Polaris
Freedom happens now.

The Future of Human Rights on web3

A Vision for a Survivor-Centered Internet

JUNE 2022

By Anjana Rajan and Frederick Reynolds

EXECUTIVE SUMMARY

The internet, currently in its second generation, is a double-edged sword. It has driven transformative innovation that has greatly benefited society, but it has also been a conduit for the darkest sides of human behavior. In the last few years, the latter has become increasingly true; web2 platforms have enabled human trafficking, conspiracy theories, genocide, repression, and domestic terrorism. While employee whistleblowing, calls for federal regulation, and formation of oversight bodies have drawn long-overdue attention to this crisis, these alone will not address the root cause of the problem. These challenges are not just the actions of a few unethical individuals or companies; they are a result of a fundamentally flawed internet architecture and broken incentive structures that enable the powerful to exploit the vulnerable for profit. If we maintain the status quo, we will continue to see economic equality, human rights, and democracy crumble while financial crimes, violence, and fascism rise.

A new internet architecture is emerging that has the potential to solve these problems. This third generation of the internet, called web3, is built upon principles of decentralization, transparency, and cryptography. Unsurprisingly, the debate over web3 is already fraught. Advocates of web3 believe that it offers several key benefits: privacy and security of data, resilience against cyber threats, broader inclusion and trust in the global economy, and an opportunity to shift power back to the participants of a democratic society. Critics of web3 see it as a breeding ground for unregulated crime and get-rich-quick Ponzi schemes that can harm vulnerable consumers. Still, others fervently hope that if they just ignore it — or worse yet, ban it — web3 will just go away.

The truth about web3 lies somewhere in between these viewpoints. As human trafficking and financial crimes experts, we believe the critiques are valid because in its current state, web3 only offers the possibility, not the promise, of an open and equal society. Without native design principles of **ownership with consent, speaking truth to power,** and **privacy with accountability,** web3 will suffer the same fate of web2 and will fail to protect the most vulnerable people in our communities, including those who have suffered from sexual abuse, labor exploitation, racial and ethnic discrimination, government-sponsored brutality, or violent extremism. Survivors of these forms of trauma have always faced, and will continue to face, challenges to restoring their freedom and holding their perpetrators accountable for their crimes. Even as new web3 technologies emerge, there is a high risk that exploitation and abuse will persist unless we intentionally design the technologies to natively protect them and a free and open society.

But as cryptographers and technologists, we believe there is an opportunity to build a web3 that lives up to the proposed values and virtues that it can — and in fact, must — achieve. We can build web3 the right way if we combine a systems-level understanding of human rights with a technical understanding of cryptography. Building this correctly will not only empower vulnerable communities, but it can also serve as a barrier that blocks authoritarian regimes and criminal actors from weaponzing web3 technologies to oppress and abuse others. This ultimately keeps these malicious factions operating in the internet of the past, thus widening the future competitive advantage for democratic freedom, prosperity, and security.

The fight against human trafficking can serve as an inductive proof for all other human rights issues because it is deeply complex, highly pervasive, and historically nonpartisan. If we can design web3 for this particular vulnerable population, then we can scale this to other issues, such as terrorism, climate change, gender-based violence, public health, or election security. Polaris has created an ambitious ten-year vision to combat human trafficking that has three focus areas: sex trafficking, labor trafficking, and money laundering. By building off of this strategy and utilizing our expertise on applied cryptography, Polaris wants to contribute to the innovative discourse on this emerging technology.

Since web3 is still in its infancy, this paper offers three thought experiments that identify new and native design principles:

- **Ownership with Consent:** Fighting sex trafficking by building NFTs using threshold signature schemes

- **Speaking Truth to Power:** Fighting labor trafficking by building information escrows using secure multiparty computation

- **Privacy with Accountability:** Fighting money laundering by building web3 marketplaces using zero-knowledge proofs

A survivor-centered internet will benefit everyone. Entrepreneurs can still profit off of their innovation. Users can communicate with their friends and family with ease. But it will also ensure that our values to protect others are codified into the new internet economy. We invite technologists, humanitarians, investors, entrepreneurs, and policymakers to join us in our vision.

BACKGROUND

Human trafficking is the illicit business of exploiting vulnerable people for profit. It is a $150 billion criminal industry with 25 million victims worldwide. Human trafficking does not happen in a vacuum; it is the end result of a range of other persistent injustices and inequities in our society and our economy. Not surprisingly, all available data suggests that the majority of trafficking victims identified in the United States are people who have historically faced discrimination and the resulting political, social, and economic consequences: people of color, indigenous communities, immigrants, and people who identify as LGBTQ+ are disproportionately victimized. People living in poverty or foster care, as well as those struggling with addiction, trauma, abuse, or unstable housing, are all at comparatively higher risk for trafficking. Preventing human trafficking at the scale of the problem means changing the underlying systems — particularly the racial, gender, and economic injustices mentioned above — that make people vulnerable and ultimately make trafficking possible.

Polaris is a nonpartisan NGO that has been a leader in the fight against sex and labor trafficking for 20 years. We power our work through survivor-informed strategies, technology-enabled programs, and data-driven policy change. As creators and operators of the National Human Trafficking Hotline, we have answered over 340,000 signals, identified nearly 74,000 cases of human trafficking, and assisted over 30,000 victims and survivors 24 hours a day, 7 days a week, 365 days a year. As a result, we have generated the single largest data asset on human trafficking and have used our insights to shape federal policy.

One of the consistent truths about human trafficking is that it is profoundly adaptable. If you shut down one venue, traffickers will find a new one. Wherever there are vulnerable people and communities, there will be someone who finds a way to exploit them. The impact of COVID-19 was no exception; as the pandemic spread across the globe, traffickers adapted how they used force, fraud, and coercion to recruit, groom, and victimize vulnerable people. 2020 data from the National Human Trafficking Hotline shows that online recruitment into trafficking situations increased significantly by 22%. During the lockdowns, recruitment of victims from more common sites, like strip clubs, foster homes, and schools went down drastically while the internet was reported as the top recruitment location for all forms of sex and labor trafficking. Most notably, our analysis found that Facebook and Instagram, specifically, were the primary methods of trafficking recruitment in 2020. There was a 125% increase in reports of recruitment on Facebook over the previous year and a 95% increase on Instagram. If we want to prevent trafficking from happening in the first place, we must also address the future environment in which it may flourish — web3.

Over the next decade, Polaris is turning its attention to new approaches to end trafficking, including a commitment to address the underlying racial, gender, and economic root causes of sex and labor trafficking. This programmatic vision, entitled The Big Fights, has three pillars:

- **Big Fight on Sex Trafficking:** Reduce sex trafficking in 25 cities by expanding the social safety net available to vulnerable populations, shifting legal accountability for trafficking away from sex workers and towards sex buyers, and changing the norms around sex buying to remove victim stigmatization

- **Big Fight on Labor Trafficking:** End labor trafficking of migrant guestworkers in the United States by ending the current tied H-2 visa system, empowering migrant workers to demand fair recruitment, and changing the standards of behavior for employers and recruiters

- **Big Fight on Financial Systems:** Equip the global financial sector to disrupt sex and labor trafficking at scale by assisting anti-money laundering investigations, shifting financing practices to reward businesses with good labor practices, and building new financial inclusion initiatives for vulnerable populations

Polaris aims to complement its current strategy on the Big Fights to include a future-looking technology thesis, focusing on novel applications of core cryptographic primitives and protocols in the web3 economy. This paper introduces these ideas at a high level as a way to engage members of the private sector, civil society, and government who are working to shape the architecture, policies, and business models of web3.

EMERGING CONCEPTS AND PREVIOUS RESEARCH

Each of the ideas in the three forthcoming chapters — **ownership with consent, speaking truth to power,** and **privacy with accountability** — is predicated on web3 reaching a mature state, obtaining a critical mass of user adoption to drive network effects, and developing new technologies that can drive more sophisticated use cases. It is important to acknowledge that the feasibility of these proposed ideas depends heavily on these foundational concepts being in place. Therefore, these three chapters should be viewed as thought experiments of what is possible, rather than blueprints for how these concepts could be built.

Our ideas are built upon two key bodies of work, as described below.

Decentralized Identity

As web3 economies continue to grow, consumers will want to be able to control and select which personal data they want to share with online systems and businesses. Conversely, online systems and businesses will want to decrease fraud and be able to gain trust that consumer information shared with them is accurate and pertains to the person they are transacting with. Decentralized identity (or, self-sovereign identity) helps address this problem.

Decentralized identity, however, is a complex problem to solve. Decentralized identity systems are difficult to bootstrap to gain enough user traction and do not leverage user networks built on existing web2 services. Decentralized identity also struggles to prevent sybil attacks, which does not protect against users creating multiple fraudulent accounts. Decentralized identity struggles to both maintain privacy *and* meet regulatory and compliance standards, such as know-your-customer (KYC) and anti-money laundering (AML), making it very hard to detect criminal and malicious activity. And finally, it is very challenging for users to manage their private keys, making it very likely for them to lose their credentials which creates the risk of losing valuable assets.

A team of cryptographers has worked to address this problem in a paper entitled "CanDID: Can-Do Decentralized Identity with Legacy Compatibility, Sybil-Resistance, and Accountability" (Maram, et al.). The proposed decentralized identity platform solves for the following challenges, as quoted directly from their paper:

- ***Legacy-compatible credential issuance:*** CanDID leverages oracle systems to construct users' credentials based on data with existing, unmodified web services

- ***Sybil-resistance:*** CanDID enforces deduplication of identities, meaning that it issues credentials in a manner that is unique per user

- ***Accountability:*** The CanDID committee can identify credentials associated with users who should be prevented from using the system, e.g. appearing on a sanctions list, for further action such as blacklisting. This process involved privacy-preserving fuzzy matching of identifier strings using new techniques in a multiparty computation (MPC) setting

- ***Key recovery:*** CanDID allows a user to store their key with the CanDID committee to facilitate recovery, should a user lose their private key. They may leverage *existing* online accounts according to any policy they desire in order to recover their key in a manner that provides privacy for account identifiers

- ***Implementation and evaluation:*** The authors describe and report on the performance of a basic implementation of CanDID. They also report on experiments of the most computationally demanding system component: MPC-based privacy-preserving fuzzy matching for sanctions screening.

The ideas described in our paper build upon the foundational design of decentralized identity described in the CanDID solution. Readers are encouraged to refer to the CanDID paper for references and bibliographic information.

Cryptographic Information Escrows

Information escrows, as noted in the titular paper (Ayers, et al.), “offer the ability to remove the first-mover disadvantage that can deter people with socially valuable private information from disclosing that information to others. Information escrows allow users to transmit sensitive information to a trusted intermediary — the escrow agent — who only forwards the information under prespecified conditions.” Information escrows can serve as an ideal solution for situations when a victim is reluctant or afraid to report a more powerful perpetrator out of fear that they may face punishment for whistleblowing.

A tech nonprofit, Callisto, was the first organization to build a cryptographic reporting escrow that helps victims of sexual assault report their assailant. In a paper entitled ‘Callisto: A Cryptographic Approach to Detect Serial Predators of Sexual Misconduct” (Rajan, et al.), they outline how they architected and deployed this tool.

Cryptographic information escrows build trust in a fundamentally new way. Four key principles (Rajan) define such systems:

- ***Threshold-Based:*** one victim’s record stays locked until a threshold of risk is met by one or more people

- ***Zero-Trust Network:*** the data stored in the escrow is protected from both outside and inside threats

- ***Human Legal Firewall:*** the record is unlocked by a person (ideally, an attorney) who can establish privilege and block false accusations or spoofing

- ***Multiple Calibrated Options:*** victims are presented with several options for how they choose to take action

The ideas described in this paper build upon the foundational design of information escrows described in the Callisto solution. Readers are encouraged to refer to the Callisto paper for references and bibliographic information.

CHAPTER 1: OWNERSHIP WITH CONSENT

Fighting sex trafficking by building NFTs using threshold signature schemes

Non-fungible tokens (NFTs) use smart contracts on the Ethereum blockchain to represent ownership of unique digital items (such as images, videos, songs, and other digital media) and physical items (such as collectible items, physical art, or real estate). In web2, the content created online is owned and monetized by the technology platform that distributes it. This often puts the content creator in a disadvantageous position because they can neither fully own nor fully profit from their own creations. NFTs challenge this paradigm in two ways: first, NFTs are universally compatible and can be distributed across any Ethereum application, thus reducing the asymmetrical power of a third-party middleman; and second, NFTs have a native cryptocurrency built in that drives monetary value, thus allowing individuals to prove their ownership of the content, determine its scarcity, and earn royalties.

Even without a third-party middleman controlling the distribution, the concept of digital content ownership is still a complex issue. Should a photo be owned by the person who created the photo, or the person who is the subject of the photo? Many would agree that the subject should also have ownership and profit-sharing rights of their own image; in fact, after artist Richard Prince appropriated model Emily Ratajkowski's Instagram posts without her permission, she created an NFT of her own portrait and auctioned it at Christie's as a way of reclaiming financial power of her image. But what if Richard Prince had created this NFT first? If he had, it would have been nearly impossible for Ratajkowski to reclaim her image because the blockchain would have proven Prince's immutable ownership and protected his royalties. The current state of NFTs rewards the fastest mover,[1] not necessarily the subject of the digital asset, which could prove to be very problematic for future victims of sex trafficking.

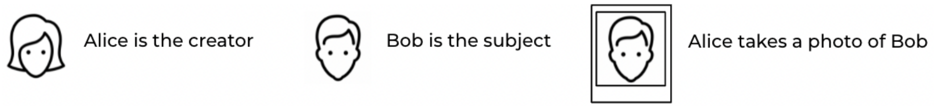
Sex trafficking comes in many forms, and these typologies evolve as technologies change. Traffickers use force, fraud, and coercion to create illicit sexual abuse material of their victims and then sell and distribute it online against their will. The National Human Trafficking Hotline has documented cases of family members, intimate partners, and individual sex traffickers earning profit from distributing a victim's non-consenting appearance in pornographic material. This includes child sexual abuse material (which is considered human trafficking even without the use of force, fraud, and coercion) and non-consensual sexual images (which is very prevalent within intimate partner violence and is a high risk factor for sex trafficking). Traffickers will inevitably move their illicit enterprises to web3, and it is easy to imagine how they can weaponize the immutable ownership properties of NFTs against their victims.

New and innovative technology architectures could significantly reduce the spread of online sexual abuse materials. **Could we architect NFTs such that both the creator and the subject must mutually consent[2] in order for photos or videos to be monetized?** A cryptographic solution — threshold signature schemes — can help make NFTs with consent a reality.

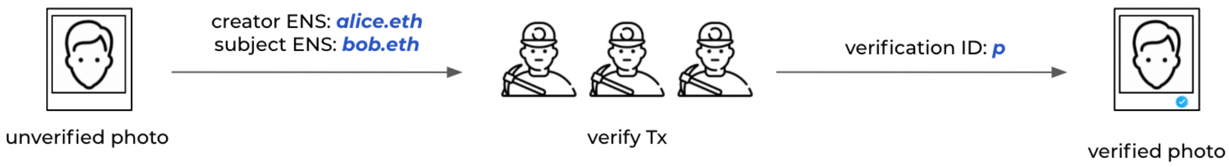
Imagine a web3 entrepreneur wants to build a new NFT marketplace, called OpenOcean, that allows users to distribute and monetize their photos and videos. OpenOcean wants to entice users to the platform by giving them full financial ownership of their photos while making sure they are not enabling the spread of online sexual abuse material. Alice wants to post a photo she took of Bob on OpenOcean's platform. In order for Alice to monetize the image, she will need Bob's consent.

[1] There have been several examples of non-consensual NFTs, including here and here.

[2] There is an important nuance that needs to be addressed: not all photos are used in the same way and therefore should be handled with the appropriate governance. There is a big difference between an abuser posting pornography of their intimate partner to seek revenge versus a citizen posting a video of police brutality to seek accountability. This consent algorithm should support precision and exceptions to this rule. For example, the algorithm could be applied exclusively to sexually explicit content, while waiving consent requirements for high-profile government officials.

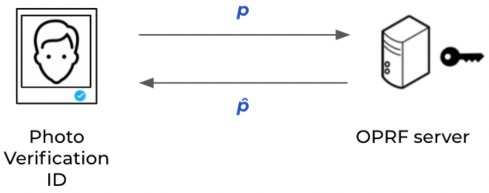


OpenOcean will first need to verify the image. As inputs to the verification algorithm, Alice will submit the image file, the creator's Ethereum Name Service (ENS) name (alice.eth), and the subject's ENS name (bob.eth). Using a combination of a zero-knowledge facial recognition algorithm[3] and the blockchain consensus voting protocol, OpenOcean can confirm if Alice is telling the truth about Bob's identity in the photo. If she is telling the truth, the verification algorithm will return an ID *p* back to Alice to signify that the photo has been properly verified.

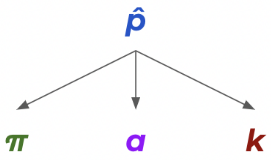


OpenOcean needs to create a mechanism for both Bob and Alice to give permission to post the photo via digital signatures. To prove true consent, OpenOcean can use a collection of cryptographic tools, including Shamir secret sharing[4] and oblivious pseudorandom functions[5] (OPRF).

OpenOcean takes the photo verification ID *p* and runs it through an OPRF hosted on its own centralized key server. This will return a high-entropy value, $\hat{p}$, that they can then use to create the appropriate threshold of secret shares.



From the value $\hat{p}$, OpenOcean runs a deterministic key derivation function to get three values: π (a photo index), a (the slope of a line), and k (the key).



[3] One example of privacy-preserving facial recognition is described in the following paper, entitled "SCiFI - A System for Secure Face Identification" by Margarita Osadchy, Benny Pinkas, Ayman Jarrous, and Boaz Moskovich (2010).

[4] Shamir secret sharing is a technique that lets us split secret key s into many shares s_1, s_2...s_n so that (1) a single share s_i reveals nothing about s, and (2) when all n shares become public, anyone can reconstruct the secret s. To create shares of s we generate a random nth-degree polynomial in a plane of possible secret shares whose y-intercept is the secret s. The shares of s are points on this polynomial. A single point reveals nothing about the polynomial, but n-1 points reveal the polynomial and thus enable computing its y-intercept.

[5] An oblivious pseudorandom function (OPRF) uses a secret key k_s to map a value p to a pseudorandom value $\hat{p}$. This secret key k_s is stored on OpenOcean's centralized key server. A client who has an input p can interact with the key server to obtain $\hat{p}$, an "entropy-boosted" encoding of p. The "oblivious" property refers to the fact that in this process, the key server learns nothing about p, yet the client learns $\hat{p}$. It is emphasized that this process is deterministic: evaluating the OPRF at the point p using the key k_s always results in the same value $\hat{p}$.

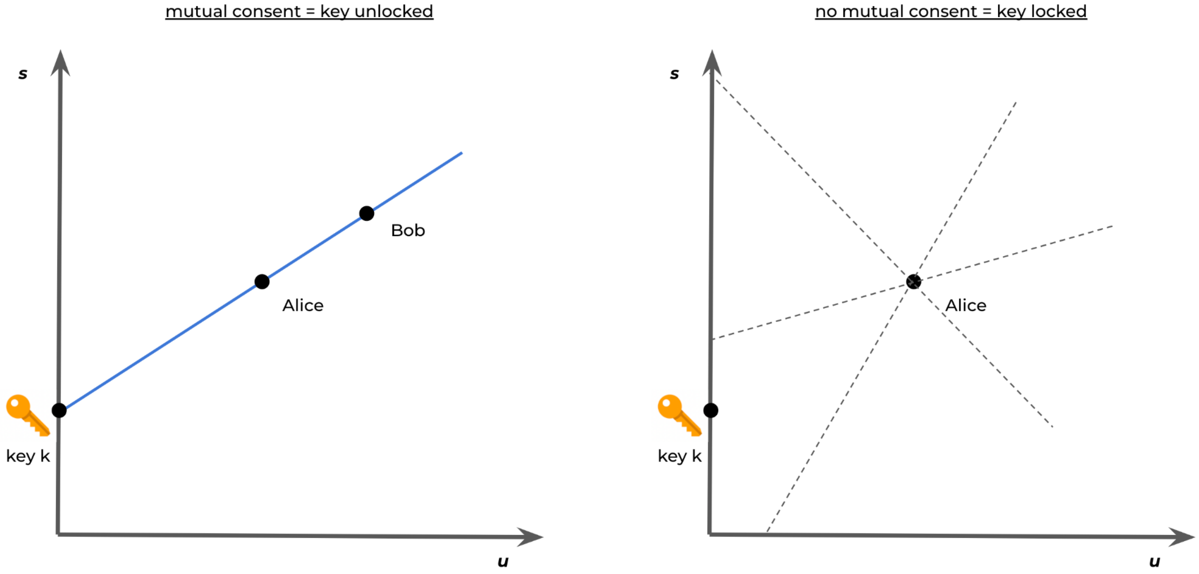
Using the line formula and the derived values *a* and *k*, we calculate a secret share s_i defined as the following:

$$\underset{\text{for all } i \text{ creators and subjects}}{s_i} = \overset{\text{slope}}{a}\underset{\text{hashed user ENS name}}{u_i} + k \rightarrow \text{key located at y-intercept}$$

The hashed coordinates, $H(u_i, s_i)$, of Alice and Bob respectively, are stored on the blockchain.

Now, OpenOcean can seek out permission from all parties to post the photo. For any photo with the index π, OpenOcean will ask both Alice and Bob to digitally sign with their secret share s_i. With both shares, OpenOcean can interpolate the key *k* hidden at the y-intercept and allow the photo to be posted.

If Bob does not consent to have the picture posted, OpenOcean cannot interpolate the correct y-intercept, which means they cannot recover key k, and therefore Alice cannot monetize the photo.



LIMITATIONS

NFT platforms may not initially see the economic incentive to implement such a solution because they are optimizing for rapid growth on their platforms and may not want to build in infrastructure that might slow this down. There is an opportunity for smart regulation to create these incentives. In March 2022, the Biden Administration passed an executive order on ensuring responsible development of digital assets. Federal and state governments should explore ways to ensure that ownership with consent is a key part of a regulatory framework.

CHAPTER 2: SPEAKING TRUTH TO POWER

Fighting labor trafficking by building information escrows using secure multiparty computation

In December 2019, Dr. Li Wenliang, an ophthalmologist at Wuhan Central Hospital, learned about a possible disease outbreak and warned his colleagues on a private group chat to take protective measures. Days later, Dr. Li was censured by hospital leaders and summoned to the Public Security Bureau in Wuhan, where he was forced to sign a statement in which he was accused of spreading false rumors and disturbing the public order. A few weeks later, this novel coronavirus disease, called COVID-19, had spread throughout Wuhan, instigating the worst global pandemic in a century. On February 7, 2020, Dr. Li died of this disease.

A month later, the COVID-19 pandemic swept the United States. As businesses and borders began to close, the U.S. government quickly declared that migrant agricultural workers who help feed our nation were considered “essential workers” and were therefore allowed into the country. But an analysis of data from the National Human Trafficking Hotline suggests that while the government theoretically recognized these workers as vital to our nation’s well-being, they were not protected from trafficking and abuse. Agricultural workers who come to the United States on legal, temporary work visas known as H-2As have long been vulnerable to labor trafficking,[6] but during the pandemic, things got dramatically worse. A Polaris study shows that there was a 70% increase in calls to the National Human Trafficking Hotline of labor trafficking among migrant agricultural workers who held an H-2A visa during the pandemic, which included 66% of victims having their wages withheld or stolen and 34% of victims being denied healthcare. Some of the methods of force, fraud, and coercion that labor trafficking victims often report include withholding of earnings, needs, or wants; threats to blacklist or report to police or immigration enforcement; threats to harm the victim, their family, or others; fraud or misrepresentation of the job; excessive working hours; having a debt or quota; withholding or destroying important documents; isolation and intimidation; and emotional and physical abuse.

The challenges that both Dr. Li and these agricultural workers faced can best be described as a game theory problem — namely, that it is extremely difficult to speak truth to power because there is a first-mover disadvantage with very high consequences for the whistleblower. But, once that first whistleblower comes forward, it opens up the gates for the other whistleblowers to take collective action and hold powerful institutions accountable. Such was the case with Dr. Li — the Chinese authorities emotionally isolated him by attacking his credibility and honor, making it difficult for him to be believed by other members of society. Only after immense pressure from a mass global outrage, the Chinese authorities ultimately revoked their admonishment posthumously of Dr. Li. Similarly, many labor traffickers will intentionally create information asymmetry by isolating their victims as a way to more easily control them.

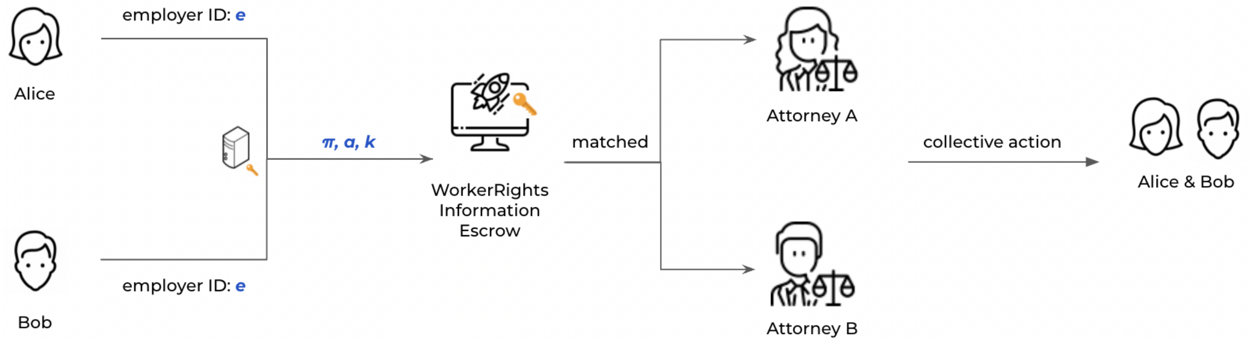
One way to solve this game theory problem is to increase the flow of knowledge among workers about exploitative or dangerous employment situations. Armed with this information, they are more likely to make complaints without fear of retaliation and ultimately feel empowered to unitedly advocate for justice. **Could we build information escrows that reduce information asymmetry among workers while protecting them from negative consequences?** A cryptographic solution — secure multiparty computation — can make speaking truth to power a reality.

Imagine a web3 entrepreneur wants to launch a new nonprofit, called WorkerRights, that builds a whistleblowing platform for vulnerable workers. Workers can enter sensitive details of their working conditions to learn if their situation of trafficking, exploitation, or abuse is happening to other people

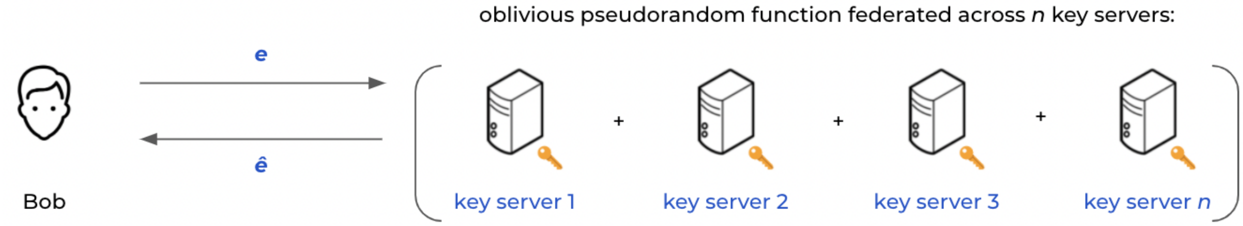
[6] Labor trafficking is the recruitment, harboring, transportation, provision, or obtaining of a person for labor or services, through the use of force, fraud, or coercion for the purpose of subjection to involuntary servitude, peonage, debt bondage, or slavery (as defined by the Trafficking Victims Protection Act of 2000, 22 U.S. Code § 7102).

without compromising their privacy. This can be done through an information escrow. Information escrows use secure multiparty computation to enable a group of individual parties who do not trust neither each other nor a common third party to jointly compute a function that depends on all of their private inputs.

Here is how this escrow would work: Alice, a migrant agricultural worker, can create a timestamped, encrypted record that names her employer using a series of unique identifiers. WorkerRights holds that record in escrow so that nobody (not even WorkerRights) can see that information unless another worker names the same employer. A second worker, Bob, also creates a timestamped, encrypted record that names his employer. If the identities of the employer match, WorkerRights can unlock the records and connect each worker with a third-party human rights attorney. The attorney can protect these conversations under attorney-client privilege and help each worker navigate all of their options to pursue justice in a way that makes sense for them, including enabling matched workers to take coordinated and informed action together.



The escrow is built using very similar cryptographic primitives described in the previous chapter. The employer ID *e* is randomized using an oblivious pseudorandom function (OPRF). This high-entropy value is then used to create a Shamir secret sharing scheme with values π (the matching index), *a* (the slope of a line), and *k* (the key). And, the key *k* to unlock these records is stored as the y-intercept of a linear function. However, the difference in this example is that WorkerRights is not part of Alice or Bob's trust model. Therefore, the trust in the escrow is federated by splitting the OPRF server key across multiple different parties. To reconstruct the OPRF key, an additive secret sharing scheme[7] is used. WorkerRights can distribute this key across a coalition of partner organizations, including other workers rights and anti-human trafficking NGOs, to ensure that no one party can decrypt this data by themselves.


oblivious pseudorandom function federated across n key servers:
e
ê
Bob
key server 1
+
key server 2
+
key server 3
+
key server n

LIMITATIONS

While information escrows are an effective way to empower vulnerable communities, these systems also need to be designed to prevent fraudulent use. A malicious actor may attempt to create duplicate user

[7] Additive secret sharing involves breaking a secret key into fragments that add up to the original secret. Once divided into shares, each share is distributed to different participants.

accounts in order to submit multiple reports about the same perpetrator in an attempt to falsely unlock the escrow. In order to prevent this type of sybil attack, the system must onboard users on an invite-only basis. For example, a user can be verified through a whitelisted identifier (such as a phone number, email address, or unique access code) to ensure they work for a specific employer. This will help ensure that these types of systems cannot be abused.

CHAPTER 3: PRIVACY WITH ACCOUNTABILITY

Fighting money laundering by building web3 marketplaces using zero-knowledge proofs

In order for there to be a democratic and safe web3 economy, cryptocurrency needs to be largely clean from criminal activity. Money laundering, which is the processing of criminal proceeds to disguise their illegal origin, is the engine that fuels many criminal enterprises, including human trafficking. In the 1970s, the United States passed a collection of regulations called the Bank Secrecy Act. This laid the foundation for the U.S. Anti-Money Laundering (AML) framework, which enables financial institutions to detect and report criminal transactions. The AML field is a multi-disciplinary community of financial institutions, regulatory bodies, human rights organizations, technology companies, and law enforcement agencies who work together to both prevent and respond to money laundering.

With the adoption of cryptocurrency, the AML community has evolved to include this new technology in their frameworks. In 2013, the Financial Crimes Enforcement Network (FinCEN) — a bureau in the Department of Treasury — issued guidance[8] that virtual currency exchanges were money transmitters, and therefore were subject to AML requirements. This meant that they must identify their customers using know-your-customer (KYC) tools and submit suspicious activity reports (SAR) to regulators. KYC practices often require users to provide proof of their legal identity, such as a driver's license or a passport. These measures have helped with both the prevention and response side of combating money laundering in cryptocurrency.

However, a new type of cryptocurrency has recently emerged. Privacy coins, such as ZCash, offer the same principles of decentralization as other cryptocurrencies, but also offer the ability to encrypt transactions so that nobody can see the sender, recipient, or amount without the user's permission. This means that the current KYC tools and anti-money laundering tools, which require full knowledge of a user's legal identity and make cryptocurrency pseudo-anonymous, may be at risk of becoming ineffective. More concerningly, current financial controls are predicated on the assumption that users maintain a touchpoint with traditional financial institutions, eventually convertering cryptocurrency to fiat currency at virtual exchanges. But as the web3 economy grows, the need to convert to fiat will decrease, meaning that our current tools for KYC and SAR may be turning to sand.

Unsurprisingly, this technology has spawned fraught ideological controversy: are privacy coins inherently good or bad? Technologies absorb the values of its users and can be used for both benevolence and malice; take, for example, end-to-end encrypted (E2EE) messaging platforms. While E2EE messaging is certainly used by criminals, it is also an essential part of the human rights toolkit. Privacy coins are likely to follow a similar adoption pattern — a human trafficker may use privacy coins to obfuscate their illicit profits, but a survivor may find freedom from their trafficker's financial coercion with the same tool.

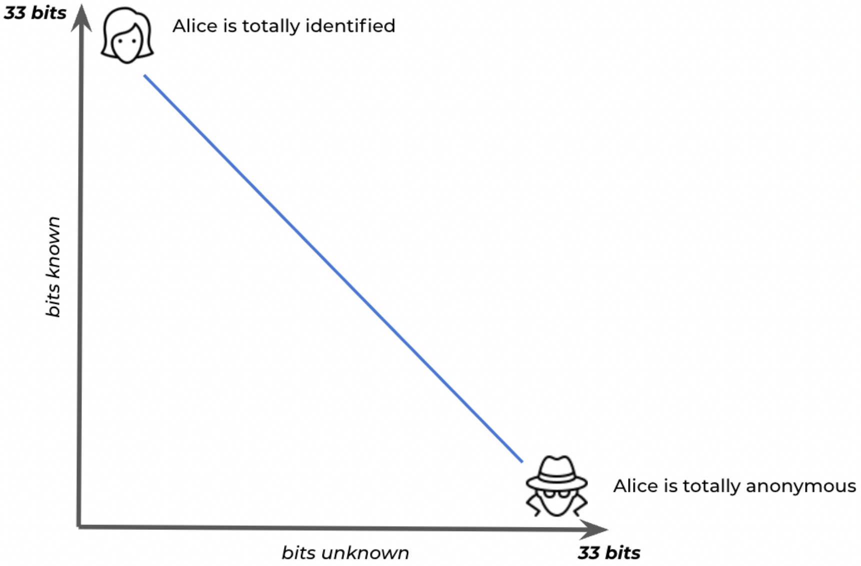
Instead of fighting the adoption of privacy coins, our focus should be on building web3 economies that make these coins less liquid for criminal activity and more liquid for law-abiding citizens. **Could we architect new web3 marketplaces that require users to prove their absence of malice instead of their full legal identity?** A cryptographic solution — zero-knowledge proofs — can help make privacy with accountability a reality.

Imagine a web3 entrepreneur wants to build an online lodging marketplace, called AirBnBlockchain, that allows users to book homestays. AirBnBlockchain wants to entice users to their platform by accepting privacy coins, but also wants to make sure they are not abetting criminal activity or money laundering.

[8] The FinCEN guidance on virtual currencies was updated in 2019.

Alice wants to book a homestay on AirBnBlockchain's platform using a pseudonym, but she will first need to prove absence of malice.[9]

Pseudonymity exists on a continuum between total identification and total anonymity. A person's pseudonymity can be described by its entropy, which is measured in bits. Everytime a bit of entropy is added, the possible combinations double. Since there are about 7.8 billion people in the world, the identity of a random, unknown person contains just under 33 bits of entropy — in other words, 2^{33} ≈ 8 billion. As we learn a new fact about a person, that information reduces the entropy of their identity by a certain amount.

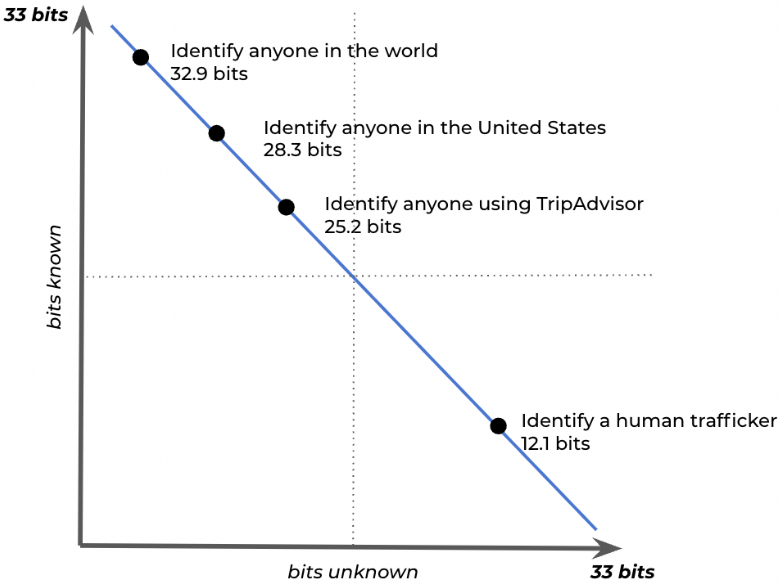


Take for example, the following populations. As a malicious actor becomes a member of a narrower population, the entropy reduces. Conversely, if a person is not part of this narrow population, fewer bits of their identity are revealed:

Population	Total Number	Entropy
The number of people in the world	7.753 billion	32.9 bits
The number of people in the United States	330 million	28.3 bits
The number of monthly active users on TripAdvisor in 2018	38 million	25.2 bits
The number of human traffickers identified by the National Human Trafficking Hotline in 2019	4,384	12.1 bits

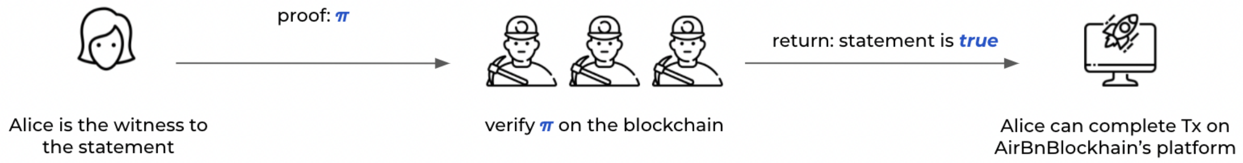
Therefore, AirBnBlockchain doesn't need Alice to disclose all 32.9 bits of her identity, they only need her to disclose 12.1 bits, which still gives her a reasonable amount of pseudonymity.

[9] Hotels, motels, and homestays are used in a variety of ways, in both sex and labor trafficking types, by both traffickers and victims. From 2007 to 2017, the National Human Trafficking Hotline identified 3,596 cases of human trafficking involving a hotel or motel. According to a survey run by Polaris, 75% of survivors reported coming into contact with hotels at some point during their trafficking situation.



A zero-knowledge proof is a cryptographic method that allows a witness party to prove to a verifying party that a given statement is true, without revealing any other information. Zero-knowledge proofs have several defining properties. First, if the proof is complete, then the verifier will be convinced the statement is true; conversely, if the statement is false, the verifier cannot not be convinced the proof is complete. Second, the proof has to be efficient; it must be logarithmic in size and linear in speed. Finally, the verifier must learn nothing about the witness.

Alice, the witness, can send a zero-knowledge proof *π* to AirBnBlockchain, who can use the blockchain consensus voting protocol to prove an absence of malice. If the blockchain's consensus voting protocol verifies this proof, then she can complete her transaction with AirBnBlockchain and purchase her homestay.


proof: π
return: statement is true
Alice is the witness to the statement
verify π on the blockchain
Alice can complete Tx on AirBnBlockhain's platform

But how does Alice prove an absence of malice? We can use the expertise of the anti-human trafficking and anti-money laundering community to create *attestations*. These financial crimes professionals can use their expertise to identify known criminal actors, suspicious enterprises, and precise financial patterns that are high risk of human trafficking behavior, which can then be codified into a zero-knowledge proof used by AirBnBlockchain.

LIMITATIONS

Building proactive detection systems that can flag potential human trafficking activity is challenging because the crime of human trafficking is very nuanced. This is why having subject matter expertise about human trafficking is so important in the anti-money laundering field because without a trauma-informed approach, these detection systems could accidentally harm victims and survivors.

The Polaris Financial Intelligence Unit brings together the anti-money laundering, banking, and law enforcement communities with the expertise of human trafficking survivors and others in the anti-trafficking field. As web3 builders aim to implement an attestation-based solution described above, they should engage Polaris on defining these properties in a survivor-centered way.

CONCLUSION

In the public debate around cryptography and encryption, we often only see two sides represented: one side that says we should protect victims and survivors at all costs, even if that means we break encryption to do it, and the other side that says we should protect encryption at all costs, even if that means victims and survivors get hurt. Predictably, these fault lines are spilling into the debate around web3.

This is a false dichotomy. There is a third way that can optimize for both virtues because cryptography has the potential to protect and empower victims and survivors. And, while we protect the integrity of the technology, we can still hold perpetrators (and the platforms that enable them) accountable for their abuse and exploitation. But doing so will require innovative thinking and an accurate understanding of how these technologies work.[10] It is imperative that we hold both of these virtues with equal importance as we debate the merits and criticisms of web3. Polaris offers the following recommendations to stakeholders across Silicon Valley, Capitol Hill, and Wall Street.

RECOMMENDATIONS

There is a high sense of urgency to build web3 in a way that protects human rights. If we act now, we can build an internet economy that is prosperous, safe, and inclusive. If we fail to do so, the opportunity to design this correctly will close and human rights will be abused. We propose the following recommendations:

1. As the United States government and other intergovernmental organizations deepen their understanding of web3, it should commit to a degree of regulatory certainty and create regulatory sandboxes that enable and encourage innovation. Regulators should analyze different web3 components and understand both the risk posed by malicious actors as well as the opportunity presented to law-abiding citizens. By doing so, it will help ensure that the government is creating regulatory solutions that are targeted and risk calibrated and provide regulatory incentives for founders to include solutions that protect human rights.

2. The anti-money laundering sector must also keep up with the pace of innovation in web3 and challenge its own understanding of how to think about digital identity and law enforcement. By proactively understanding and engaging with these technologies, it will help ensure that AML professionals are able to build modern and more effective techniques to fight financial crimes.

3. In return for this, entrepreneurs and investors in Silicon Valley must commit to building solutions that protect human rights, defend national security, and fight financial crime. We encourage them to use the design principles listed in this paper — **ownership with consent, speaking truth to power,** and **privacy with accountability** — as initial guidance for an ESG framework.

4. As web3 develops, there should be close collaboration between the private sector and NGOs to understand how these technologies can be built and used to support fundamental human rights. By doing so, it can help web3 live up to the promise and potential its supporters believe it can have.

ACKNOWLEDGEMENTS

We would like to thank Dr. Chris Mattmann of NASA Jet Propulsion Laboratory, Prof. Ari Juels of Cornell Tech, and many members of the Polaris Financial Intelligence Unit, Executive Team and Board of Directors for reading drafts of this paper and providing invaluable comments and feedback.

[10] Anjana Rajan, "The Role of Technology in Countering Trafficking in Persons," The United States House of Representatives Committee on Science, Space, and Technology Subcommittee on Investigations & Oversight and Subcommittee on Research & Technology, July 28, 2020.