I Wasn't There: Applications of Blockchain to Privacy Preserving Reality Protection

Jacob N. Shapiro,[1,*] Kaya Alpturer,[1] Aadityan Ganesh,[1] Xilin Yang,[1]

[1]Princeton University

[*]To whom correspondence should be addressed; E-mail: jns@princeton.edu.

This version April 15, 2024

1 Executive Summary

Advances in the capabilities of generative artificial intelligence (GenAI) to create realistic inauthentic content turn digital media records into an unreliable evidence of actual events. We organized a workshop with experts from diverse backgrounds assembled for a day-long workshop to think through how blockchain technologies could help address the resulting challenges. The main points emerged from that workshop are summarized below:

- Generative AI models such as DALL-E, StableDiffusion, and Midjourney have advanced significantly and can now produce realistic media that challenges content provenance techniques. The misuse of generative AI introduces serious risks, including the creation of non-consensual intimate imagery and fabricated scenarios that affect public discourse and individual privacy.

- Some efforts currently exist to address generative AI challenges. The Content Authenticity Initiative (CAI) and complementary Coalition of Content Provenance and Authenticity (C2PA) focus on setting the technical standards to provide greater media transparency. CAI was launched to build a system that attaches attributional details to digital content. C2PA is a broader initiative which aims to establish technical standards for certifying the origin of a media.

- Blockchain technologies introduce the possibility of providing a decentralized platform for claim authentication, paralleling legal processes by allowing claimants to present evidence and validators to adjudicate, thus standardizing proof submission. Digital signatures play a crucial role in verifying message senders' identities on the blockchain, employing public and secret keys to authenticate transactions and contracts securely.

- Two primary proof types can be utilized for blockchain authentication: cryptographic proofs for immutable verification based on security assumptions, and game-theoretic proofs that rely on economic incentives to discourage false claims.

- Two potential strategies to combat GenAI-produced deepfakes and false media include content provenance and content disprovenance. Content provenance focuses on measures that verify the content as genuine. However, such solutions may unintentionally result in digital surveillance states or digital surveillance capitalism. Content disprovenance focuses on debunking or question the authenticity of generated content, which will face technological limits as AI content generation tools advance.

- Addressing deceptive generated media and enhancing content integrity in digital media necessitate a blend of human and algorithmic verification, relevant education, and the development of resilient technologies to protect people against sophisticated manipulation tactics.

2 Introduction

This white paper reports on and extends the discussions held at a workshop organized at the Princeton Center for the Decentralization of Power Through Blockchain Technology (DeCenter) on *Decentralized Reality Provenance*. Organized on December 4, 2023, the event brought together experts from the fields of computer science, public policy, and philosophy to discuss what comes after the advances in the capabilities of generative artificial intelligence (GenAI) that make digital media unreliable evidence of actual events, given GenAI's ability to create realistic inauthentic content.

Given the multifaceted nature of the problem, the event encouraged an interdisciplinary dialogue on (1) what kinds of tools are needed to differentiate inauthentic, GenAI-produced media from authentic non-GenAI media, and what the current capabilities of blockchain are to enable people to disprove claims about their activities, (2) what processes for deploying such solutions might look like, and lastly (3) what challenges exist for current and potential solutions and what these challenges imply for society in general.

2.1 Meeting agenda

The workshop was organized in three sessions. Each session involved three concurrent 45 minute discussions by a mix of experts in three separate tables. At the end of each session, the discussions were reported to everyone and the experts rotated to form three new groups at each table. The sessions themes were chosen to be the following.

Session 1. What are the threats from manipulated media which content provenance will not solve?

- Table 1. How manipulated media is contributing to epistemic decay.
- Table 2. On-chain evidence and off-chain truth.
- Table 3. Coming challenges to media integrity.

Session 2. How can we leverage decentralized technologies to stave them off?

- Table 1. Limits of media forensics and the role of semantic information.
- Table 2. Decentralized approaches to truth telling.
- Table 3. Human-machine interaction in detection.

Session 3. What are the business, legal, and technological enablers we need?

- Table 1. Federal and Private Sector Roles in Enabling Information Markets.
- Table 2. Approaches to manipulation resistance.
- Table 3. Technological enablers of democratic discourse.

The discussions were held under Chatham house rule and the workshop concluded with a reception.

3 The threats from manipulated media

3.1 Capabilities of generative AI

GenAI includes tools that are designed to create new digital content based on natural language prompts. Such tools have scaled rapidly when it comes to the quality of the generated content. Specifically, most of these tools are now convincingly creating seemingly authentic images, audio and natural language that require expert review to differentiate from authentically produced media. Moreover, these tools are also quite capable when it comes to manipulating media by generating content that builds on real images/recordings. It is highly likely that this technology will continue to advance in its capabilities, making media provenance more challenging. We briefly list below and elaborate on the state-of-the-art capabilities of these tools:

- *Image/video/audio generation:* Most recently the transformer models that led to the training of models such as DALL-E, StableDiffusion, Midjourney, and many others have been able to make the task of generating convincing realistic images from natural language prompts and existing image/video readily accessible to the public. Similar tools have been trained to allow voice synthesis through text-to-speech models. A particular use case that warrants a lot of concern is the ability of these tools to create deepfakes, defined as false content designed to imitate the appearance/voice of a targeted individual.

- *Recent developments in natural language processing:* Another frontier of generative AI is the development of large language models (LLMs) that are designed to generate language content in any context. Given the speed at which these tools can generate language that appears to be written by humans, this technology makes automated content generation at unprecedented scales possible. By themselves, LLMs do not specifically verify that what they write is true - similar that what they write is the best fit between data they have been trained upon relative to the prompt provided. This raises similar use case concerns that LLM generated text could mislead or intentionally deceive humans into thinking what they are reading is authoritative information.

3.2 Potential harms

There are several potential malicious applications of GenAI, some of which have started to have a noticeable impacts both on societies and on public discussions. As noted, the main avenues for harmful content generation currently include the creation of image/video or sound of an individual doing or saying things that did not actually happen. While this can be used to target high-profile individuals such as politicians and celebrities, it also can be used to target average individuals most commonly so far in the form of non-consensual intimate imagery. Non-consensual deepfake imagery disproportionately affects women. Some examples of harmful GenAI content uses include the following examples below, with a list of specific incidents is provided in Appendix B.

- Attacks on civic or corporate leaders;
- Attacks on teachers/professors;
- Bullying/harassment of individuals in schools/workplaces;
- Election interference;
- Misappropriation of identity for advertising, brand damage; and
- Insurance fraud

3.3 Need for a solution

From there, the workshop turned to a discussion of blockchain technologies' potential for providing a decentralized platform for authentication. The majority of the work on this topic so far has revolved around content provenance and the creation of tools/policies to confirm that a given piece of content was created and/or edited at a specific time or place. The Content Authenticity Initiative (CAI) and complementary Coalition of Content Provenance and Authenticity (C2PA) exist to provide tools and methodologies for watermarking media and storing a publicly accessible record of its creation and edits. These efforts are led by companies that work with digital media. Content disprovenance, which we define as proving that a given event recorded in digital media specifically does not represent a true event, has attracted much less attention in this discussion section. Any potential solution will likely require a transition period of adoption, and the solution will be a result of an evolving effort and will have limitations itself during this period. There will be many challenges involved in this process, including and not limited to how the authentication process is going to work and what will need to be communicated to the public with regard to the limits of such technology. We cover these in more detail in Section 5.

4 What can blockchain do for truth?

We have discussed the efforts currently exist to address generative AI challenges. From this section, our workshop report dives into the nuts and bolts of what blockchain can do to deploy these efforts. In 4.1, we use a running example to analyze how blockchains authenticate or disprove claims. In 4.2, we introduce two general strategies of proofs that can be used to prove validity. We then discuss the examples of finance in 4.3 and the examples of media in 4.4, considering the use of these strategies.

4.1 How can blockchain authenticate a claim

The blockchain plays a dual role in authenticating the validity of a claim by acting as a channel for the claimant to showcase proofs supporting their claims (similar to the role of lawyers in a court of law) and act as a platform for validators to weigh the evidence presented by the claimant and rule on the validity of the claim (similar to the role of the jurors in a court of law). Apart from creating a decentralized platform to facilitate authentication of claims, it further makes the process accessible to claimants by standardizing the protocols required to furnish the associated proofs.

As a running example, we use the problem of establishing the identity of the sender of a message. The sender would like to establish that it was not impersonated while sending the message. Such a claim is fundamental to all the decentralized finance tools built on top of the blockchain. If two parties are signing a contract on a blockchain, the parties should not be able to void the contract by claiming they were impersonated later on. The blockchain lifts the burden of creative lawyering in establishing such basic claims through standardizing the protocol to prove correctness of the claim.

The above challenge is tackled through the use of digital signatures. A digital signature scheme consists of a public key (which is made available to everyone, by printing on a public directory) and a secret key (known only to the sender). To establish identity, the sender encrypts a message using its secret key and sends both the message and the encrypted text, the so-called 'signature' of the sender for the message. The validity of the sender of the message can be verified by decrypting the signature using the public key and checking if it matches with the original message that was sent along with the signature.

In most of the applications that we would be discussing (including the running example), the proofs submitted through the standardized procedures prescribed by the blockchains is assumed to be true unless proven otherwise. Thus, it should be extremely straightforward for the validators to disprove a false claim. In the running example, to disprove a claim, the validator would have to decrypt the signature sent by the sender and check that the decrypted text matches the message, which can be done efficiently even by a standard laptop. Additionally, the protocols should be such that generating a false proof that would give a false positive in the verification protocol followed by the validators should be infeasible. For instance, it is impossible for an impersonator to compute the signature of a message without knowing the secret key except with a negligible probability, even if it has seen the signatures of lots of other messages signed previously by the sender. This way, generating a false claim that would throw the verification protocol haywire (in this case, generating the signature that produces a false positive from the identity authentication protocol) is extremely improbable.

Even if the proofs are submitted on-chain, in the situation where there is no centralized organization paid to run the ledge the validators still might not run the verification protocol and produce the necessary certificate verifying the authenticity of the claim. Thus, for decentralized solutions the validators need to be incentivized through embedding various payoffs for highlighting false proofs submitted by claimants. As we will see in Section 4.3.2, designing the protocol so that the incentives of the validators is aligned with verifying claims truthfully is an interesting challenge on its own.

Even for non-standard claims without a standardized protocol prescribed by the blockchain (see Section 4.4.2 for such an example), assigning the burden of proof to the claimants makes the authentication process much more streamlined. For instance, validators can hire impartial experts with access to the proofs, which would be impossible if the burden of proof lies with the validators.

The claims can range from being completely off-chain, to being a mix of off-chain and on-chain information, to being completely on-chain. Proving that a wallet has a residual balance larger than some threshold would entail tracking all recorded transactions performed by the wallet, all of which are recorded on-chain (4.3.1). On the other hand, proving the validity of a video posted on the blockchain would involve corroborating some form of off-chain evidence like the advise of a video forensic analyst (4.4.2).

4.2 Cryptographic and game-theoretic proofs

In almost all the examples considered below, a mix of two kinds of proofs are used to prove validity. The first is cryptographic in nature. As the name suggests, the guarantees on the validity of the proof holds as a consequence of the cryptographic tools involved. Voiding the validity of such proofs would also imply voiding of the guarantees given by the underlying cryptographic tools. However, under some mild computational assumptions, like the hardness of solving various intractable problems in efficient time (e.g, the discrete log), the underlying cryptographic tools are secure and thus, the guarantees carry-over to the authentication procedure on the blockchain. The second kind of proofs are game-theoretic in nature (optimistic proofs in the blockchain parlance). In a game-theoretic proof, all validators and claimants are assumed to be utility-maximizing. Even though the submission of false claims is quite easily possible when the guarantees are game-theoretic, the claimant trying to publish a false proof is penalized in the equilibrium induced by utility-maximizing agents.

We discuss applications in the following section.

4.3 Examples from Finance

4.3.1 Balance verification

We begin by discussing an example that admits a proof that is purely cryptographic in nature. Suppose that the claimant wants to establish that its wallet has a residual balance of at least some threshold. Such a claim is necessary whenever two parties conduct a transaction, where the receiver has to ensure the sender has the necessary balance in its wallet to process the transfer.

The proof for this example is simple. The sender would just have to broadcast its wallet ID and the balance that it would want to establish. Validators can go through all transactions recorded in the previous blocks to track the flow of money from the sender's accounts and confirm the claim if the balance in the sender's wallet is at least the claimed threshold.

The contents of the blocks in a blockchain are immutable since the hash of a block is stored in its successor. If the contents of a block are changed, so would its hash, flagging a contradiction to the hash stored in its successor. It is computationally impossible to compute an alternate set of contents to go on the block without altering the hash. Therefore, the claimant would not be able to change the set of previous transactions recorded on the blockchain and thus, will not be able to force false positives through fake proofs.

4.3.2 Verifying the market price of a cryptocurrency

Next, we discuss an example of a proof that is primarily game-theoretic. Consider constructing a proof to establish the market price of a particular cryptocurrency. Note that organizing a majority vote would not be fruitful, since voters in a blockchain typically tend to hold a non-trivial amount of the cryptocurrency and would find it in their best interest to inflate the price of the currency in the vote. Further, relying solely on cryptography is impossible since establishing the price requires information exogenous to the blockchain.

The following proof establishes the price of a cryptocurrency assuming that the claimant is strategic. The claimant posts a claim on its willingness to trade a non-trivial quantity of the cryptocurrency at the believed market price. If the claimant quotes a price above or below the true market price, an arbitrage opportunity is created, and the claimant faces a significant loss when the arbitrage is liquidated by arbitrageurs. Indeed, a layer of cryptography is required on top of the economic incentives so that the agent does not back out of the trade when someone tries to capitalize on the arbitrage. This is done through a smart contract (a contract signed by the claimant on-chain). The smart contract locks the amount of cryptocurrency or the equivalent USD that the claimant promises to trade. It further makes sure that the potential opportunity to arbitrage stays live for some time by not releasing the capital that has been locked up until a certain number of blocks have been proposed since the smart contract was set up. For example, a new block in Ethereum is proposed every 12 seconds. Locking up capital for 300 blocks roughly corresponds to keeping the arbitrage opportunity alive for an hour, and would provide sufficient time for arbitrageurs to liquidate the opportunity if the price is indeed skewed.

This is very similar to the approach used in *automated marketmakers* like UniSwap. Automated marketmakers are on-chain currency exchanges. The price of a cryptocurrency is determined by a publicly specified on-chain exchange rate. The price might not match the true (market)price, creating an arbitrage opportunity. Liquidating the arbitrage on-chain would drive the on-chain price of the cryptocurrency towards the off-chain marketprice, paving way to efficient price discovery on-chain.

4.4 Examples from Media

4.4.1 Format of storage as a method of authentication

As a consequence of standardized protocols for submission of proofs, proofs not in the required format can be safely ignored by validators, which could rule out another class of false claims. For instance, suppose a claimant want to establish knowledge of some information before a particular date.

This can be achieved by storing the information on the blockchain prior to the date. However, this would reveal the information to the world. To maintain secrecy, the claimant can hash the message and store the hash on-chain. Conditional on knowing only the hash, no additional knowledge of the underlying message can be learnt. To prove knowledge when finally revealing the message, the claimant has to revel the message along with a pointer to the block with the hash of the message. It is computationally impossible for the claimant to construct another message that matches the hash of the original message. Thus, revealing the message of a hash in a block published prior a date can be considered proof of knowledge prior to the date. The proof is fully cryptographic and needs no game-theoretic assumptions on the behaviour of claimants and validators. Further, any message without a published hash can be considered ineligible for a "proof-of-knowledge" claim.

This could be particularly be useful to show the existence of a piece of media without actually revealing the contents of the media. Wikileaks for instance, would commit to the hashes before revealing the entire document. Similarly, suppose a contractor wants to prove to a contractee that the agreed upon job has been completed, but is constrained in the ability to meet the contractor. Posting the hash on-chain would not reveal the nature of the job to someone other than the contractor and the contractee. Revealing the image later on, when meeting the contractee and verifying that the hash matches would be sufficient proof that the contracted job was completed on time.

Posting such hashes can also be used to generate loose alibis. For instance, posting the hash of an image of performing a task on the blockchain on a particular day is a sufficient proof of performing the task before the given day. To battle a claim of doing something malicious on the day the hash was posted, using the hash posted on the blockchain as a proof is not sufficient. Even though the hash was posted on the given day, the task could have been performed on some earlier day, and thus for full disprovenance another method of auditing the timing of the recording would be required. Even without that, such a proof may be helpful in the battle of public opinion.

4.4.2 Potential approaches to tackle deepfakes

In this section, we discuss a challenge for which accommodating a standard protocol to authenticate truthfulness is extremely hard. Consider establishing a piece of media (image or a video) is true and not a deepfake.

Establishing deepfakes as true has heavy economic incentives. For example, deepfakes exaggerating the extent of an accident could get the claimants a larger sum from the insurance agent and deepfakes projecting politicians in a wrong light would have non-trivial consequences in divesting votes from their parties. Cryptography, by its own, cannot be of much help, since the opinion of a forensic analyst is required in establishing the validity of a video. An economic solution is feasible if validators are paid to hire an expert on behalf of the blockchain. However, this is an instance of the free-rider problem, where validators might only pretend to hire an expert and vote randomly, thereby pocketing the fee that they were supposed to pay the expert. Setting up a reputation system with a higher payout of votes from a reputed voter would align the interests of voters to that of the blockchain – vote with an expert opinion. The payouts perform a dual role, ensuring a larger turnout in the majority votes (that increases the credibility of such elections) and aligning the interests of the voters and the blockchain in hiring experts. The protocol can also slash the funds of voters whose recommendation ended in the minority. In some extreme examples, false or irresponsible voting could be considered perjury and the threat of jail time could also align the incentives of the voters and the blockchain. However, care requires to be taken in such scenarios so that voters are not disincentivized to participate in the vote.

Another approach could be through the ideas suggested in Section 4.4.1. Additional information can be required to be stored alongside the media to prove its validity, failing which, validators need not even consider authentication. Time-stamping and source-stamping pictures could help in tackling the deepfake challenge. Suppose all pictures are tagged with a signature from the camera which captured the image. A deepfake would lack such a signature. Thus, checked the validity of signatures in such trusted hardware used to capture images could be a criterion for claiming truthfulness of pictures. In particular, pictures submitted for authentication without such a valid signature could be ignored by validators. Note that C2PA standards use a similar idea of binding timestamp and geographical location stamp and source-stamp along with the media, but is centralized. Thus, the records stored for content provenance can be compromised in C2PA if the storage entity becomes faulty.

A downside of this approach is that it threatens revealing the source of every picture requiring authentication. This could stifle the number of pictures that would be submitted to the blockchain for authentication. For instance, a whistle-blower would not want to have their signature identified and thus, would not publish pictures for validation. A careful trade-off between claims that require a specific format of data storage and the potential drop in the number of users due to the difficulties in preparing the data in the required form needs to be weighed in.

We summarize various approaches that can be taken to use in a evidence backed claim framework below in Table 1.

Claim	Evidence	Connection
Person A owns a token m	Transaction capturing the transfer of the token is recorded on the blockchain	Slashing validators if they voted for an invalid transaction (game-theoretic)
Person A sent a file m	Digital signature of m	Cryptographic
The price of a cryptocurrency is $\$X$	Smart contract signed by the claimant trading the cryptocurrency at $\$X$	Game-theoretic
A knew X before a particular date	The hash of X was on a block published before the said date	Cryptographic
A picture is not a deepfake	The camera capturing the picture has watermarked the picture	Cryptographic
A generic claim X	A majority vote on X	1. Slashing (game-theoretic) 2. Jail time (if perjury laws apply)

Table 1: Claims, evidences validating the claims and the nature of the guarantees on the evidence.

5 Content provenance, disprovenance and corresponding challenges

We open up this section for a wide discussion reflecting on the uses of technology and the implications of its uses. In combating content manipulation, we can take approaches either to prove that the content is true or to disprove the manipulated content. Content provenance is difficult in that it requires tying the creation to a certain identity, place and time. On the other hand, content disprovenance does not require disproving all of the information about the claimed event. It only requires showing the inconsistency of parts of the information. Each approach comes with certain challenges which we discuss below.

5.1 Content provenance: getting the audience to pay attention

In proving a content is true, workshop participants discussed a standard for assuring authenticity, which is to timestamp the creation. “In the content provenance system, you can record the creation at each point in time—stamp the time using C2PA protocol.” An additional approach is to promote cameras that timestamp the moment of photo-taking. “We want cameras that timestamp the moment of capturing an image, and these cameras may appeal to wartime journalists.” However, there is more than just developing a product. Thinking about applications, we may want content provenance for social media: If the newspaper won’t publish unless it sees it on a blockchain, then it is an added assurance for authenticity for public news.”

However, timestamping is essentially tying one’s identity to the creation. Participants pointed out that “over-reliance on identity will stifle the truth in certain contexts.” Tying identity to creations has a political risk if such creation threatens governing authorities. For example, if we have protesters recording suppression of dissent by autocratic government, by tying their identities to the creations, the protesters are exposed to high political risk.

Besides timestamping, participants mentioned that another approach in content provenance: adopt standardized data labeling. In AI-generated arts, for example, one could insert a watermark saying “this content is manipulated by model X.” In other types of content, the notions may come in different forms.

With these approaches plausible, a series of questions arise: When it comes to conveying messages to the end users, participants raised a range of educational challenges: “We need to get the education to both the young and the old.” But the challenge in content provenance is not just conveying information to users, but also figuring out whether users care about these labels: “To what extent is it clear to people and matters to people – do people care whether this is fake or not?” For example, users usually don’t want to hear “This is an automated climate report” at the beginning of a climate report. In addition, a discussant pointed out: “Adding signals, labeling data may cause distrust for the things unlabeled. (In some settings, this can be a positive thing.) Also, If I see labeled data but feel differently, I may think the label is not working.”

5.2 Content disprovenance: What do we need to to disprove claims made with manipulated or generated content?

There are multiple real-life examples where content disprovenance is strongly needed. Recently deepfakes imitating the current British prime minister Rishi Sunak were used in advertisements to interfere with public opinion. Similarly, AI-generated content proved itself to be an effective tool for market manipulation. Solutions that enabled rapid and persuasive disprovenance could have mitigated these harms.

We brainstormed a variety of ways we can distinguish any audio video or image as real or fake. “I would identify the sources to see whether it is coming from a reliable site.” One discussant brought up an interesting skill: “Experts in audio or photographs have built intuitions – they listen to so many real or fake audios that they can identify with their instincts. They identified the subtle differences and immediately recognized the authenticity of the image, but they had difficulty explaining this intuition. With fake audio or photos getting more and more advanced, this intuition will run out very soon.”

But is there a way for us to output these intuitions into research methods? There are many things exposing the fakeness of videos, such as biometric details, shadows, and cultural features. A model can be deceived, but people have senses. “Yet, people are responsible, and machines are not. With humans, you can always ‘shoot up the chain’ by contacting the person who is involved in the video. The problem with the algorithm is that it does not have responsibility.” We brought up the need for human verification to complement the algorithmic models in detecting whether a content has been manipulated. “We need a way to check humans’ work, and so far, we cannot do this without humans.”

Putting the challenge of content disprovenance in a larger social setting, participants discussed how social media companies grapple with the question of defining what is ‘fake’. “We can combine human detection with C2PA and add more context into tools like Community Notes. But there is an extra challenge for fact-checking platforms: sometimes the content is not totally fake, but misleading.” Participants also discussed developing capacity for building manipulation resistance in society. “To add authenticity and build trust in the market, we need to hire bigger teams. We want to enable a good information ecosystem to have someone point at a picture and say: ‘this is wrong,’ and have someone come out to correct this issue. The social media accounts of respectable people should ring the bell for us. We also want to introduce moral offsetting: have the companies be responsible and find solutions for the misinformation they caused.”

5.3 What technology cannot do right now

There are technical challenges for content provenance. In terms of using watermarks to label which images are AI-generated, Google DeepMind was the first big tech company to launch such a tool. This watermarking tool called SynthID embeds an invisible pattern in generated images. However, a recent study done by researchers at the University of California – Santa

Barbara and Carnegie Mellon University proposed a series of regeneration attacks and showed that “all invisible watermarks are vulnerable to the proposed attack” (Zhao et al. 2023).

A Further discussions from the workshop

A notable and optimistic thread in the dialogue is the application of AI in government initiatives, where it's being leveraged in cities and states to streamline processes from waste reduction to health benefits enrollment: "Look at Boston, New Jersey, and California, the government is using LLM for waste reduction. We can increase efficiency in the government agency by introducing process optimization. Think about the paperwork for Medicaid: how can language models alleviate some bureaucratic burden to help people enroll in a plan more suitable for their health demands? The number of people being supported by health benefits is an indication of democracy, and people can benefit from the process of registration for being more efficient."

Such use cases showcase AI's capacity to enhance the efficiency of government agencies, counteracting the stereotype of big, inefficient government structures. "The point is: that trying to change people always results in catastrophic consequences. Trying to make things more tailored to users is what this technology can bring to the table. This is the shortcut to resolving public trust." The conversation also touches on the dualistic influence of AI in different political regimes, contrasting its alignment with democratic values in some contexts against its use in autocratic settings. This raises broader concerns about the democratization of AI and its ability to empower citizens versus control them. "Promoting democracy using civic technology, we can use LLM to align AI with democracy."

Speaking of democracy, we brought up the implications for spam reduction. "Two commonly used ways to reduce spamming are adding permission and expenses – both have anti-democratic implications. Social media does both for us quite extensively. Spamming detection is not very subjective. For example, Google uses spam classifiers, and the machine detects common combinations of words and tones." Here, we suggest an important baseline in spam detection, which is to return the decision to the end users: "If I get a lot of emails from junk folders and see that they use spam-me content, I might still open them. I see certain content and decide to look at it. Later, it became my decision as a reviewer. To me, that should be an objective for information detection or filtering. Moving eyeballs off the centralized content and letting users decide. Giving the decision back to the end users is a way to decentralize."

B Known examples of deepfakes

UC Berkeley professor Hany Farid has created a tracker to trace election deepfakes. Recent news on New York Times also reported the instance of A.I. scamming that uses people's voices.

C Agenda for future work

A consensus was reached amongst the participants of the workshop on the importance of understanding the frontiers of decentralized verification of claims. The discussions provide an opportunity to consider a plethora of follow-ups on each of the challenges highlighted in the discussions.

For instance, understanding the role and limitations of technology in circulation and verification of news and media would be an interesting direction to explore. If a piece of media has to satisfy extremely hard constraints before being accepted as not-fake by the society, it places a strain on free speech. It would also make the use of such technology for purposes like whistle-blowing that require privacy quite challenging. On the other hand, making the constraints extremely lax could cause a widespread circulation of fake media, which needs to be avoided. Therefore, regulating digital media poses both technical and policy challenges.

The economics associated with provenance is another interesting direction. As discussed in the paper, game theoretic guarantees of validity are a non-trivial addition to the claims that can be authenticated only through cryptographic means. Apart from use of economics in provenance, the economics associated with the creation of false proofs and insuring against such claims are interesting directions to explore.