Who Watches the Watchmen? A Review of Subjective Approaches for Sybil-Resistance in Proof of Personhood Protocols

Divya Siddarth[1], Sergey Ivliev[2], Santiago Siri[3] and Paula Berman[3]*

[1] RadicalXChange Foundation, New York City, NY, United States, [2] Department of Economics, Perm State University, Perm, Russia, [3] Democracy Earth Foundation, San Francisco, CA, United States

Most self-sovereign identity systems consist of strictly objective claims, cryptographically signed by trusted third-party attestors. Lacking protocols in place to account for subjectivity, these systems do not form new sources of legitimacy that can address the central question concerning identity authentication: "Who verifies the verifier?" The legitimacy of claims is instead derived from traditional centralized institutions such as national ID issuers and KYC providers. This architecture has been employed, in part, to safeguard protocols from a vulnerability previously thought to be impossible to address in peer-to-peer identity systems: the Sybil attack, which refers to the abuse of a digital network by creating many illegitimate virtual personas. Inspired by the progress in cryptocurrencies and blockchain technology, there has recently been a surge in networked protocols that make use of subjective inputs such as voting, vouching, and interpreting to arrive at a decentralized and Sybil-resistant consensus for identity. In doing so, these projects illustrate that the best technologies do not abstract away subjectivity but instead embrace it as a necessity and strength. In this review, we will outline the approaches of these new and natively digital sources of authentication—their attributes, methodologies, strengths, and weaknesses—and sketch out possible directions for future developments.

Keywords: decentralized identity, sybil-protection, crypto-governance, decentralized governance, blockchain voting, universal basic income (UBI)

INTRODUCTION

For blockchain networks to move from strictly providing financial services into enabling social and political applications, decentralized protocols for verifying unique human identities must be devised. Lacking reliable means to do so, most blockchain governance practices currently validate membership by employing Proof of Stake (requiring ownership of a given cryptocurrency) or Proof of Work (requiring ownership and use of mining hardware). These resource-based membership systems have rendered most crypto-governance practices into plutocracies, with a few powerful players able to control choices according to their own interests (De Filippi, 2019).

Clearly, this is antithetical to democratic principles. If blockchains are to become a significant public infrastructure, particularly in the space of civic engagement, then Proof of Work's "one-CPU-one-vote" or Proof of Stake's "one-dollar-one-vote" systems will not suffice: in order to enable democratic governance, protocols that signal unique human identities to enable "one-person-one-vote" systems must be created.

At the center of this question of identity is the Sybil attack problem. Previously thought to be impossible to address in distributed identity systems, Sybil attacks describe the abuse of a digital network by creating many illegitimate virtual personas (Douceur, 2002; Ford, 2012; Swathi et al., 2019). When applied to the identity space, this challenge has also been defined as the "unique human" (or, more realistically, "semi-unique human") problem (Buterin, 2019)[1]. However, there has recently been a surge in networked protocols that make use of subjective inputs—such as voting, vouching, and interpreting—to arrive at a decentralized and Sybil-resistant consensus for identity. Driven by the goal of creating a digital layer for humanity, with on-chain and off-chain governance structures (de Filippi and McMullen, 2018), where peers are able to vote, organize, and transact freely, recent projects have been contributing significant learnings to the domain. The present article aims to provide a cohesive review of these learnings and of the projects currently being built to solve for this type of consensus protocol.

We begin by briefly reviewing previous approaches to natively digital identity protocols. We cover both decentralized approaches, with the resource-based mechanisms of Proof of Work and Proof of Stake, and centralized approaches, with private Internet platforms that function as credential providers. In section Proof of Personhood Protocols, we develop a formal definition of Proof of Personhood as a new class of identity protocols aimed at providing a solution for the Sybil attack via subjective approaches. We then outline the established primary properties of these protocols as well as their applications. In section A Taxonomy of Approaches, we develop a taxonomy of approaches that have previously been employed to address the seminal question "How can we distinguish a human from a machine?" and serve as the underlying primitives of Proof of Personhood protocols. In section A Review of Existing Efforts, we outline the leading implementations in the space—their attributes, methodologies, strengths, and weaknesses—and sketch out possible directions for future developments. In section Discussion, we discuss these projects holistically, in the context of strengths, weaknesses, and tradeoffs as conceptualized in the space. We then outline the research gaps evident in each protocol, with the hope of providing a path forward to addressing and solving these issues.

The information below is compiled through a review of the academic literature in the space as well as a compilation of secondary research on each of the approaches discussed, many of which have not yet been studied by the academy. We note that, while the presented protocols are by no means an exhaustive

[1]Buterin, V. (2014). *Problems*. Retrieved from: https://github.com/ethereum/wiki/wiki/Problems/89fd07ffff8b042134e4ca67a0ce143d574016bd (accessed October 15, 2020).

summary of all existing implementations, they do provide a diverse and relevant base from which to understand the state of current work and to identify relevant tradeoffs and gaps in the space.

PREVIOUS APPROACHES TO IDENTITY VERIFICATION

First, we briefly review previous approaches to identity, both decentralized and centralized. In the blockchain space, this begins with Bitcoin's Sybil-protection mechanism, coined Proof of Work. This solution employs a resource-based membership model by proposing a challenge to nodes that require computational work. The CPU that solves this challenge first obtains the right to add the next block of transactions to the chain and also wins a reward in Bitcoin. If the other CPUs computationally agree with the validity of the event, they add that block to their own chains and turn toward solving the challenge for the next block. The majority decision is represented by the longest chain, which has the greatest computational resources effort invested in it. Thus, Proof of Work is a "one-CPU-one-vote" system (Nakamoto, 2009) and therefore centered around machine attributes rather than subjective, human-centered inputs. Similarly, with governance practices relying on Proof of Stake—the main Sybil-protection alternative to Proof of Work—membership is established by the ownership of an asset, in this case, financial stake.

The lack of a robust notion of personhood that could sustain a democratic governance model for blockchain networks has led to the development of plutocracies (De Filippi, 2019; De Filippi et al., 2020): voting power is always relative to stake ownership (clearly so in the Proof of Stake case, and the problem remains even if stake consists of Proof of Work mining rigs). This results in a fat-tailed distribution of voting power in those systems, which reflects the Pareto distribution of wealth in society and financial markets (Klass et al., 2006; Benhabib et al., 2014).

The relevance of formalizing identities through natively digital proofs can also be inferred from centralized networks: major Internet platforms such as Facebook, Twitter, and Google established themselves partially by achieving a sufficient level of consensus over their identity credentials, thus creating a trust layer on top of which a myriad of social applications could be built. This has in turn facilitated the emergence and the spread of multiple borderless political and social movements as demonstrated by the role of social media in both national and international politics over the past few years (Bennett, 2012; Bruns et al., 2015; Tufekci, 2017). The only former alternative able to reach such widespread use pertained to nation-states: passports, licenses, and national ID cards. The creation of a global identification system outside of the strict control of nation-states has accelerated communication and knowledge creation, forming a networked social infrastructure that has allowed for a new kind of participative politics.

However, there are major vulnerabilities with this system, most pertinently (i) privacy concerns and data misuse

and (ii) the risk of creating exclusions to the system, with significant adverse social and political effects. The underlying architecture and ownership structure of these current centralized protocols exposes society to surveillance, political manipulation, and data theft; this is particularly relevant in our current global environment, marked by receding democratic freedoms and rising digital authoritarianism (Freedom House, 2018, 2020).

It is also important to note that officially recognized forms of ID are problematic for an estimated 1.1 billion people around the globe (Vyjayanti, 2017). Therefore, there is a clear need to create a protocol for identity consensus that can operate outside of centralized structures, whether they be nation-states or centralized and privately owned platforms, while enabling the governance of blockchain networks to prevent the concentration of power and influence present in current Proof of Stake or Proof of Work systems. The efforts concerning the creation of a distributed and human-centered protocol coalesce into a third denomination: Proof of Personhood (Borge et al., 2017).

PROOF OF PERSONHOOD PROTOCOLS

Research in the Proof of Personhood (henceforth PoP) ecosystem aims to extend and improve upon Proof of Work and Proof of Stake approaches by focusing on methods capable of creating an analogous decentralized protocol to enable one-person-one-vote systems over blockchain networks. In order to lead to a Sybil-resistant consensus for human identification, such a system needs to ensure that every identity within their domain is (i) *unique*, so that no two people should have the same identifier, and (ii) *singular*, so that one person should not be able to obtain more than one identifier (Wang and De Filippi, 2020). The different protocols reviewed here aim to achieve Sybil-resistance while also maintaining self-sovereignty (anybody can create and control an identity without the involvement of a centralized third party) and being privacy-preserving (one can acquire and utilize an identifier without revealing personally-identifying information in the process). Those three requirements, Sybil-resistance, self-sovereignty, and privacy preservation, compose the "Decentralized Identity Trilemma"[2,3]. Proof of Personhood approaches aim to achieve those three requirements, to different degrees, by establishing the following:

- Subjective substrate: some form of "human entropy" that can act as a substitute for the computational work employed by the Proof of Work protocol or the financial stake employed by the Proof of Stake protocol. This substrate can be expressed in the form of voting, interpreting, being present in a specific place (physical space or cyberspace) and time, or interacting

with others. Typically, the kind of substrate provided needs to be easy for humans to produce, but difficult for Artificial Intelligence to replicate, thus diminishing the ability of computer-generated false identities to take over the protocol. Additionally, this substrate needs to be relatively easy for humans to produce once but relatively difficult for humans to produce two times or more, thus placating the ability of human-generated false identities to take over the protocol. A salient feature of these substrates is that they will typically involve minimal to zero personally identifiable information, thus preserving the privacy of authenticated individuals.

- Objective incentive: an incentive for nodes to join the network and continuously maintain its legitimacy. Ideally, this incentive needs to be strong enough to ensure that it is more valuable to be a part of the network as a legitimate entity than selling one's membership as a Mechanical Turk. With the exception of Upala, all of the protocols described in this report employ or aim to employ some form of Universal Basic Income in cryptocurrency, associated with the protocol and distributed equitably to all members. This incentive can serve as a way to employ a system of behavioral economics, where one loses currency through misbehavior (by somehow attacking the legitimacy of the protocol) or earns more by behaving in ways that make the protocol stronger. Additionally, there may be other incentives, such as the desire for partial privacy or full anonymity in online spaces and transactions.

Primary Properties

We further outline the following desired primary properties of the Proof of Personhood protocols, consolidated from the literature, that allow for comparison:

- Decentralization: Decentralization is a multifaceted measure, analyzing how many independent parties have effective control over various components of the distributed system (Srinivasan, 2017). In the case of PoP, the components of interest are the identity registry, graph of connections and vouching links, software code and releases, operation systems, blockchains, and hardware. Sybil-resistant identity systems must have controls or incentives that would prevent control by bad actors, either through collusion or through purchasing verified identities *en masse*. Decentralization minimizes trust by eliminating third parties and thus also maximizes collusion resistance. It is particularly important to identify who has permissions to write to the registry of identities, that is, whether the registry is permissionless or instead permissioned and controlled by an organization or a consortium. If the latter is the case, the protocol is not decentralized as the trust on a registry manager is required, even if the identity information is stored on a decentralized ledger.

- Privacy preservation: We analyze levels of privacy preservation through measures of anonymity, pseudonymity, unlinkability, unobservability, and plausible deniability (Beckers and Maritta, 2012). A brief outline: Anonymity means that individuals are not identifiable. Pseudonymity denotes the usage of an identifier rather than a real name.

---

[2]Howitt, A., Burnett, D., Meyer, F., Wagner, K., Yuan, Z., and Micheli, F. (2019). *"Not a Sybil!": Exploring the Path to Non-Dystopian Approaches to Digital Personhood* [Github post]. Available online at: https://github.com/WebOfTrustInfo/rwot9-prague/blob/master/draft-documents/proof_of_personhood.md (accessed October 15, 2020).

[3]Maciek, L. (2019). *Decentralized Identity Trilemma*. Retrieved from: http://maciek.blog/dit/?cookie-state-change=1574327093444 (accessed October 15, 2020).

Unlinkability refers to the fact that attackers cannot determine linkages between items such as transactions and addresses. Unobservability implies the anonymity of the persons involved in the transactions. The final aspect is plausible deniability—the ability to convincingly deny the possession of a certain identity and impossibility for authorities to prove the opposite.

- Scalability: The identity system should have the capacity to onboard and provide service to a significant fraction of the global population. In addition, the system should be socially scalable, have sufficient incentives for people to join the network, and have a low barrier to entry (Szabo, 2017), including technical (private key management, interacting with specialized software, e.g., Metamask), financial (paying blockchain fees, staking value tokens or cryptocurrencies), and physical (offline proof of presence ceremonies, peer to peer physical vouching meetings).

Applications

In this section, we outline existing and theorized applications of successful Proof of Personhood protocols. Each of these requires Sybil-resistance to succeed, particularly at scale:

- Universal basic income: A successful program of cryptocurrency-based universal basic income (UBI) must ensure that recipients cannot create fake identities to receive multiple UBI allotments and defraud the system[4]. PoP protocols serve as protection against this kind of attack, enabling the equitable distribution of cryptocurrencies to anyone who is a participant within their networks. Currently, the Idena Network, a Proof of Person blockchain reviewed in this article, is distributing mining rewards that amount to the equivalent of USD $50–60/month[5] for over 4,000 nodes. Another example is that of the Duniter protocol, also reviewed in this article, where participation is incentivized with universal dividends (UD) in Ğ1 cryptocurrency, where each member receives one UD per day. The value of UD in Ğ1 units changes every 6 months as a predefined proportion of the total monetary mass of Ğ1. Currently 1 UD = 10.16 Ğ1[6]. The Ğ1 economy is completely independent from other currencies, which causes its value to vary with the level of adoption on different localities. Rough community estimations have placed it at around $30/month, although the theme is subject to continuous debate[7].
- Peer-to-peer governance: As discussed in section Previous Approaches to Identity Verification, a robust mechanism for signaling unique identities is a requirement for democratic governance in peer-to-peer systems. In the absence of such a protocol, existing Proof of Work and Proof of Stake systems establish resource-based membership mechanisms,

which result in a plutocratic model that, in most cases, makes voting expendable (De Filippi, 2019). Thus, PoP is a requirement for meaningful governance practices within open networks.

- Public goods funding: Addressing global-scale challenges, such as climate change, pandemic, refugee crisis, wealth inequality, etc., requires the ability to deftly coordinate across territorial or institutional boundaries to support supranational public goods. So far, most public goods have been supported by centralized entities, but this nation-state system has thus far proven largely ineffective in addressing global challenges. Traditional Web 2.0 platforms attempted to bridge this gap, facilitating voluntary individual adherence to movements (e.g., "Extinction Rebellion," "Fridays for Future") and aiming to provide a response to these challenges (Bennett, 2012). However, lacking reliable means to formalize unique identities, these movements are rarely able to transform the *noise* of social-media-organizing into a clear *signal* represented by votes or financial commitments. If adopted at a large scale, blockchain-based PoP protocols could contribute to a networked social infrastructure to enable this transition.
- Quadratic voting and quadratic funding: These algorithmic mechanisms for collective decision-making and resource allocation reward diversity of support more strongly than they reward the individual intensity of preference. In other words, a group of people voting or allocating resources to an option has a higher impact than a single individual expressing that same amount of support. Such designs imply that splitting one's votes or funds across multiple accounts increases the impact one is able to exert, thus creating a high incentive for Sybil attacks. Therefore, an anti-Sybil identity system is a requirement for their application within the context of open, peer-to-peer networks (Buterin et al., 2018; Lalley and Glen, 2018). Gitcoin, a crowdfunding platform for open source projects which employs quadratic funding, is currently planning to implement two of the solutions reviewed in this article, BrightID and the Idena Network, by December 2020[8].
- Social Media: Social media signaling methods (impressions, likes, upvotes, etc.) are prone to manipulation by bots (Ferrara et al., 2016). Strong Sybil protection of social media accounts could help address the spread of fake news and fake impressions as well as digital advertisement-related frauds. Currently, the Idena Network is implementing their Proof of Personhood solution for their internal forum[9] and Discord channel[10].
- Airdrops: A popular way to advertise new blockchain projects is to distribute ("airdrop") a fraction of cryptocurrency tokens to a wide distribution list (Harrigan et al., 2018). However, it is very common for those systems to suffer Sybil attacks, even when requiring different forms of identification such as a telegram account or passports. This has led to a switch toward airdrops proportional to user balance of a specific coin and lockdrops (where users need to lock

---

[4]Howitt, A. (2019). *Proposal for a Decentralized Unique Identity Seeding Protocol.* Retrieved from: https://ubiresearch.org/proposal-for-a-decentralized-unique-identity-seeding-protocol (accessed October 15, 2020).

[5]Duniter (2020). Duniter's Ğ1 currency dashboard.

[6]Duniter's Ğ1 currency dashboard, footnote 5.

[7]Ğ1 Forum, https://forum.monnaie-libre.fr/t/donnez-votre-estimation-du-taux-de-change-g1/1987

[8]*Roadmap*, Gitcoin

[9]Idena Forum, https://forum.idena.website/

[10]Idenauth, https://github.com/iyomisc/idenauth

some coins and receive tokens proportionally to signaling). Those techniques, despite being Sybil resistant, privilege users already having a high amount of crypto-holdings. Thus, Proof of Personhood protocols are a key component in enabling egalitarian airdrops.

- Decentralized oracles: Oracles are used as a mechanism to provide factual offchain data to blockchains and smart contracts (Teutsch, 2017). To prevent collusion and coordinated majority attacks on oracles, in addition to Skin-in-the-Game staking, many existing protocols employ KYC for identity verification, which leads to centralization of their services. Proof of Personhood networks are able to provide an alternative solution for decentralized oracles: witnesses can be randomly sampled from the set of human participants to reach a consensus on arbitrary evidence (fact, proof of event outcome, resolution of a dispute, or even opinion poll, etc.).
- Peer-to-peer economy: A Sybil-resistant network utilizing the same currency can have peers directly engage with each other on transactions, exchanges, and co-production of goods and services without the need of intermediaries to establish trust (Selloni, 2017). Currently, Duniter, one of the Proof of Personhood protocols reviewed in this article, utilizes Gchange[11], an active internal market for exchanges denominated in their Ğ1 currency, and gannonce[12], a crowdfunding platform for ğ1 projects.

A TAXONOMY OF APPROACHES

Before reviewing existing solutions, in this section, we will outline the different theoretical primitives that underpin Proof of Personhood approaches. Throughout the past few decades, different methods have been outlined in order to address one fundamental question: how can we distinguish a human from a machine? We describe recent approaches below.

Reverse Turing Tests

In the opening of his 1950 paper, "Computing Machinery and Intelligence," Alan Turing asked, "Can machines think?" In order to narrow this question down to one with an objective answer, Turing created the "Imitation Game," in which an evaluator having a conversation with another entity through a text-only channel attempts to determine whether the entity in question is a human or a computer. Known as the Turing test, to this day it is applied at an annual competition in artificial intelligence, The Loebner Prize, that rewards the most human-like computer programs based on subjective assessment from human judges (previously a panel, and as of 2020 evaluated by the public).

This method created the base for a reverse test, the CAPTCHA, a "Completely Automated Public Turing test to tell Computers and Humans Apart," widely used to elicit proof from humans that they are not bots. It does so by requiring humans to parse through distorted words and images, a class of tasks known as "AI-hard" (von Ahn et al., 2003): difficult for an algorithm to

perform, simple for a human. However, in addition to serving the purpose of authenticating humans, the input from a CAPTCHA test is also used to calibrate the pattern recognition capacities of artificial intelligence algorithms. Thus, machine learning presents an evolving threat to the functioning of these CAPTCHAs.

New approaches are being developed in order to address this challenge. The Idena Network, outlined in "section Idena Network," has shown that, in order for CAPTCHAs to resist the dynamic development of AI connected with neural networks and deep learning, they must not be generated algorithmically, but instead created by humans (Idena Network, 2019). Only then will those tests move out of the class of "recognition" tasks, solvable by neural networks and instead be classified as AI-hard problems, requiring an understanding of implied meaning or the use of common sense reasoning.

Recent AI-hard tests extend principles from the Winograd Schema Challenge (Levesque et al., 2012), which would pose implied-meaning questions like the one below:

"The trophy would not fit in the brown suitcase because it was too big. What was too big?"

1. The trophy
2. The suitcase

However, due to its reliance on textual representation, WSCs may be vulnerable to new advances in natural language processing such as GPT-3 (the accuracy of the state-of-the-art models in WinoGrande challenge currently reaches 0.7–0.85 AUC compared to 0.94 AUC human performance[13]). Additionally, this approach requires specific language knowledge" and therefore fails to create a standard that can be applied internationally. Thus, the use of images is more likely to remain robust in the long term.

As demonstrated on **Figure 1**, the Idena Network builds on these previous approaches by creating an AI-hard test that both requires common sense reasoning and is based on visual representation. Named FLIP, it asks users to choose between two orderings of images, with only one set conveying a logical and meaningful story. Human accuracy in solving FLIPs is at 95%, while AI teams have been able to reach 60–76% (Moritz, 2020). Alternative AI-hard reverse Turing tests are VCR (Zellers et al., 2018), ROPES[14] (Reasoning Over Paragraph Effects in situations), ALFRED[15] (Action Learning From Realistic Environments and Directives), and others, although to our knowledge these are not being employed by any Proof of Personhood solution at the moment.

One important note is that, while reverse Turing tests may prevent automated and bot attacks, they fail to address human-generated attacks, in which one individual passes the test multiple times and creates multiple different identities. To address this second aspect of the challenge, approaches such as the FLIP created by the Idena Network or ATUCAPTS (Andersen and Vincent, 2016) restrict participation via employing elements of a Pseudonym Party, as described below.

---

[11]Gchange website, https://www.gchange.fr/#/app/market/lg?last
[12]Gannonce website, https://gannonce.duniter.org/#/

[13]WinoGrande Challenge, Allen Institute for AI
[14]Reasoning Over Paragraph Effects in situations, Allen Institute for AI
[15]Action Learning From Realistic Environments and Directives, Allen Institute for AI
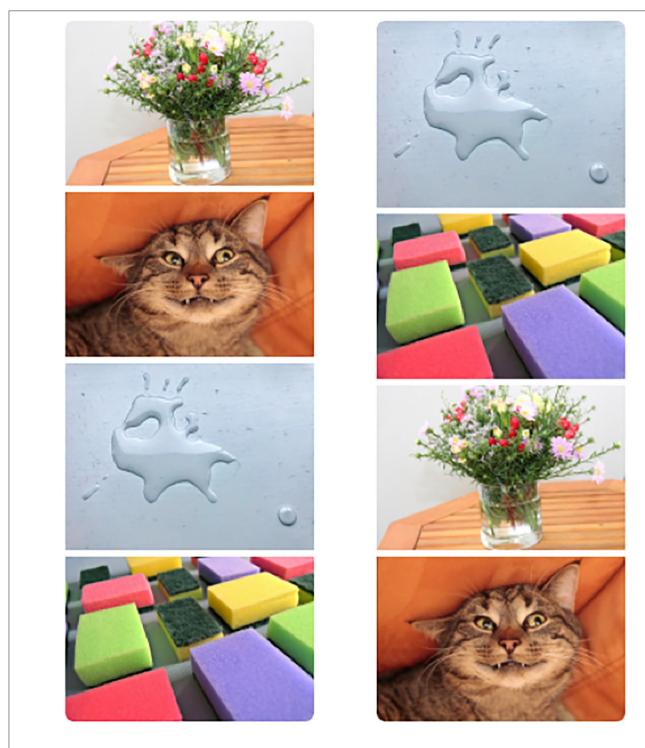
FIGURE 1 | Example of a flip: a meaningful story **(Left)** and a meaningless sequence of pictures **(Right)**. Idena Network, 2020.

Pseudonym Parties

Pseudonym parties are an effective mechanism to avoid a tradeoff between accountability and anonymity in digital domains. This is a "back to basics" approach that builds upon a simple security foundation: real humans can only be in one place at a time (Ford, 2012). In this method, authentication comes from being physically present at a specific place and time. In this physical space, attendees will formalize procedures to register their presence, such as individuals scanning each other's QR codes and, by that act, generating an anonymous credential or token (Borge et al., 2017). These credentials can then be utilized to establish membership in online communities. In essence,

pseudonym parties act as a framework for individuals to conduct reverse Turing tests on each other. For as long as hyper-realistic digital forgeries (deep fakes) remain in the Uncanny Valley (Mori et al., 2012), virtual pseudonym parties may be feasible (Nygren, 2019). This is especially true if utilized in combination with AI-hard tests, which require human interpretation. In these combined approaches, users willing to be authenticated need to perform the test simultaneously. Thus, due to the fact that the same person cannot perform two actions at the same time, the protocol ensures the singularity of authentication.

Pseudonym parties provide significant accountability since membership rights are limited and therefore can be revoked while preserving a relatively high degree of anonymity since credentials do not need to contain any personally identifiable information.

However, the requirement of significant engagement for authentication is a clear downside, particularly since credentials are not permanent: all "nodes" must be synchronized at a certain frequency so that new individuals may join the protocol. There are also concerns regarding the authentication of users in remote or faraway locations who may not be able to attend physical gatherings. This friction may be minimized by leveraging existing gatherings such as conferences, rituals, and civic ceremonies for identity authentication (Ford, 2020) as well as through virtual pseudonym parties.

Web of Trust

A Web of Trust consists of identity certificates that can be digitally signed by other users who, by that act, declare the certificate valid and thus provide Proof of Personhood. Through this process, the nodes of the network are effectively partitioned into Sybils and non-Sybils (Viswanath et al., 2010). The Web of Trust paradigm relies on the fact that, while an attacker may be able to arbitrarily create Sybil identities in social networks, it will be much harder to substantiate said identities with an arbitrarily large number of false connections to trusted nodes (Viswanath et al., 2012). Sybil nodes will thus be poorly connected to the trusted network and easily identified. Web of trust schemes may be further reinforced by a reputation system that serves to track trust levels and prevent deception (Dunphy and Petitcolas, 2018).

We note that there have been attempts to create a web of trust framework through automated graph analysis of existing social trust networks, particularly social media networks. One such approach is SybilRank, which aims to identify fake accounts within bounded social media networks and has met some success (Cao et al., 2012). However, it is unlikely that widely used online social networks are good candidates for large-scale identity approaches, particularly for sensitive applications like civic engagement. This is due not only to the ease with which attackers can create "false" nodes with real relationships and connections to other nodes (Ferrara et al., 2016; Ford, 2020) but also because re-orienting an identity program around privately owned, centralized social network platforms is antithetical to the project of self-sovereign identity solutions.

Despite the long-term interest in Web of Trust, with the first limited-scope version set out in the second PGP manual in 1992, several inherent issues with the approach have prevented its large-scale adoption. First, a combination of different claims and credentials may not entirely guarantee Sybil-resistance (Wang and De Filippi, 2020). Furthermore, levels of trust cannot be easily quantified, and only first-degree relationships can be fully trusted, which can constrain the network. Similar to pseudonym parties, these issues can also prevent users from low-infrastructure or remote locations from acquiring key signs or building in-network credibility (Wilson and Giuseppe, 2015). To correct for these issues, the web of trust paradigm has been adapted, extended, and paired with other approaches, notably in the form of mutual surety graphs as well as graphs with other topological features (Shahaf et al., 2019). Two of the implementations that we will discuss also aim to extend the subjectivities of the web of trust approach, moving toward a more intersectional paradigm, as described below (Immorlica et al., 2019).

Intersectional Identity

Intersectional Identity is a framework that aims to bridge formal verification methodologies and the informal mechanisms through which individuals check the validity of identity-related claims. It builds upon traditional Web of Trust schemes by expanding the scope of markers that can be taken into account, such as one's name, age, address, gps history, interactions, skills, work, education, etc. All of these different markers can be translated into bits, so any given individual is associated with an exponentially large number of potential bits that may be useful for authentication.

This framework achieves uniqueness or Sybil-resistance by drawing from three aspects of identity highlighted in the classical sociology of Georg Simmel: sociality, intersectionality, and redundancy (Schützeichel, 2013). Here sociality refers to the fact that every aspect of identity is shared. Intersectionality implies that the set of others with whom the identity markers of any given individual are shared differs for each marker, thus no individual or group can serve as a central "chokepoint" for identity verification. Redundancy denotes that the uniqueness of an individual is over-determined by the countless unique intersections of groups or sources of trust that each person finds themselves in through the course of their lives. With data architectures put in place to record intersectional markers, Sybil-resistant identities can be established by tracking just a few characteristics that uniquely identify an individual while keeping sensitive information private (Immorlica et al., 2019).

Token Curated Registry

The Token Curated Registry (TCR), in contrast to the schemes outlined above, was not originally devised as a method for identity verification. In essence, TCRs draw from work on incentive systems designed to replace list owners through the creation of economic incentives for decentralized list curation. In this scheme, members of a registry hold tokens associated with the list, which may increase in value if they are able to maintain its quality, legitimacy, or popularity, thus attracting more list applicants who want to add their data to it (Asgaonkar and Bhaskar, 2018). Members can establish trust through different mechanisms, such as staking a certain amount of funds, voting, or

vouching for other members accurately. TCRs have successfully been applied toward curating professional profiles, media content, and other services and are particularly instrumental in enabling decentralized courts for blockchain-based dispute resolution frameworks (Lasaege et al., 2019). Building upon these successes, different identity solutions employ this mechanism to create an incentive for members of an identity registry to go through the effort of verifying each other's uniqueness and singularity.

Decentralized Autonomous Organizations

The DAO acronym refers to Decentralized Autonomous Organizations. DAOs are a class of smart contract (Norta, 2015)[16] devised to automate the execution of organizational governance and fund allocation. In that sense, these contracts may be thought of as an automated constitution. This organizational framework emerged as a possibility due to the creation of Ethereum, a blockchain network that permits Turing-complete computations[17] (Minks, 2017), leading to the growth of smart contract development. By deploying DAO contracts into the Ethereum blockchain, organizations allow their participants to pool funds (denominated in cryptocurrencies), maintain real-time control of resources, and vote on resource allocation to different projects with governance rules that are formalized, automated, and enforced by the conditions encoded into the smart contract.

This type of organizational framework is employed, in different ways, by several of the Proof of Personhood protocols described in this review. In contrast to the majority of smart contracts, which serve strictly financial purposes, DAOs are highly likely to entail human decision-making in their functioning. Thus, their activities may be thought of as "human entropy," observable on-chain, serving as a meaningful substrate for different aspects of Proof of Personhood solutions.

A REVIEW OF EXISTING EFFORTS

We now outline the approaches of these new and natively digital sources of authentication—their attributes, methodologies, strengths, and weaknesses—and sketch out possible directions for future developments.

Idena Network

Idena is an open-source project created by an anonymous group of engineers and computer scientists. It has its own blockchain, which is driven by a proof-of-person consensus, with every node linked to a cryptoidentity with equal voting power —thus, it is a fully decentralized solution. The Idena Network implements a novel way of achieving Sybil-resistance by combining human-generated reverse Turing tests (Idena's FLIP test is described in

detail in "section Reverse Turing Tests" —Reverse Turing Tests) with elements of a virtual pseudonym party.[18]

To join the network, participants must attend live authentication ceremonies held simultaneously for the entire network. During these synchronous events, one must complete a set of FLIP tests within a limited amount of time. Afterwards, users are required to create new tests. This is an important element, given that in order for the FLIP test to resist machine learning and truly belong to the category of "AI-hard," it was designed to not be fit for automated, algorithmic generation. FLIPs are instead always created by humans.

The frequency of those authentication ceremonies is determined by the size of the network—currently they are conducted around once every 2 weeks. Given that tests cannot be solved by existing AI, Idena successfully provides a proof of personhood. However, it is not strictly anti-Sybil, with a probabilistic margin of error: although highly unlikely, a person with exceptional ability could solve more than one set of FLIPS within their allotted time, thus earning more valid identities within the network.

As an additional layer of security, Idena requires new members to present an invite code to be able to join their first authentication ceremony. This code can only be obtained through existing members, thereby creating a Web of Trust. This also extends into a reward-based system: at every validation ceremony, Idena rewards all of its members with its $IDNA cryptocurrency; by inviting members who consistently attend validation ceremonies, one may gain compounded rewards.

Launched in August 2019, to date the Idena Network has been able to validate 4,012 identities[19]. Their approach presents a significant advance for the research and development concerned with natively digital identity protocols. This network demonstrates that combining human-generated AI-hard tests with "liveness" —a synchronous event—can play a critical role in Sybil prevention: the time constraint prevents a single entity from solving more than one set of FLIPs, while the human-generation aspect provides a defense against machine learning. Furthermore, the protocol protects privacy, as it involves no data point except that of proof of conscious cognitive ability.

The most salient tradeoff in this system is the significantly high coordination cost to achieve recurrent, simultaneous solving of FLIPS: all nodes must continuously participate in the synchronous events; otherwise, their identities expire. This reduces the incentive for nodes to join the network, although the relative value of the rewards paid by the protocol for successful validation and participation in block producing, currently at ~$50–60 USD/month, may succeed in creating a new habit for users. Additionally, it remains to be seen whether Idena's Sybil-resistance strategy will be able to weather the dynamic development of AI connected with neural networks

---

[16]Szabo, N. (1994). *Smart Contracts* [Website]. Available online at: https://www.fon.hum.uva.nl/rob/Courses/InformationInSpeech/CDROM/Literature/LOTwinterschool2006/szabo.best.vwh.net/smart.contracts.html (accessed October 15, 2020).

[17]Buterin, V. (2014). *Ethereum Whitepaper* [Blog Post]. Retrieved from https://ethereum.org/en/whitepaper/ (accessed October 15, 2020).

[18]The Idena Network (2019). *How Idena Works* [Blog]. Available online at: https://medium.com/@idena.network/how-idena-works-4f4d19aabbb1#:$\sim$\$:text=Idena%20is%20a%20truly%20decentralized,identity%20for%20the%20consequent%20epoch.&text=There%20was%20no%20human%20being%20in%20the%20blockchain%20so%20far (accessed October 15, 2020).

[19]Idena Network (2020). *Epoch 54*. Retrieved from: https://scan.idena.io/epoch/54 (accessed October 15, 2020).

and deep learning. Furthermore, the robustness and long-term effectiveness of their incentive systems may also be tested in the future by the creation of markets that sell false identities and/or attacks by mechanical turks.

Humanity DAO

Humanity DAO is an Ethereum-based protocol. It was designed to incentivize a set of economic actors to maintain a registry of unique human identities without a central authority[20] and leveraged existing work on Token Curated Registries (Asgaonkar and Bhaskar, 2018). In Humanity DAO's case, holders evaluate candidate identities and deem them legitimate through consensus-based voting. The protocol consists of the following steps:

1. Applicants made a request to join the list using their social media profile information.
2. Applicants staked a fee on their candidacy. If the applicant got rejected, the application fee was ceded.
3. Members of the list voted on whether the new applicant should be included based on the submitted profile. Members were incentivized to curate the list honestly in order to generate demand from new applicants, leading to a long-term sustainability of the project.

The registry had a method called *is Human* that any smart contract could query to see whether a given Ethereum address had been confirmed as a unique human. Humanity DAO also deployed a Universal Basic Income smart contract with 2,500 Dai (∼$2,500), which early applicants could claim at a rate of 1 Dai per month, until supply ran out (Chen and Ju-Chun, 2019).

Launched in May 2019, the project quickly gained rapid traction, reaching around 640 approved members and being adopted by many influential figures within the Ethereum network; however, growth stagnated after the initial community of early adopters from the network was saturated.[21] Furthermore, as a fully decentralized solution, the creators had very little ability to change the protocol after it was launched. As related to us by the founder, this resulted in Humanity DAO suffering various forms of attacks, including one in which a change to the smart contract made it prohibitively expensive for new applicants to join. These repeated attacks led to the eventual termination of the project in January 2020.

Kleros

Kleros is an Ethereum-based protocol for decentralized dispute resolution. Their successful experiments with TCRs for distributed courts led the team to propose "Proof of Humanity": a solution for identity based on TCRs combined with a web of trust and based on submitted photos, bios, and video recordings. This information will be stored using the IPFS (InterPlanetary File System). Kleros' approach distinguishes itself by appending

to the functioning of its protocol a recourse to adjudicate cases of faulty or duplicate users. This is done through distributed dispute resolution systems such as the Kleros Court[22] (or if decided by members through the registry's internal governance, other alternatives such as Aragon's courts[23]).

Within the Proof of Humanity protocol, users can vouch for each other with a certain amount of financial stake. To incentivize the maintenance of the registry, vouching deposits serve as a bounty, available for anyone able to correctly identify false positives in the registry. If a member vouches for users that are later determined to be duplicate or false by the distributed court, they are punished in the form of being removed from the registry and losing their vouching deposit, thus discouraging such attacks (Lasaege et al., 2019).

While this protocol has significant promise in building an effective reputation-based web of trust with tools in place to adjudicate cases in which the singularity of an identity is disputed, it compromises the biometric information of members by requiring a video selfie and other additional information, which may de-incentivize potential users.

Upala

Upala is an Ethereum-based protocol designed to be interoperable with DAOs[24]. Its social graph consists of verification groups that assign a score for each member, denominated in currency; this gives members the right to steal from the shared pool of the group they belong to, the amount of their score. The act of stealing (a "bot explosion" in Upala terms) automatically deletes their identity.

Thus, this model implements the social responsibility concept, in which groups are incentivized to develop approval mechanisms that lead toward having highly trusted members. Any existing DAO may fit into the Upala protocol, given that members are willing to trust each other by collateralizing funds in exchange for distributing reputation. Groups can be also be composed by direct end users or other groups—thus combining uniqueness scores into larger pools. This framework generates a market for identity authentication where on the supply side groups are trying to gather as many users as possible (through subgroups or directly), with the highest reputation (i.e., lowest risk of explosion) and the maximum deposits, and on the demand side users are trying to get the highest scores for the lowest investment of reputation or money.

The Upala model expands on the principles behind a Token Curated Registry (where members are incentivized to maintain a high-quality list) and also employs an intersectional lens by enabling different schemes to be created and combined within its protocol.

Given that the uniqueness scores are, to a certain extent, relative to pooled funds, this may lead to capital-rich users having ease in obtaining higher scores—although groups may establish different verification mechanisms capable of placating this vulnerability. Another major vulnerability encoded within

---

[20]McAteer, R. (2019). *Introducing HumanityDAO* [Blog Post]. Retrieved from: https://medium.com/marbleorg/introducing-humanity-90ddf9ead235 (accessed October 15, 2020).

[21]Wikiel, K. (2019). *HumanityDAO Post Mortem*. Available online at: https://trackato.substack.com/p/humanity-dao-post-mortem (accessed October 15, 2020).

[22]Kleros Court: https://blog.kleros.io/kleros-court-revitalised/

[23]Aragon Court, https://anj.aragon.org/

[24]Upala's Documentation: https://upala-docs.readthedocs.io/en/latest/

this model would be an avalanche user exit: if an event leads to loss of trust in Upala, an avalanche of individuals may explode their identities in panic to seize assets, ignoring the reputation consequences. However, it is possible that the probability of such a scenario materializing decreases as trust consolidates within the system, with usage by third parties for scoring users increasing.

Upala has launched its first working prototype on the Kovan test net of Ethereum in June 2020.

BrightID

BrightID operates an intersectional web of trust protocol, built through graphing social connections, with the additional input of trusted seed identities. The purpose of this protocol is to allow users to provide proof that they are not using multiple accounts on a single application, and it is thus designed to be interoperable with Web 2.0. social media platforms. The interconnectivity of its graph is designed to identify true identities and Sybil identities based on node position in relation to trusted seeds.

Thus, BrightID is the solution most in line with the Intersectional Identity paradigm, formalizing social connections in order to allow for a variety of nodes to join the system and customize their own evaluation criteria. In that sense, there are no obvious limits to the number of trusted seeds in the BrightID graph: any application utilizing their authentication solution may establish its own BrightID node with different trusted seeds. Each BrightID node runs its own instance of ArangoDB to store the graph of Web of Trust connections. Every verification can be broadcasted to a specified isolated smart contract on Ethereum or another blockchain.

The social graph serves as a common base across all nodes, but the analysis of that same graph can be distinct, so the protocol does not require consensus across nodes. Applications may either run their nodes in a centralized or closed manner, sharing their analysis and verification outputs only with themselves, or they can provide a greater level of decentralization, allowing any user to run the verification and sample the output from a large number of nodes. To control for Sybil attacks, BrightID runs GroupSybilRank, a modification of the SybilRank algorithm, to estimate the anti-Sybil score of the network participants based on affinity between groups. Proposed to be used as the official BrightID anti-Sybil algorithm[25], the effectiveness of this algorithm in the presence of multiple attack vectors remains to be proved.

BrightID's open Web of Trust architecture is robust and promising. That said, at this early stage, BrightID's social experiment has significant challenges to overcome in terms of Sybil-resistance, decentralization, self-sovereignty, and privacy. As of July 2020, its solution is limited to a small seed network, so there are no established paths for individuals or groups who are completely independent from the existing network to self-authenticate—thus, it is not yet a fully self sovereign solution. This is not an intractable limitation as new nodes can potentially define new verification methods that would allow for islands of users to be verified. However, scaling this process is far from

trivial. One possible solution is through establishing partnerships with existing social media platforms that reach a wide net of users, but this would largely defeat the aim of the initial motivators of Proof of Personhood solutions. Therefore, the crucial challenge for the success of this experiment is finding a path forward for scalability while maintaining decentralization.

To this end, BrightID's whitepaper encourages the creation of new seeding DAOs (Decentralized Autonomous Organizations) and establishes that the BrightID Main DAO will promote research of different seed selection methods as well as the creation of tools that can make seed selection scalable. In that sense, BrightID's success may be in tandem with an increase in the adoption of decentralized governance frameworks. Another possible pathway to scale is through BrightID's weekly pseudonym parties, during which prospective members can meet the existing community and form new links to obtain verification.

A new blockchain, IDChain, was recently introduced to implement BrightID DAO governance[26]. IDChain is a fork of the geth Ethereum node software, at which BrightID participants can self register via a web service to receive a lifetime supply of Eidi (the native gas token on IDChain). Hedge for Humanity, a US-based, tax-deductible 501(c) (3) charitable organization, plans to start distributing US$1/month to each of BrightID's users as a Universal Basic Income, as a way to incentivize attacks that can provide greater visibility into the vulnerabilities of the identity system. Currently, BrightID has 556 users with a positive anti-Sybil rank[27].

Duniter

The Duniter project, originally named uCoin, was started in June 2013[28]. The project is a technical implementation of the relative theory of money, developed by Stéphane Laborde, where a Universal Dividend is described as having its value relative to its monetary mass. Duniter is an independent blockchain utilized to mint the Ğ1 cryptocurrency as a Universal Dividend available to unique human participants. Authentication within the Duniter protocol is done through a Web of Trust type of scheme.

In order to become a part of Duniter's Web of Trust, one must receive five different vouches from existing members[29]. Duniter members are required to check a statement where they agree to vouch solely for new applicants who they have met in the physical world or know enough to be able to contact remotely through different channels, such as social network, forums, email, video conferences, and phone calls[30]. For each new member, a cryptographic key pair is created. Furthermore, any newcomer must be at a maximum distance of five different connections from "referent members," which can be thought of as more central and highly trusted seed identities. Referent members are defined by a graph property that intends to mimic the

---

[25]BrightID. (2020). *BrightID Anti-Sybil*. [Github post] Retrieved from https://github.com/BrightID/BrightID-AntiSybil.

[26]Stallard, A. (2020a). *Introducing IDChain* [Blog]. Retrieved from: https://medium.com/brightid/introducing-idchain-392c76c31d73 (accessed October 15, 2020).

[27]BrightID data, BrightID, September 2020

[28]Duniter License, https://duniter.org/en/wiki/g1-license/

[29]Web of Trust settings, footnote 5

[30]Duniter Forum, https://forum.duniter.org/t/nombre-didentites-dexclusions-de-revocations-etc/7428
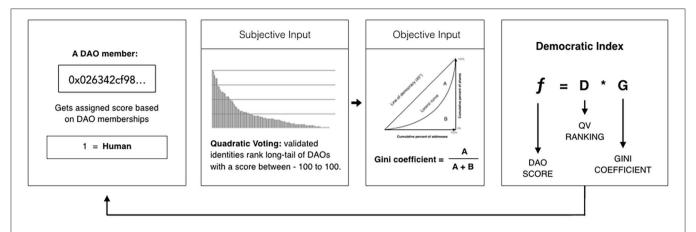
FIGURE 2 | Step 1: Every address that belongs to a DAO will be weighted by the Equality protocol with a percentage of the Democratic Index obtained from the intersection of DAOs that constitute the identity of such address. Step 2: In order to counterweight false positives on the Gini Coefficient, addresses that achieve a high score will be granted the right to rank the different DAOs analyzed by the Equality Protocol oracle, according the their corresponding ability to ensure that no single Human controls more than one identifier within its domain. Step 3: The Gini coefficient for the democracy ranges from 0 to 1, with 0 representing perfect totalitarianism and 1 representing perfect democracy. It computes a score that measure the share distribution by each segment of addresses belonging to a DAO. Step 4: A Democratic Index is calculated for each DAO, as a function of their position in the Quadratic Voting rank and their Gini Coefficient.

trust parameters of the real world, human, and social graphs[31]. These requirements imply that it entails significant time for new certifications to be emitted, allowing the network to manually monitor and mitigate attacks. By doing so, the Duniter protocol delegates technical governance to human governance while still maintaining a relatively high degree of decentralization. This Web of Trust is seeing a slow but steady growth in France and nearby countries. As of September 2020, the Ğ1 currency had 2,801 holders[1].

Equality Protocol

The Equality Protocol approach creates a meta-protocol against which other identity protocols can measure their legitimacy[32]. As outlined in **Figure 2**, it is designed to create an intersubjective space able to account for measurements of both collective intentionality and objective facts by combining a subjective function that provides legitimacy to the score based on Quadratic Voting and an objective function that measures the Gini Coefficient of any DAO existing on the Ethereum blockchain. It will create a Democratic Index, as shown in the figures below, and assign a score to every Ethereum address relative to the intersection of DAOs in which it belongs as a member or its position in the social graph of blockchain-based transactions.

The current interface provides a basic personhood calculation for each DAO member of v1 Moloch DAOs, with 529 addresses receiving a probabilistic human score[33]. Examples of DAO contracts with parameters applicable as inputs for a Probabilistic Anti Sybil Score oracle include MolochDAO, DAOstack, Kleros, and Aragon DAOs. Additional sources of

trust can be included: different credentialing mechanisms could instantiate their protocols through a DAO, while other non-EVM protocols such as the Idena Network could be made compatible by allowing token swaps.

This approach intends only to be a meta-protocol contemplating existing DAO members rather than forming a Proof of Personhood substrate in and of itself. However, in addition to requiring that new applicants get voted in, DAOs are usually joined through the staking of resources. This system may favor capital-rich users, who could find their way into several DAOs with more ease. This is not an intractable challenge, as it is possible to earn shares through non-financial contributions, and different membership protocols could potentially make the environment more inclusive over time. Furthermore, while the Quadratic Voting function brings a desirable governance component to identity verification, it remains to be seen whether it can serve as a rigorous substrate to signal the legitimacy of DAOs, given the inherent conflicts of interest that emerge due to the impact of results on the probabilistic score of voters.

DISCUSSION

The seven reviewed projects in the digital identity space have made previously unimaginable progress in creating robust, repeatable paradigms to construct a PoP solution. They approach the problem in a diversity of ways and use a variety of different substrates in order to successfully authenticate human users: reverse Turing tests (Idena Network), social graph data emerging from a Web of Trust type of scheme (Duniter, BrightID, Humanity DAO, and Kleros), and intersectional approaches that combine an analysis of objective financial value and its distribution within different domains, with some other forms of human entropy or social signaling that can be detected online (Equality Protocol and Upala).

---

[31]"Deep-dive into the Web of Trust", (Duniter, 2018).

[32]Democracy Earth (2020). *Equality Protocol* [Github Post]. Retrieved from https://github.com/DemocracyEarth/paper (accessed October 15, 2020).

[33]Democracy Earth (2019). *Sovereign Dapp* [Github Repository]. Retrieved from https://github.com/DemocracyEarth/dapp (accessed October 15, 2020).

TABLE 1 | Summary of Approach Properties and Characteristics.

Approach	Governance	Size	Blockchain	Protocol	Substrate	Decentralization	Privacy	Scalability
Idena network	Public network	4,012	Idena	Synchronous reverse turing-test ceremonies	FLIPS	Decentralized identity registry management. Every participant can run a full validator/mining node. Change of the protocol requires network consensus	No personally identifying information (PII) sharing required. Node IP address observable	Node install, getting invite code and regular participation in the validation ceremonies required. Participation incentivized with mining and ceremony rewards (~$1.5–2/day)
Humanity DAO	DAO	640	Ethereum	Web of trust + token curated registry	Social media information	P2P vouching. Decentralized identity registry management	Social media account (Twitter) sharing required. Ethereum address observable	Web3 dapp interaction, paying Ethereum fees and identity stake required. Participation incentivized with a UBI (1 Dai/month)
Kleros	DAO + legal entity	–	Ethereum	Web of trust + token curated registry	Personal information, photos, and video selfies	P2P vouching. Decentralized identity registry management	Video selfie sharing required. Ethereum address observable	Web3 dapp interaction, paying Ethereum fees and identity stake required. Participation incentivized indirectly with court rewards
Upala	DAO	–	Ethereum	Price of forgery	Identity + controlled proof of stake	Decentralized identity registry management	No additional PII sharing required. Ethereum address observable	Web3 dapp interaction, paying Ethereum fees and identity stake required
BrightID	DAO	556	Ethereum/ IDChain	Intersectional web of trust	Existing social connections + weekly online meetups	P2P vouching. Semi-decentralized identity registry management	No PII sharing required. Social graph and IDChain address observable	Mobile app vouching required. Participation incentivized with IDChain native token airdrop that can be used to pay fees
Duniter	Public network	2,801	Duniter	Web of trust	Contact information + 5 connections to existing members	P2P vouching. Decentralized identity registry management	No PII sharing required. Social graph observable	Vouching meeting with 5 members required. Participation incentivized with universal dividends in Ğ1 cryptocurrency (~$1/day)
Equality protocol	DAO	529	Ethereum	Meta protocol: democracy index	Participation in DAOs	Decentralized identity registry management	No PII sharing required. Ethereum address observable	Web3 dapp interaction and paying ethereum fees required

Despite the clear shortcomings present in each of these methods, their creative uses of subjectivity can point toward interesting, hybrid approaches to verify Proof of Personhood. In fact, the distinction between methods for Sybil-protection may be overdrawn: most of the solutions outlined in this review employ a combination of them in order to secure their networks. In some cases, this combination of tactics is formalized, such as with Idena Network—predicated on reverse Turing tests but accessible through their invite code system, an instance of a Web of Trust approach. In other cases, additional protocols are appended tangentially and informally, such as BrightID's employment of weekly Pseudonym Parties to welcome new members. In this sense, the theoretical primitives have almost false separations: when it comes to implementation, they truly co-occur and build on one another rather than being contained by the distinctions we see explicated in the academy.

Below is a comparison of the different properties of PoP solutions as established in the literature. We consider each approach in light of not only its base-level attributes—governance structure, size, blockchain, protocol, and substrate—but also its performance on the primary desired properties of decentralization, privacy, and scalability.

As outlined in **Table 1**, each project entails significant tradeoffs, with each substrate forming or leading to a possible weakness in the system. Here we outline the research gaps evident in each protocol, with the hope of providing a path forward to addressing and solving these issues. We begin with the Idena Network, the only fully decentralized and privacy-preserving solution. Currently, the synchronous reverse Turing test model of the network requires a significant commitment of time and effort on the part of its participants, who must participate in regular validation ceremonies approximately every 2 weeks. While its Sybil-resistance strategy is currently effective, it remains to be

seen whether AI-hard tests will be able to resist the dynamic development of AI connected with neural networks and deep learning. Furthermore, it is not certain that the current incentive model put in place will be sufficient to disincentivize the creation of a marketplace for false identities with mechanical turk attacks.

Humanity DAO, while extremely promising, required the use of privately owned identity information from social networks like Twitter to verify identity, again exposing users to the vulnerabilities of Internet monopolies and largely defeating the aim of the initial motivators of such consensus identity proofs. The system also fell prey to attack due to its necessarily fixed protocol. Kleros requires that users submit a range of personal information and video proof—effectively biometrics, which is likely to prevent many from using the service, and it remains to be seen whether their system of reward and punishment will be sufficient to prevent dishonest vouching. Upala's social responsibility concept shows promise in preserving trust but may be more accessible to capital-rich users, given that uniqueness scores are in part relative to stake—although this may be contemplated by different authentication methodologies or governance rules established by groups adopting the protocol. Upala's protocol also runs the risk of suffering an "explosion" avalanche, with users exiting the protocol *en masse*.

Duniter's requirement of at least five vouching links and a maximum distance of five different connections from the referent members imply that it entails significant time for new certifications to be emitted. This deliberately slower rate exhibits good Sybil-protection properties but significantly restricts the network growth. This is not currently particularly problematic, as there is not yet a plurality of implementations that would allow for the creation of a proper benchmark for "adoption rate." What is already possible to infer, from the set of projects reviewed here, is that the adoption of subjective Proof of Personhood protocols is distinct from that of other authentication technologies since they grow more slowly, at a "human rate." However, Duniter's model is also restricted to its community geographics, which limits its application to local use cases as opposed to those intended for a global audience.

The Equality Protocol, an intersubjective consensus protocol to evaluate other protocols, does not form a substrate of identity verification in and of itself, and it is currently fairly restrictive in its scope, as it solely contemplates members of decentralized autonomous organizations. However, as discussed above, while participation rights can be earned through non-financial means, this system is particularly vulnerable to a concentration of power with capital-rich users, who can purchase the ability to join and participate in these organizations with more ease. In addition, the use of Quadratic Voting, while clearly appropriate as a consensus mechanism, has not yet been fully substantiated as a method of signaling the legitimacy of DAOs. The key point here is around a possible conflict of interest: the protocol takes as ground truth the user rating of DAOs according to their Sybil-resistance; however, there does not yet exist a verifiable mechanism within the protocol to confirm whether this *perception* of Sybil-resistance from the user is true or not. This may lead to misaligned incentives and inaccurate assessments.

Finally, BrightID is currently the most intersectional solution explored in this paper and thus may have significant scalability potential. However, its current reliance on establishing trust through connections to a small, trusted seed network makes it difficult for independent groups to self-authenticate. BrightID has a certain degree of centralization, as it relies on privately configured nodes to manage identity registries, selected by BrightID founding team, although there is promising potential for improvement with the introduction of the IDChain and integration of the seed selection and vouching process into IDChain-based DAO. One more possible hindrance to the adoption of BrightID stands in its reliance on a public social graph, which may compromise the privacy of authenticated users if the real-world identity of some of the participants is revealed. Finally, the Sybil-resistance of BrightID's GroupSybilRank algorithm has yet to be proven.

As four out of the seven solutions analyzed in this review rely primarily on a Web of Trust, it is important to note that presently there is no evidence of the Web of Trust schemes' effectiveness for Sybil-resistance in the presence of multiple attack vectors. Bad actors may forge multiple real relationships under different names in different groups: if there are enough non-intersecting small groups, an attacker may be able to grow a significant amount of Sybils over time. The prevention of such attacks often requires sophisticated data processing and modeling techniques: a notable example is Facebook's periodic take-down of, on average, two billion fake accounts per quarter using machine learning algorithms like SybilEdge, which employ behavioral and content classifiers to flag an account as abusive (Adam, 2020).

Thus, we see that there is still significant work to be done. One possible mode of inquiry is to look to PoP systems that are not directly blockchain-based but instead use more intersectional approaches. A theoretical approach to such a project was outlined by Nicole Immorlica et al. proposing a protocol of verifying identity through proofs of social intersection, extending the Web of Trust approach (Immorlica et al., 2019). This system would allow users to check the claims of others, with varying levels of trust, or credit, assigned to each user in relation to others; this system of credit could also extend to groups of users, as relevant, to further prevent false claims. Such a system has been partially implemented by Identiq, which has created a providerless, peer-to-peer network that allows for companies to collaborate to validate users[34]. However, Identiq is not only itself privately-owned and closed-source, it also puts validation power in the hands of corporations and thus does not provide a fully decentralized solution, particularly one that could be leveraged for civic engagement purposes.

Protocols that focus directly on social interaction are also relevant here. Consider Nomqa[35], an upcoming solution that verifies humanity by scoring interactions between users based on subjective meaning. This approach brings in the much-needed subjectivity component to identity solutions, considering collective, rather than purely individual, approaches to identity.

---

[34]Website: identiq.com
[35]Website: nomqa.com

Markedly more offline solutions have also been proposed through the use of "pseudonym parties": Personhood.online, a project developed at the École Polytechnique Fédérale de Lausanne[36], integrated physical gatherings with a DID architecture and next-generation blockchain technologies aimed at scalability and privacy. However, this ambitious effort has been inactive since 2018. Another insightful proposal was to produce a temporary proof of personhood based on physical attendance, forming a "seed set" comprised of said attendees (Howitt, 2020). These seeds can then validate other identities, creating trusted clusters, which can fan out and validate larger and larger sets and communities.

Additional possible directions of future inquiry include explorations of blind research into social networks—expanding the possibilities of establishing trust between nodes while maintaining their privacy[37] as well as anticollusion systems[38]. A prominent use case for Proof of Personhood solutions is in the context of blockchain-based voting. However, by generating a record of transactions, blockchains can facilitate bribery, with smart contracts created to reward users if they are able to demonstrate a certain voting pattern through a publicly verifiable transaction (Daian et al., 2018). Minimal Anti-collusion Infrastructure is a scheme, currently employed by BrightID[39], that aims to address these types of attacks by allowing voters to switch their voting keys at any time: thus, one may provide a voting receipt but can never guarantee that said vote had not been formerly invalidated by a key switch. While there are still possible vectors of attack (one could sell their private key), the Minimal Anti-collusion Infrastructure outlines a promising approach to address on-chain privacy for identities being used in voting mechanisms. Another proposal to "make honesty the optimal strategy" is to have each edge within a network acting as a prediction market: in case the legitimacy of a node is challenged, reputation flows from the losers to the winners.[40] This proposal is in line with Klero's approach of appending a distributed dispute resolution system into its protocol and using vouching stakes as a bounty available for network policing.

Any endeavor to create functioning digital democracies can be undermined by exploitation of identity, from the automated creation of false identities to corruption by third parties controlling a voter registry. Determining who has the right to participate cannot be an afterthought of democracy: it is its elemental task. However, it must also be noted that democratic governance is possible even with bounded Sybil penetration, meaning that a small amount of error within a system can be

forgiven, which opens up possibilities for more intersectional and subjective approaches (Shahaf et al., 2019).

Finally, the steady advancement of machine learning and artificial intelligence makes the question of formalizing identity frameworks particularly urgent. Trustworthy and high-quality information is the foundation of a functioning democracy—and yet from deep fakes to language model outputs, machine-generated information is becoming easier to generate and spread. In the future, there may be a need for cryptographic signatures on selected media or information pieces to establish trust and authenticity (Ford, 2020).

Thus, in many senses, governance, democracy, and identity are strictly correlated. Structuring communication architectures anchored on decentralized, privacy-preserving, self-sovereign, and Sybil-resistant identity protocols that can reach all humans with an Internet connection can open the path for new, radically participative peer-to-peer political movements and economies.

CONCLUSIONS

Identity is one of our most fundamental human attributes. However, in the age of surveillance capitalism, identity itself has become a part of a new, digital political frontier[41] (Zuboff, 2019). As Edward Snowden, one of the most prominent activists for the end of surveillance practices in the world, recently warned during a videoconference at the 2019 Web3 Summit in Berlin[42]: "The one vulnerability being exploited across all systems is Identity."

If the "State is the monopoly on violence" as Max Weber once defined it (Weber, 1919), then the Surveillance State (or Surveillance Capital) is the monopoly on identity. Consolidated credential mechanisms today all verify humans by implementing practices that require the disclosure of personal and private information to an identifier. Eventually, this wealth of information accrues into credential monopolies, which are a prominent force in the perilous drift toward democratic deconsolidation now threatening Western democracies. While there is significant space for action in advancing effective public policies that contemplate those threats, approving and enforcing them is often extremely challenging in the face of the powerful market forces they stand against. In that sense, the alternative technological paradigms that may arise from Proof of Personhood systems could provide a relevant path toward guaranteeing privacy and participation rights.

Furthermore, surveillance capitalism bears a worldview that downgrades human value and dignity in favor of machine learning systems. Proof of Personhood systems counter that logic by creating the building blocks of a human-centered economy, where individuals directly control and have governance rights over the networks, communities, and organizations they belong to. These systems invert the current logic of capitalism, creating the base for solidarity economies that can safeguard and elevate the role of human consciousness, choice, and agency.

Yes, the approaches explored in this review fall short of this goal in several ways, some still relying on existing sources

---

[36]Website: personhood.online

[37]Whitehat, B., and Gurkan, K. (2020). *Private Social Network Search* [Website]. Available online at: https://ethresear.ch/t/blind-find-private-social-network-search/6988 (accessed October 15, 2020).

[38]Buterin, V. (2019). Minimal Anti-collusion Infrastructure.

[39]Stallard, A. (2020b). *Anonymous Participation Using BrightID* [Blog]. Available online at: https://medium.com/brightid/anonymous-participation-e-g-voting-using-brightid-42a13b4d1c94 (accessed October 15, 2020).

[40]Watts, P. (2018). *Towards Proof-of-Person* [Github post]. Available online at: https://github.com/WebOfTrustInfo/rwot7-toronto/blob/master/topics-and-advance-readings/towards-proof-of-person.md (accessed October 15, 2020).

[41]"The Social Smart Contract", Democracy Earth, 2017

[42]Edward Snowden, Web3 Summit, 2019

---

of centralized information, others on small networks or high-friction synchronous tasks. Nonetheless, Proof of Personhood projects present one of the few viable alternatives capable of addressing these problems at their root. In doing so, they illustrate that the best technologies do not abstract away subjectivity. Instead they embrace it, seeing subjectivity for what it is: not just a necessity, but a strength.

AUTHOR CONTRIBUTIONS

PB and DS have co-written and contributed to the manuscript in its entirety. SI has primarily contributed to the Proof of Personhood section and the analysis of the decentralization, privacy, and scalability aspects of each protocol while also reviewing the manuscript in its entirety. SS has primarily contributed to the Introduction and Conclusion sections while also making significant contributions to the analysis of protocols and reviewing the manuscript in its entirety. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

REFERENCES

Adam, B. (2020). *Detecting Fake Accounts on Social Networks with SybilEdge. Facebook Research (blog).* Available online at: https://research.fb.com/blog/2020/04/detecting-fake-accounts-on-social-networks-with-sybiledge/. (accessed on April 21, 2020).

Andersen, G., and Vincent, C. (2016). "ATUCAPTS: automated tests that a user cannot pass twice simultaneously," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence IJCAI'16,* (New York, NY: AAAI Press), 3662–3669.

Asgaonkar, A., and Krishnamachari, B. (2018). Token curated registries - a game theoretic approach. *arXiv [Preprint].* Available online at: https://arxiv.org/abs/1809.01756

Beckers, K., and Maritta, H. (2012). "A foundation for requirements analysis of privacy preserving software," in *Multidisciplinary Research and Practice for Information Systems Lecture Notes in Computer Science,* eds. G. Quirchmayr, J. Basl, I. You, L. Xu, and E. Weippl (Berlinn, Heidelberg: Springer), 93–107. doi: 10.1007/978-3-642-32498-7_8

Benhabib, J., Alberto, B., and Shenghao, Z. (2014). The distribution of wealth in the blanchard–yaari model. *Macroecon. Dyn.* 20, 466–481. doi: 10.1017/S1365100514000066

Bennett, W. L. (2012). The personalization of politics: political identity, social media, and changing patterns of participation. *Ann. Am. Acad. Polit. Soc. Sci.* 644, 20–39. doi: 10.1177/0002716212451428

Borge, M., Kokoris-Kogias, E., Philipp, J., Linus, G., Nicolas, G., and Bryan, F. (2017). "Proof-of-personhood: redemocratizing permissionless cryptocurrencies," in *2017 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW),* (Paris: IEEE: Paris, France), 23–26. doi: 10.1109/EuroSPW.2017.46

Bruns, A., Gunn, E., Eli, S., Anders Olof, L., and Christian, C. (2015). *The Routledge Companion to Social Media and Politics.* New York, NY: Routledge. doi: 10.4324/9781315716299

Buterin, V. (2019). *Hard Problems in Cryptocurrency: Five Years Later.*

Buterin, V., Hitzig, Z., and Weyl, E. G. (2018). *Liberal Radicalism: A Flexible Design For Philanthropic Matching Funds.* Available online at SSRN: https://ssrn.com/abstract=3243656~doi:10.2139/ssrn.3243656 (accessed October 15, 2020).

Cao, Q., Sirivianos, M., Yang, X., and Pregueiro, T. *Aiding the Detection of Fake Accounts in Large Scale Social Online Services.* (2012). 15-15.

Chen, Y. P., and Ju-Chun, K. (2019). "CryptoAR wallet: a blockchain cryptocurrency wallet application that uses augmented reality for on-chain user data display," in *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services MobileHCI'19,* (Taipei, Taiwan: Association for Computing Machinery.), 1–5. doi: 10.1145/3338286.3344386

Daian, P., Kell, T., Miers, I., and Juels, A. (2018). *On-Chain Vote Buying and the Rise of Dark DAOs.*

De Filippi, P. (2019). *Blockchain Technology and Decentralized Governance: The Pitfalls of a Trustless Dream. SSRN Scholarly Paper ID 3524352.* Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.3524352

de Filippi, P., and Mcmullen, G. (2018). *Governance of Blockchain Systems: Governance of and by Distributed Infrastructure.* Research Report. Blockchain Research Institute; COALA.

De Filippi, P., Morshed, M., and Wessel, R. (2020). Blockchain as a confidence machine: the problem of trust & challenges of governance. *Technol. Soc.* 62:101284. doi: 10.1016/j.techsoc.2020.101284

Douceur, J. R. (2002). "The sybil attack," in *Revised Papers from the First International Workshop on Peer-to-Peer Systems IPTPS'01,* Berlin, Heidelberg: Springer-Verlag. 251–260. doi: 10.1007/3-540-45748-8_24

Duniter (2018). *Deep Dive into the Web of Trust.* Retrieved from https://duniter.org/en/deep-dive-into-the-web-of-trust/ (accessed October 15, 2020).

Dunphy, P., and Petitcolas, F. A. P. (2018). A first look at identity management schemes on the blockchain. *IEEE Secur. Privacy* 16, 20–29. doi: 10.1109/MSP.2018.3111247

Ferrara, E., Onur, V., Clayton, D., Filippo, M., and Alessandro, F. (2016). The rise of social bots. *Commun. ACM* 59, 96–104. doi: 10.1145/2818717

Ford, B. (2012). *Pseudonym Parties: An Offline Foundation for Online Accountability (PRELIMINARY DRAFT)*, March.

Ford, B. (2020). *Technologizing Democracy or Democratizing Technology? A Layered-Architecture Perspective on Potentials and Challenges.* Unpublished draft – final version to appear in forthcoming book from the University of Chicago Press.

Freedom House, (2018). *The Rise of Digital Authoritarianism. Freedom House 2018.* Available online at: https://freedomhouse.org/report/freedom-net/2018/rise-digital-authoritarianism (accessed October 15, 2020).

Freedom House, (2020). *A Leaderless Struggle for Democracy. Freedom House 2020.* Available online at: https://freedomhouse.org/report/freedom-world/2020/leaderless-struggle-democracy (accessed October 15, 2020).

Harrigan, M., Shi, L. and Illum, J. (2018). "Airdrops and privacy: a case study in cross-blockchain analysis," in *2018 IEEE International Conference on Data Mining Workshops (ICDMW),* (Singapore), 63–70. doi: 10.1109/ICDMW.2018.00017

Howitt, A. (2020). Proposal for a decentralized unique identity seeding protocol.

Idena Network, (2019). *AI-Resistant CAPTCHAs: Are they really possible?*

Idena Network, (2020). What is a flip. Available online at: https://idena.io/?view=flip_challenge (accessed October 15, 2020).

Immorlica, N., Jackson, M. O., and Weyl, E. G. (2019). *Verifying Identity as a Social Intersection. SSRN Scholarly Paper ID 3375436.* Rochester, NY: Social Science Research Network. doi: 10.2139/ssrn.3375436

Klass, O. S., Ofer, B., Moshe, L., Ofer, M., and Sorin, S. (2006). The forbes 400 and the pareto wealth distribution. *Econ. Lett.* 90, 290–95. doi: 10.1016/j.econlet.2005.08.020

Lalley, S., and Glen, W. E. (2018). "Quadratic voting how mechanism design can radicalize democracy (December 24, 2017)," in *American Economic Association Papers and Proceedings.* 1, 1–5. doi: 10.2139/ssrn.2003531

Lasaege, C., Ast, F., and George, W., (2019). *Kleros: Short Paper.*

Levesque, H., Davis, E., and Morgenstern, L. (2012). "*The winograd schema challenge,*" in *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. Available online at: http://www.aaai.org/ocs/index.php/KR/KR12/paper/download/4492/4924

Minks, T. L. (2017). Ethereum and the SEC: why most distributed autonomous organizations are subject to the registration requirements of the securities act of 1933 and a proposal for new regulation. *Tex. AM L Rev.* 5:405. doi: 10.2139/ssrn.3124544

Mori, M., MacDorman, K. F., and Kageki, N. (2012). The uncanny valley [From the Field]. *IEEE Robot. Autom. Mag.* 19, 98–100. doi: 10.1109/MRA.2012.2192811

Moritz, J. (2020). The "IDENA AI" with 60% accuracy at solving FLIPS.

Nakamoto, S. (2009). *Bitcoin: A Peer-to-Peer Electronic Cash System. Cryptography Mailing List*. Available online at: https://Metzdowd.Com, March

Norta, A. (2015). "Creation of smart-contracting collaborations for decentralized autonomous organizations," in *International Conference on Business Informatics Research*, (Springer), 3–17. doi: 10.1007/978-3-319-21915-8_1

Nygren, J. (2019). Pseudonym Pairs : A foundation for proof-of-personhood in the web 3.0 jurisdiction. doi: 10.5281/zenodo.3380889

Schützeichel, R. (2013). "Georg simmel: soziologie. untersuchungen über die formen der vergesellschaftung," in *Hauptwerke der Emotionssoziologie*, eds. K. Senge and R. Schützeichel (Wiesbaden: Springer Fachmedien), 311–32. doi: 10.1007/978-3-531-93439-6_45

Selloni, D. (2017). "New forms of economies: sharing economy, collaborative consumption, peer-to-peer economy," in *CoDesign for Public-Interest Services. Research for Development*. (Cham: Springer, Cham). doi: 10.1007/978-3-319-53243-1_2

Shahaf, G., Ehud, S., and Nimrod, T. (2019). *Genuine Personal Identifiers and Mutual Sureties for Sybil-Resilient Community Formation*.

Srinivasan, B. S. (2017). *Quantifying Decentralization. Medium*. Available online at: https://news.earn.com/quantifying-decentralization-e39db233c28e (accessed on October 31, 2017).

Swathi, P, Chirag, M., and Dhiren, P. (2019). "Preventing sybil attack in blockchain using distributed behavior monitoring of miners," in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–6. doi: 10.1109/ICCCNT45670.2019.8944507

Szabo, N. (2017). Money, Blockchains, and Social Scalability. *Unenumerated*. Available online at: http://unenumerated.blogspot.com/2017/02/money-blockchains-and-social-scalability.html (accessed October 15, 2020).

Teutsch, J. (2017). *On Decentralized Oracles for Data Availability*. Available online at: https://people.cs.uchicago.edu/~teutsch/papers/decentralized_oracles.pdf (accessed October 15, 2020).

Tufekci, Z. (2017). *Twitter and Tear Gas: The Power and Fragility of Networked Protest*. Yale University Press.

Viswanath, B., Gummadi, K. P., Ansley, P., and Alan, M. (2010). "An analysis of social network-based sybil defenses," in *SIGCOMM '10: Proceedings of the ACM SIGCOMM 2010 conference August 2010*, 363–374. doi: 10.1145/1851182.1851226

Viswanath, B., Mainack, M., Gummadi, K. P., Ansley, P., and Alan, M. (2012). "Canal: scaling social network-based sybil tolerance schemes," in *EuroSys '12:*

Proceedings of the 7th ACM European Conference on Computer Systems April 2012, 309–322.

von Ahn, L., Blum, M., Hopper, N. J., Langford, J. (2003). "CAPTCHA: Using hard ai problems for security," in *Advances in Cryptology — EUROCRYPT 2003 Lecture Notes in Computer Science*, ed E. Biham (Berlin, Heidelberg: Springer.), 294–311. doi: 10.1007/3-540-39200-9_18

Vyjayanti, D. (2017). *Counting the Uncounted: 1.1 Billion People without IDs. World Bank 2017*. Available online at: https://blogs.worldbank.org/digital-development/counting-uncounted-11-billion-people-without-ids (accessed October 15, 2020).

Wang, F., and De Filippi, P. (2020). *Self-Sovereign Identity in a Globalized World: Credentials-Based Identity Systems as a Driver for Economic Inclusion. SSRN Scholarly Paper ID 3524367*. Rochester, NY: Social Science Research Network. Available online at: https://papers.ssrn.com/abstract=3524367. doi: 10.3389/fbloc.2019.00028

Weber, M. (1919). "*Politics as a Vocation*". Originally a speech at Munich University, 1918, published in 1919 by Duncker & Humblodt, Munich as "Politik als Beruf". Munich.

Wilson, D., and Giuseppe, A. (2015). From pretty good to great: enhancing PGP using bitcoin and the blockchain. *Netw. Syst. Secur.* 9408, 368–375. doi: 10.1007/978-3-319-25645-0_25

Zellers, R., Yonatan, B., Ali, F., and Yejin, C. (2018). From recognition to cognition: visual commonsense reasoning. *arXiv.* 6713–6724. doi: 10.1109/CVPR.2019.00688

Zuboff, S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs.