White Paper: DEEP FAKERY—An Action Plan

Authors: David Chu[1], Ilke Demir[2], Kristen Eichensehr[3], Jacob G. Foster[4], Mark L. Green[5], Kristina Lerman[6], Filippo Menczer[7], Cailin O'Connor[8], Edward Parson[3], Lars Ruthotto[9], Amit Sahai[10], Jose Sotelo[11], Luca Venturi[12]

[1]Institute for Defense Analyses; [2]Deep Scale; [3]UCLA, School of Law; [4]UCLA, Department of Sociology; [5]UCLA, Department of Mathematics; [6]USC, Information Sciences Institute; [7]Indiana University, School of Informatics, Computing, and Engineering; [8]UC-Irvine, Department of Logic and Philosophy of Science; [9]Emory University, Department of Mathematics and Computer Science; [10]UCLA, Department of Computer Science; [11]Université de Montréal, Laboratoire d'Informatique des Systèmes d'Apprentissage; [12]New York University, Department of Mathematics

INTRODUCTION

This white paper came out of an exploratory workshop held on November 15-16, 2019 at the Institute for Pure and Applied Mathematics at UCLA. Represented at the workshop were members of the mathematics, machine learning, cryptography, philosophy, social science, legal, and policy communities. Discussion at the workshop focused on the impact of deep fakery and how to respond to it. **The opinions expressed in this white paper represent those of the individuals involved, and not of their organizations or of the Institute for Pure and Applied Mathematics.**

"Deep fake" technology represents a substantial advance on earlier technologies of image, audio, and video manipulation like photoshopping. It emerged from the recent deep learning revolution, especially the development of generative adversarial networks. It enables the efficient, computer-assisted production of highly believable audio and video in which real people appear to be saying things they never said and doing things they never did.

The application of deep learning to audio and video synthesis is not exclusively harmful; beneficial use cases arise in artistic, scientific, or even therapeutic contexts. We therefore propose the term **deep fakery** for a specific socio-technical configuration in which deep fake technology is used for malicious or anti-social purposes.

More broadly, the term deep fakery includes *any* use of this technology to produce deceptive, apparently authentic representations, whether text, images, audio, video, or online profiles. The term therefore encompasses not only fabricated videos but also fabricated online actions.

Fundamentally, deep fakery is a technology of human augmentation that enhances our capacity to produce alternate realities and pass them off as real.

When referring to a specific instance of deep fakery, we will use the popular term "deep fake." We emphasize, however, that we are only focusing on anti-social or malicious applications. The reader should parse "deep fake" as "malicious deep fake."

Combating deep fakery will be an arms race. On the one side stand those who want to unmask malicious deep fakes and prevent their spread. On the other side are actors who want to make malicious deep fakes more difficult to detect and contain.

What is a reasonable goal in this battle? What would victory look like? And who should be involved?

By analogy with many societal problems, e.g. crime, we propose two goals:
(1) Reduce the incidence of deep fakery and the harms it causes to a manageable level; and
(2) Provide redress for individuals, groups, and organizations harmed by the persisting level of deep fakery, and contain the adverse societal impacts.

We emphasize at the outset that deep fakery is a socio-technical problem that cuts across many disciplines. As our workshop confirmed, it is not merely a challenge for the machine-learning community. It demands a transdisciplinary research agenda with input from cryptographers, social scientists, and legal and policy experts.

We have organized this white paper into four sections: (1) A taxonomy of the harms caused by deep fakery; (2) an overview of incentives that affect deep fakery; (3) a catalogue of promising research directions needed in the battle against deep fakery; and (4) a brief action plan for researchers and policy-makers.

HOW DEEP FAKERY HARMS INDIVIDUALS, ORGANIZATIONS, GROUPS, AND SOCIETY

To assess the consequences of deep fakery and motivate an appropriate response, it is essential to develop a detailed taxonomy of its harms. These harms strike at multiple levels of social organization; we discuss each of these levels in turn.

The harms caused by deep fakery vary in their time profile. Some harms occur at breakneck speed (e.g., reputational harm from a widely-circulated deep fake); others may build slowly but

be long-lasting in character (e.g., the erosion of trust in video evidence). Recognizing these differences is pertinent to effective responses, though we do not explore the issue further.

Harms to individuals and organizations:
On the individual level, we have divided the harms of deep fakery into three categories: those that arise from being ***portrayed*** within a deep fake; those that arise from being ***deceived*** by deep fakery; and those that arise from others who ***have been deceived*** by deep fakery.

When an individual is ***portrayed*** in a deep fake, a variety of harms can follow. In some cases, other people come to believe an individual has done things they have not in fact done. This could lead to reputational, legal, monetary, and personal consequences. In other cases, even if the video is not believed or is exposed as a fake, it can nevertheless change attitudes towards the targeted individual. For example, in the case of deep fakes that involve non-consensual pornographic depiction, viewers could variously see the person portrayed as a victim, a sex object, or inherently tied to a salacious context. In either scenario, there can be emotional harms like shame, distraction, and distress, as well as harms to self-identity. There can also be related financial costs, such as legal costs to protect reputation, or therapy costs for the harmed individual.

Individuals can be harmed when they are ***deceived*** by deep fakery. As a result of this deception, they may fail to act in their own best interests. This can lead to reputational harms and harms to relationships. To take one recent case, a UK-based CEO was fooled by the faked voice of his superior requesting a money transfer, incurring both reputational and financial harm. In addition to these direct harms of deception, there may also be psychological harm or shame that results from the experience of being deceived.

Finally, individuals can be harmed by the actions of those who ***have been deceived*** by deep fakery. In extreme cases, a crowd convinced an individual has committed a crime could directly harm the targeted individual. In other cases, inflamed passions may play out online. On the flip side, legitimate audio recordings and videos might be distrusted because of the existence of deep fakes; this could lead to evidence being discounted by jurors, harming one or more parties in criminal or civil suits.

Many of these harms to individuals can also occur to organizations, like businesses, political groups, etc. For example, misinformation disseminated via deep fakery may lead an organization to take actions that are not in its best interests.

Harms to groups:
Special harms can accrue to groups or categories of individuals as a result of deep fakery; these include racial or ethnic groups, religious groups, people with specific sexual orientation or gender identity, children, teenagers, or the elderly.

There are several issues here. Some groups may be targeted because of their historically disadvantaged status. Prevalent stereotypes may be reinforced and other reputational harms may accrue to a targeted group as a result of deep fakery; imagine the harms inflicted by a 21st century *Protocols of the Elders of Zion* supported by video "evidence" created with deep fake technology. Such reputational attacks may lead to direct physical, social, monetary, and legal harms.

There may be extra harms to groups that are especially vulnerable to particular types of deep fakery. Consider a few examples. Women are differentially targeted by pornographic deep fakes; this leads to the individual harms described above, but also harms women as a group. Extrapolating existing trends, we note that teenagers may be distinctly vulnerable to cyber-bullying and reputational harms enhanced by deep fakery, with social, psychological, and physical consequences, while the elderly may be distinctly vulnerable to financial fraud enabled by deep fakery.

Lastly, certain social groups who have not been treated as trustworthy or expert, including women and some racial minority groups, have benefited from the ability to deploy video evidence to support claims that might have been arbitrarily dismissed in the past. If trust in video evidence is degraded by deep fakes, such groups may be especially harmed.

Harms to society:
Beyond the damage to individuals, organizations, and specific groups, deep fakery harms society at large. Individual and social harms are interrelated, and many social harms piggyback on individual harms. Below we will discuss five of the more serious possibilities of this sort.

The first possibility arises from the ***social erosion of trust*** in video and photographic evidence as a result of deep fakery. This harm would manifest most obviously in a courtroom setting, but would by no means stop at the courthouse door. More generally, if people understand media as potentially fake, they receive less information from it, and will tend to update their beliefs less strongly in response to such evidence. In the long run, this can lead to a more poorly informed public, who are less trustful of good evidence. In addition, social knowledge biases, such as conformity bias (a bias toward the most prevalent belief), tend to be stronger in the

absence of good evidence; if all evidence (even good evidence) is downgraded by the existence of deep fakery, social knowledge biases may play a larger role in belief formation, with detrimental societal effects.

Second, ***political polarization*** has been an increasing worry in the social media era. Deep fakery may contribute to this problem in several ways. The erosion of trust in evidence may impact reasoning about evidence by polarized groups (especially evidence marshaled by the "other" side). Second, deep fakery may be deployed by actors that want to increase partisan anger or division, either to advance their electoral prospects or to undermine trust in democracy.

Third, there may be distinct harms as deep fakery affects the ***workings of electoral systems***. These systems function to aggregate judgments across many individuals. Once those judgments are contaminated (directly or indirectly) by deep fakery, electoral systems may fail to accurately reflect aggregate preferences. They may also lose their legitimacy. Given the importance of government in shaping the lives of constituents, serious harms may arise.

Fourth, there are substantial ***national security harms*** that can arise from deep fakery. For instance, in Gabon an (apparently) fake video of the president was used to justify an attempted military coup. In the past, wars have been sparked by misinformation ("Remember the Maine" and the Spanish-American War, or the Gulf of Tonkin incident and the Vietnam War). Deep fakery could contribute to future such events.

A final harm from deep fakes arises from the ***need to suppress*** or avoid them. In similar cases, we have seen arms races between groups attempting to deceive and those attempting to protect public belief; information warfare and enemy propaganda provide obvious historical examples. Such arms races can be costly — including time costs and money costs. Furthermore, the new status quo can be less efficient as a result of the arms race—consider the costs of increased security at airports in terms of time and money.

INCENTIVES RELEVANT TO CREATING, SPREADING, AND DETECTING DEEP FAKES

The incentives that affect the spread of malicious deep fakes and the degree to which they are mutable vary considerably by both the type of activity and the type of actor. Here, we roughly distinguish between three main activities: creating, spreading, and detecting malicious deep

fakes. We also distinguish between three main types of actor: (1) individual people or organizations; (2) platforms (especially social media companies)[1]; and (3) governments.

We decompose incentives into two broad types: costs and benefits. These can take many specific forms (financial, social norms, time, computing resources, or legal sanction). When the benefits of a behavior outweigh the costs, then agents are more likely to engage in that behavior. The incentives facing actors vary by region, especially when it comes to regulation. Our discussion is mostly focused on the US, though developments in other regions will be noted.

Creating:

The democratization of technology for creating deep fakes maps onto a substantial decrease in the direct costs (in time, computing resources, and expertise) of creating malicious deep fakes. There are also very few negative sanctions for having created a malicious deep fake. For example, there are few legal frameworks for punishing anti-social creators, and the exposure to social sanctions is low. On balance, the costs are low and getting lower.

By contrast, there can be substantial socio-economic benefits for individuals, companies, or other organizations to create deep fakes, such as the power to sway public discourse, promote products, and get revenge on a targeted individual (e.g. in the case of deep fake porn). While it may be possible to limit the force of these incentives, they cannot be eliminated wholesale. In general, the benefits of creating a malicious deep fake can be decreased by changing the individual reception of deep fakes, so that they are believed and acted upon less frequently.

There are also substantial benefits for developing the technology that generates deep fakes. The technology has many appealing and amusing applications; it can be used to insert Nicholas Cage into every movie! Its scientific applications are also quite profound; they are intimately connected to building better generative models of natural phenomena of interest. This is not the case for scientific applications of the technology involved in detection of deep fakes.

Since the direct costs of generation are falling and the various benefits are less easily manipulated, the easiest way to reduce creators' activity may be stronger and more effective

[1] We use this term in its broad sense to refer to a particular business model and accompanying socio-technical configuration, rather than the more particularized legal or regulatory term of art. By platform we mean the corporations that provide the technical and social infrastructure for deep fakes to spread; prominent examples include Facebook, Twitter, YouTube, etc.

sanctioning and legal accountability (i.e., increasing the costs of having created malicious deep fakes). Modifying such incentives ultimately falls to platforms and government actors; the former might ban users who create malicious deep fakes, and the latter could create laws encouraging action by platforms or directly sanctioning individual or corporate creators.

Spreading:
There are benefits for individual, corporate, and political actors who spread deep fakes. For example, spreading deep fakes can increase the number of followers or 'likes' on social media platforms, which can improve social status or (in the case of influencers) translate to financial rewards. It can also suppress voter turnout for a political candidate. Insofar as deep fakes promote user engagement, there are also incentives for platforms to permit their spread.

Given that there is no firm legal framework that regulates the sharing of malicious deep fakes, there are not many incentives to discourage their spread. For example, currently, there are only weak rewards (e.g., peer approval) for individuals that label or report deep fakes.

Creating stronger social and professional norms that favor high-quality and authentic information over information that is emotionally charged or confirms our biases could disincentivize people from spreading deep fakes. Platforms could tweak their ranking algorithms to decrease the reach of information posted by accounts that have previously shared or reshared deep fakes, creating an incentive against spreading. Financial incentives could also be created, aimed at individuals or platforms; individuals might receive a small "bounty" from platforms for reporting a fake instead of spreading it, while government actors could fine platforms for not taking reasonable action to discourage the spread of malicious deep fakes (given an appropriate regulatory framework). If information about reports of deep fakes were publicly available, then tech companies would face social pressure to respond quickly, lest they be accused of inaction.

Detecting:
Incentives to detect deep fakes are comparably low. Automated detection is difficult and manual detection is laborious, so costs are high. At the same time, there are meager social or financial rewards for detecting a malicious deep fake. Platforms could provide incentives to experts and non-experts to detect and flag false content (another version of the "bounty" mentioned above), while governments and platforms could periodically sponsor competitions to develop better algorithms for detecting deep fakes (as, in fact, they are currently doing).

Who watches the watchers? Platforms and governments:

Platforms (especially social media companies) are key actors in deep fakery. The relevant incentives vary tremendously. Traditional media, such as television or newspaper, have strong legal and financial incentives not to spread inauthentic information. Under current regulation, however, social media companies and other platforms are assigned a distinct legal status[2] that shields them from liability with respect to the content that spreads using their infrastructure. At the same time, a large portion of their revenue is obtained from targeted ads, whose value depends on the number of active users; anything that promotes engagement (which malicious deep fakes may do) is in their financial interest. These incentives could be changed, however, through government regulation, encouraging platforms to combat the spread of deep fakes. In addition to changing platforms' legal status to expose them to liability, other proposals include taxing targeted ads, which may result in movement toward other streams of revenues (e.g., subscriptions). If appropriately incentivized, platforms could in turn change incentives for individuals and organizations. For example, through well-designed mechanisms, they could incentivize users to label potential deep fakes. Because of their access to large amounts of data, platforms could also substantially improve our ability to detect deep fakes (as some are currently doing in the DeepFake Detection Challenge).

Governments can play an important role in combating the creation and spread of deep fakes through their power to shape incentives for individuals and organizations as well as platforms. Yet governments also face incentives that complicate their role. Some governments engage in propaganda and information warfare, and deep fake technology can substantially enhance these activities. Within their borders, however, governments usually have incentives to reduce the spread of false information, including malicious deep fakes, insofar as it is disruptive or destabilizing to the status quo. To this end, they can impose legislation to change the incentives for creators, spreaders, and media platforms. This can be challenging, as illustrated by proposals to amend CDA 230 while preserving freedom of expression and digital content creation.

Incentives and future research:

This discussion of incentives is deeply connected to questions in the field of mechanism design, and represents a promising target for research. Insofar as we better understand the

[2] This legal status is set out in Section 230 of the Communications Decency Act (CDA 230), under which "[n]o provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider." 47 U.S.C. § 230. Explication of who does (not) fall within these statutory definitions is beyond this white paper's scope, which references Section 230 to point out the affordances of certain entities' legal status as platforms.

constellation of incentives that lead individuals to create deep fakes, other individuals to spread them, and platforms to permit their dissemination, we can design alternative institutions and systems of incentives that curtail deep fakery.

PROMISING RESEARCH DIRECTIONS

We foresee three general domains of potentially impactful research: technical, socio-economic, and legal. Note that in all three domains, while certain disciplinary perspectives may be in the foreground, transdisciplinary efforts are required.

On the **technical** side, one may work toward improved capacity to detect deep fakes, whether by internal characteristics of the deep fake (e.g., identifying subtle features of the audio or video that reliably indicate manipulation) or by recognizing tell-tale signs of how it spreads (e.g., discovering distinctive spreading properties of synthesized as opposed to authentic media). Because these research streams are inherently subject to an arms race dynamic, it is also important to proactively develop cryptographic methods that can be used to produce audio and video recordings that can be cryptographically authenticated. Further research is needed into how to do this seamlessly and efficiently, while also preserving user privacy. This is especially true for authenticating processed video: while a device could potentially authenticate a raw video file, such a file would be too large for use. New efficient cryptographic authentication mechanisms are needed to efficiently authenticate compressed video files.

On the **socio-economic** side, one can try to understand better the dynamics of user behavior that impels users to participate in the spread of deep fakes (whether by creating them, sharing them, or believing them). One can study how to design mechanisms that change the incentives for individuals, groups, platforms, and governments to prevent this spread. Research is also needed to better understand and quantify the various harms produced by deep fakes.

On the **legal** side, one can try to get a better handle on which regulatory and legal policies will be effective in curbing deep fakery, and anticipate unintended negative consequences of such policies. Legal research is needed to see if it is possible to address issues raised by deep fakery within existing legal frameworks, or whether new domestic legislation or international law is called for. The existence of deep fakes raises issues about the admissibility and credibility of evidence in a courtroom setting, and these need to be addressed by bringing together technical expertise and legal knowledge.

We will now discuss these domains in greater detail.

Technical research directions: (Note: This section assumes a higher level of technical knowledge than the rest of the white paper.)

Background:
Recent advances in computing power, coupled with powerful deep learning methods, have enabled increased photorealism in computer vision and computer graphics. In particular, advances in the generative power, realism, and efficiency of variational autoencoders (VAE) and generative adversarial networks (GAN) for facial reconstruction and video synthesis have resulted in the emergence of image/video/text deep fakes and deep fakery. For example, the mainstream deep fake algorithm consists of two autoencoders, trained on source and target videos. Learning algorithms use data to train the parameters ("weights") of the autoencoders. Keeping the encoder weights similar for both source and target ensures that general features can be embedded by the encoder, while face-specific features can be integrated by the decoder. This setup allows expressions to be transferred from the source image to the target, even if the target has never been observed making that expression. Other approaches (e.g., Deep Video Portraits and face reenactment) follow similar methodologies but employ GANs or blendshapes instead. The results of these approaches can be very realistic; there may, nevertheless, be defects like skipped frames, tone mismatch, boundary artifacts, and face misalignments due to illumination changes, occlusions, video compressions, and sudden motions. Such imperfections, while imperceptible to the human eye, provide a potential basis for detection methods.

We will discuss three general categories of detection methods, based on content, patterns of sharing, and anticipation of new generative approaches.

Content-based detection:
Image forensics is an active area of research, which has developed several approaches for detecting inauthentic content that exploit compression artifacts, distortions, and image quality. Unfortunately, for synthetic images and videos produced by deep fakery, such generative artifacts are harder to exploit, due to the non-linearity and complexity of the learning process. For detecting deep fakes, there are pure machine-learning-based approaches (i.e., training classifiers); approaches based on detecting facial attribute manipulation (e.g., blinks, symmetry); and approaches based on biological signals (e.g., heartbeat). Because of the intrinsic arms race dynamic, more research effort must be spent on detection methods as ever more realistic deep fakes emerge. There are many promising directions here: using different signals (biological, physiological, temporal, spectral, residual, generative, etc.), different

features of these signals, different priors (symmetry, artifacts, facial attributes, etc.), and complex classification algorithms.

Sharing-based detection:
Another technical research direction explores the spread (or virality) of content; in particular, it asks whether the way deep fakes spread is different in character from how other content spreads. An obvious hypothesis would be that "cheap" fakes (content modified using simple methods) spread misinformation by exploiting cognitive and social vulnerabilities of humans, whereas photorealistic deep fakes follow a more sophisticated path, similar to authentic content. This direction also includes research on developing algorithms that slow down the spread of deep fakes. Second, researchers should investigate why particular contexts are more suitable for the rapid spread of deep fakes. For example, deep fakes are rarely used for financial fraud (so far), yet they are widely used for pornographic content, which can spread rapidly. Third, further research is needed on indicators that allow automated recognition of accounts that are producing or promulgating deep fakes. Lastly, methods that automatically assign and update a credibility score for media sources would quantify the expected authenticity of content shared by those sources. To the extent that such a score could be computed for any spreader (e.g., any node in the social network), detection algorithms could be augmented by considering the credibility scores of those sharing or resharing a piece of information. Spreading deep fakes (either malicious deep fakes, or benign ones that are not flagged as such) would damage this score. Notice that many of these research directions dovetail with ongoing research into other forms of online misinformation or disinformation.

Prediction-based detection:
We also suggest that more powerful generative models will lead to more powerful methods of detection. Although it may sound paradoxical, research devoted to making better deep fakes and more complex deep learning architectures will be essential to developing countermeasures. "White hat" actors must be at the cutting edge of research into generating deep fakes. Even if existing detectors perform with high accuracy, we must be in a position to anticipate, detect, and take countermeasures for the next generation of deep fakes.

Prevention via authentication:
We expect that there will be an ongoing arms race between those who want to detect and curb (malicious) deep fakes and those who create and spread them. As a first step, it will be essential to make use of existing cryptographic methods to embed a signature or watermark that allows for authentication of an audio or video recording at the time it is produced. It is also essential to develop these methods further.

In practice, authentication and watermarking of original content at capture time needs to be deployed proactively at the hardware level; for that reason, hardware vulnerabilities must be taken into consideration. Authentication processes must also be carefully designed so that it is computationally infeasible to create a deep fake that bears an "authentic" watermark. The watermarking process should also be computationally efficient. This poses an important research question in cryptography since the current computational overhead is not realistic for everyday use and widespread deployment. At the same time, authentication techniques should be developed to protect user privacy, so that total (and undesirable) transparency isn't the price of victory in the war on deep fakery. Finally, it is essential to develop authentication approaches that limit the vulnerability of trusted authorities, perhaps through a distributed approach like the blockchain.

To leverage the full power of cryptography, serious effort should be put into creating a taxonomy of threat models, including type of fake media, type of threat, type of vulnerabilities, type of technical specifications that can be exploited, and parameters and evaluation metrics to measure the effects of these attacks. Prevention that works for one threat model may not work well for another.

Socio-economic research directions:
There is also a substantial research agenda for the social, behavioral, and economic sciences around the spread and consequences of deep fakes.

Cognitive vulnerabilities to deep fakery:
First, cognitive scientists and computational social scientists should explore the socio-cognitive biases and heuristics that support the spread of deep fakes. Many of the biases and heuristics that have been explored in the spread of misinformation are likely to apply in this case; it would be valuable, however, to understand whether deep fakes interact more or less strongly with these biases and heuristics. In particular, information presented in visual rather than textual form may trigger different cognitive processes. Connecting earlier work on socio-cognitive biases and misinformation to scholarship on the processing and memory of video media provides one promising angle. It would also be valuable to know if there are group characteristics (e.g., age, education, or socio-economic status) that increase vulnerability to deep fakes. Finally, it will be important to explore the ways that deep fakery changes individual behaviors, beliefs, and preferences. This research stream would require substantial collaboration with experts on the creation of deep fakes, as experimental investigations should be as realistic as possible and hence leverage best-in-class deep fake technology while manipulating relevant variables.

Societal reaction to deep fakery:
Deep fakery has the potential to impact society in multiple ways, many of which are only beginning to be assessed. There is a broad research agenda here, which might include survey research and interviews (to capture awareness of and response to the existence of deep fakery, or the effect of deep fakery on trust in evidence); analysis of the demographic characteristics of those targeted (or fooled) by deep fakery; and estimates of the electoral consequences of specific instances of deep fakery.

Issues in mechanism design:
Deep fakery can also be addressed from the perspective of mechanism design. How can principals design incentives for agents at multiple levels (e.g., governments for platforms, platforms for individual users) so that desirable behavior is favored? As noted in the earlier section on incentives, research is needed into what mechanisms are likely to be effective in reconfiguring incentives of users, platforms, and governments to reduce deep fakery to a more manageable level. Experimentation with possible mechanisms is needed before they are deployed on a widespread basis; this provides a rich vein for researchers in experimental computational social science to mine.

Consider one interesting set of questions about mechanisms: Can professional standards for individuals, companies, and institutions be effective in curbing behavior that promotes deep fakes? Can one foster social norms that would create expectations to, for example, identify bots, label deep fakes, or eliminate inauthentic accounts? Could there be peer or institutional pressure to nudge people to act against creating or spreading harmful content? Are there existing norms or professional standards in this space useful for individuals and institutions to adopt or adapt?

Historical evolution of deep fakery:
Scholars of science, technology, and society can contextualize deep fakery by comparing it to earlier cases of technological change, particularly change surrounding the production and spread of media. Examples include the advent of radio, telegraph, and television. Work along these lines could explore cases where negative social impacts have been induced by technical change, versus cases where they have been mitigated by societal adaptation.

Education for deep fakery:
Building on the substantial literature on correcting misinformation, scholars can explore strategies for more effectively educating the public about deep fakery and alerting users to deep fakes when they occur. Is it possible to deploy this educational campaign without eroding

trust in evidence more generally? Can interventions be designed that avoid "backlash" effects, where notification about individual cases may actually increase belief in the flagged content?

Understanding the harms caused by deep fakery:
We have given an overview of the harms caused by deep fakery. We need to understand and when possible quantify these harms. We also need to understand what interventions can help individuals, organizations, groups, and society to recover. For example, it is clear that being portrayed in deep fake pornography is profoundly harmful, but we need a better understanding of how this affects an individual over the longer term, and what interventions and assistance might help those affected in moving forward with their lives.

Legal research directions:
The US legal system is highly adaptable, but it is clear that legal research is needed to address the issues raised by deep fakery.

Some of the work in this domain is descriptive and translational. For example, research is needed to describe in legal terms how the harms from deep fakes are either analogous to, or distinct from, existing legal categories.

Legal research might be especially fruitful in the context of trials. Research is needed for how to discuss the admissibility of evidence (in the legal, not colloquial sense) in court in the era of deep fakery. It would be helpful to create roadmaps for attorneys with the questions that should be asked in discussing whether something is, or is not, a deep fake, and to provide a better sense of what was technologically feasible at the time an alleged deep fake was created. Systemically, it is prudent to consider whether the Federal Rules of Evidence themselves should be adjusted to account for deep fakery, or whether it is possible to educate legal actors and thereby fold any developing technologies into existing rules and procedures.

Legal processes will also require input from technical experts. For example, court proceedings may require affidavits from technical experts explaining the technical feasibility of creation of deep fakes or other forms of manipulation, and also the state of the art in detecting deep fakes.

Deep fakes touch on areas of law as diverse as election law, consumer law, torts, criminal law, administrative law, and international law. The implications of deep fakery for these legal areas, as well as the law's role in addressing deep fakes and the harms they cause, are ripe targets for legal scholarship.

A PLAN OF ACTION

Any plan of action will involve taking some actions immediately, while ramping up research into methods that can be deployed as they are developed. A successful plan will involve a mix of various toolkits: technical, socio-economic, and legal.

1. Across these domains, the research described in this white paper needs appropriate levels of funding.

2. It is important to rapidly and widely deploy recording devices for images, audio, and video that have cryptographic authentication built into them. As research proceeds, it will be possible to do this more efficiently and in ways that dovetail better with existing means for information to flow onto and around the internet. It is also essential to develop approaches to authentication that preserve privacy and limit dependence on centralized, trusted authorities; hence, distributed solutions should be explored.

3. Methods of detection should be made a priority, and researchers should experiment with a wide variety of approaches. Contests between algorithms and modalities have been effective in other arenas in accelerating progress, and should be used here as well (e.g., the recently launched DeepFake Detection Challenge).

4. Novel ways to modify incentives should be tried, as described in this white paper. Such incentives might be social, economic, or legal. Initial theoretical work in mechanism design should be followed by experimental deployment (on academic or actual platforms) before widespread use.

5. The tools of social science and network science should be employed to understand and curtail the spread of deep fakes.

6. It is important that any regulatory and legal approaches that are pursued be informed by a thorough understanding of how deep fakes are created, detected and spread, and should involve input from experts. Legal issues related to evidence and torts should be addressed in a context that is informed by input from experts. This should be initiated before deep fakes become too frequent.

7. The internet has changed how the public accesses information and news. Education should include instruction in how to be media-savvy and avoid being duped by deep

fakery. Because detection of deep fakes may require extensive technical expertise, the public should be educated to trust non-partisan experts rather than simply distrust all media, or decide whether to trust a piece of evidence on the basis of their previous beliefs. This will require education about confirmation bias and other cognitive biases that increase our vulnerability to deep fakes.