carries long-term consequences, many of which are similar to the older fears about deepfakes.

That is, the very fact that deepfakes have rapidly become a routine part of political commentary—whether as art or as a means to subtly manipulate and reinforce narratives—should give us significant pause. Even when intended as satire, deepfakes can very much contribute to an atmosphere where the lines between fact and synthetic fabrication blur. They condition us to accept that seeing is not necessarily believing. And they allow us to view political emotion and expression as orthogonal to concerns of truth or good faith. Who cares if an image portrays a false or mean-spirited stereotype: technically, it's just a satire, right? So what if it spurs emotions based on false narratives as long as the images are only politically–not technologically—deceptive.

An older concern is also that—when audiences are repeatedly exposed to content that plays fast and loose with the truth, we might gradually lose our trust in our institutions and sources of information. Relatedly, some have feared that skepticism might evolve into cynicism and disengagement—dangerous shifts that undermine democratic discourse and flourishing. We think this remains very much true even in a world full of positive, satirical, or artistic deepfakes.

It's also worth observing that even a satirical video or image can be screenshotted, shared widely, stripped of its context, and mistakenly accepted as genuine. (Ironically, in some of our efforts tracking political deepfakes, we found that journalists are amongst the individuals most likely to share this information, presumably to fact check and inoculate the public.) At this point, researchers know very little about which people believe or are persuaded by deepfakes that careful viewers would casually write off as clearly satirical, expressive, or symbolic. What begins as a humorous jab could be repurposed to serve more nefarious ends in future elections, or could simply come to confuse and mislead people unintentionally as an image travels down the social media grapevine.

In the end, while the recent U.S. election may have defied the apocalyptic predictions of deepfake doom, we suggest this should not lull us into complacency. The worst harms may have been avoided by the outcomes of the election, such that postelection deepfakery related to the political process was not called for. This is a counterfactual we cannot examine. Negative impacts may also have very much been avoided because of the significant attention devoted to deepfakes by advocates, policymakers, researchers, and more. If the FTC, FCC, FTC, and a couple dozen state governments had not raised alarms and proposed or enacted transparency requirements or bans, what might have occurred? If social media companies had not faced pressure to incorporate better fact checking, limitations on political

advertising and misinformation detection and labeling strategies, how much more might have been spread? Finally, we also cannot clearly account for the impact of deepfakes on more local elections that do not benefit from as much scrutiny, or for elections all around the world, where civil society, journalistic, and fact checking institutions may be weaker.

As such, we suggest that political deepfakes remain a potent area for concern: The usage of deepfakes to spread misinformation is just beginning. And, in fact, they may be used to spread misleading narratives, foster partisan anger, and more, even when the content is not pretending to be authentic! The challenges are, indeed, more complex, not less.

To us, they call for ongoing vigilance, and continue to demand improved (and more conceptually sophisticated) fact-checking mechanisms and efforts to bolster digital literacy. The absence of a decisive, outright deepfake scandal this time around is not a signal to lower our guard; rather, it is an opportunity to address the slow-burning erosion of trust in our information ecosystem before it can lead to (even) more significant harm.

Sandra M. Stevenson–

The challenges of image manipulation have always been a part of conversation in the world of photojournalism. Photo editors are the first line of defense in being able to root out images that falsely purport to show the truth. Manipulation ranges from recognizing that the image is "set up" based on composition of various elements in a frame, to reviewing the entire take to understand the scene progression, to photoshop, even traditional darkroom manipulation.

The introduction of AI poses a new set of challenges for visual editors. Technology continues to evolve at breakneck speed and is now included as part of the camera, including mobile phones, hardware and software – and not by choice. While there will be rogue players who will use this technology to their advantage, there are people who don't understand the technology, use it and can cause the same harm. Furthermore, the photojournalism world is interconnected. News publications share and/or syndicate content generated by freelancers and staff with wire services, freelancers, companies and organizations, governments and UGC also contribute to news publications and wire services, and all are reliant upon wires for supplementing their news coverage. One bad image can have a ripple effect on an already fragile ecosystem.

An example of how vulnerable the media ecosystem is was on display when the British Royals released an image of Princess Kate in March 2024: . In the end, the image was proven to not be AI generated but poor Photoshop skills, however it shows how images can make their way into news coverage under the guise of being factual. Fortunately, most news organizations turned the error into a teachable moment for audiences, and internally, a reexamination of trust between news organizations and a government entity, in this case with Royal handouts, unfolded.

There is so much at stake if media outlets aren't clear about their AI policies, internally and externally, and if they are not transparent with audiences on how AI is used in the journalism produced. It is imperative that newsrooms revisit AI technology and incorporate into their ethics policy frequently to get ahead of issues or pitfalls. Editors should also continue to maintain their ethics policies and remind contract visual journalists periodically.

Media should continue to report on the implications of AI and societal impacts. Furthermore, news organizations should invest in a concerted effort to engage in ongoing media literacy through social media, especially leading up to elections.

And while news organizations such as The Washington Post may go through great lengths to be transparent with readers, it is imperative for the media to know that it may not be enough to combat spreading mis/disinformation. Furthermore, media must be mindful of how the use or reporting of AI is handled, as there are times where media has inadvertently further amplified false information.

Victoria Szabo– "Artistic Co-Production with Generative AI"

My presentation at the Convening was entitled "Artistic Co-Production with Generative AI." My talk focused on questions and issues raised in the creative media application space around generative AI, and how these insights might apply to present and future education and governance issues. I began with the question of how to define the locus of visual misinformation in light of media effects that gesture towards certainty, truthfulness and authenticity, through how they have been used in the past. In both photo video documentation, an implicit Indexicality to the real has been challenged by increasingly convincing deep fakes, generative image creation, and increasingly video. Fakery has always been present, as has technological mediation of whatever impression or trace is created by processes of imaging and capture. We have relied upon AI "tells" in image and video, but the field is evolving rapidly, with generative AI productions becoming increasingly convincing. As we move away from any assumption that we can recognize a fake when we see it - and, crucially, that we care whether it is fake or not - the implications of this twofold shift in perception and its significance reverberate across numerous knowledge domains. I've used Stephen Colbert's notion of Truthiness to talk about VR representations of the real as an asymptotic approach to a convincing simulation. It applies here too, to media that don't claim to tackle over our sensorium, yet insinuate themselves into our archives and cultural memory as building blocks that masquerade as media artifacts, traces, and ultimately data exhaust inhaled by an increasingly capacious remix engine.

I began my presentation by recounting recent controversies in the art and film worlds, with notable examples from 2022 when an AI artist won a prestigious award, and Hollywood actors struck in part over the use of their likenesses in future productions. I presented these controversies as early warning signs of disruptions that will affect how we think about not only the nature of evidence going forward, but also

about how artistic control, authorship,identity, and intellectual property are evolving as they relate to both intentional productions and unintentional. These considerations affect the artists themselves, and also pose questions around what the purpose of art may be in our society, and whether artists are necessary to produce them. We discussed the case of Billy Joel's new music video, which pulls from his personal media archives to create a video for a new song. This is a case in which the artist is in seeming control of his own media legacy, and raised the question of whether or not it matters if you "know" what is going on. For my students, the video was a novelty, and though they first thought it was genuine footage, the revelation that it was AI-produced created only mild surprise. For me, significantly older and also a longtime fan of Joel's work, the video evoked nostalgia, anxiety, and unease with the uncanny valley aspects of Joel's appearance.

How then do we begin to think about what the future holds, and how to engage with it? At this point we turned to another example drawn from the recent Ars Electronica 2024 exhibition in Linz. This year, the prestigious Golden Nica prize was awarded for the first time to an artist whose work included substantial generative AI components. As a group we considered how a generative video untethered to an individual artist, but harnessed by a different music video content creator for an aesthetic end, might be understood within the new media landscape we inhabit today. We watched Paul Trillo's "Washed out 'The Hardest Part'" music video and together deconstructed its media effects. First we reflected upon its impact, and then speculated on how it was made. The film is a continuous time-tunnel story of a couple's relationship over the years in multiple settings, and some of our group guessed it had been produced in the generative video tool Sora, which was in pre-release to selected users. We looked at some of the prompts that had been used by the artist to explore its prompt engineering as a form of authorship, and to reflect upon whether enough prompt complexity and specificity = artistic creativity. For a film director this correlation makes sense; but at the level of content, the recycled material used to produce the media, the analogy was more ambiguous. Looking next to Sora's website, we noted how it is being advertised as a way to bring historical imagery to life; a clear next case use of filling in the historical record. Afterwards I shared some examples of generative images and AI based on the UVA campus, intentionally choosing a photo I had taken the day before of a campus monument as a test case resonant with past associations, given the recent history of campus-based protests and responses to them. We discussed these examples in light of Ted Chiang's recent New Yorker essay, "Why AI Isn't Going to Make Art." We built upon Chiang's criteria for human art production - decision, iteration, control - and discussed whether AI-collaborative art practice is continuous with prior movements and techniques. In thinking through the implications of these critiques, I referenced conceptual art in relation to iterative prompt generation; collage and remix alongside principles of selection; video effects to tools like stable diffusion; motion capture and visualization to data art and interactive installation. We explored more case study examples, and critiqued AI systems from aesthetic and co-productive, agential stances.

Ultimately I concluded the conversation with a call for Generative AI Media Literacy, with attention to these keywords: decision, iteration, control, aesthetics, and agency, and how they played out in the contemporary media landscape. Essentially, this approach is an empowered user/consumer/creator. My

prescription, in sum, includes calls to: complement deepfake and appropriation discussions with critical and creative application examples; to engage in foundational critique of media indexicality as ever being unquestionably valid; to explore through-lines from prior forms of media misappropriation and recontextualization (ie to historicize the moment); to call for renewed attention to source attribution and trust; to acknowledge democratization of fakery, but also concede AI as potentially equity-producing; to consider provenance, authority and process questions as essential to any media production, including generative AI; to investigate cross-over interests between media arts and archives/documentary; to explore and perhaps embrace AI glitch-aesthetics, even as the bugs blend into to machinery; and finally, to consider the profound affective responses engineered by emerging generative media productions, regardless of how they are made, and whether we agree with the ethics of their production, circulation, and effects.

Samuel Woolley– "Manipulating Public Opinion in Private: Visual Generative AI, Propaganda, and Encrypted Messaging Applications"

When experts discuss the problem of propaganda online, they often focus upon how it flows over large, public, social media platforms. Because of this, problems associated with deepfakes or disinformation spreading across Facebook, Instagram, TikTok, YouTube and X are relatively well-documented.

But computationally enhanced propaganda also spreads across a range of other spaces, from partisan websites to messaging boards (Benkler et al., 2018). Research on the political manipulation of information flows in countries such as Brazil (Evangelista and Bruno, 2019), India (Farooq, 2018), and Ukraine (ISD, 2022) reveals that closed-off spaces—including applications that offer privacy in the form of end-to-end encryption—also play a critical role in spreading propaganda today. In the United States, apps like Telegram, Messenger, WeChat, and WhatsApp have been on the rise in terms of both general usage and as vehicles for sophisticated influence operations (Gursky et al, 2020).

Some researchers argue that visual propaganda, from memes to videos, is more potent than its written cousin (Sundar et al., 2018). Others suggest that automation and generative AI can be used in tandem with images to make manipulation efforts both more influential, more widespread and more targeted (Woolley and Howard, 2018; Goldstein et al., 2024). Private messaging services, and particularly encrypted messaging applications (EMAs), also seem to amplify and enhance propaganda campaigns (Rosenblat et al., 2024). This is because people using them not only have an expectation of security due to platform design, but also because they tend to follow and communicate with people they already know (and trust) in such spaces.

My collaborators and I have noted that EMAs offer a greater deal of relational potency to propagandists (Trauthig et al., 2024). The challenge, which partisan groups like Narendra Modi's Bharatiya Janata Party (BJP) seem to have handily overcome, is that such influence efforts require getting deceptive content to seep into more intimate, trusted, communication channels. The BJP and others have achieved this by building a massive ecosystem of both messengers and chats on WhatsApp. This allows, for instance,

disinformation purposefully spread by government actors to metastasize into misinformation spread by friends and family members. This content is then often spread by these same trusted actors across other communication channels, online and off.

We also note that "cascade logic" is central to how propaganda flows over EMAs and other chat apps (Gursky et al., 2022). On platforms like WhatsApp "as information is trafficked upstream (making its way from private conversations to the mainstream) as well as downstream (allowing information to withdraw from the public eye), it can get distorted, decontextualized, and thereby, transport false information.

Over the last six years we have completed a range of mixed methods analyses on the spread of propaganda over EMAs and other chat apps, including 150+ global interviews with producers of computational propaganda and those working to track and uncover digital disinformation. Memes and short-form videos—including AI-doctored and AI-generated visuals—are critical currency for communication on platforms like Telegram, WeChat, and WhatsApp, perhaps even more so than on public social media platforms such as Facebook and X.

To build understandings of how propaganda flows across the internet and beyond, researchers must do more work on its use and consumption on these more private, more intimate, spaces online. To build better understandings of how it effects people today, we must begin to zero in on visual, generative-AI produced, influence campaigns. Increasingly, these are the tools used by propagandists working to scale and focus their efforts. We cannot allow the malign persuasive power of such technologies and strategies to remain hidden. (References available in Appendix E.)

5. CONCLUSION

The Visual Generative AI and Propaganda Convening provided a platform for multidisciplinary dialogue among experts from industry, academia, and policymaking. Discussions highlighted the urgent need to address the rapidly evolving impact of AI-generated visual content, which pose risks to public trust in information, institutions, and democratic processes.

Participants emphasized the accelerating evolution of generative AI technologies, which, while offering transformative potential, also presents substantial challenges when misused for manipulation, deception, and propaganda. The discussions underscored the growing complexity of countering deepfakes and manipulated media.

Key takeaways from the convening included the need for:

- The development of ethical and regulatory frameworks to govern the use of generative AI
- Regulatory frameworks that integrate technological solutions, industry collaboration, and legal enforcement
- Agile policies capable of adapting to the evolving landscape of AI-driven threats, with an emphasis on international cooperation to globally address the issues

- A focus on transparency, accountability, and the protection of individual autonomy in the digital space
- Enhanced media literacy and education
- Cross-disciplinary collaboration across sectors to create effective and sustainable solution

6. ACKNOWLEDGMENT

I extend my sincere gratitude to Laurent Dubois, Academic Director of the Karsh Institute of Democracy, Jane Kulow, Labs Coordinator of the Karsh Institute of Democracy, and the entire staff for their facilitation of the event, their support, and their willingness to host the convening. I also thank Kemi Jona, Vice Provost for Online Education and Digital Innovation, and Megan Barnett, Vice Provost for Pan University Initiatives, for being a part of the event despite their busy schedules. I am grateful to my colleagues at the Digital Technology for Democracy Lab for their participation and involvement in the event.

7. APPENDICES

__Appendix A: List of Participants

Invited Participants from outside UVA:

- Bilva Chandra (Google Deepmind)
- Renée DiResta (Social Media Researcher and Author)
- Josh Goldstein (Researcher)
- Claire Leibowicz (Head Of Ai & Media Integrity, Partnership on AI)
- Santiago Lyon (Head of Advocacy and Education, Content Authenticity Initiative)
- Sophie Nightingale (Lecturer in Psychology, University of Lancaster)
- Kaylyn Jackson Schiff (Assistant Professor in the Department of Political Science and Co-Director of the Governance and Responsible AI Lab, Purdue University)
- Daniel Schiff (Assistant Professor in the Department of Political Science and Co-Director of the Governance and Responsible AI Lab, Purdue University)
- Sandra Stevenson (Deputy Director of Photography, Washington Post)
- Victoria Szabo (Research Professor of Visual and Media Studies, Duke University)
- Samuel Woolley (Dietrich Chair of Disinformation Studies and Associate Professor in the Department of Communication, University of Pittsburgh)

Participants from UVA:

- Mona Kasra (Organizer)
- Laurent Dubois
- Steven L. Johnson
- David Nemer
- Andre Sobral
- Lori Young
- Megan Wiessner

__Appendix B: Additional Resources and Readings

The materials included in this section consist of a curated collection of research papers, case studies, and other relevant readings, intended to serve as a resource for further reading: https://virginia.box.com/s/w3f25l9xfy2ngb3iuh0polgqpgztw3je

__Appendix C: Panelists

This section provides biographical information about the panelists:

Renée DiResta, Social Media Researcher and Author–

Renée DiResta is a social media researcher and the author of *Invisible Rulers: The People Who Turn Lies into Reality*. She studies adversarial abuse online, ranging from state actors running influence operations, to spammers and scammers, to issues related to child safety. From 2019–2023 she was the technical research manager at the Stanford Internet Observatory, a cross-disciplinary program of research, teaching, and policy engagement for the study of abuse in current information technologies. DiResta has advised Congress, the executive branch, and academic, civic, and business organizations on issues related to technology and policy, including information operations, generative AI, election security, researcher transparency, child safety, and more. DiResta is a contributor at The Atlantic.

Santiago Lyon, Head of Advocacy and Education, Content Authenticity Initiative–

Santiago Lyon is the head of advocacy and education for the Adobe-led Content Authenticity Initiative, working to combat misinformation through digital content provenance. He is an award-winning photojournalist, media executive, and educator, with a photography career spanning over 40 years. He was a Nieman Fellow at Harvard University (2003–2004) and later became VP/director of photography at the Associated Press until 2016. Under his direction, the AP won three Pulitzer Prizes for photography. In 2012, he was a Sulzberger Fellow at Columbia University and was chair of the jury for the 2013 World Press Photo contest. He serves on the board of Eddie Adams Workshop and on the advisory board for the VII Foundation.

Samuel Woolley, Dietrich Chair of Disinformation Studies, Department of Communication, University of Pittsburgh–

Samuel Woolley is the inaugural William S. Dietrich II Endowed Chair in disinformation studies and an associate professor in the department of communication at the University of Pittsburgh. His internationally recognized work on computational propaganda has revealed the ways in which a wide variety of groups around the world leverage automation, artificial intelligence, and coordinated armies of users to control the flow of information during pivotal events. He is the author of four books and numerous articles and essays on propaganda, disinformation, and emerging technologies. His most recent book, *Manufacturing Consensus: Understanding Propaganda in the Age of Automation and*

Anonymity, is an in-depth exploration of the people behind modern propaganda campaigns. His PhD is from the University of Washington.

Mona Kasra (Moderator), Faculty Co-Lead, Digital Technology for Democracy Lab, Associate Professor of Digital Media Design, University of Virginia–
Mona Kasra is a faculty co-lead at the University of Virginia's Digital Technology for Democracy Lab. She is a new media artist, interdisciplinary researcher, and associate professor of digital media design. Her work explores the political and theoretical implications of visual media technologies within our culture and cross-culturally. Kasra has exhibited in galleries and film festivals throughout the United States and internationally and has received two Helen Hayes Award nominations for her media design work for live performances. She serves as the chair of ACM SIGGRAPH (2023–2024) and is a board member for the New Media Caucus.

__Appendix D: Video and Transcript of the Public Panel Discussion

A full video recording and a downloadable transcript of the panel discussion is available at: https://youtu.be/sKS1DCq4VyI?si=BkF6KK0yf4U3_ugr.

__Appendix E: References

References for for Samuel Woolley's Essay "Manipulating Public Opinion in Private: Visual Generative AI, Propaganda, and Encrypted Messaging Applications":

Benkler, Y., Faris, R., & Roberts, H. (2018). Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford University Press.

Evangelista, R., & Bruno, F. (2019). WhatsApp and political instability in Brazil: targeted messages and political radicalisation. Internet Policy Review, 8(4). https://doi.org/10.14763/2019.4.1434

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? PNAS Nexus, 3(2). https://doi.org/10.1093/pnasnexus/pgae034

Gursky, J., Glover, K., Joseff, K., Riedl, M.J., Pinzon, J., Geller, R., & Woolley, S. C. (2020, October 26). Encrypted propaganda: Political manipulation via encrypted messages apps in the United States, India, and Mexico. Center for Media Engagement. https://mediaengagement.org/wp-content/uploads/2020/10/Encrypted-Propaganda-Political-Manipulation-Via-Encrypted-Messages-Apps-in-the-United-States-India-and-Mexico.pdf

Gursky, J., Riedl, M., & Woolley, S. (2021, March 21). The disinformation threat to diaspora communities in encrypted chat apps. The Brookings Institution. https://www.brookings.edu/articles/the-disinformation-threat-to-diaspora-communities-in-encrypted-chat-apps/

Gursky, J., Riedl, M. J., Joseff, K., & Woolley, S. (2022). Chat apps and cascade logic: A multi-platform perspective on India, Mexico, and the United States. Social Media + Society, 8(2).

Institute for Strategic Dialogue. (2022, October 26). A false picture for many audiences: How Russian-language pro-Kremlin Telegram channels spread propaganda and disinformation about refugees from Ukraine. Institute for Strategic Dialogue.
https://www.isdglobal.org/digital_dispatches/a-false-picture-for-many-audiences-how-russian-language-pro-kremlin-telegram-channels-spread-propaganda-and-disinformation-about-refugees-from-ukraine/

Rosenblat, M., Trauthig, I., & Woolley, S. (2024). Safeguarding Encrypted Messaging Platforms from Voter Manipulation. New York University, Stern Center for Human Rights & University of Texas at Austin, Center for Media Engagement.
https://bhr.stern.nyu.edu/wp-content/uploads/2024/10/NYU-CBHR-Covert-Campaigns_FINAL-FINAL-Sep29.pdf

Sundar, S., Molina, M., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps?. Journal of Computer-Mediated Communication, 26(6), 301-319.

Trauthig, I. K., Martin, Z. C., & Woolley, S. C. (2023). Messaging Apps: A Rising Tool for Informational Autocrats. Political Research Quarterly, 77(1), 17–29. https://doi.org/10.1177/10659129231190932

Woolley, S., & Howard, P. (2016). Political communication, computational propaganda, and autonomous agents: Introduction. International Journal of Communication, 10.