A New Identity and Financial Network

Disclaimer

This crypto-asset white paper has not been approved by any competent authority in any Member State of the European Union. The person seeking admission to trading is solely responsible for the content of this crypto-asset white paper.

This cryptoasset white paper complies with Title II of Regulation (EU) 2023/1114 and, to the best of the knowledge of the management body, the information presented in the crypto-asset white paper is fair, clear and not misleading and the crypto-asset white paper makes no omission likely to affect its import.

Introducing Worldcoin

Worldcoin was founded with the mission of creating a globally-inclusive identity and financial network, owned by the majority of humanity. If successful, Worldcoin could considerably increase economic opportunity, scale a reliable solution for distinguishing humans from AI online while preserving privacy, enable global democratic processes, and show a potential path to AI-funded UBI.

Worldcoin consists of a privacy-preserving digital identity network (World ID) built on proof of personhood and, where laws allow, a digital currency (WLD). Every human is eligible for a share of WLD simply for being human. World ID and WLD are currently complemented by World App, the first frontend to World ID and the Worldcoin Protocol, developed by the contributor team at Tools for Humanity (TFH).

“Proof of personhood” is one of the core ideas behind Worldcoin, and refers to establishing an individual is both human and unique. Once established, it gives the individual the ability to assert they are a real person and different from another real person, without having to reveal their real-world identity.

Today, proof of personhood is an unsolved problem on a global scale, making it difficult to vote online or distribute value on a large scale. The problem is even more pressing as

increasingly powerful AI models will further amplify the difficulty of distinguishing humans from bots. If successful as part of Worldcoin, World ID could become a global proof of personhood standard.

Some of the core assumptions behind Worldcoin are:

1. Proof of personhood is a missing and necessary digital primitive. This primitive will become more important as increasingly powerful AI models become available.
2. Scalable and inclusive proof of personhood, for the first time, allows aligning the incentives of all network participants around adding real humans to the network. Bitcoin is issued to secure the Bitcoin network. Worldcoin is issued to grow the Worldcoin network, with security inherited from Ethereum.
3. In a time of increasingly powerful AI, the most reliable way to issue a global proof of personhood is through custom biometric hardware.

The following dynamic Whitepaper shares the reasoning behind the implementation of the project as well as the current state and roadmap.

World ID

World ID is privacy preserving proof of personhood. It enables users to verify their humanness online via a custom biometric device called the Orb. The Orb has been designed based on the realization that custom biometric hardware might be the only long term viable solution to issue AI-safe proof of personhood verifications. World IDs are issued on the Worldcoin protocol, which allows individuals to prove that they are human to any verifier (including web2 applications) while maintaining their privacy through zero-knowledge proofs. In the future, it should be possible to issue other credentials on the protocol as well.

World ID aspires to be personbound, meaning a World ID should only be used by the individual it was issued to. It should be very difficult to use by a fraudulent actor who stole or acquired World ID credentials.

Worldcoin Token

While network effects will ultimately come from useful applications being built on top of the financial and identity infrastructure, the token is issued to all network participants to align their incentives around the growth of the network. This is especially important early on to bootstrap the network and bypass the "cold start problem".

World App

World App is the first frontend to World ID: it guides individuals through the verification with the Orb, custodies an individual's World ID credentials and implements the cryptographic protocols to share those credentials with third parties in a privacy preserving manner. It is designed to provide frictionless access to global decentralized financial infrastructure. Eventually, there should be many different wallets integrating World ID.

How does Worldcoin Work?

Worldcoin revolves around World ID, a privacy-preserving global identity network. Using World ID, individuals will be able to prove that they are a real, unique human to any platform that integrates with the protocol. This will enable fair airdrops, provide protection against bots/sybil attacks on social media, and enable the fairer distribution of limited resources. Furthermore, World ID can also enable global democratic processes and novel forms of governance (e.g., via quadratic voting), and it may eventually support a path to AI-funded UBI.

To engage with the Worldcoin protocol, individuals must first download World App, the first wallet app that supports the creation of a World ID. Individuals visit a physical imaging device called the Orb to get their World ID *Orb-verified*. Most Orbs are operated by a network of independent local businesses called Orb Operators. The Orb uses multispectral sensors to verify humanness and uniqueness to issue an Orb-verified World ID.

Potential Applications

Worldcoin could significantly increase equality of opportunity globally by advancing a future where everyone, regardless of their location, can participate in the global digital economy through universally-accessible decentralized financial and identity infrastructure. As the network grows, so should its utility.

Today, many interactions in the digital realm are not possible globally. The way humans transact value, identify themselves, and interact on the internet is likely to change fundamentally. With universal access to finance and identity, the following future becomes possible:

Finance

Owning & Transferring Digital Money: Sending money will be near instant and borderless, globally. Available to everyone. The world could be connected financially and everyone would be able to interact economically on the internet. The COVID relief fund for India, where over $400 million was raised in a short period of time by individuals around the world to support the country is a hint at what can be possible. Overall, this has the potential to connect people on a global scale unlike anything previously seen in human history.

Digital money is safer than cash, which can be more easily stolen or forged. This is especially important in crisis situations where instant cross-border financial transactions need to be possible, such as during the Ukrainian refugee crisis, where USDC was used to distribute direct aid. Additionally, digital money is an asset that individuals can own and control directly without having to trust third parties.

Identity

Keep the Bots Out: Bots on Twitter, spam messages, and robocalls are all symptoms of the lack of sound and frictionless digital identity. These issues are exacerbated by rapidly advancing AI models, which can solve CAPTCHAs and produce content that is convincingly "human". As services ramp up defenses against such content, it becomes essential that an inclusive and privacy-preserving solution for proof of personhood is available as public infrastructure. If every message or transaction included a "verified human" property, a lot of noise could be filtered from the digital world.

Governance: Currently, collective decision making in web3 largely relies on token-based governance (one token, one vote), which excludes some people from participating and heavily favors those with more economic power. A reliable sybil-resistant proof of personhood like World ID opens up the design space for global democratic governance mechanisms not just in web3 but for the internet. Additionally, for AI to maximally benefit all humans, rather than just a select group, it will become increasingly important to include everyone in its governance.

Intersection of Finance and Identity

Incentive Alignment: Coupons, loyalty programs, referral programs and more generally sharing value with customers are traditionally prone to fraud as the incentives for fraudulent actors are high. Frictionless and fraud resistant digital identity helps to align incentives and benefit both consumers and companies. This could even incept a new wave of companies owned in part by their users.

Equal Distribution of Scarce Resources: Crucial elements of modern society, including subsidies and social welfare, can be rendered more equitably by employing proof of personhood. This is particularly pertinent in developing economies, where social benefit programs confront the issue of resource capture—fake identities employed to acquire more than a person's fair share of resources. In 2021, India saved over $500 million in subsidy programs by implementing a biometric-based system that reduced fraud. A decentralized proof of personhood protocol can extend similar benefits to any project or organization globally. As AI advances, fairly distributing access and some of the created

value through UBI will play an increasingly vital role in counteracting the concentration of economic power. World ID could ensure that each individual registers only once and to guarantee equitable distribution.

Proof of Personhood (PoP)

Different applications have different requirements for PoP. For high-stakes use cases such as global UBI, the democratic governance of AI and the Worldcoin project, a highly secure and inclusive PoP mechanism to prevent multiple registrations is needed. Therefore, the Worldcoin developer community with the Worldcoin Foundation is laying the foundations for a high-assurance PoP mechanism with World ID. A World ID is issued upon World App installation and is fully verified when the holder visits an Orb and conducts the biometric verification. The following sections walk through the fundamental building blocks of PoP and how those are implemented in the context of World ID.

Building Blocks

On a high level, there are several building blocks that are required for an effective PoP mechanism. These include “deduplication” to ensure everyone can only verify once, “authentication” to ensure only the legitimate owner of the proof of personhood credential can use it and “recovery” in case of lost or compromised credentials. This section discusses these building blocks on a high level.

A proof of personhood mechanism consists of three different actors and the data that they exchange.

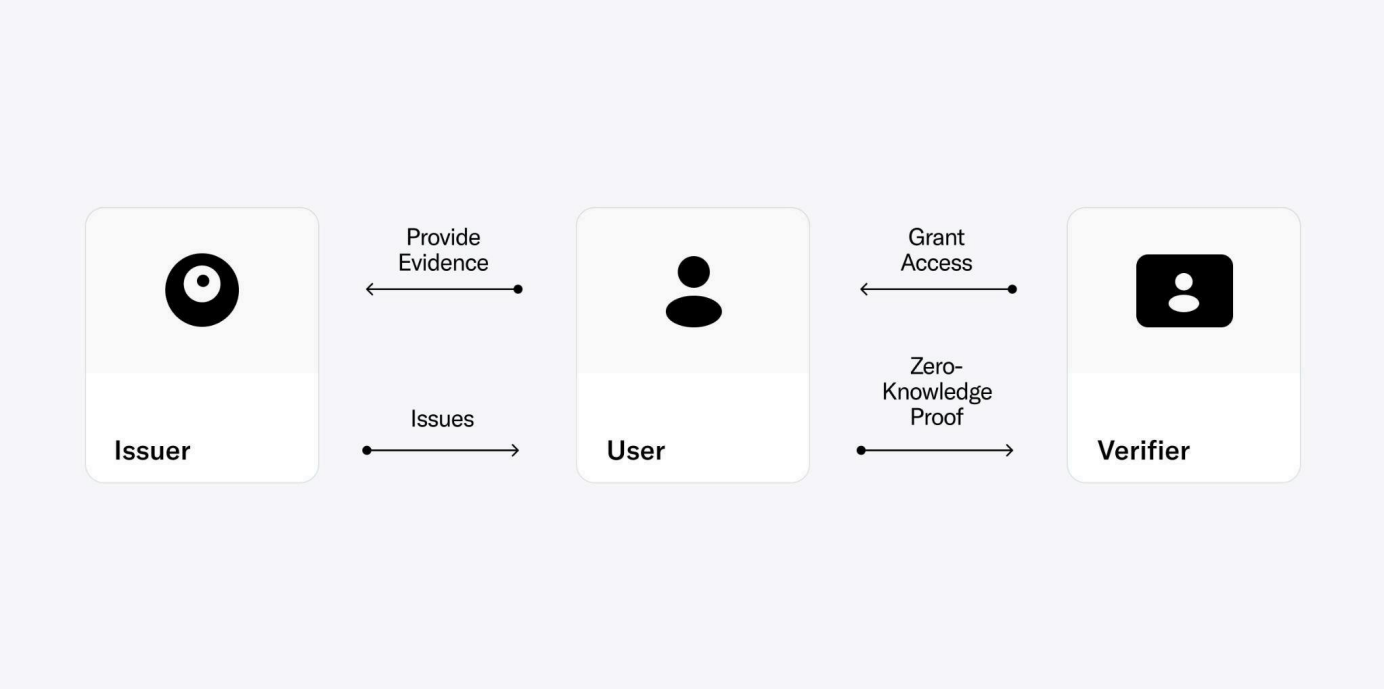

Fig. 1

Highly simplified diagram describing the interaction of the different actors of a proof of personhood ecosystem that are required for a user to authenticate as human.

For the context of this section, these terms are defined as follows:

- **User**: An individual seeking to prove specific claims about herself in order to access certain resources or more generally qualify for certain actions. Within the context of a PoP protocol those claims are related to proving uniqueness and personhood.
- **Credential**: A collection of data that serves as proof for particular attributes of the user that indicate the user is a human being. This could be a range of things, from the possession of a valid government ID to being verified as human and unique through biometrics.
- **Issuer**: An trusted entity that affirms certain information about the user and grants them a PoP credential, which enables the user to prove their claims to others.
- **Verifier**: An entity that examines a user's PoP credential and checks its authenticity as part of a verification process to grant the user access to certain actions.

Certain interactions between users, issuers and verifiers, like deduplication, recovery and authentication are important building blocks for a functional PoP mechanism. This section gives a high level overview of the building blocks of a general PoP mechanism.

Detailed explanations on how those are implemented with World ID follow in later sections.

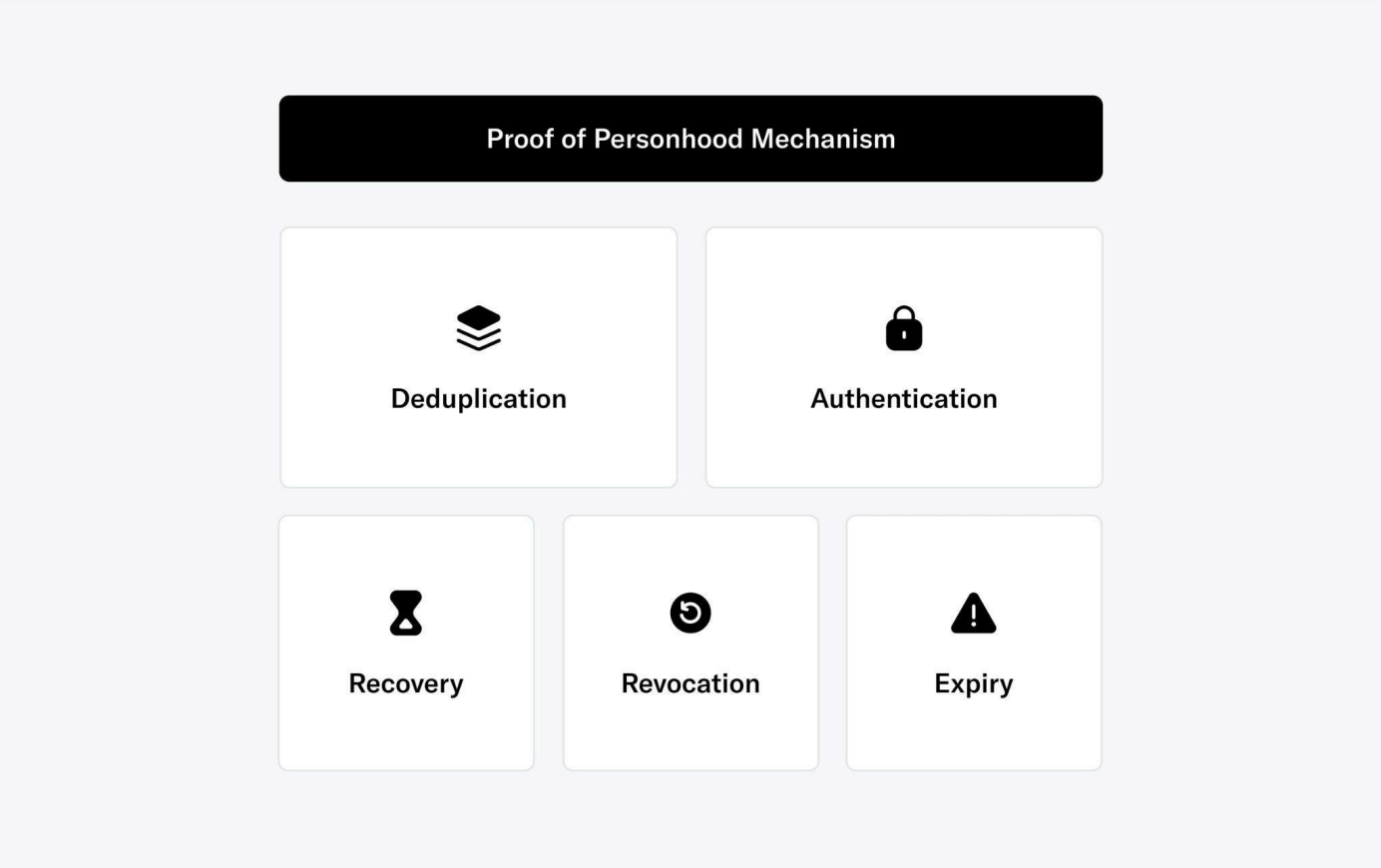

Fig. 2

2: Visualization of the different building blocks that make up an effective proof of personhood mechanism

Deduplication

For a PoP to be useful, it needs to have a notion of uniqueness. If the PoP can be acquired multiple times and transferred to fraudulent actors or bots, it cannot be trusted and fails to serve its purpose. Therefore, a PoP mechanism needs to deduplicate between the users that are issued a proof of personhood credential. This is the hardest challenge for any PoP mechanism.

Authentication

To make PoP credentials useful it needs to be hard to transfer credentials to someone else (e.g. bots) and for them to use the credentials to prevent fraud. This is especially important to protect individuals who may be unaware of the consequences of selling their credentials. This challenge is inherent in identity systems as a whole. Authentication can prevent fraudsters from using credentials, even if the respective user is unaware or attempts to collaborate with the fraudster.

When issuing PoP credentials, issuers only need to validate that someone is indeed a unique person. Beyond that, no additional personal information is required. However, each PoP credential needs to be uniquely tied to a specific person. Even if credentials are not transferable, wallets and phones can be transferred. Therefore, for high-integrity use cases, it is crucial to authenticate the user as the rightful owner of the PoP credential. This prevents the unauthorized use of credentials. A similar approach is followed during e.g. airline boarding, where an airline gate assistant verifies both the possession of a valid travel document and the consistency of the individual's identity with the document.

Recovery

If the user has lost access to their credentials or their credentials have been compromised, effective recovery mechanisms are needed. However, in setups where users are responsible for managing their own keys, this is a significant challenge. In the context of a PoP protocol, there are multiple mechanisms that can be used:

- **Restoring a User-Managed Backup**: The simplest method for credential recovery involves storing encrypted user-managed backups of their credentials. This allows users to restore their credentials, such as on a new device when their previous one is lost.
- **Social Recovery**: If no user-managed backup exists, but the user has set up social recovery, the credentials can be recovered through the help of friends and family.
- **Recover Keys**: If neither backups nor social recovery are available, the user needs to return to the issuer to regain access to their original credential. The user needs to

prove to the issuer that they are the legitimate owner of a certain credential. Upon successful authentication, the issuer grants access to the credential again. This process is similar to obtaining a new government ID after losing the previous one. The user can get a new ID with the same information on it[1]. This process may not be viable for some credentials: for example, if a private key was generated by the user and only the public key is recorded by the issuer (e.g. World ID).

- **Re-Issuance**: In situations where regaining access to the original credential through the issuer is not possible or undesirable (e.g. due to identity theft). In that case, re-issuance provides a way to invalidate the previous credential and issue a new credential. This can be compared to freezing a credit card and ordering a new one. Importantly, the availability of a re-issuance mechanism to rotate keys makes the illegitimate acquisition of other individuals' PoP credentials financially unviable from a game-theoretic perspective. The true holder of the credential can always recover their credentials and invalidate the bought/stolen credential. However, this does not protect against all cases of identity transfer, especially those that involve collusion or coercion.

Two other properties add to the integrity of a PoP mechanism:

Revocation

While the hope is that all participants act with integrity, this cannot be assumed. In instances where an issuer is found to be compromised or malicious, the impact can be mitigated by issuers or developers removing affected PoP credentials from their list of accepted credentials. If the issuance of a credential is decentralized across multiple issuing locations and only a subset is affected, the respective subset could be revoked by the issuing authority itself. An example in terms of today's credentials could be a university granting a diploma to a person who hasn't met all the criteria. If the fraud is identified, the diploma is revoked.

Expiry

The efficacy of security mechanisms degrades over time and new mechanisms are continuously being developed. As a result, many identity systems incorporate a predefined expiry date to credentials at the point of issuance. An example are passports. Although expiry is not required for a PoP mechanism to work, its inclusion can increase the PoP's integrity.

The combination of the mentioned building blocks make up for a functional proof of personhood mechanism. An exemplary smartphone App is shown in the following figure.

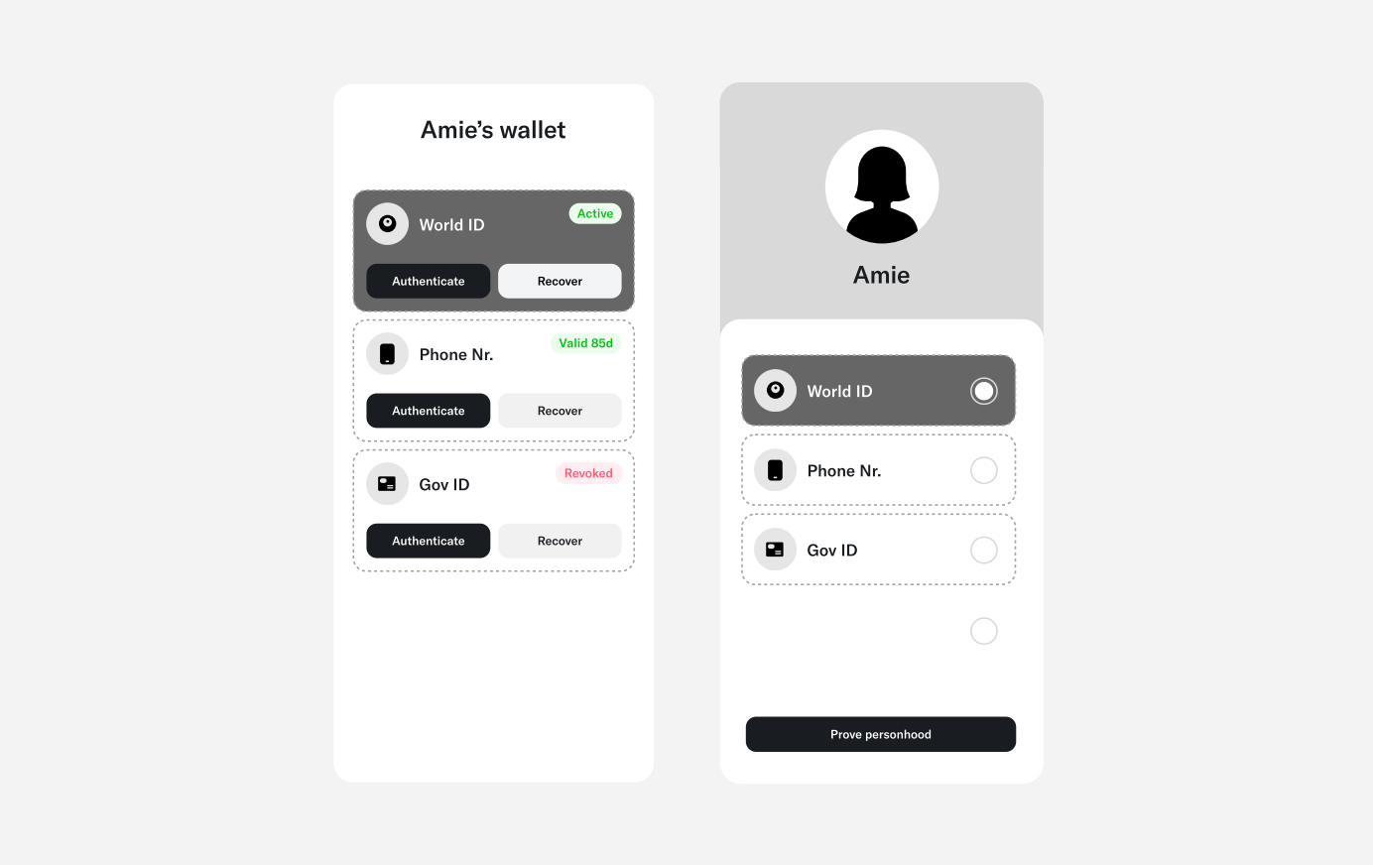

Fig. 3

Illustrated is a wallet that holds various proof of personhood credentials granted by different issuers. The credentials can be used to provide assurance to a verifier that a given user is indeed a human in order for the verifier to accept and perform a transaction.

Solving PoP at Scale

Based on these high level building blocks, several requirements can be deduced to evaluate different approaches to a global PoP mechanism:

- **Inclusivity and scalability**: A global PoP should be maximally inclusive, i.e. available to everyone. This means the mechanism should be able to distinguish between billions of people. There should be a feasible path to implementation at a global scale and people should be able to participate regardless of nationality, race, gender or economic means.
- **Fraud Resistant:** For a global proof of personhood, the important part is not "identification" (i.e. "is someone who they claim they are?"), but rather negative identification (i.e."has this person registered before?"). This means that fraud prevention, in terms of preventing duplicate sign-ups, is critical. A significant amount of duplicates would severely restrict the design space of possible applications and make it impossible to treat all humans equally. This would have severe implications for use cases like a fair token distribution, democratic governance, reputation systems like credit scores, and welfare (including UBI).
- **Personbound:** Once a proof of personhood is issued, it should be *personbound:* it should be hard to sell or steal (i.e. transfer) and hard to lose. Note that if the PoP mechanism is designed properly, this wouldn't prevent pseudonymity. This leads to the requirement that the PoP mechanism should allow for authentication in a way that makes it hard for fraudsters to impersonate the legitimate individual. Further, even if the individual lost all information, irrespective of any past actions, it should always be possible for them to recover.

Those cover the requirements that can be deduced from the required building blocks of a proof of personhood mechanism. However, there are further important requirements that can be deduced from the values inherent to the Worldcoin project:

- **Decentralization**: The issuance of a global PoP credential is foundational infrastructure that should not be controlled by a single entity to maximize resilience and integrity.

- **Privacy:** The PoP mechanism should preserve the privacy of individuals. Data shared by individuals should be minimized. Users should be in control of their data.

Mechanisms to Verify Uniqueness Among Billions

Based on the above requirements, this section compares different mechanisms to establish a global PoP mechanism in the context of the Worldcoin project.

Proof of Personhood Mechanisms

	Online Accounts	KYC	Web of Trust	Social Graph Analysis	Biometrics
Privacy	Possible	Possible	Possible	Possible	Possible
Fraud Resistance	No	Possible	No	No	Possible
Inclusivity & Scalability	Possible	No	Possible	Possible	Possible
Decentralization	Possible	No	Possible	Possible	Possible
Personbound	No	Possible	Possible	Possible	Possible

Fig. 4

An overview of proof of personhood mechanisms. Worldcoin contributors' research concluded that biometrics is the only method that can fulfill all essential requirements, provide the system is implemented appropriately

Online accounts

The simplest attempt to establish PoP at scale involves using existing accounts such as email, phone numbers and social media. This method fails, however, because one person can have multiple accounts on each kind of platform. Further, accounts aren't personbound i.e. they can be easily transferred to others. Also, the (in)famous CAPTCHAs, which are commonly used to prevent bots, are ineffective here because any human can pass multiple of them. Even the most recent implementations[2] that basically rely on an internal reputation system, are limited.

In general, current methods for deduplicating existing online accounts (i.e. ensuring that individuals can only register once), such as account activity analysis, lack the necessary fraud resistance to withstand substantial incentives. This has been demonstrated by large-scale attacks targeting even well-established financial services operations.

Official ID verification (KYC)

Online services often request proof of ID (usually a passport or driver's license) to comply with *Know your Customer* (KYC) regulations. In theory, this could be used to deduplicate individuals globally, but it fails in practice for several reasons.

KYC services are simply not inclusive on a global scale; more than 50% of the global population does not have an ID that can be verified digitally. Further, it is hard to build KYC verification in a privacy–preserving way. When using KYC providers, sensitive data needs to be shared with them. This can be solved using zkKYC and NFC readable IDs. The relevant data can be read out by the user's phone and be locally verified as it is signed by the issuing authority. Proving unique humanness can be achieved by submitting a hash based on the information of the user's ID without revealing any private information. The main drawback of this approach is that the prevalence of such NFC readable IDs is considerably lower than that of regular IDs.

Where NFC readable IDs are not available, ID verification can be prone to fraud—especially in emerging markets. IDs are issued by states and national governments, with no global system for verification or accountability. Many verification services (i.e. KYC

providers) rely on data from credit bureaus that is accumulated over time, hence stale, without the means to verify its authenticity with the issuing authority (i.e. governments), as there are often no APIs available. Fake IDs, as well as real data to create them, are easily available on the black market. Additionally, due to their centralized nature, corruption at the level of the issuing and verification organizations cannot be eliminated.

Even if the authenticity of provided data can be verified, it is non-trivial to establish global uniqueness among different types of identity documents: fuzzy matching between documents of the same person is highly error-prone. This is due to changes in personal information (e.g. address), and the low entropy captured in personal information. A similar problem arises as people are issued new identity documents over time, with new document numbers and (possibly) personal information. Those challenges result in large error rates both falsely accepting and rejecting users. Ultimately, given the current infrastructure, there is no way to bootstrap global PoP via KYC verification due to a lack of inclusivity and fraud resistance.

Web of Trust

The underlying idea of a “web of trust” is to verify identity claims in a decentralized manner.

For example, in the classic web of trust employed by PGP, users meet for in-person “key signing parties” to attest (via identity documents) that keys are controlled by their purported owners. More recently, projects like Proof of Humanity are building webs of trust for Web3. These allow decentralized verification using face photos and video chat, avoiding the in-person requirement.

Because these systems heavily rely on individuals, however, they are susceptible to human error and vulnerable to sybil attacks. Requiring users to stake money can increase security. However, doing so increases friction as users are penalized for mistakes and therefore disincentivized to verify others. Further, this decreases inclusivity as not everyone might be willing or able to lock funds. There are also concerns related to privacy (e.g. publishing face images or videos) and susceptibility to fraud using e.g. deep

fakes, which make these mechanisms fail to meet some of the design requirements mentioned above.

Social graph analysis

The idea of social graph analysis is to use information about the relationships between different people (or the lack thereof) to infer which users are real.

For example, one might infer from a relationship network that users with more than 5 friends are more likely to be real users. Of course, this is an oversimplified inference rule, and projects and concepts in this space, such as EigenTrust, Bright ID and soulbound tokens (SBTs) propose more sophisticated rules. Note that SBTs aren't designed to be a proof of personhood mechanism but are complementary for applications where proving *relationships* rather than *unique humanness* is needed. However, they are sometimes mentioned in this context and are therefore relevant to discuss.

Underlying all of these mechanisms is the observation that social relations constitute a unique human identifier if it is hard for a person to create another profile with sufficiently diverse relationships. If it is hard enough to create additional relationships, each user will only be able to maintain a single profile with rich social relations, which can serve as the user's PoP. One key challenge with this approach is that the required relationships are slow to build on a global scale, especially when relying on parties like employers and universities. It is a priori unclear how easy it is to convince institutions to participate, especially initially, when the value of these systems is still small. Further, it seems inevitable that in the near future AI (possibly assisted by humans acquiring multiple "real world" credentials for different accounts) will be able to build such profiles at scale. Ultimately, these approaches require giving up the notion of a unique human entirely, accepting the possibility that some people will be able to own multiple accounts that appear to the system as individual unique identities.

Therefore, while valuable for many applications, the social graph analysis approach also does not meet the fraud resistance requirement for PoP laid out above.

Biometrics

Each of the systems described above fails to effectively verify uniqueness on a global scale. The only mechanism that can differentiate people in non-trusted environments is their biometrics. Biometrics are the most fundamental means to verify both humanness and uniqueness. Most importantly, they are universal, enabling access irrespective of nationality, race, gender or economic means. Additionally, biometric systems can be highly privacy-preserving if implemented properly. Further, biometrics enable the previously mentioned building blocks by providing a recovery mechanism (that works even if someone has forgotten everything) and can be used for authentication. Therefore, biometrics also enable the PoP credential to be personbound.

Different systems have different requirements. Authenticating a user via FaceID as the rightful owner of a phone is very different from verifying billions of people as unique. The main differences in requirements relate to accuracy and fraud resistance. With FaceID, biometrics are essentially being used as a password, with the phone performing a single 1:1 comparison against a saved identity template to determine if the user is who they claim to be. Establishing global uniqueness is much more difficult. The biometrics have to be compared against (eventually) billions of previously registered users in a 1:N comparison. If the system is not accurate enough, an increasing number of users will be incorrectly rejected.

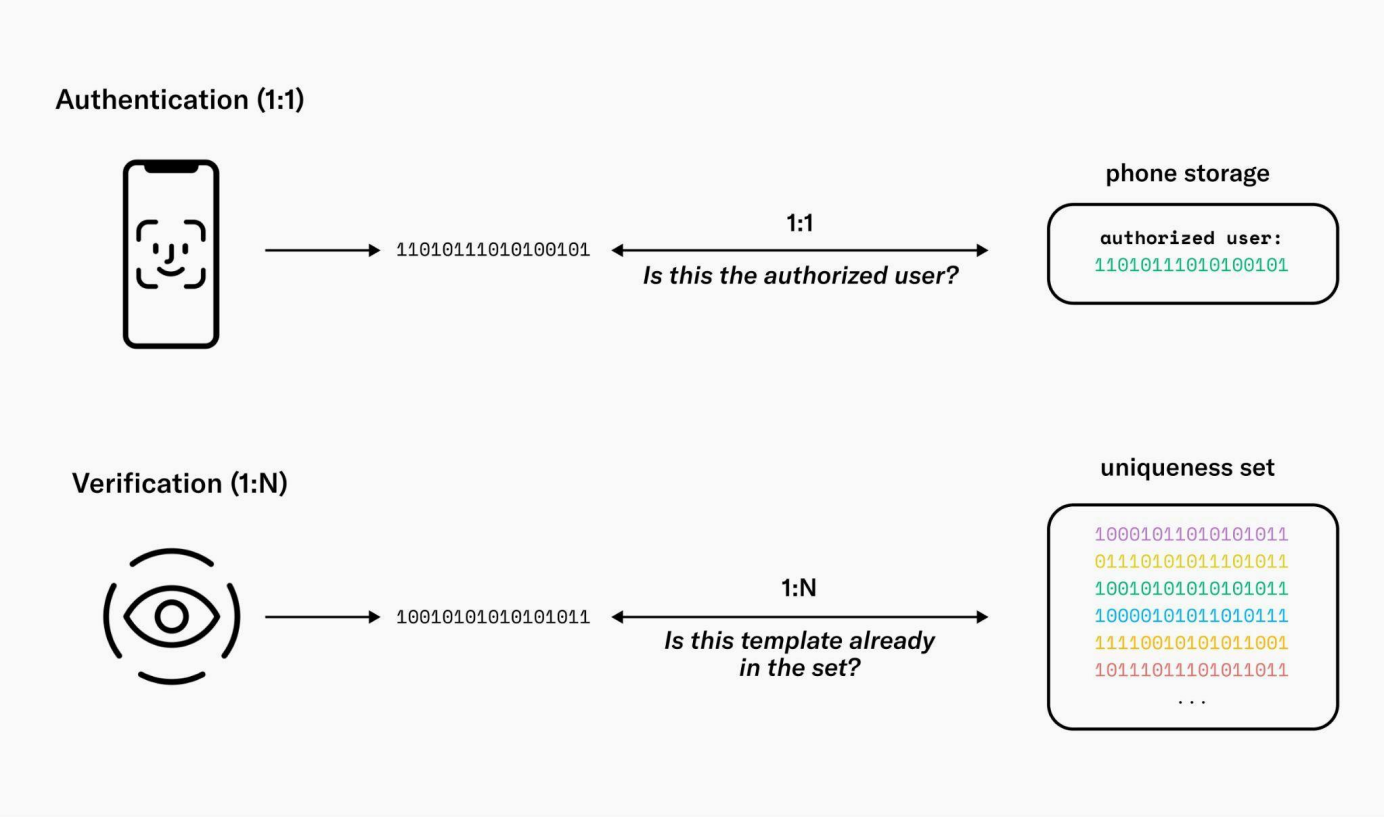

Fig. 5

Regarding biometrics, there are two modes to consider. The simpler mode is 1:1 authentication, comparing a user's template against a single previously enrolled template (e.g., Face ID). For global proof of personhood, 1:N verification is needed, comparing a user's template against a large set of templates to prevent duplication registrations.The error rates and therefore the inclusivity of the system are majorly influenced by the statistical characteristics of the biometric features being used. Iris biometrics outperform other biometric modalities and can achieve false match rates beyond 2.5×10^{-14} (or one false match in 40 trillion). This is several orders of magnitude more accurate than the current state of the art in face recognition. Moreover, the structure of the iris exhibits remarkable stability over time.

The error rates and therefore the inclusivity of the system are majorly influenced by the statistical characteristics of the biometric features being used. Iris biometrics outperform other biometric modalities and can achieve false match rates beyond 2.5×10^{-14} (or one false match in 40 trillion). This is several orders of magnitude more accurate than the current state of the art in face recognition. Moreover, the structure of the iris exhibits remarkable stability over time.

Biometric Modalities

	Fingerprint	Face	DNA	Iris
Privacy	Possible	Possible	Hard	Possible
Accuracy for global scale	Not enough	Not enough	Sufficient	Sufficient
Scalability	High	High	Low	High
Modification	Easy	Medium	Hard	Hard
Liveness detection	Hard	Good	Hard	Good

Fig. 6

An overview of different biometrics modalities reveals that iris biometrics is the only modality that can fulfill all essential requirements. While each modality has its advantages and disadvantages, iris biometrics stands out as the most reliable and accurate method for verification of humanness and uniqueness on a global scale.

Furthermore, the iris is hard to modify. Modifying fingerprints through cuts is easy, while imaging them accurately can be difficult, as the ridges and valleys can wear off over time. Moreover, using all ten fingerprints for deduplication or combining different biometric modalities is vulnerable to combinatorial attacks (e.g. by combining fingerprints from different people). DNA sequencing could in theory provide high enough accuracy, but DNA reveals a lot of additional private information about the user (at least to the party that runs the sequencing). Additionally, it is hard to scale from a cost perspective and implementing reliable liveness detection measures is hard. Facial biometrics offers significantly better liveness detection compared to DNA sequencing. However, compared to iris biometrics, the accuracy of facial recognition is much lower. This would result in a growing number of erroneous collisions as the number of registered users increases. Even under optimal conditions, at a global scale of billions of

people, over ten percent of legitimate new users would be rejected, compromising the inclusivity of the system.

Therefore, based on the outlined trade-offs of different biometric modalities, iris recognition is the only one which is suitable for global verification of uniqueness in the context of the Worldcoin project.

World ID: Implementing PoP at Scale

Based on the conclusion that the only path to verify uniqueness on a global scale is iris biometrics, Tools for Humanity built a custom biometric device, called the Orb. This device issues an AI-safe[3] PoP credential called World ID. The Orb is built from the ground up to verify humanness and uniqueness in a fair and inclusive manner.

Fig. 7

The Orb which verifies a person's humanness and uniqueness to issue a person's World ID.

The issuance of World ID is privacy-preserving, as the humanness check happens locally and no images need to be saved (or uploaded) by the issuer. Using World ID reveals minimal information about the individual, as the protocol employs zero-knowledge proofs. The vision for the device is for its development, production and operation to be decentralized over time such that no single entity will be in control of World ID issuance.

The following section explains the previously mentioned building blocks for an effective proof of personhood mechanism:

- Deduplication
- Authentication
- Recovery
- Revocation
- Expiry

and how they are implemented in the context of World ID.

Deduplication

The hardest part for an inclusive yet highly secure PoP mechanism is to make sure every user can receive exactly one proof of personhood. Based on the previous evaluation iris biometrics are the best means to accurately verify uniqueness on a global scale (see limitations).

The other potential error inherent to biometric algorithms is the false acceptance of a user. The false acceptance rate is largely dependent upon the system's capacity to detect presentation attacks, which are attempts to deceive or spoof the verification process. While no biometric system is entirely impervious to such attacks, the important metric is the effort required for a successful attack. This consideration was fundamental to the conception of the Orb. Developing the Orb was a decision that did not come lightly. It represented a high-cost endeavor. However, from first principles, it was required to build the most inclusive yet secure verification of humanness and uniqueness. The Orb is designed to verify uniqueness with high accuracy, even in hostile contexts where the presence of malicious actors cannot be excluded. To accomplish this, the Orb is

equipped with every viable camera sensor spanning the electromagnetic spectrum, complemented by suitable multispectral illumination. This enables the device to differentiate between fraudulent spoofing attempts and legitimate human interactions with a high degree of accuracy. The Orb is further equipped with a powerful computing unit to run several neural networks concurrently in real-time. These algorithms operate locally on the Orb to validate humaneness, while safeguarding user privacy. While no hardware system interacting with the physical world can achieve perfect security, the Orb is designed to set a high bar, particularly in defending against scalable attacks. The anti-fraud measures integrated into the Orb are refined constantly.

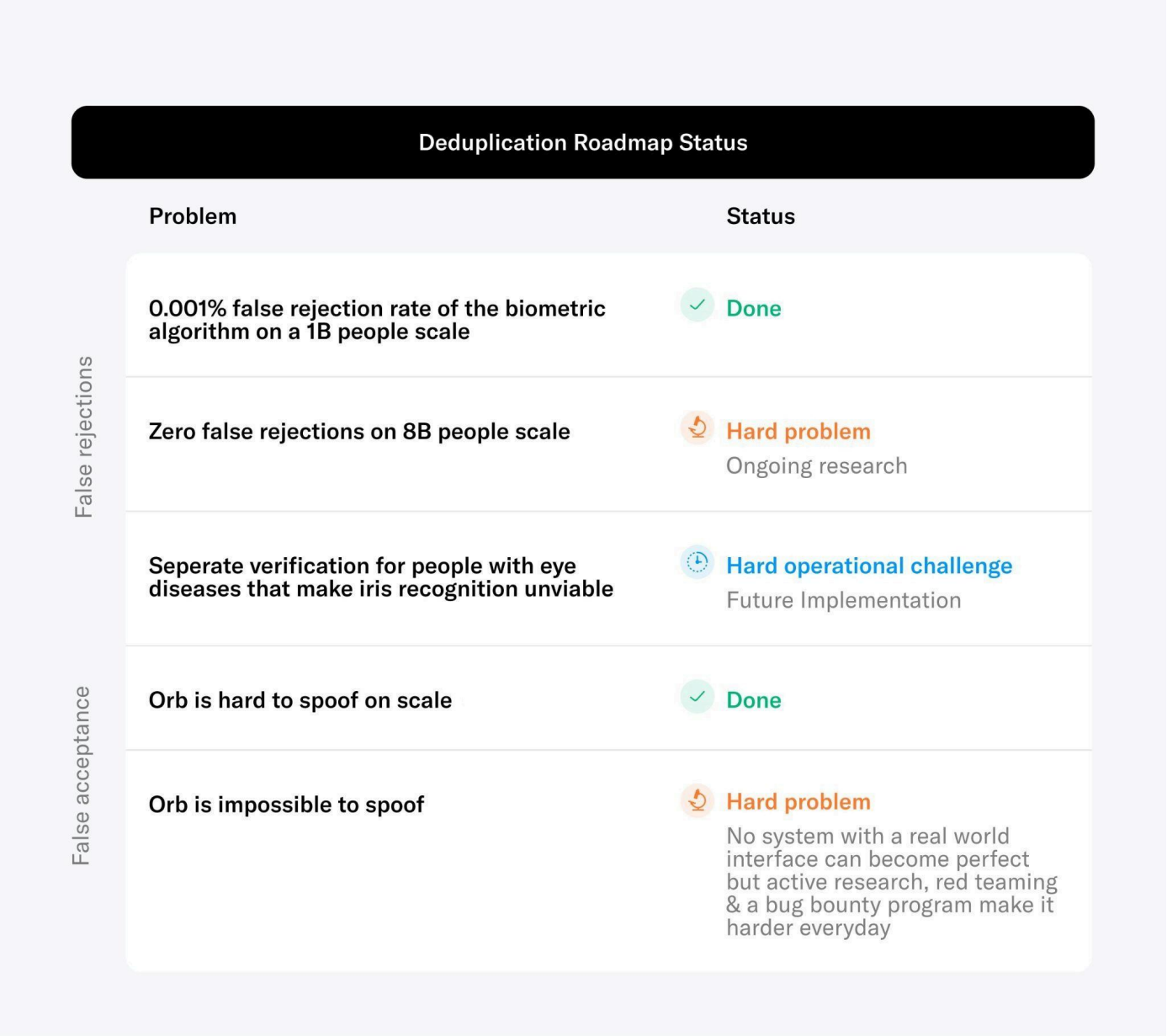

Fig. 8

The minimum required functionality with respect to deduplication to roll out a proof of personhood mechanism to one billion people has been reached. However, there is ongoing research to increase the inclusivity and security of the proof of personhood mechanism.

Authentication

Authentication seeks to ensure that only the legitimate owner of a World ID issued by the Orb is able to authenticate themself beyond proving that they own the keys. This plays a critical role in preventing the selling or stealing of World IDs. Within the scope of World ID, there are two primary mechanisms at one's disposal. Selecting the appropriate mechanism is up to the verifier, as each mechanism offers varying degrees of assurance and friction.

Face Authentication

Face-based authentication is similar to Apple's Face ID. Authentication involves a 1:1 comparison with a pre-existing template that is stored on the user's phone, which requires considerably lower levels of accuracy in contrast to the 1:N global verification of uniqueness[4] that the Orb is performing. Therefore, the entropy inherent to facial features is sufficient. To enable this feature, an encrypted embedding of the user's face, signed by the Orb, is end-to-end encrypted and transmitted to the World ID wallet on the user's mobile device. Subsequently, facial recognition, performed locally on the user's device in a fashion similar to Face ID, could be used to authenticate users, thereby ensuring that only the person to whom the World ID was originally issued can use it for authentication purposes.

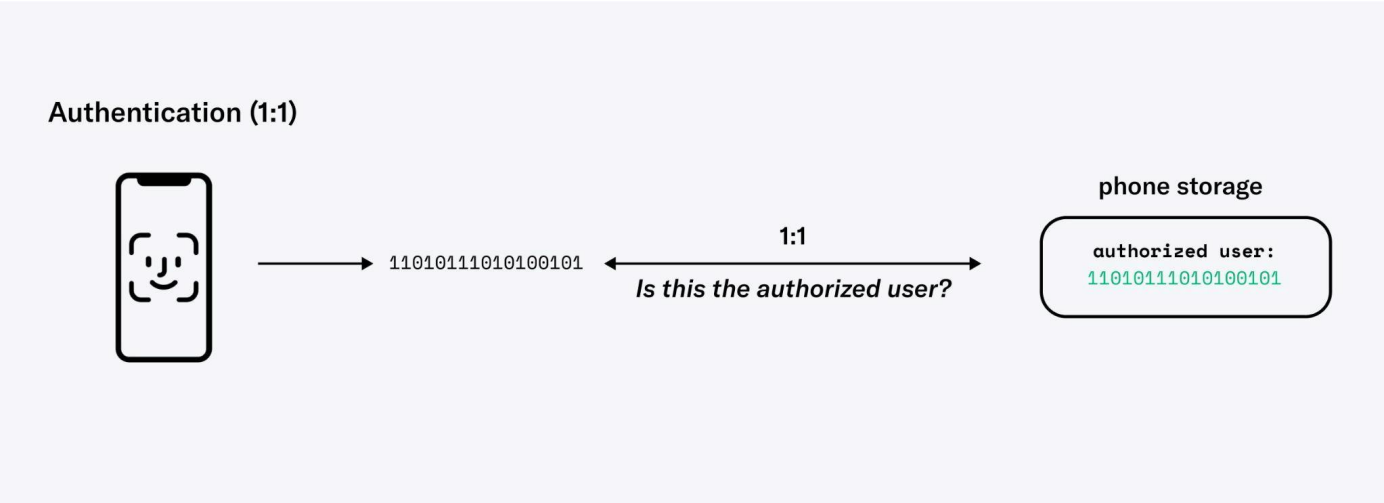

Fig. 9

Visualization of face authentication on a user's phone which compares a selfie with the face image captured by the Orb. This can help make it very difficult to use somebody else's World ID.

This mechanism facilitates the extension of the secure hardware guarantees from the Orb to the user's mobile device. However, given that the user's device is not intrinsically trusted, there is no absolute assurance that the appropriate code is being executed nor that the camera input can be trusted. To increase security, ongoing research is investigating Zero Knowledge Machine Learning (ZKML) on mobile devices. Nevertheless, in the absence of custom hardware, this approach cannot provide the same security guarantees as the Orb. Therefore, face authentication on the user's device should be reserved for applications with lower stakes.

While this feature is not yet implemented, it is expected to be released later this year. The first step for the implementation is for the Orb to send an end-to-end encrypted face embedding to the user's phone where it can later be compared against a selfie. The self-custody of face images is a requirement for face authentication and therefore determines who can later on participate in face authentication. Therefore, this feature has a high priority on the roadmap.

Iris Authentication

This is conceptually similar to face authentication with the difference that a user needs to return to an Orb, presenting a specific QR code generated by the user’s World ID wallet. This process validates the individual as the rightful owner of their World ID. Using iris authentication through the Orb increases security.

This authentication mechanism can be compared with, for example, physically showing up to a bank or notary to authenticate certain transactions. Although inconvenient, and therefore rarely required, it provides increased security guarantees. This feature is under active development and is expected to be released in the coming months.

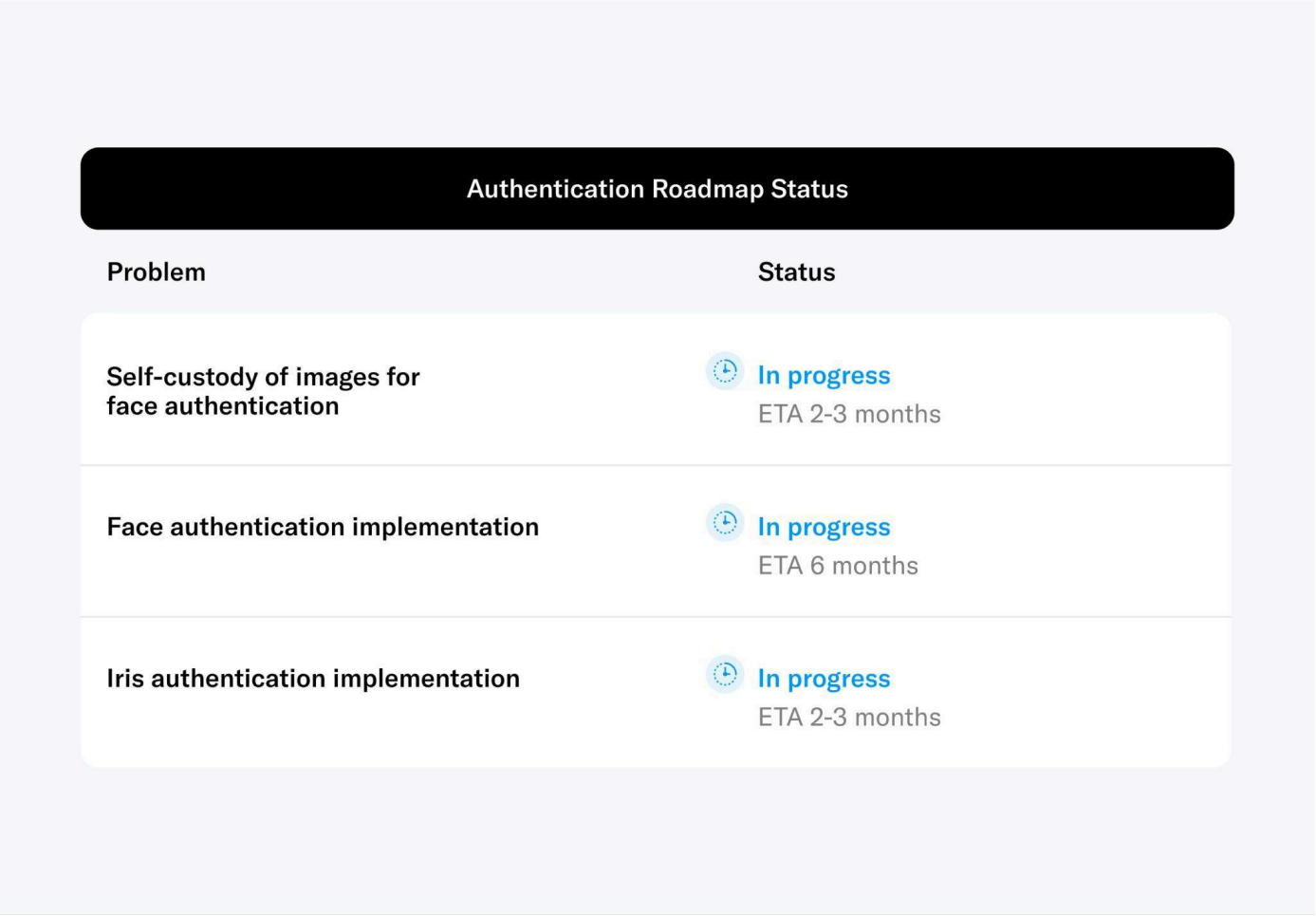

Fig. 10

Authentication is a high priority to make the trading of World ID hard and thereby increase the integrity of the Orb based proof of personhood. Self custody of images is required for a retroactive rollout of face authentication to users who have been previously verified.

Recovery

The simplest way to restore World ID is via a backup. Social recovery is not implemented today but is likely to be explored in the future. The most important recovery mechanism for Orb-based proof of personhood is reissuance. If the user has lost access or the World ID has been compromised by a fraudulent actor, individuals can get their World ID re-issued by returning to the Orb, without the need to remember a password or similar information.

It is critical to understand, however, that the recovery facilitated by biometrics exclusively refers to the World ID. Neither other credentials held by the user's wallet nor the wallet itself can be recovered, due to security considerations.

The initial implementation is planned to be realized through key rotation, which will be released soon. Notably, use cases that require long-lasting nullifiers[5] such as reputation or single-claim rewards will be limited due to the nullifier's potential reset through recovery. This is also discussed in the limitations section. However, this limitation does not impact the 'humanness' attestation; for instance, the verification of an account on a continuous basis through sessions, or time-bounded votes where only participants whose latest recovery preceded the beginning of the voting period are allowed. To enable key recovery requires solving hard research challenges to preserve privacy.

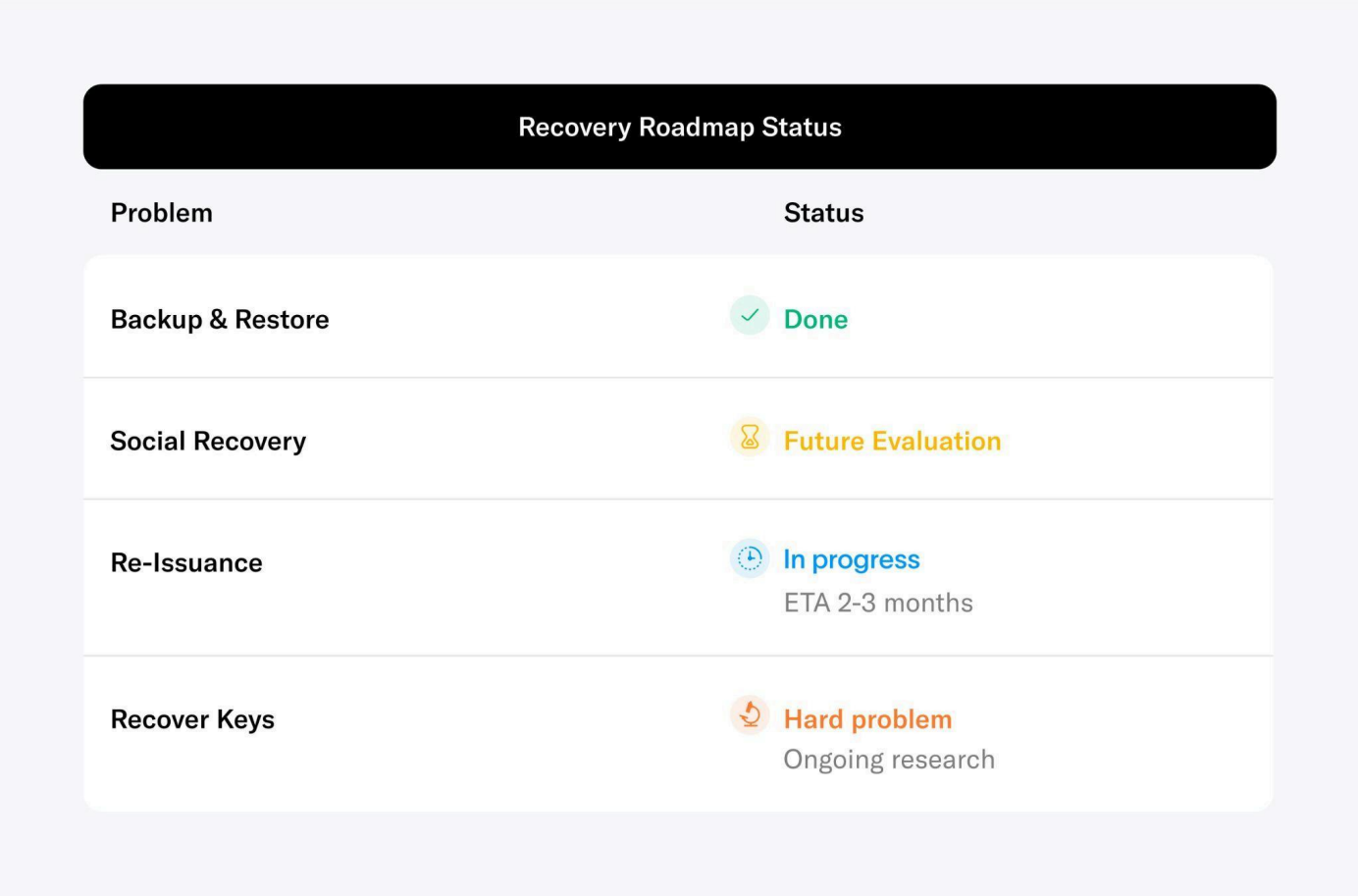

Fig. 11

There are several ways to recover someone's World ID. The easiest way is to create and restore a backup. If no backup is available, the World ID can be restored via re-issuance which is on the roadmap for the next 2-3 months. To implement biometric key recovery in a safe and privacy-preserving manner, several open research questions would need to be solved. It is therefore currently unclear if biometric key recovery will be possible.

Revocation

In the event of a compromised Orb, malicious actors could theoretically generate counterfeit World IDs[6]. If it is determined by the community that an issuer is acting inappropriately or a device is compromised, the Worldcoin Foundation, in alignment with the prevailing governance structure, can "deny list" World IDs linked to a specific issuer or device for its own purposes, while other application developers can implement their own measures. Users who inadvertently find themselves impacted can simply get their World ID re-issued by any other Orb. More details around the mechanism can be found in the decentralization.

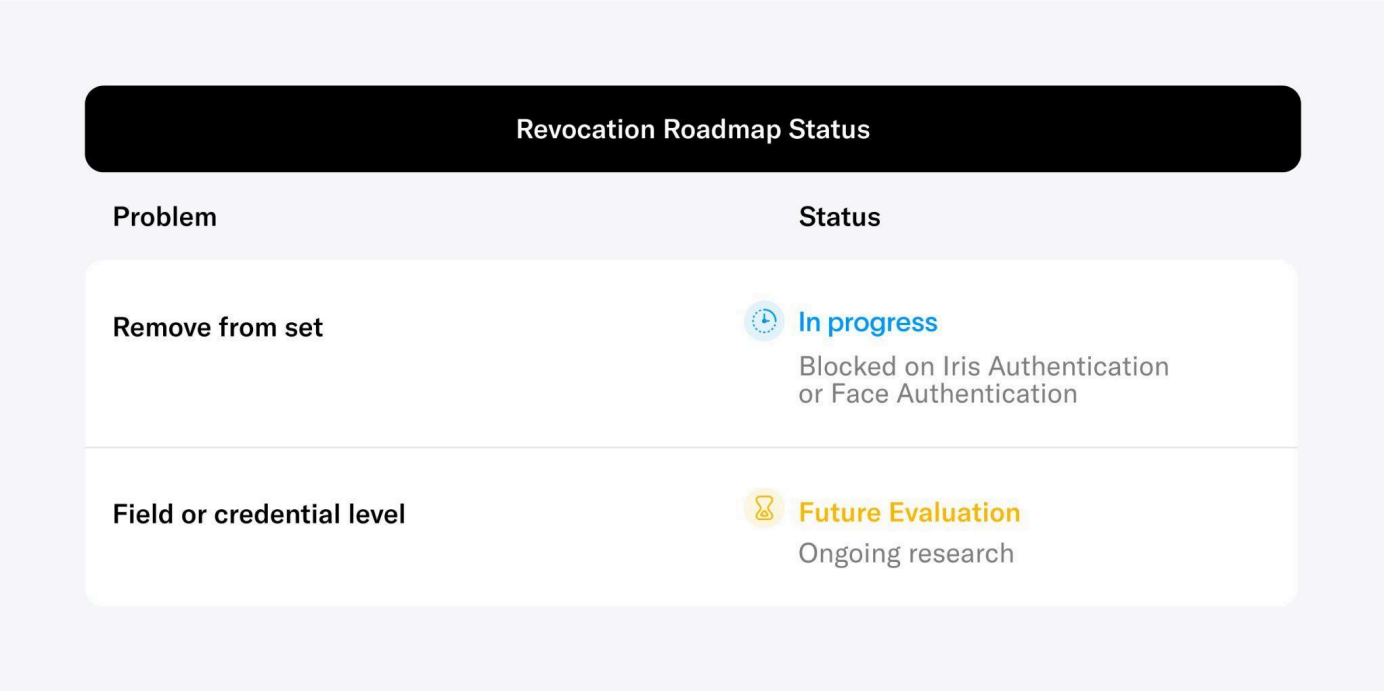

Fig. 12

Revocation will at first be implemented as by creating a set on chain with all credentials that are still active i.e. not revoked. Later, this will likely transition to a field on the credential level.

Expiry

Even in the absence of tangible fraudulent activities, a device could retrospectively be identified by the community as vulnerable, or simply as having outdated security standards. In such instances, in line with the governing principles of the Foundation, World IDs can be subjected to a set expiry. This essentially amounts to a revocation process but with a predefined expiry period that affords individuals ample time for re-verification, such as one year. Further, in accordance with its governance, the Foundation could eventually decide to expire verifications after a set period of time to further strengthen the integrity of the PoP mechanism in the interest of all participants.

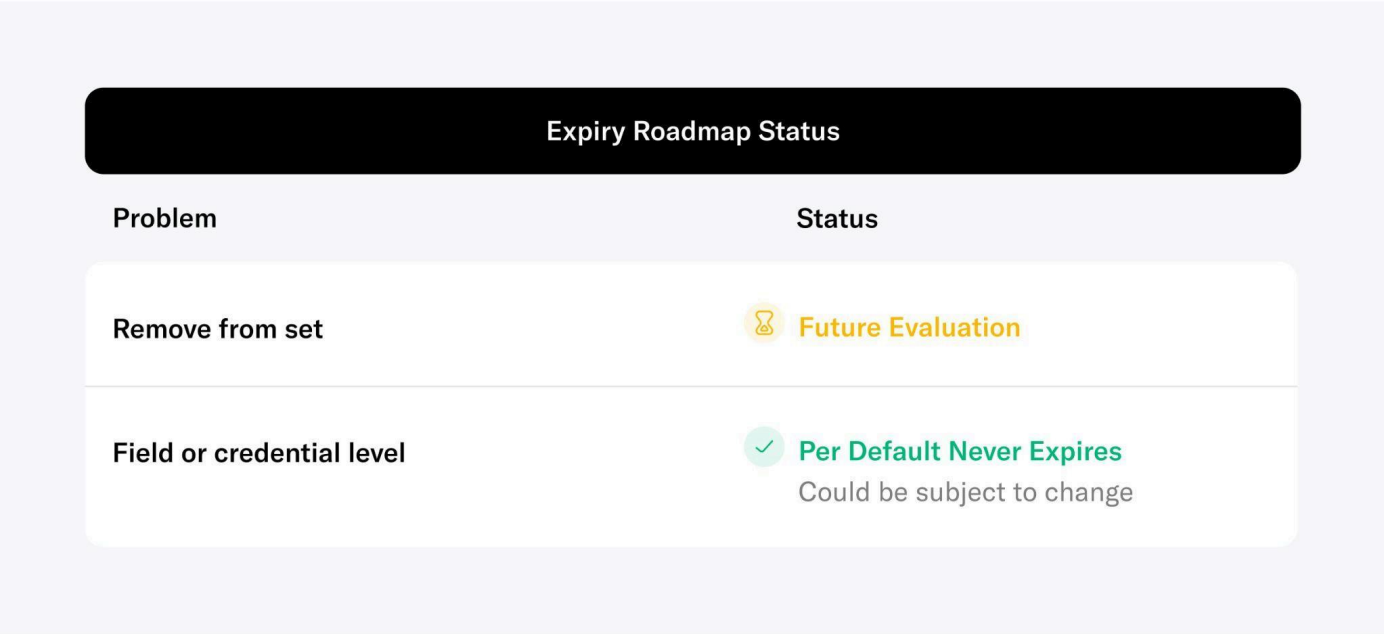

Fig. 13

Retroactive expiry will likely be needed but has a lower priority compared to other features and will be evaluated in the future. It is not yet decided if default expiry of World IDs i.e. assigning them a default validity period after which users have to return to the Orb will be needed. As of today, the World ID is valid forever as long as it is not revoked. Based on learnings in the coming years this could change.

Further Research

Despite the defensive measures outlined in this section, which significantly raise the threshold for fraudulent activities and can likely limit its impact beyond any existing scalable proof of personhood verification mechanism, it is important to recognize their inability to completely protect against all threats, such as collusion or other attempts to circumvent the one-person-one-proof principle (i.e. bribing others to vote a particular way). To further raise the bar, innovative ideas and research in mechanism design will be necessary.

Footnotes

1. Possibly except for the validity date ↩
2. In recent implementations virtually all major providers switched from “labeling traffic lights” to the so-called *silent* CAPTCHAs (e.g. reCaptcha v3) ↩
3. In this context, AI-safe refers to a process that’s hard for AI models. It’s assumed, for example, that spoofing the Orb is significantly harder for AI than performing a CAPTCHA. ↩
4. where N is the total number of previously verified users ↩
5. In the context of World ID, each holder has a unique nullifier for themselves in each application. This nullifier is what enables sybil resistance while preserving privacy as verifiers can use such nullifiers to prevent multiple registrations. ↩
6. the Orb's secure computing environment was designed to make such compromises extremely difficult ↩

Technical Implementation

The preceding sections explained the necessity for a universal, secure, and inclusive proof of personhood mechanism. Additionally, they discussed why iris biometrics appears to be the sole feasible path for such a PoP mechanism. The realization via the Orb and World ID has also been explained on a high level. The subsequent section dives deeper into the specifics of the architectural design and implementation of both the Orb and World ID.

Architecture Overview

To get a World ID, an individual begins by downloading the World App. The app stores their World ID and enables them to use it across multiple platforms and services. The World App is user-friendly, particularly geared towards crypto beginners, and offers simple financial features based on decentralized finance: allowing users to on- and off-ramp, subject to the availability of providers, swap tokens through a decentralized exchange, and connect with dApps through WalletConnect. Importantly, the system allows other developers to create their own clients without seeking permission, meaning there can be various apps supporting World ID.

Once verified through the Orb, individuals are issued a World ID, a privacy-preserving proof-of-personhood credential. Through World ID, they can claim a set amount of WLD periodically (Worldcoin Grants), where laws allow. World ID can also be used to authenticate as human with other services (e.g., prevent user manipulation in the case of voting). In the future, other credentials can be issued on the Worldcoin Protocol as well.

To make World ID and the Worldcoin Protocol easy to use, an open source Software Development Kit (SDK) is available to simplify interactions for both Web3 and Web2 applications. The World ID software development kit (SDK) is the set of tools, libraries, APIs, and documentation that accompanies the Protocol. Developers can use the SDK to leverage World ID in their applications. The SDK makes web, mobile, and on-chain integrations fast and simple; it includes tools like a web widget (JS), developer portal, development simulator, examples, and guides.

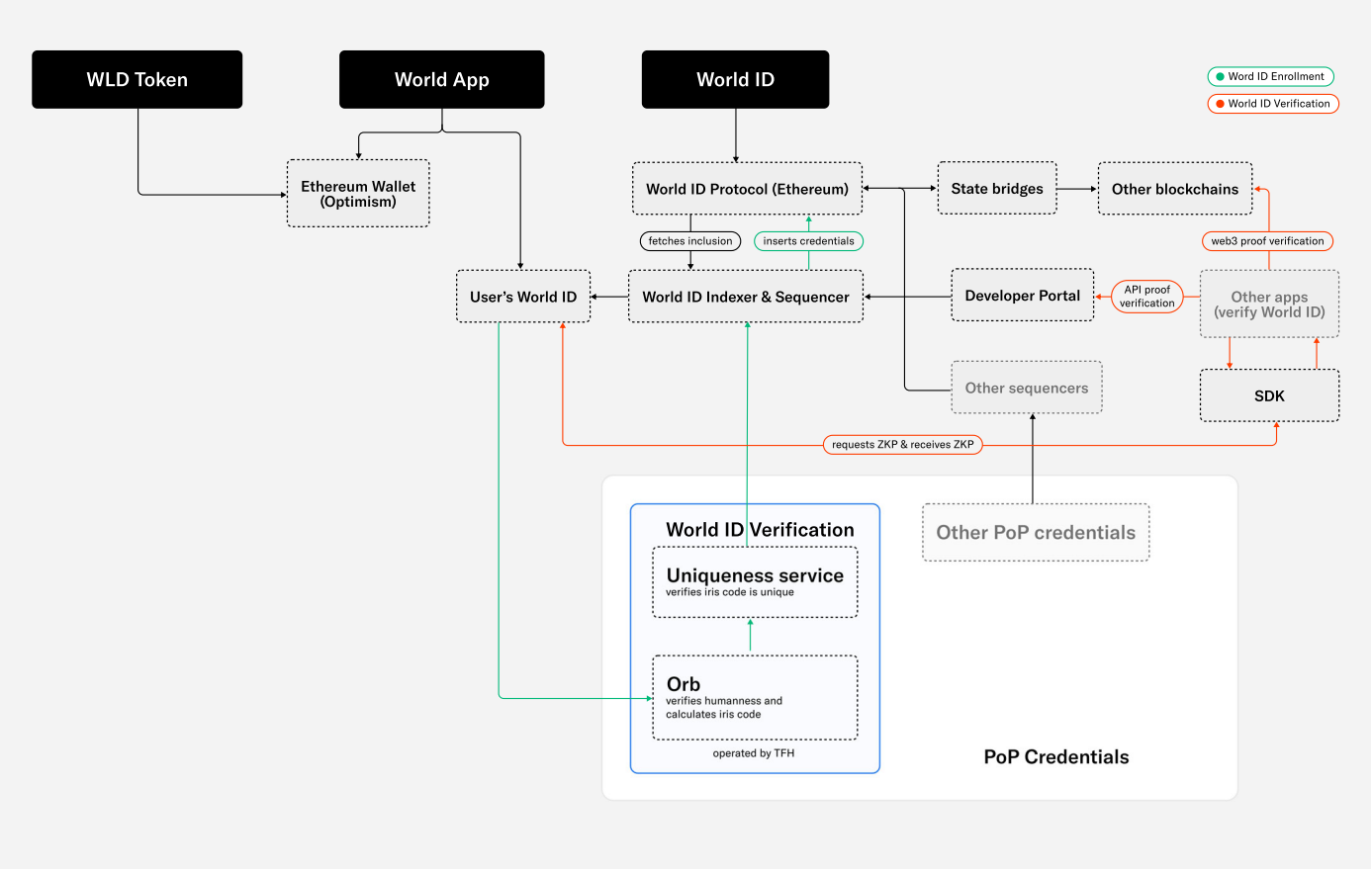

Fig. 1

High level overview of the Worldcoin system and the connection between individual products and protocols.

The Orb

Previous sections discussed why a custom hardware device using iris biometrics is the only approach to ensure inclusivity (i.e. everyone can sign up regardless of their location or background) and fraud resistance, promoting fairness for all participants. This section discusses the engineering details of the Orb, which was first prototyped and developed by Tools for Humanity.

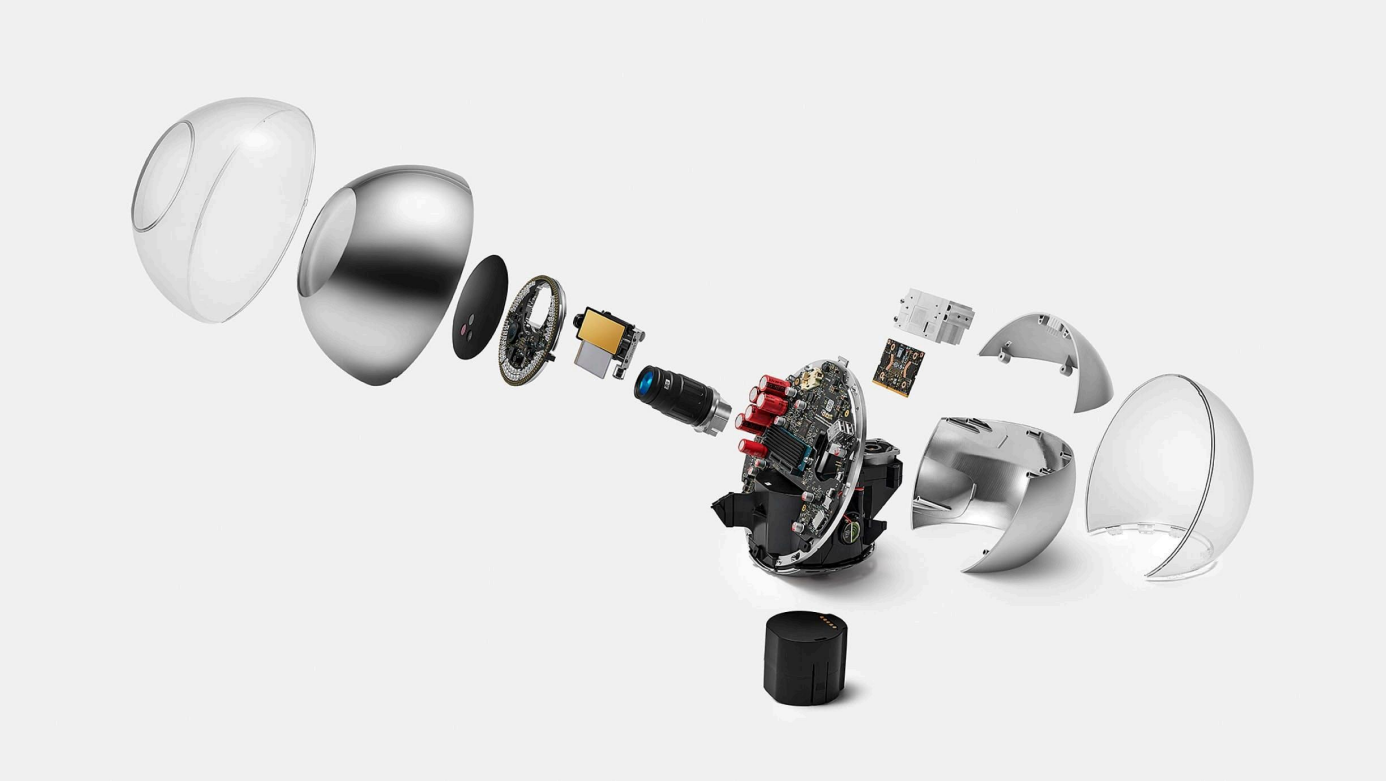

Fig. 2

All relevant components of the Orb visible.

Why Custom Hardware is Needed

It would have been significantly easier to use off the shelf available devices like smartphones or iris imaging devices. However, neither is suitable for uncontrolled and adversarial environments in the presence of significant incentives. To reliably distinguish people, only iris biometrics are suitable for this globally scalable use case . To enable maximum accuracy, device integrity, spoof prevention as well as privacy, a custom device is necessary. The reasoning is described in the following section.

In terms of the biometric verification itself, the fastest and most scalable path would be to use smartphones. However, there are several key challenges with this approach. First, smartphone cameras are insufficient for iris biometrics due to their low resolution across the iris, which decreases accuracy. Further, imaging in the visible spectrum can result in specular reflections on the lens covering the iris and low reflectivity of brown eyes (most of the population) introduces noise. The Orb captures high quality iris images with more than an order of magnitude higher resolution compared to iris recognition standards.

This is enabled by a custom, narrow field-of-view camera system. Importantly, images are captured in the near infrared spectrum to reduce environmental influences like different light sources and specular reflections. More details on the Orb's imaging system can be found in the following sections.

Second, the achievable security bar is very low. For PoP, the important part is not identification (i.e. "Is someone who they claim they are?"), but rather proving that someone has not verified yet (i.e. "Is this person already registered?"). A successful attack on a PoP system does not necessitate the attacker's impersonation of an existing individual, which is a challenging requirement that would be needed to unlock someone's phone. It merely requires the attacker to look different from everyone who has registered so far. Phones and existing iris cameras are missing multi-angle and multi-spectral cameras as well as active illumination to detect so-called presentation attacks (i.e. spoof attempts) with high confidence. A widely-viewed video demonstrating an effective method for spoofing Samsung's iris recognition illustrates how straightforward such an attack could be in the absence of capable hardware.

Further, a trusted execution environment would need to be established in order to ensure that verifications originated from legitimate devices (not emulators). While some smartphones contain dedicated hardware for performing such actions (e.g., the Secure Enclave on the iPhone, or the Titan M chip on the Pixel), most smartphones worldwide do not have the hardware necessary to verify the integrity of the execution environment. Without those security features, basically no security can be provided and spoofing the image capture as well as the enrollment request is straightforward for a capable attacker. This would allow anyone to generate an arbitrary number of synthetic verifications.

Similarly, no off-the-shelf hardware for iris recognition met the requirements that were necessary for a global proof of personhood. The main challenge is that the device needs to operate in untrusted environments which poses very different requirements than e.g. access control or border control where the device is operated in trusted environments by trusted personnel. This significantly increases the requirements for both spoof

prevention as well as hardware and software security. Most devices lack multi-angle and multispectral imaging sensors for high confidence spoof detection. Further, to enable high security spoof detection, a significant amount of local compute on the device is needed, without the ability to intercept data transmission, which is not the case for most iris scanners. A custom device enables full control over the design. This includes tamper detection that can deactivate the device upon intrusion, firmware that is designed for security to make unauthorized access very difficult, as well as the possibility to update the firmware down to the bootloaders via over the air updates. All iris codes generated by an Orb are signed by a secure element to make sure they originate from a legitimately provisioned Orb instead of, for example, an attacker's laptop. Further, the computing unit of the Orb is capable of running multiple real-time neural networks on the five camera streams (mentioned in the last section). This processing is used for real time image capture optimization as well as spoof detection. Additionally, this enables maximum privacy by processing all images on the device such that no iris images need to be sent to the verifier.

While no hardware system interacting with the physical world can achieve perfect security, the Orb is designed to set a high bar, particularly in defending against scalable attacks. The anti-fraud measures integrated into the Orb are constantly refined. Several teams at Tools for Humanity are continuously working on increasing the accuracy and sophistication of the liveness algorithms. An internal red team is probing various attack vectors. In the near future, the red teaming will extend to external collaborators including through a bug bounty program.

Lastly, the correlation between image quality and biometric accuracy is well established, and it is expected that deep learning will benefit even more from increased image quality. Given the goal of reducing error rates as much as possible to achieve maximum inclusivity, the image quality of most devices was insufficient.

Since commercially available iris imaging devices did not meet the image quality or security needs, Tools for Humanity dedicated several years to developing a custom

biometric verification device (the Orb) to enable universal access to the global economy in the most inclusive manner possible.

Hardware

Three years of R&D, including one year of small-scale field testing and one year of transition to manufacturing at scale, have led to the current version of the Orb, which is being open sourced. Feedback for design improvements is welcome and highly encouraged. The remainder of this section will go through a teardown of the Orb, with a few engineering anecdotes included.

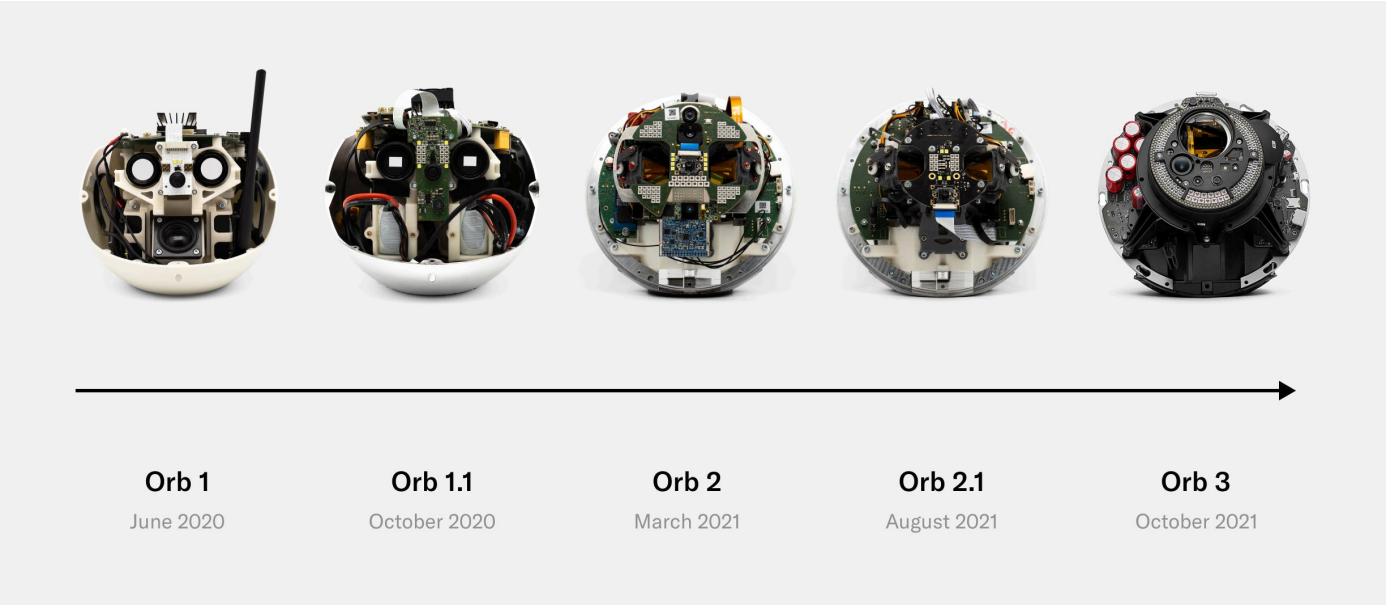

Fig. 3

Three years of Orb R&D

Today's Orb represents a precise balance of development speed, compactness, user experience, cost and at-scale production with minimal compromise being made on imaging quality and security. There will likely be future versions that are optimized even further both by Tools for Humanity and other companies as the Worldcoin ecosystem decentralizes. However, the current version represents a key milestone that enables scaling the Worldcoin project.

The following takes the reader through some of the most important engineering details of the Orb, as well as how the imaging system works. For security purposes, only tamper detection mechanisms that are meant to catch intrusion attempts are left out.

Design

Fundamental to the development of the Orb was its design. A spherical shape is an engineering challenge. However, it was important for the design of the Orb to reflect the values of the Worldcoin project. The spherical shape stands for Earth, which is home to all. Similarly the Orb is tilted at 23.5 degrees, the same degree at which the Earth is tilted relative to its orbital plane around the sun. There's even a 2mm thick clear shell on the outside of the Orb which protects the Orb just like the atmosphere protects Earth. The resemblance of Earth symbolizes that the Worldcoin project is meant to give everyone the opportunity to participate, regardless of their background and the Orb and its use of biometrics is a reflection of that since nothing is required other than being human.

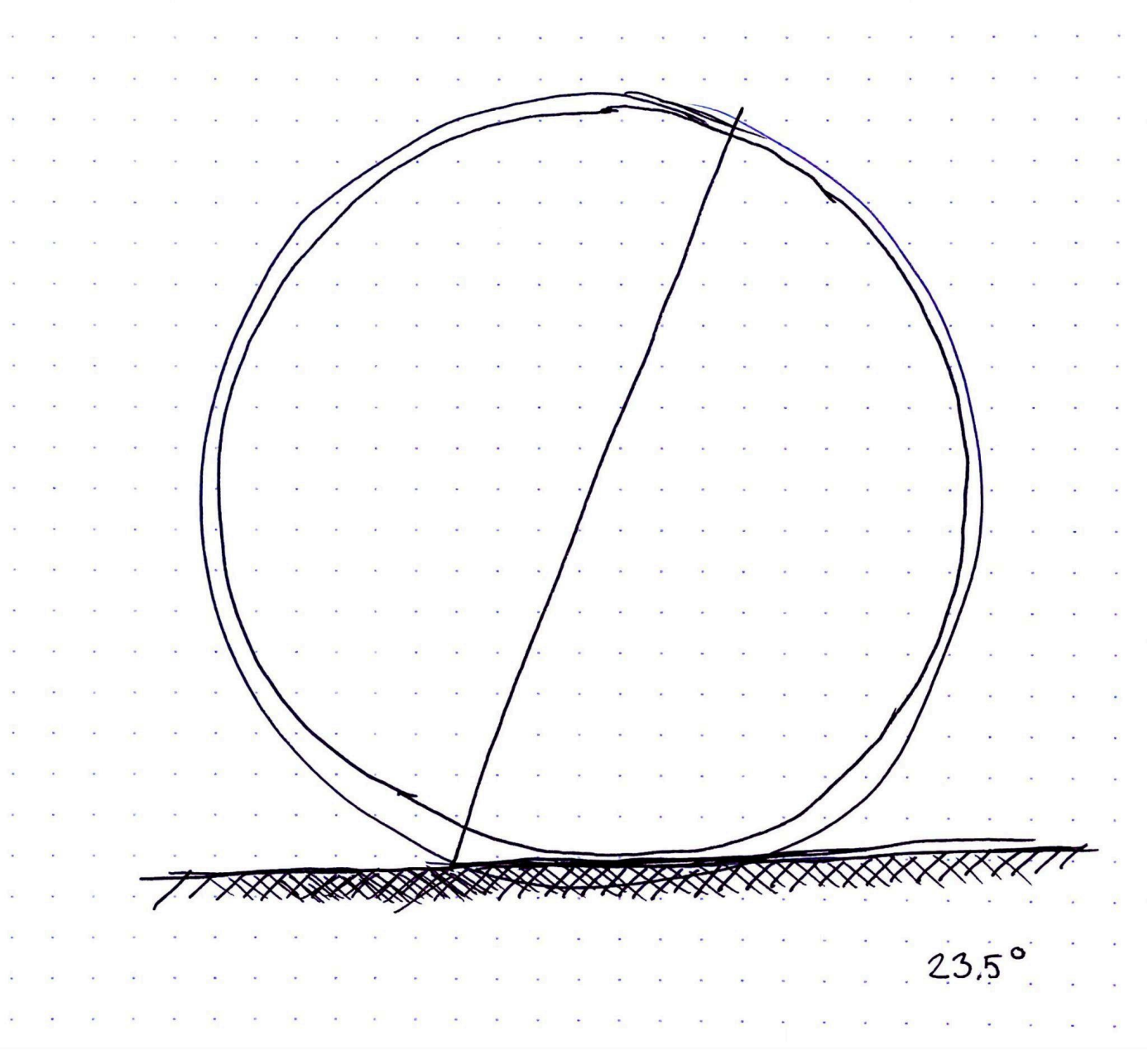

Fig. 4

A sketch of the Orb

Mechanics

When removing the shell, the mainboard, optical system and cooling system become visible. Most of the optical system is hidden in an enclosure that, together with the shell, forms a dust- and water-resistant environment to enable long-term use even in challenging environments.

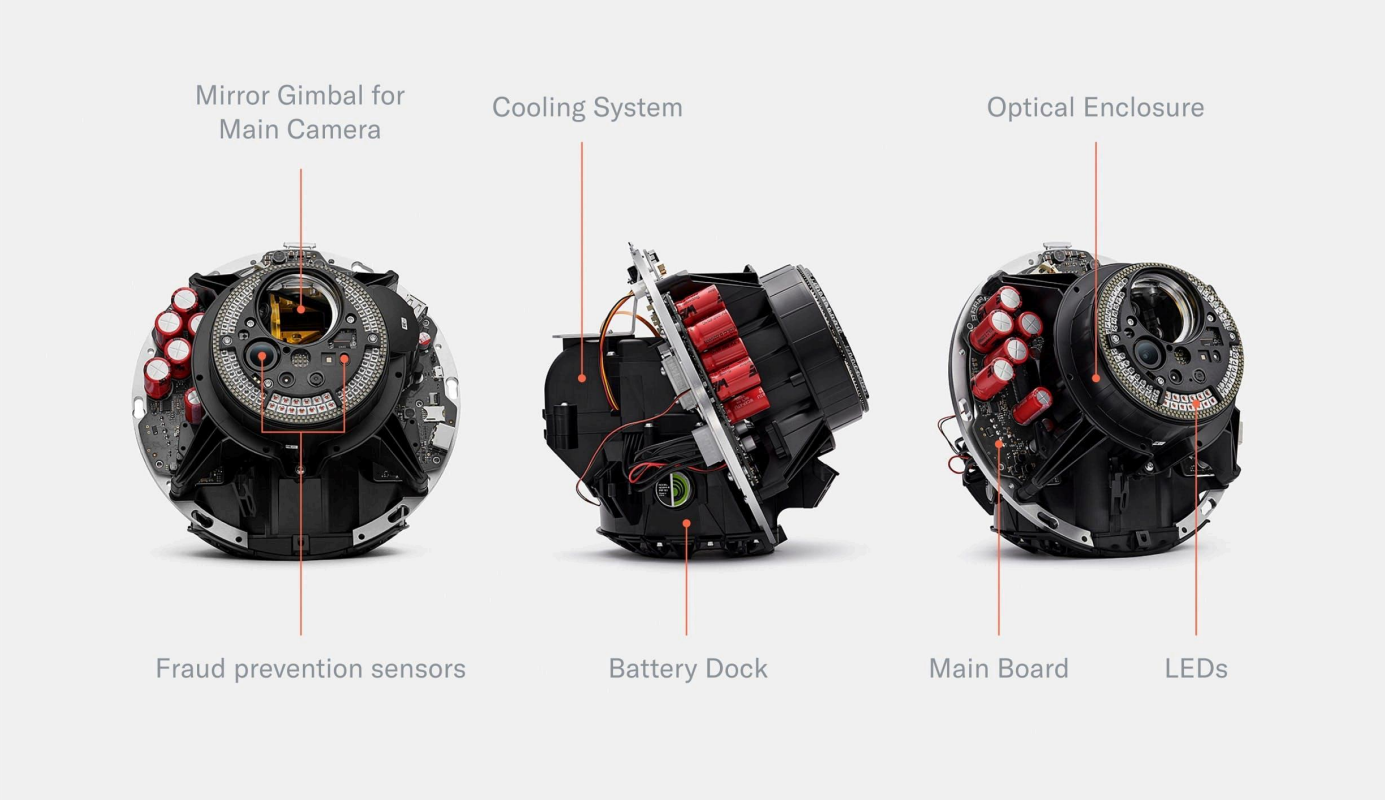

Fig. 5

Orb underneath the shell

The Orb consists of two hemispheres separated by the mainboard which is tilted at 23.5° —the angle of the rotational axis of the earth. The mainboard holds a powerful computing unit to enable local processing for maximum privacy. The frontal half of the Orb is dedicated to the sealed optical system. The optical system consists of several multispectral sensors to verify liveness and a 2D gimbal-enabled narrow field of view camera to capture high resolution iris images. The other hemisphere is dedicated to the cooling system as well as speakers. An exchangeable battery can be inserted from the bottom to enable uninterrupted operation in a mobile setting.

Once the shell is removed, the Orb can be divided into four core parts:

- Front: The optical system
- Middle: The mainboard separates the device into two hemispheres
- Back: The main computing unit as well as the active cooling system
- Bottom: An exchangeable battery

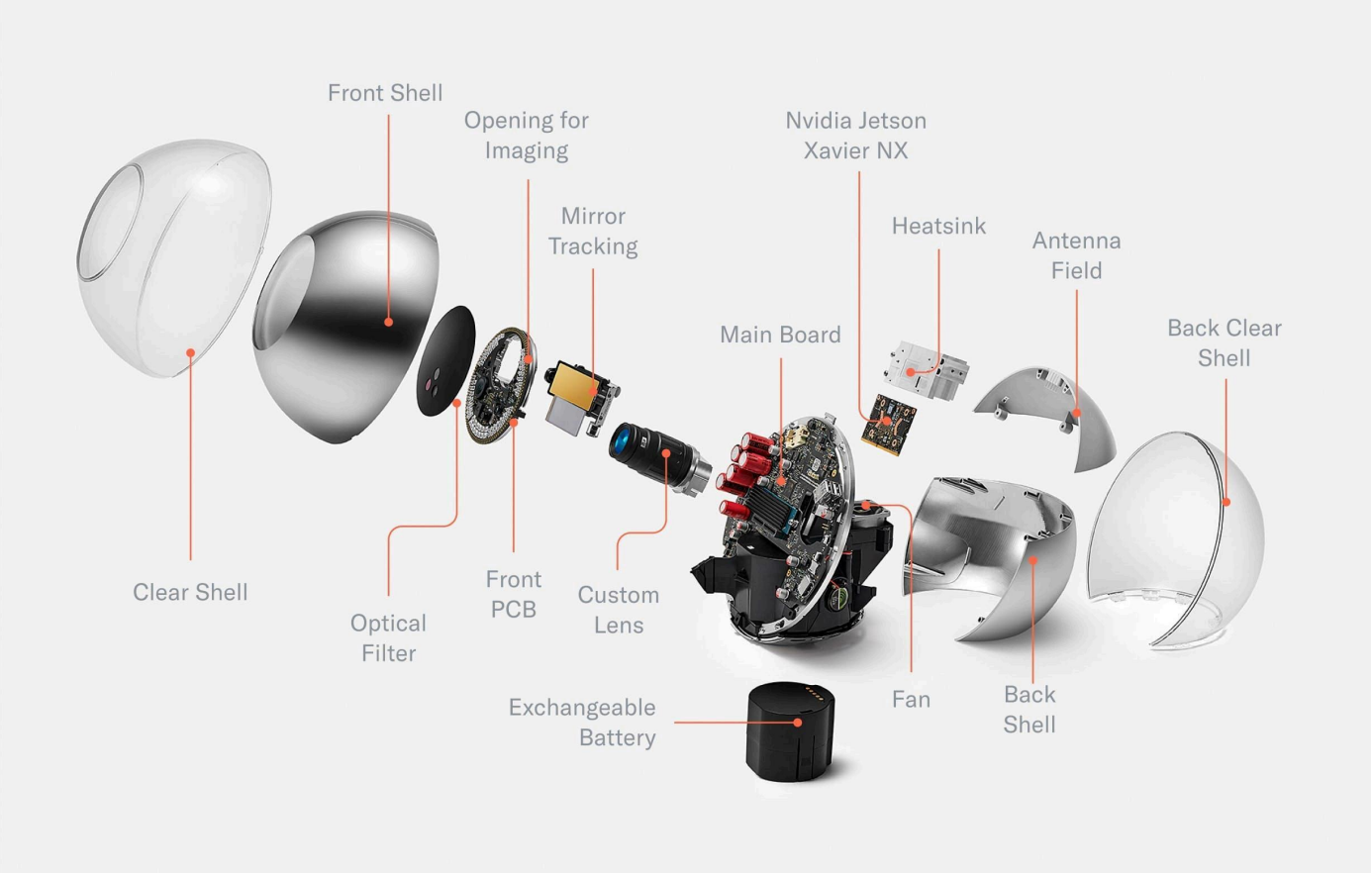

Fig. 6

Explosion CAD of all relevant components

With the housing material removed (e.g. the dust-proof enclosure of the optical system), all relevant components of the Orb become visible. This includes the custom lens, which is optimized for both near infrared imaging and fast, durable autofocus. The front of the optical system is sealed by an optical filter to keep dust out and minimize noise from the visible spectrum to optimize image quality. In the back, a plastic component in the otherwise chrome shell allows for optimized antenna placement. The chrome shell is covered by a clear shell to avoid deterioration of the coating over time.

First prototypes were tested outside the lab as early as possible. Naturally, this taught the team many lessons, including:

Optical System

With the first prototype, the signup experience was notoriously difficult. Over the course of a year the optical system was upgraded with autofocus and eye tracking such that alignment becomes trivial when the person is within an arm's length of the Orb.

Battery

No off-the-shelf battery would last for a full day on a single charge. A custom exchangeable battery was designed based on 18650 Li-Ion cells—the same form factor as the cells used in modern electric cars. The battery consists of 8 cells with 3.7V nominal voltage in a 4S2P configuration (14.8V) with a capacity of close to 100Wh, which is a limit imposed by regulations related to logistics. Now there's no limit to Orb uptime.

Fig. 7

Custom exchangeable battery

The Orb's custom battery is made of Li-Ion 18650 cells (the same cells used in many electric cars). With close to 100Wh, the capacity is optimized for battery lifetime while complying with transportation regulations. A USB-C connector makes recharging convenient.

Shell

The coating of the shell sometimes deteriorated in the handheld use case. Therefore, a 2mm clear shell was added to both optimize the design as well as protect the chrome coating from scratches and other wear.

UX LEDs

To make the user experience more intuitive, especially in loud environments where a person might not be able to hear sound feedback, an LED ring was added to help guide people through the sign-up process. Similarly, status LEDs were exposed next to the only button on the Orb to indicate its current state.

Optical System

Early field tests showed that the verification experience needed to be even simpler than anticipated. To do this, the team first experimented with many approaches featuring mirrors that allowed people to use their reflection to align with the Orbs imaging system. However, designs that worked well in the lab quickly broke down in the real world. The team ended up building a two-camera system featuring a wide angle camera and a telephoto camera with an adjustable ~5° field of view by means of a 2D gimbal. This increased the spatial volume in which a signup can be successfully completed by several orders of magnitude, from a tiny box of 20x10x5mm for each eye to a large cone.

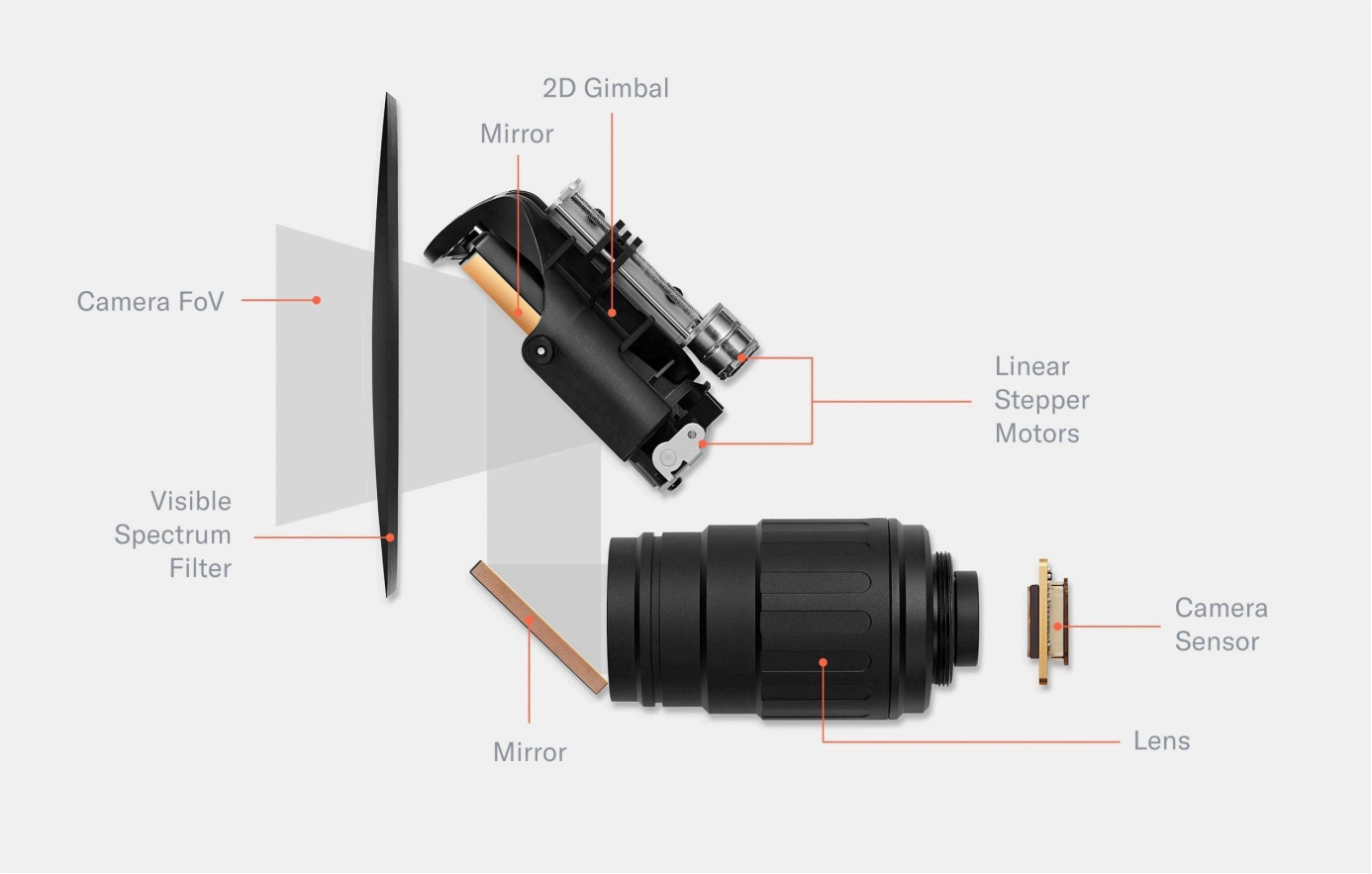

Fig. 8

Telephoto lens and 2D gimbal

The main imaging system of the Orb consists of a telephoto lens and 2D gimbal mirror system, a global shutter camera sensor and an optical filter. The movable mirror increases the field of view of the camera system by more than two orders of magnitude. The optical unit is sealed by a black, visible spectrum filter which seals the high precision optics from dust and only transmits near infrared light. The image capture process is controlled by several neural networks.

The wide angle camera captures the scene, and a neural network predicts the location of both eyes. Through geometrical inference, the field of view of the telephoto camera is steered to the location of an eye to capture a high resolution image of the iris, which is further processed by the Orb into an iris code.

Beyond simplicity, the image quality was the main focus. The correlation between image quality and biometric accuracy is well established.

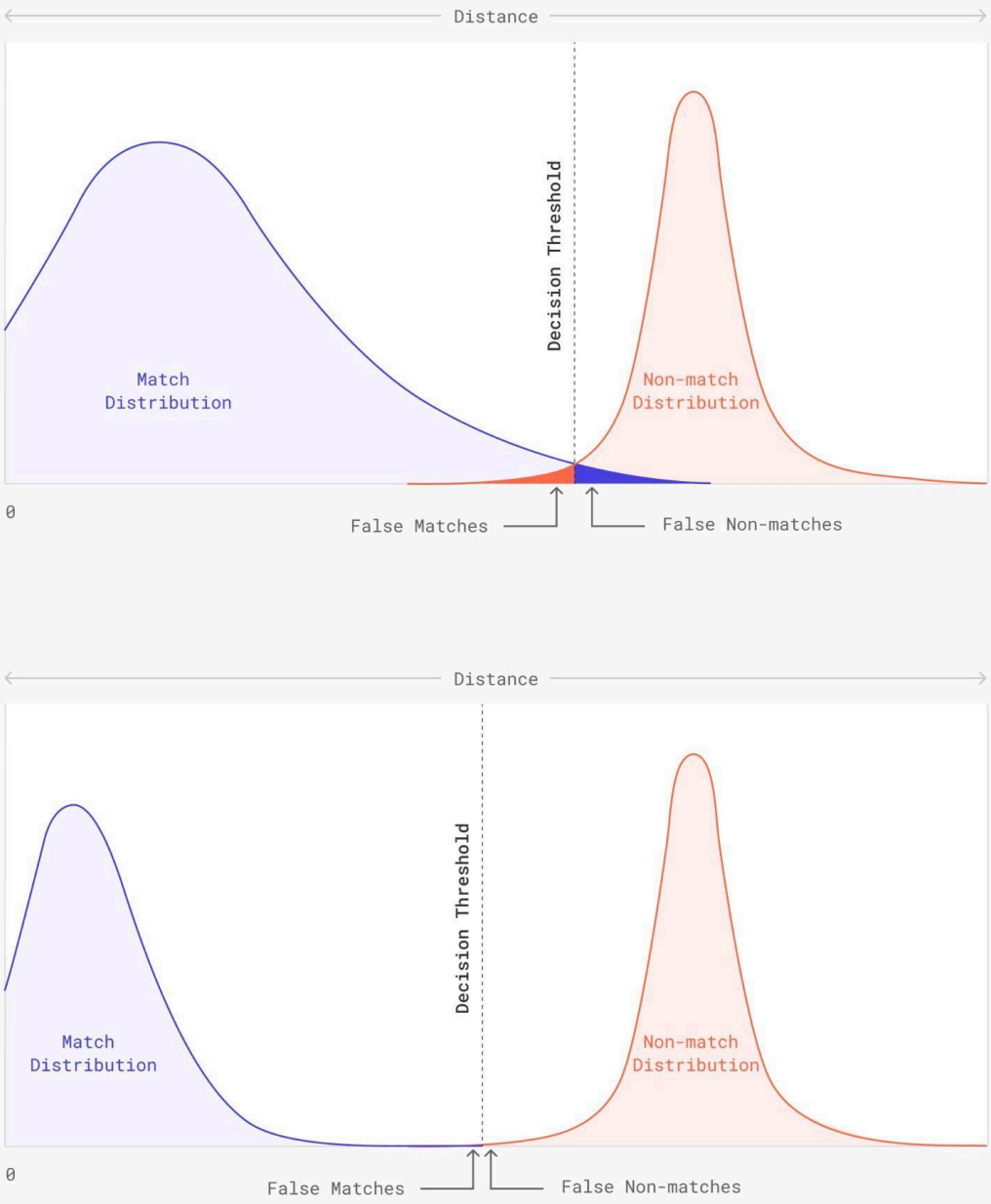

Fig. 9

Schematic representation illustrating the importance of high-quality imaging for decreasing error.

Here pairwise comparisons are plotted: the match distribution for pairs of the same identity (blue) and non-match distribution for pairs of different identity (red). In a perfect system, the match-distribution would be a very narrow peak at zero. However, multiple sources of error widen the distribution, leading to more overlap with the non-match

distribution and therefore increasing False Match and False Non-Match rates. High quality image acquisition narrows the match-distribution significantly and therefore minimizes errors. The width of the non-match distribution is determined by the amount of information that is captured by the biometric algorithm: the more information is encoded in the embeddings the narrower the distribution.

Many off-the-shelf products have been tested but there wasn't any lens compact enough to meet the imaging requirements while still being affordable. Therefore, the team partnered with a well known specialist in the machine vision industry to build a customized lens. The lens is optimized for the near infrared spectrum and has an integrated custom liquid lens which allows for neural network controlled millisecond-autofocus. It is paired with a global shutter sensor to capture high resolution, distortion free images.

Fig. 3.10:

a) Custom telephoto lens. The telephoto lens was custom designed for the Orb. The glass is coated to optimize image capture in the near infrared spectrum. An integrated liquid lens allows for durable millisecond autofocus. The position of the liquid lens is controlled by a neural network to optimize focus. To capture images free of motion blur, the global shutter sensor is synchronized with pulsed illumination.

b) A comparison of the image quality of the Worldcoin Orb vs. the industry standard clearly show the advancements made in the space. The camera and the corresponding pulsed infrared illumination are synchronized to minimize motion blur and suppress the influence of sunlight. This way, the Orb creates lab environment conditions for imaging, no matter its location. Needless to say, the infrared illumination is compliant with eye safe standards (such as EN 62471:2008).

Image quality was the one thing never compromised no matter how difficult it was. In terms of resolution the Orb is orders of magnitude above the industry standard. This provides the basis for the lowest error rates possible to, in turn, maximize the inclusivity of the system.

Electronics

When disassembling the Orb further, several PCBs (Printed Circuit Boards) are visible, including the front PCB containing all illumination, the security PCB for intrusion detection and the bridge PCB which connects the front PCB with the largest PCB: the mainboard.

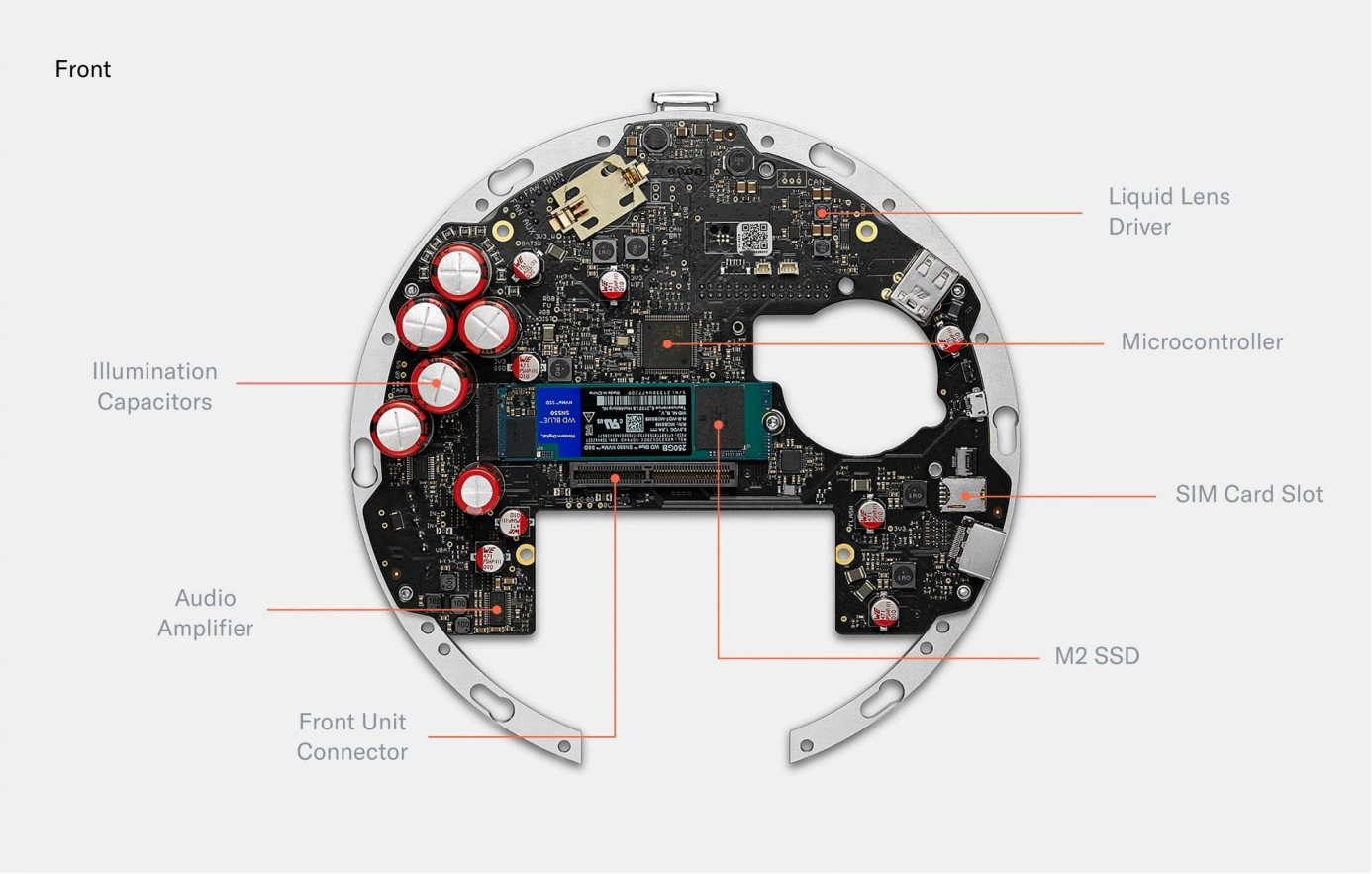

Fig. 11

The front of the mainboard

The front of the mainboard holds capacitors to power the pulsed, near infrared illumination (certified eye safe). There are also drivers to power the deformation of the liquid lens in the optical system. A microcontroller controls precise timing of the peripherals. An encrypted M.2 SSD can be used to store images for voluntary data custody and image data collection. Those images are secured by a second layer of asymmetric encryption such that the Orb can only encrypt, but cannot decrypt. The contribution of data is optional and data deletion can be requested at any point in time through the World App. A SIM card slot enables optional LTE connectivity.

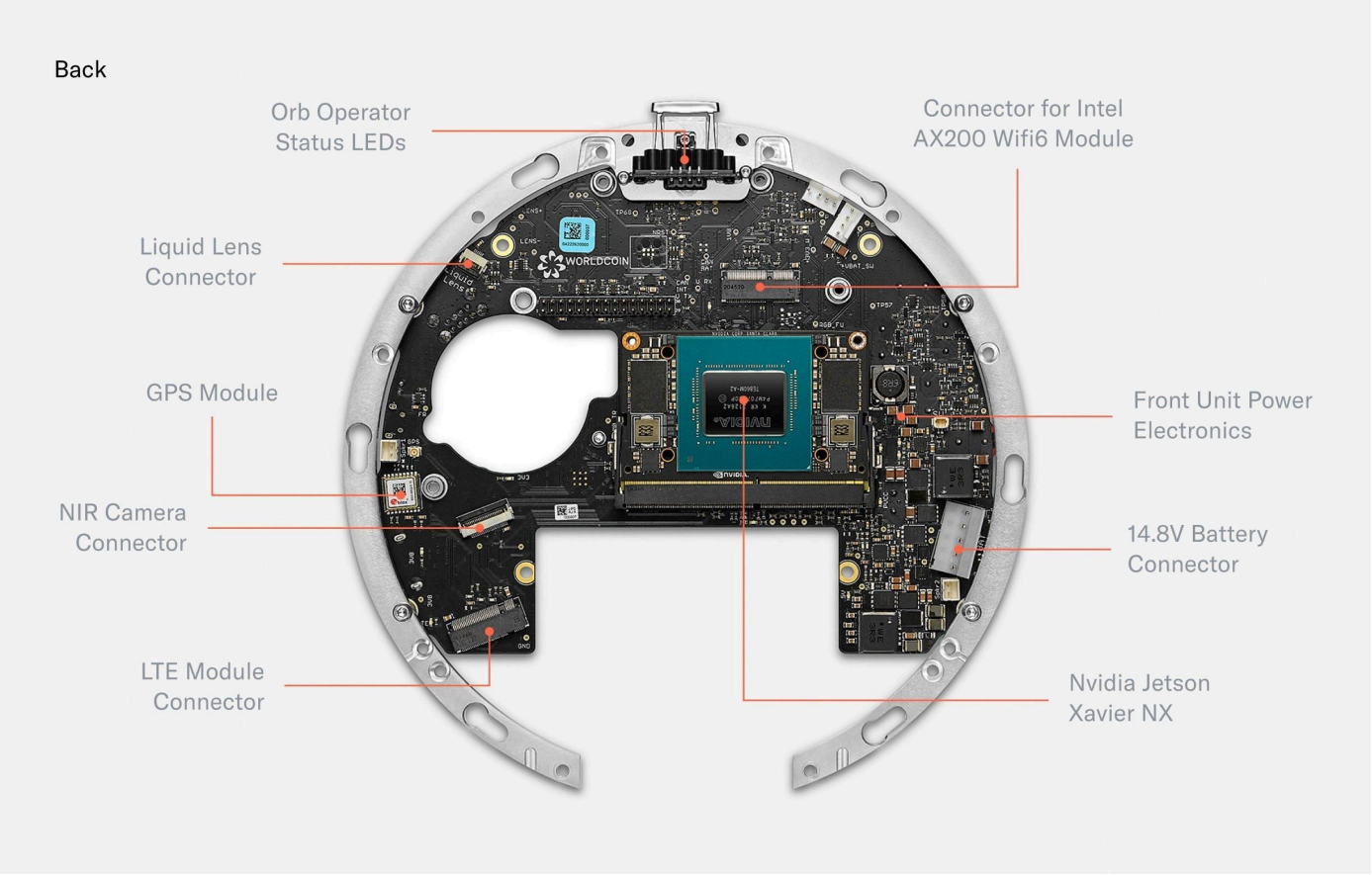

Fig. 12

The back of the mainboard

Fig. 3.

The back of the mainboard holds several connectors for active elements of the optical system. Additionally, a GPS module enables precise location of Orbs for fraud prevention purposes. A Wi-Fi Module equips the Orb with the possibility to upload iris codes to make sure every person can only sign up once. Finally, the mainboard hosts a Nvidia Jetson Xavier NX which runs multiple neural networks in real time to optimize image capture, perform local anti-spoof detection and calculate the iris code locally to maximize privacy.

The mainboard acts as a custom carrier board for the Nvidia Jetson Xavier NX SoM which is the main computing unit powering the Orb. The Jetson is capable of running multiple neural networks on several camera streams in real-time to optimize image capture (autofocus, gimbal positioning, illumination, quality checks i.e. “is_eye_open”) and

perform spoof detection. To optimize for privacy, images are fully processed on the device, and are only stored by Tools for Humanity if the user gives explicit consent to help improve the system.

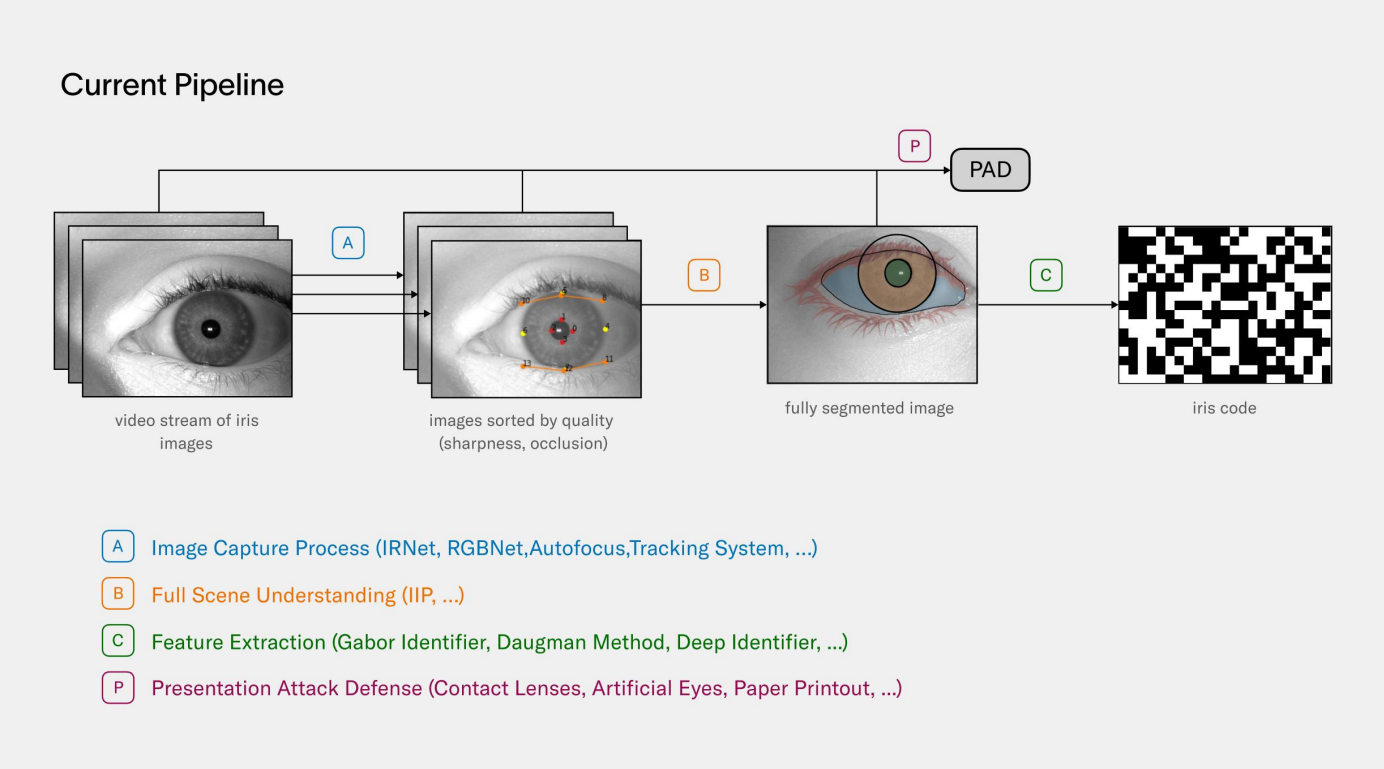

Fig. 13

A: Image capture process optimized by several neural networks in real time.

Apart from the Jetson, the other major “plugged-in” component is a 250GB M.2 SSD. The encrypted SSD can be used to buffer images for voluntary data contribution. Images are protected by a second layer of asymmetric encryption such that the Orb can only encrypt, but cannot decrypt. The contribution of data is optional and data deletion can be requested at any point in time through the app.

Further, a STM32 microcontroller controls time-critical peripherals, sequences power, and boots the Jetson. The Orb is equipped with Wi-Fi 6 and a GPS module to locate the Orb and prevent misuse. Finally, a 12 bit liquid lens driver allows for controlling the focus plane of the telephoto lens with a precision of 0.4mm.

The most densely packed PCB of the Orb is the front PCB. It mainly consists of LEDs. The outermost RGB LEDs power the “UX LED ring.” Further inside, there are 79 near infrared LEDs of different wavelengths. The Orb uses 740nm, 850nm and 940nm LEDs to capture a multispectral image of the iris to make the uniqueness algorithm more accurate and detect spoofing attempts.

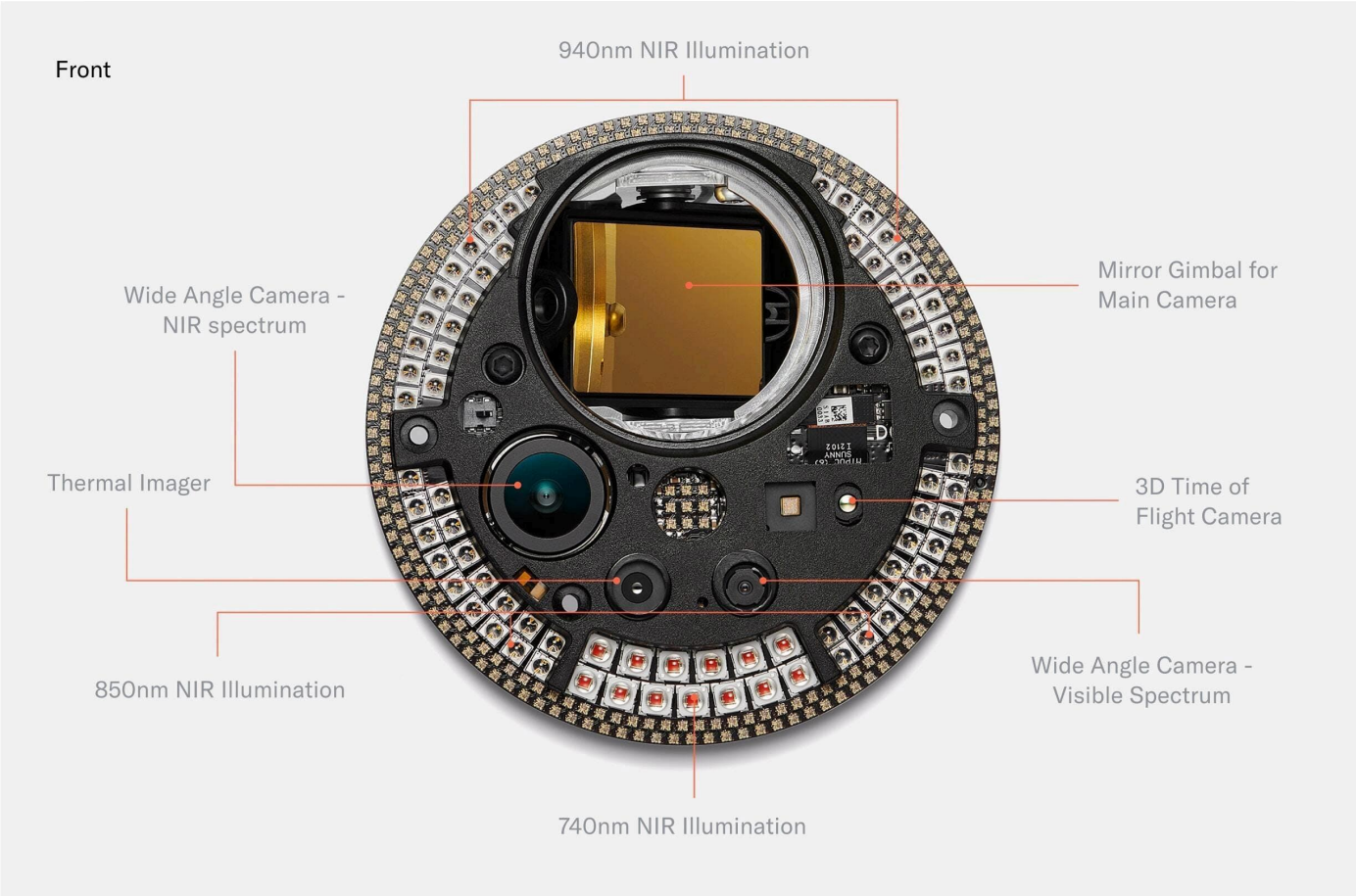

Fig. 14

Front PCB with near infrared illumination. The front PCB powers multispectral illumination as well as fraud prevention sensors. Bright illumination (which is certified eye safe) is needed for high quality image capture, like in a photography studio. Fraud prevention algorithms based on the multispectral sensors are designed to prevent spoofing and run locally on the Orb for maximum privacy. No data from those images is uploaded unless specifically requested by a person. Circular LEDs in the visible spectrum at the border of the PCB enable precise user feedback.

The front PCB also hosts several multispectral imaging sensors. The most basic one is the wide angle camera, which is used for steering the telephoto iris camera. Since every human can only receive one proof of personhood and Worldcoin is giving away a free share of Worldcoin to every person who chooses to verify with the Orb, the incentives for

fraud are high. Therefore, further imaging sensors for fraud prevention purposes were added.

When designing the fraud prevention system, the team started from first principle reasoning: which measurable features do humans have? From there, the team experimented with many different sensors and eventually converged to a set that includes a near infrared wide angle camera, a 3D time of flight camera and a thermal camera. Importantly, the system was designed to enable maximum privacy. The computing unit of the Orb is capable of running several AI algorithms in real time which distinguish spoofing attempts from genuine humans based on the input from those sensors locally. No images are stored unless users give explicit consent to help improve the system for everyone.

Biometrics

Following the exploration of iris biometrics as a choice of modality, this section provides a detailed look into the process of iris recognition from image capture to the uniqueness check:

- Biometric Performance at a Billion People Scale, addresses the scalability of iris recognition technology. It discusses the potential of this biometric modality to establish uniqueness among billions of humans, examines various operating modes and anticipated error rates and ultimately concludes the feasibility of using iris recognition at a global scale.
- Iris Feature Generation with Gabor Wavelets introduces the use of Gabor filtering for generating unique iris features, explaining the scientific principles behind this traditional method which is fundamental to understanding how iris recognition works.
- Iris Inference System explores the practical application of the previously discussed principles. This section describes the uniqueness algorithm and explains how it processes iris images to ensure accurate and scalable verification of uniqueness. This provides a comprehensive overview of the system's operation, demonstrating how theoretical principles translate into practical application.

Collectively, these sections offer a holistic overview of iris recognition, from the core scientific principles to their practical application in the Orb.

Biometric performance at a billion people scale

In order to get a rough estimation on the required performance and accuracy of a biometric algorithm operating on a billion people scale, assume a scenario with a fixed biometric model, i.e. it is never updated such that its performance values stay constant.

Failure Cases

A biometric algorithm can fail in two ways: It can either identify a person as a different person, which is called a false match or it can fail to re-identify a person although this person is already enrolled, which is called a false non match. The corresponding rates - the false match rate (FMR) and the false non match rate (FNMR) - are the two critical KPIs for any biometric system.

For the purposes of this analysis, consider three different systems with varying levels of performance.

- One of the systems, as reported by John Daugman in his paper, demonstrates a false match rate of 1.1×10^{-7} at a false non-match rate of 0.00014.
- Another system, represented by one of the leading iris recognition algorithms from NEC, has performance values as reported in the IREX IX report and IREX X leaderboard from the National Institute for Standards and Technology (NIST). These values include a false match rate of 10^{-8} at a false non-match rate of 0.045.
- The third system, conceived during the early ideation stage of the Worldcoin project, represents a conservative estimate of how well iris recognition could perform outside of the lab environment i.e. in an uncontrolled, outdoor setting. Despite these constraints, it anticipated a false match rate of 10^{-6} and a false non-match rate of 0.005. While not ideal, it demonstrated that iris recognition was the most viable path for a global proof of personhood.

A more in-depth examination of how these values are obtained from various sources is also available.

Effective Dual Eye Performance

The values mentioned above pertain to single eye performance, which is determined by evaluating a collection of genuine and imposter iris pairs. However, utilizing both eyes can significantly enhance the performance of a biometric system. There are various methods for combining information from both eyes, and to evaluate their performance, consider two extreme cases:

- The AND-rule, in which a user is deemed to match only if their irises match on both eyes.
- The OR-rule, in which a user is considered a match if their iris on one eye matches that of another user's iris on the same eye.

The OR-rule offers a safer approach as it requires only a single iris match to identify a registered user, thus minimizing the risk of falsely accepting the same person twice. Formally, the OR-rule reduces the false non-match rate while increasing the false match rate. However, as the number of registered users increases over time, this strategy may make it increasingly difficult for legitimate users to enroll to the system due to the high false match rate. The effective rates are given below:

$$\begin{aligned} \mathrm{FMR}_{\mathrm{OR}} &= 2\mathrm{FMR}(1-\mathrm{FMR}) + \mathrm{FMR}^2 \\ \mathrm{FNMR}_{\mathrm{OR}} &= \mathrm{FNMR}^2 \end{aligned}$$

On the other hand, the AND-rule allows for a larger user base, but comes at the cost of less security, as the false match rate decreases and the false non-match rate increases. The performance rates for this approach are as follows:

$$\begin{aligned} \mathrm{FMR}_{\mathrm{AND}} &= \mathrm{FMR}^2 \\ \mathrm{FNMR}_{\mathrm{AND}} &= 2\mathrm{FNMR}(1-\mathrm{FNMR}) + \mathrm{FNMR}^2 \end{aligned}$$

False Matches

The probability for the i-th (legitimate) user to run into a false match error can be calculated by the equation

$$P_{\text{FM}}(i) = 1 - P_{\text{no match with i-1 users in DB}} = 1 - (1-p)^{i-1}$$

with $p{:}\ = FMR$ being the false match rate. Adding up these numbers yields the expected number of false matches that have happened after the i-th user has enrolled, i.e the number of falsely rejected users (derivation).

$$N_{\text{FM}}(i) = \sum_{j=1}^{i} P_{\text{FM}}(j) = \frac{(1-p)^i + i \cdot p - 1}{p}$$

A high false match rate significantly impacts the usability of the system, as the probability of false matches increases with a growing number of users in the database. Over time the probability of being (falsely) rejected as a new user converges to 100%, making it nearly impossible for new users to be accepted.

The following graph illustrates the performance of the biometric system using both the OR and AND rule. The graph is separated into two sections, with the left side representing the OR rule and the right side representing the AND rule. The top row of plots in the graph shows the probability $P_{FM}(i)$ of the i-th user being falsely rejected, and the bottom row of plots shows the expected number $N_{FM}(i)$ of users that have been falsely rejected after the i-th user has successfully enrolled. The different colors in the graph correspond to the three systems mentioned earlier: green represents Daugman's system, blue represents NEC's system, and red represents the initial worst case estimate.

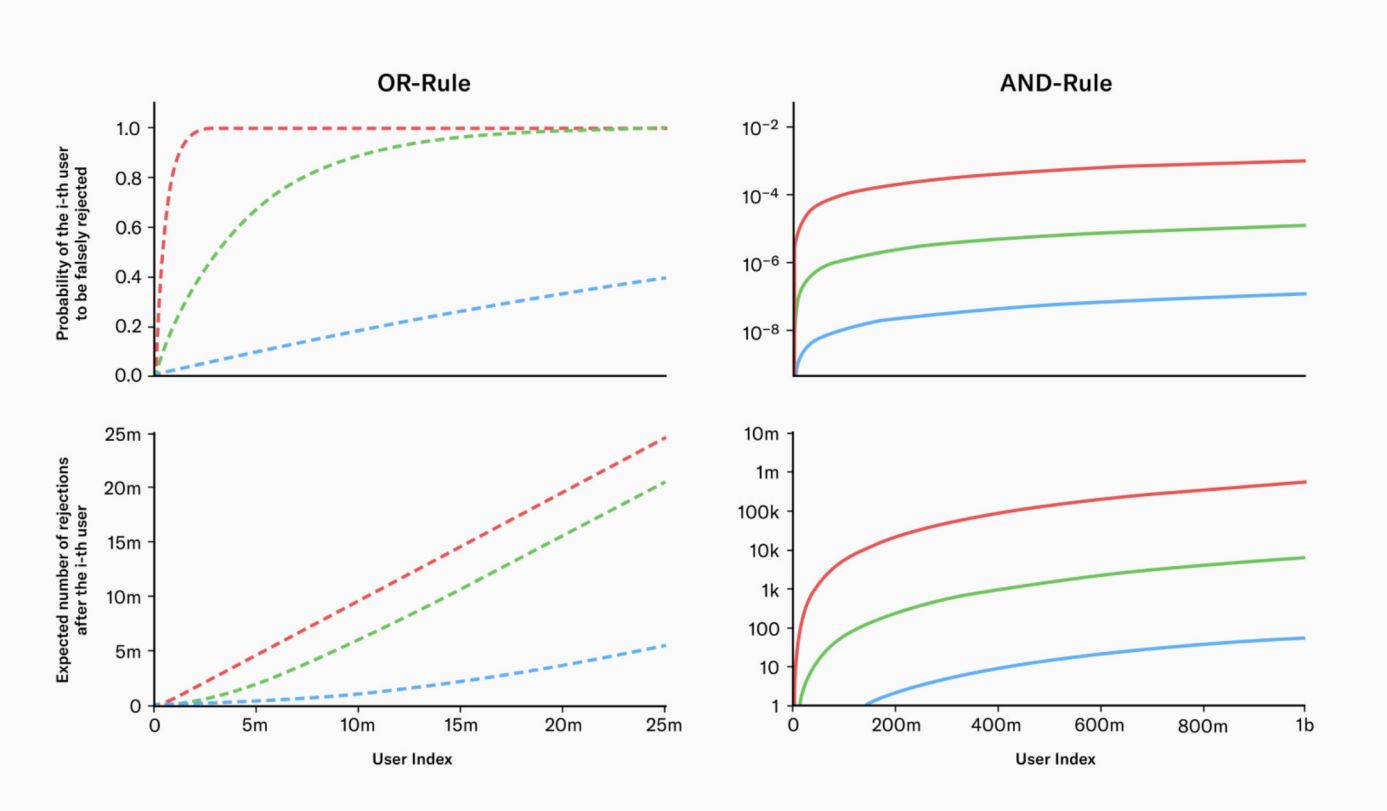

Fig. 15

Performance of biometric systems under both the OR and AND rule across three distinct scenarios: The blue line represents a highly performant system from NEC, while the green line reflects performance values as reported by John Daugman. The red line indicates a system with conservative performance values.

The main findings from the analysis indicate that when using the OR-rule, the system's effectiveness breaks down with just a few million users, as the chance of a new user being falsely rejected becomes increasingly likely. In comparison, operating with the AND-rule provides a more sustainable solution for a growing user base.

Further, even the difference between the worst case and the best case estimate of current technology matters. The performance of biometric algorithms designed by Tools for Humanity has been continuously improving due to ongoing research efforts. This has been achieved by pushing beyond the state-of-the-art by replacing various components of the uniqueness verification process with deep learning models which also significantly improves the robustness to real world edge cases. At the time of writing, the algorithm's performance closely resembled the green graph depicted in the figure above when in an uncontrolled environment (depending on the exact choice of the FNMR). This is an accomplishment noteworthy in and of itself. Nonetheless, further improvements in

the algorithm's performance are expected through ongoing research efforts. The optimum case is a vanishing error rate in practice on a global scale.

Note that for a large number of users (i≫1) and a very performant biometric system (p≪1) the equation above becomes numerically unstable. To calculate the number of rejected users for such a scenario, Taylor expand the critical part of the equation around small values of p.

$$(1-p)^i = 1 - ip + \frac{1}{2}(i-1)ip^2 + \mathcal{O}(i^3p^3)$$

The derivation of the above equation can be found here. Inserting this in the equation above yields

$$N_{\mathrm{FM}}(i) = \frac{1}{2}(i-1)ip + \mathcal{O}(i^3p^2) \approx \frac{1}{2}(i-1)ip$$

which is a valid approximation as long as $i^2p \gg i^3p^2 \leftrightarrow ip \ll 1$

False Non Matches

When it comes to fraudulent users, the probability of them not being matched stays constant and does not increase with the number of users in the system. This is because there is only one other iris that can cause a false non-match - the user's own iris from their previous enrollment. Thus, the probability of encountering a false non-match is given by

$$P_{\mathrm{FNM}} = \mathrm{FNMR}$$

The number of expected false non matches can be calculated with

$$N_{\mathrm{FNM}}(j) = j \cdot P_{\mathrm{FNM}} = j \cdot \mathrm{FNMR}$$

with j indicating the j-th untrustworthy user who tries to fool the system.

Conclusion

The conclusion is that iris recognition can establish uniqueness on a global scale. Further, to onboard billions of individuals, the algorithm needs to use the AND-rule. Otherwise, the rejection rate will be too high and it will be practically impossible to onboard billions of users.

The current performance is already beyond the original conservative estimate and the project expects the system to eventually surpass current state-of-the-art lab environment performance, even if subject to an uncontrolled environment: On the one hand, the custom hardware comprises an imaging system that outperforms typical iris scanners by more than an order of magnitude in terms of image resolution. On the other hand, current advances in deep learning and computer vision offer promising directions towards a “deep feature generator” - a feature generation algorithm that does not rely on handcrafted rules but learns from data. So far the field of iris recognition has not yet leveraged this new technology.

Iris Feature Generation with Gabor Wavelets

The objective for iris feature generation algorithms is to generate the most discriminative features from iris images while reducing the dimensionality of data by removing unrelated or redundant data. Unlike 2D face images that are mostly defined by edges and shapes, iris images present rich and complex texture with repeating (semi-periodic) patterns of local variations in image intensity. In other words, iris images contain strong signals in both spatial and frequency domains and should be analyzed in both. Examples of iris images can be found on John Daugman's website.

Gabor filtering

Research has shown that the localized frequency and orientation representation of Gabor filters is very similar to the human visual cortex's representation and discrimination of texture. A Gabor filter analyzes a specific frequency content at a specific direction in a local region of an image. It has been widely used in signal and image processing for its optimal joint compactness in spatial and frequency domain.

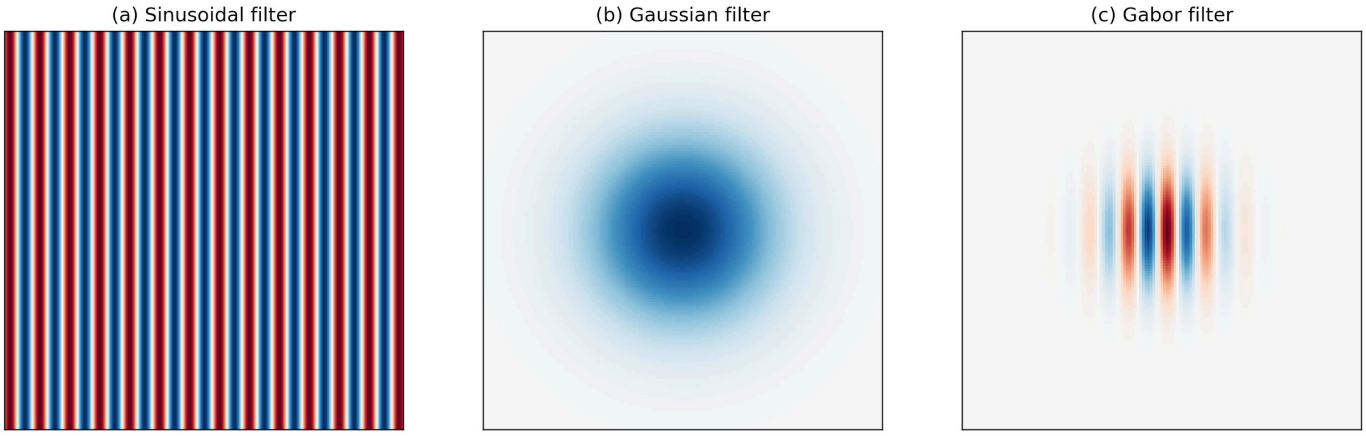

Fig. 16

Constructing a Gabor filter is straightforward. The product of (a) a complex sinusoid signal and (b) a Gaussian filter produces (c) a Gabor filter.

As shown above, a Gabor filter can be viewed as a sinusoidal signal of particular frequency and orientation modulated by a Gaussian wave. Mathematically, it can be defined as

$$G_{\lambda,\theta,\phi,\sigma,\gamma}(x,y) = \exp(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}) \exp(j(2\pi\frac{x'}{\lambda} + \phi))$$

with

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

Among the parameters, σ and γ represent the standard deviation and the spatial aspect ratio of the Gaussian envelope, respectively, λ and ϕ are the wavelength and phase offset of the sinusoidal factor, respectively, and θ is the orientation of the Gabor function. Depending on its tuning, a Gabor filter can resolve pixel dependencies best described by narrow spectral bands. At the same time, its spatial compactness accommodates spatial irregularities.

The following figure shows a series of Gabor filters at a 45 degree angle in increasing spectral selectivity. While the leftmost Gabor wavelet resembles a Gaussian, the rightmost Gabor wavelet follows a harmonic function and selects a very narrow band from the spectrum. Best for iris feature generation are the ones in the middle between the two extremes.

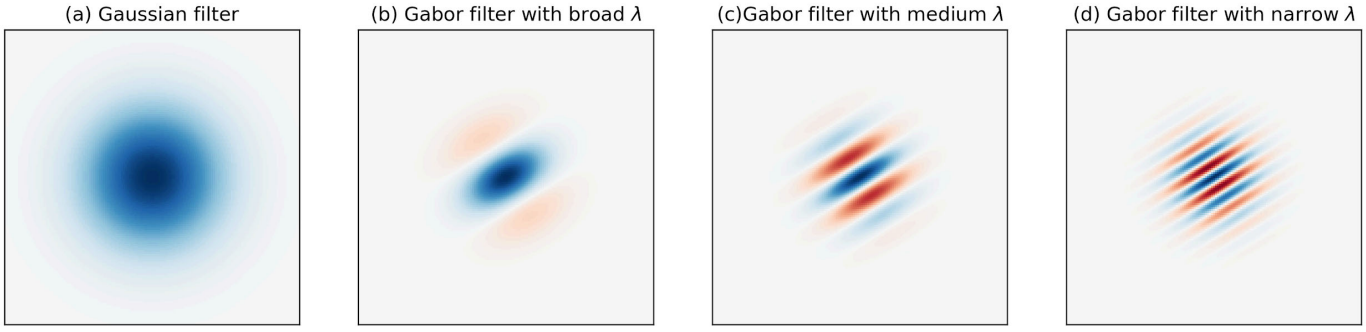

Fig. 17

Varying wavelength (a-d) from large to small can change the spectral selectivity of Gabor filters from broad to narrow.

Because a Gabor filter is a complex filter, the real and imaginary parts act as two filters in quadrature. More specifically, as shown in the figures below, (a) the real part is even-symmetric and will give a strong response to features such as lines; while (b) the imaginary part is odd-symmetric and will give a strong response to features such as edges. It is important to maintain a zero DC component in the even-symmetric filter (the odd-symmetric filter already has zero DC). This ensures zero filter response on a constant region of an image regardless of the image intensity.

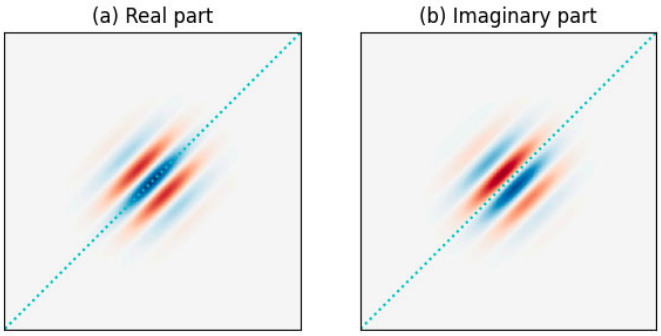

Fig. 18

Giving a closer look at the complex space of a Gabor filter where (a) the real part is even-symmetric and (b) the imaginary part is odd-symmetric.

Multi-scale Gabor filtering

Like most textures, iris texture lives on multiple scales (controlled by σ). It is therefore natural to represent it using filters of multiple sizes. Many such multi-scale filter systems follow the wavelet building principle, that is, the kernels (filters) in each layer are scaled versions of the kernels in the previous layer, and, in turn, scaled versions of a mother wavelet. This eliminates redundancy and leads to a more compact representation. Gabor wavelets can further be tuned by orientations, specified by θ. The figure below shows the real part of 28 Gabor wavelets with four scales and 7 orientations.

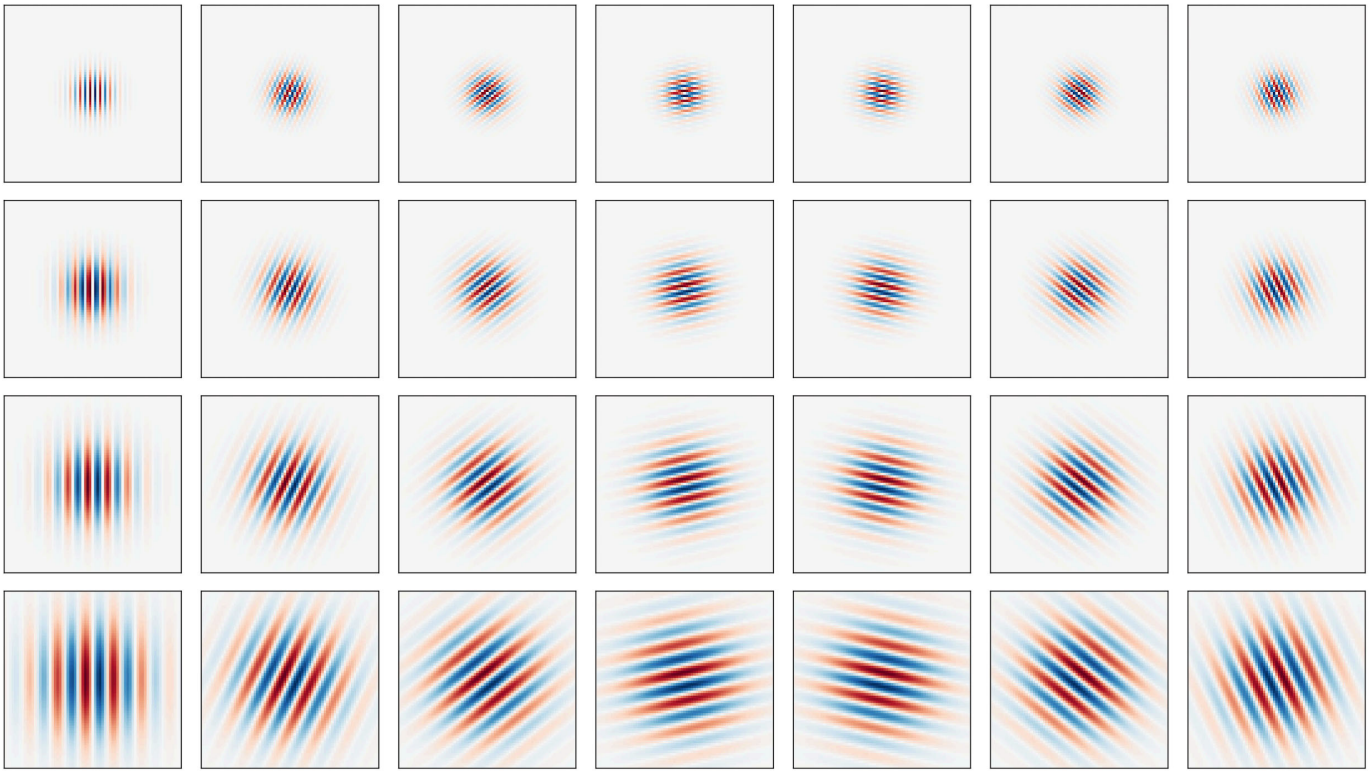

Fig. 19

Constructing Gabor wavelets with multiple scales (vertically) and orientations (horizontally) to generate texture features with various frequencies and directions. In the feature generation process, the system uses a small set of filters that concentrate within the range of scales and orientations of the most discriminative iris texture.

Phase-quadrant demodulation and encoding

After a Gabor filter is applied to an iris image, the filter response at each analyzed region is then demodulated to generate its phase information. This process is illustrated in the figure below, as it identifies in which quadrant of the complex plane each filter response is projected to. Note that only phase information is recorded because it is more robust than the magnitude, which can be contaminated by extraneous factors such as illumination, imaging contrast, and camera gain.

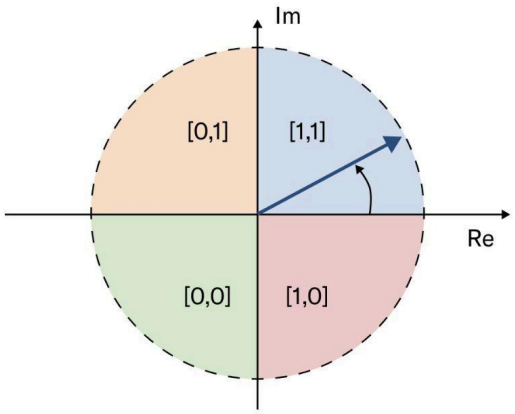

Fig. 20

Demodulating the phase information of filter response into four quadrants of the complex space. The resulting cyclic codes are used to produce the final iris code.

Another desirable feature of the phase-quadrant demodulation is that it produces a cyclic code. Unlike a binary code in which two bits may change, making some errors arbitrarily more costly than others, a cyclic code only allows a single bit change in rotation between any adjacent phase quadrants. Importantly, when a response falls very closely to the boundary between adjacent quadrants, its resulting code is considered a fragile bit. These fragile bits are usually less stable and could flip values due to changes in illumination, blurring or noise. There are many methods to deal with fragile bits, and one such method could be to assign them lower weights during matching.

When multi-scale Gabor filtering is applied to a given iris image, multiple iris codes are produced accordingly and concatenated to form the final iris template. Depending on the number of filters and their stride factors, an iris template can be several orders of magnitude smaller than the original iris image.

Robustness of iris codes

Because iris codes are generated based on the phase responses from Gabor filtering, they are rather robust against illumination, blurring and noise. To measure this quantitatively, each effect is added, namely, illumination (gamma correction), blurring (Gaussian filtering), and Gaussian noise to an iris image, respectively, in slow progression and measure the drift of the iris code. The amount of added effect is measured by the Root Mean Square Error (RMSE) of pixel values between the modified and original image, and the amount of drift is measured by the Hamming distance between the new and original iris code. Mathematically, RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{p=1}^{N}(I'_p - I_p)^2}$$

where N is the number of pixels in the original image I and the modified image I′. The Hamming distance is defined as:

$$\mathrm{HD} = \frac{1}{K}\sum_{p=1}^{K}|C'_p - C_p|$$

where K is the number of bits (0/1) in the original iris code C and the new iris code C′. A Hamming distance of 0 means a perfect match, while 1 means the iris codes are completely opposite. The Hamming distance between two randomly generated iris codes is around 0.5.

The following figures help explain the impact of illumination both visually and quantitatively, blurring and noise on the robustness of iris codes. For illustration purposes, these results are not generated with the actual filters that are deployed but nevertheless demonstrate the property in general of Gabor filtering. Also, the iris image

has been normalized from a donut shape in the cartesian coordinates to a fixed-size rectangular shape in the polar coordinates. This step is necessary to standardize the format, mask-out occlusion and enhance the iris texture.

As shown in the figure below, iris codes are very robust against grey-level transformations associated with illumination as the HD barely changes with increasing RMSE. This is because increasing the brightness of pixels reduces the dynamic range of pixel values, but barely affects the frequency or spatial properties of the iris texture.

Fig. 21

An animation showcasing the effect of varying illumination levels on the robustness of iris codes. Each frame represents an increase in illumination, portrayed through the Root Mean Square Error (RMSE) between images (blue line) and the Hamming Distance (HD) between corresponding iris codes (green line).

Blurring, on the other hand, reduces image contrast and could lead to compromised iris texture. However, as shown below, iris codes remain relatively robust even when strong blurring makes iris texture indiscernible to naked eyes. This is because the phase

information from Gabor filtering captures the location and presence of texture rather than its strength. As long as the frequency or spatial property of the iris texture is present, though severely weakened, the iris codes remain stable. Note that blurring compromises high frequency iris texture, therefore, impacting high frequency Gabor filters more, which is why a bank of multi-scale Gabor filters are used.

Fig. 22

An animation illustrating the impact of blurring on the robustness of iris codes. The blurring intensifies with each frame, as demonstrated by the Root Mean Square Error (RMSE) between images (blue line) and the Hamming Distance (HD) between corresponding iris codes (green line).

Finally, observe bigger changes in iris codes when Gaussian noise is added, as both spatial and frequency components of the texture are polluted and more bits become fragile. When the iris texture is overwhelmed with noise and becomes indiscernible, the drift in iris codes is still small with a Hamming distance below 0.2, compared to matching two random iris codes (≈0.5). This demonstrates the effectiveness of iris feature generation using Gabor filters even in the presence of noise.

Fig. 23

An animation demonstrating the impact of noise on the robustness of iris codes. With each successive frame, the level of noise is increased, shown through Root Mean Square Error (RMSE) between images (blue line) and Hamming Distance (HD) between corresponding iris codes (green line).

Conclusion

Iris feature generation is a necessary and important step in iris recognition. It reduces the dimensionality of the iris representation from a high resolution image to a much lower dimensional binary code, while preserving the most discriminative texture features using a bank of Gabor filters. It is worth noting that Gabor filters have their own limitations, for example, one cannot design Gabor filters with arbitrarily wide bandwidth while maintaining a near-zero DC component in the even-symmetric filter. This limitation can be overcome by using the Log Gabor filters. In addition, Gabor filters are not necessarily optimized for iris texture, and machine-learned iris-domain specific filters (e.g. BSIF) have the potential to achieve further improvements in feature generation and recognition performance in general. Moreover, the project's contributors are investigating novel

approaches to leverage higher quality images and the latest advances in the field of deep metric learning and deep representation learning to push the accuracy of the system beyond the state-of-the-art to make the system as inclusive as possible.

As the resilience of iris feature generation amidst external factors was showcased, it is crucial to note that even minor fluctuations in iris code variability hold significant importance when dealing with a billion people, as the tail-end of the distribution dictates the error rates, thus influencing the number of false rejections.

Iris Inference System

Building upon the theoretical foundation established in the previous sections, this section now focuses on the practical application of these principles within the Worldcoin project. Having explored the scalability of iris recognition technology and the process of feature generation using Gabor wavelets, this section explains the details of the image processing. By the end of this section, one will have a thorough understanding of how Worldcoin's iris recognition algorithm functions to ensure accurate and scalable verification of an individual's uniqueness.

Pipeline overview

The objective of this pipeline is to convert high-resolution infrared images of a human's left and right eye into an iris code: a condensed mathematical and abstract representation of the iris' entropy that can be used for verification of uniqueness at scale. Iris codes have been introduced by John Daugman in this paper and remain to this day the most widely used way to abstract iris texture in the iris recognition field. Like most state-of-the-art iris recognition pipelines, Worldcoin's pipeline is composed of four main segments: segmentation, normalization, feature generation and matching.

Refer to the image below for an example of a high resolution image of the iris acquired in the near infrared spectrum. The right hand side of the image shows the corresponding iris code, which is itself composed of $n_f = 2$ response maps to two 2D Gabor wavelets. These response maps are quantized in two bits so that the final iris code has dimensions of $n_h \times n_w \times n_f \times 2$, with n_h and n_w being the number of radial and angular positions

where these filters are applied. For more details, see check the previous section. While only the iris code of one eye is shown below, note that an iris template consists of the iris codes from both eyes.

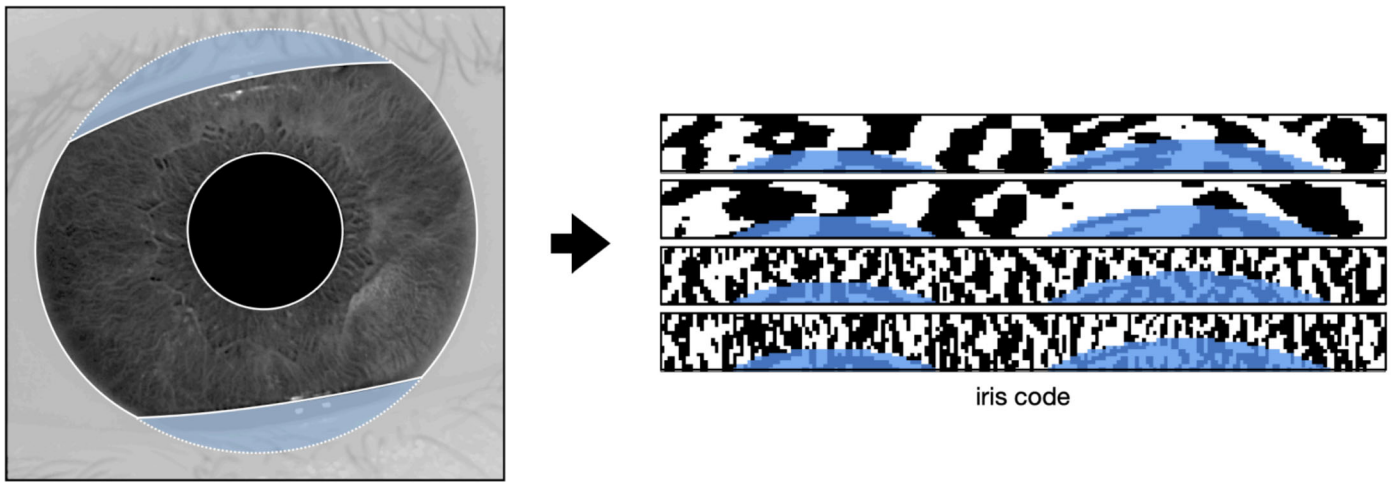

Fig. 24

Example of an input and output of the biometric pipeline. Fig. 1.a is an example of an infrared iris texture image taken by the Orb. Fig. 1.b is an example of an iris code produced from the iris texture image in Fig. 1.a, effectively aggregating the iris texture.

The purpose of the segmentation step is to understand the geometry of the input image. The location of the iris, pupil, and sclera are determined, as well as the dilation of the pupil and presence of eyelashes or hair covering the iris texture. The segmentation model classifies every pixel of the image as pupil, iris, sclera, eyelash, etc. These labels are then post-processed to understand the geometry of the subject's eye.

The image and its geometry then passes through tight quality assurance. Only sharp images where enough iris texture is visible are considered valid, because the quantity and quality of available bits in the final iris codes directly impact the system's overall performance.

Once the image is segmented and validated, the normalization step takes all the pixels relevant to the iris texture and unfolds them into a stable cartesian (rectangular) representation.

The normalized image is then converted into an iris code during the feature generation step. During this process, a Gabor wavelet kernel convolves across the image, converting the iris texture into a standardized iris code. For every point in a grid overlapping the image, two bits that represent the sign of the real and complex components of the filter response are derived, respectively. This process synthesizes a unique representation of the iris texture, which can easily be compared with others by using the Hamming distance metric. This metric quantifies the proportion of bits that differ between any two compared iris codes.

The following sections will explain each of the aforementioned steps in more detail, by following the journey of an example iris image through the biometric pipeline. This image was taken by the Orb, during a signup in the TFH lab. It is shared with user consent and faithfully represents what the camera sees during a live uniqueness verification.

The eye is a remarkable system that exhibits various dynamic behaviors, including blinking, squinting, closing, as well as the ability of the pupil to dilate or constrict and the eyelashes or any object to cover the iris. The following section also explores how the biometric pipeline can be robust in the presence of such natural variability.

Segmentation

Iris recognition was first developed in 1993 by John Daugmann and, although the field has advanced since the turn of the millennium, it continues to be heavily influenced by legacy methods and practices. Historically, the morphology of the eye in iris recognition has been identified using classical computer vision methods such as the Hough Transform or circle fitting. In recent years, Deep Learning has brought about significant improvements in the field of computer vision, providing new tools for understanding and analyzing the eye physiology with unprecedented depth.

Novel methods for segmenting high-resolution infrared iris images are proposed in Lazarski et al by the Tools for Humanity team. The architecture consists of an encoder that is shared by two decoders: one that estimates the geometry of the eye (pupil, iris, and eyeball) and the other that focuses on noise, i.e., non-eye-related elements that

overlay the geometry and potentially obscure the iris texture (eyelashes, hair strands, etc.). This dichotomy allows for easy processing of overlapping elements and provides a high degree of flexibility in training these detectors. The architecture takes into account the DeepLabv3+ architecture with a MobileNet v2 backbone.

Acquiring labels for noise elements is significantly more time-consuming than acquiring labels for geometry, as it requires a high level of precision for identifying intertwined eyelashes. It takes 20 to 80 minutes to label eyelashes in a single image, depending on the levels of blur and the subject's physiology, while it only takes about 4 minutes to label the geometry to required levels of precision. For that reason, noise objects (e.g. eyelashes) are decoupled from geometry objects (pupil, iris and sclera) which allows for significant financial and time savings combined with a quality gain.

The model was trained over a mix of Dice Loss and Boundary Loss. The Dice loss can be expressed as

$$L_D = \sum_k \left(1 - \frac{2\sum_{i,j} y_{i,j,k} \cdot p_{i,j,k}}{\sum_{i,j} y_{i,j,k}^2 \cdot \sum_{i,j} p_{i,j,k}^2}\right)$$

with $y_{i,j,k} \in \{0, 1\}$ being the one-hot encoded ground truth and $p_{i,j,k} \in [0, 1]$ the model's output for the pixel (i,j) as a probability. The third index k represents the class (e.g. pupil, iris, eyeball, eyelash or background). The Dice loss essentially measures the similarity between two sets, i.e. the label and the model's prediction.

Accurate identification of the boundaries of the iris is essential for successful iris recognition, as even a small warp in the boundary can result in a warp of the normalized image along the radial direction. To address this, a weighted cross-entropy loss was also introduced that focuses on the zone at the boundary between classes, in order to encourage sharper boundaries. It is mathematically represented as:

$$L_B = \sum_{i,j} \sum_{k} b_{i,j,k} \cdot y_{i,j,k} \cdot \log(p_{i,j,k})$$

with the same notations as before and $b_{i,j,k}$ being the boundary weight, which represents how close the pixel (i,j) is to the boundary between class k and any other class. A Gaussian blur is then applied to the contour to prioritize the precision of the model on the exact boundary while keeping a lower degree of focus on the general area around it.

$$b_{i,j,k} = G(d(i,j,S_k))$$

With $d(i,j,S_k)$ being the distance between the point (i,j) and the surface S_k as the minimum of the euclidean distances between (i,j) and all points of S_k. S_k is the boundary between class k and all other classes, G the Gaussian distribution centered at 0 with some finite variance.

Experiments were conducted with other loss functions (e.g. convex prior), architectures (e.g. single-headed model), and backbones (e.g. ResNet-101) and this setup was found to have the best performance in terms of accuracy and speed. The following graph shows the iris image overlayed by the segmentation maps as predicted by the model. In addition, landmarks are displayed calculated by a separate quality assessment AI model during the image capture phase. This model produces quality metrics to ensure that only high-quality images are used in the segmentation phase and that the iris code is generated accurately for verification of uniqueness: sharp image focused on the iris texture, well-opened eye gazing in the camera, etc.

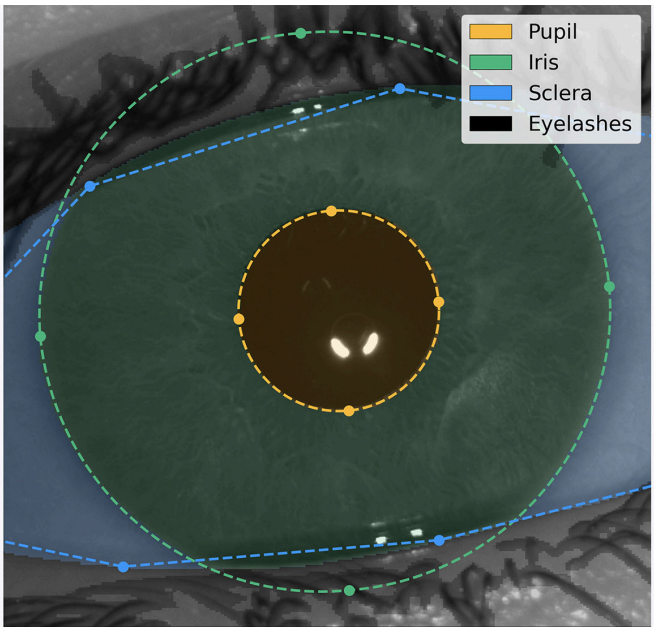

Fig. 25

Segmentation of an iris image. The AI model detects the different regions of interest of the eye in order to isolate the relevant iris texture and assess the overall image quality. This exemplifies the outcome of an employee signup conducted in the lab of Tools for Humanity.

Normalization

The goal of this step is to separate meaningful iris texture from the rest of the image (skin, eyelashes, sclera, etc.). To achieve this, the iris texture is projected from its original cartesian coordinate system to a polar coordinate system, as illustrated in the following image. The iris orientation is defined as the vector pointing from one pupil center to the other pupil center of the opposite eye.

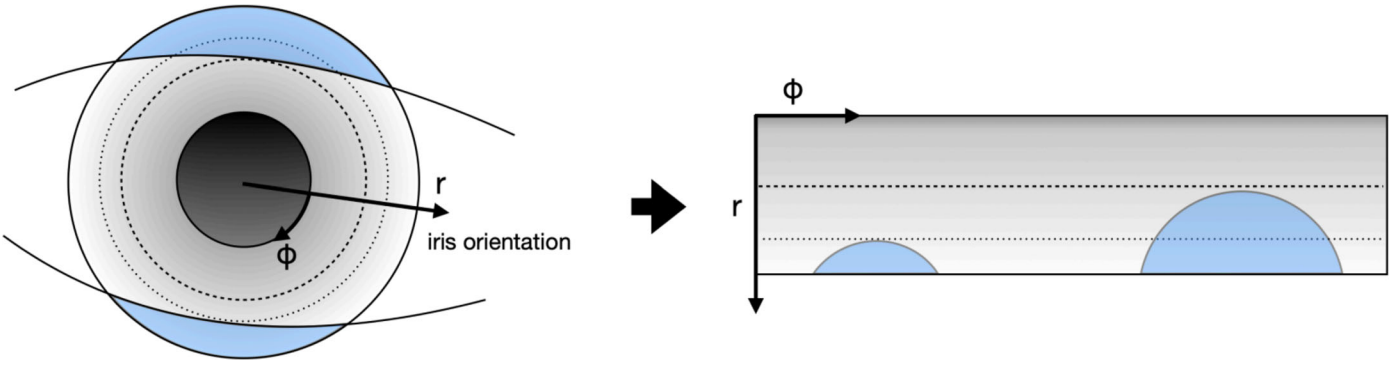

Fig. 26

Scheme of the normalization process.

This process reduces variability in the image by canceling out variations such as the person's distance from the camera, the pupil constriction or dilation due to the amount of light in the environment, and the rotation of the person's head. The image below illustrates the normalized version of the iris above. The two arcs of circles visible in the image are the eyelids, which were distorted from their original shape during the normalization process.

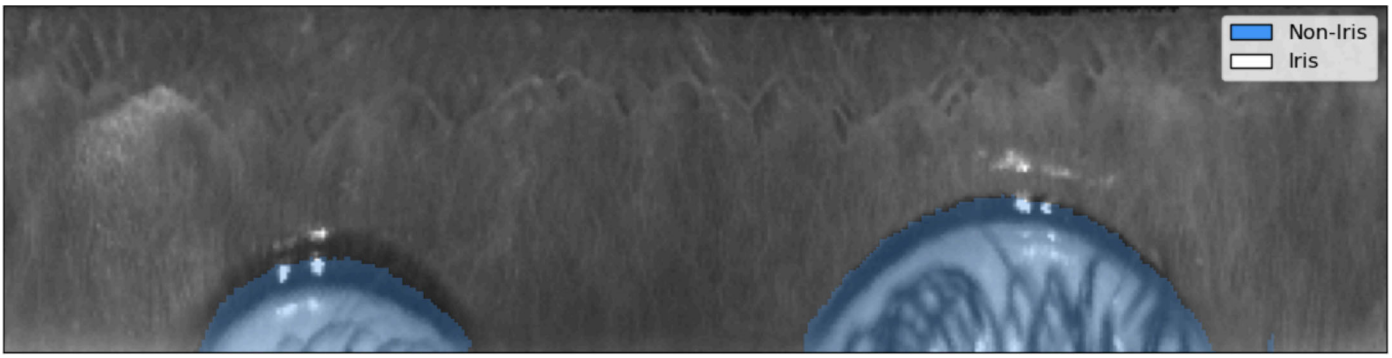

Fig. 27

Normalized iris texture. The texture is sharp and its patterns are clearly visible.

Feature generation

Now that a stable, normalized iris texture is produced, an iris code can be coded that can be matched at scale. In short, various Gabor filters stride across the image and threshold its complex-valued response to generate two bits representing the existence of a line (resp. edge) at every selected point of the image. This technique, pioneered by John Daugmann, and the subsequent iterations proposed by the iris recognition research community, remains state-of-the-art in the field.

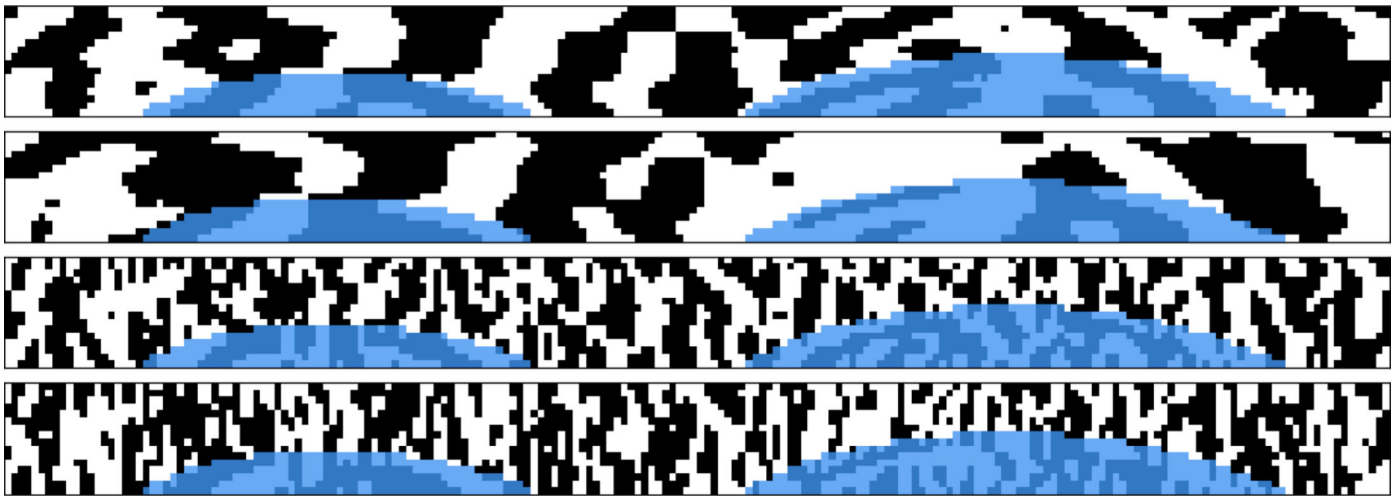

Fig. 28

Final iris code. This is the anonymized iris texture expressing one's uniqueness.

Matching

Now that the iris texture is transformed into an iris code, it is ready to be matched against other iris codes. To do so, a masked fractional Hamming Distance (HD) was used: the proportion of non-masked iris code bits that have the same value in both iris codes.

Due to the parametrization of the Gabor wavelets, the value of each bit is equally likely to be 0 or 1. As the iris codes described above are made of more than 10,000 bits, two iris codes from different subjects will have an average Hamming distance of 0.5, with most (99.95%) iris codes deviating less than 0.05 HD away from this value (99.9994% deviating less than 0.07 HD). As several rotations of the iris code are compared to find the combination with highest matching probability, this average of 0.5 HD moves to 0.45 HD, with a 1.6×10^{-7} probability of being lower than 0.38 HD.

It is therefore an extreme statistical anomaly to see two different eyes producing iris codes with a distance lower than 0.38 HD. On the contrary, two images captured of the same eye will produce iris codes with a distance generally below 0.3 HD. Applying a threshold in between allows the ability to reliably distinguish between identical and different identities.

To validate the quality of the algorithms at scale, their performance was evaluated by collecting 2.5 million pairs of high-resolution infrared iris images from 303 different

subjects. These subjects represent diversity across a range of characteristics, including eye color, skin tone, ethnicity, age, presence of makeup and eye disease or defects. Note that this data was not collected during field operations but stems from contributors to the Worldcoin project and from paid participants in a dedicated session organized by a respected partner. Using these images and their corresponding ground truth identities, the false match rate (FMR) and false non match rate (FNMR) of the system was measured.

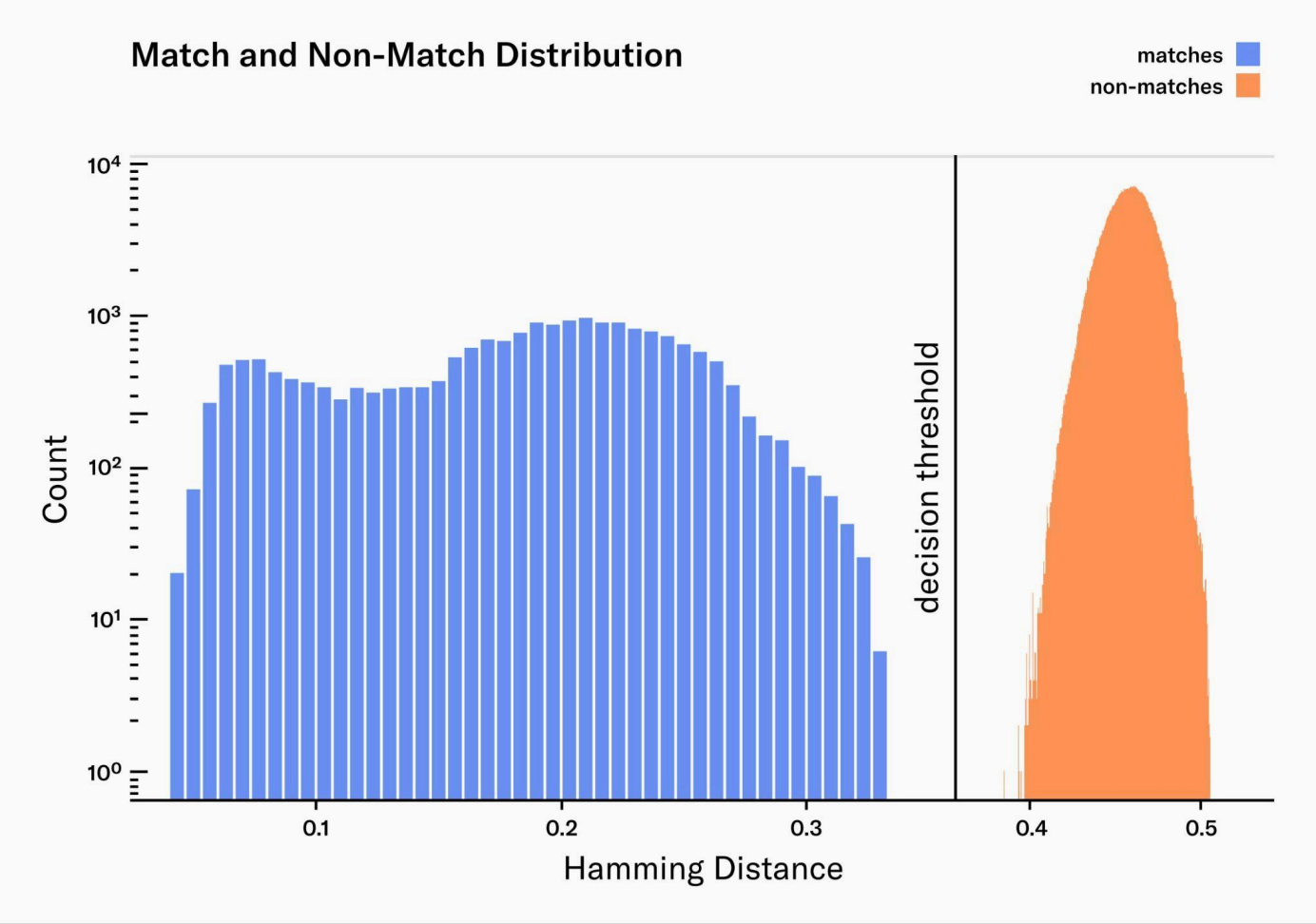

Fig. 29

Match and Non-Match Distribution

From 2.5 million image pairs, all were correctly classified as either a match or non-match. Additionally, the margin between the match and non-match distributions is wide, providing a comfortable margin of error to accommodate for potential outliers.

The match distribution presents two clear peaks, or maxima. The peak on the left (HD≈0.08) corresponds to the median Hamming distance for pairs of images taken from the same person during the same capture process. This means that they are extremely similar, as one would expect from two images of the same person. The peak on the right (HD≈0.2) represents the median Hamming distance for pairs of images taken from the same person but during different enrollment processes, often weeks apart. These are less similar, reflecting the naturally occurring variations in the same person's images taken at different times like pupil dilation, occlusion and eyelashes. Systems to narrow the matches distribution are continuously being iterated on: better auto-focus and AI-Hardware interactions, better real-time quality filters, Deep Learning feature generation, image noise reduction, etc.

As there were no misclassified iris pairs, FMR and FNMR cannot be calculated exactly. However, an upper bound for both rates can be estimated:

$$FMR = \frac{n_{FM}}{n_{TM} + n_{FM}} < \frac{1}{2.4 \cdot 10^{7}} = 4.1 \cdot 10^{-8}$$

$$FNMR = \frac{n_{FNM}}{n_{TNM} + n_{\mathrm{FNM}}} < \frac{1}{4.1 \cdot 10^{4}} = 2.4 \cdot 10^{-5}$$

With these numbers, uniqueness on a billion people scale can be verified with very high accuracy. However, also acknowledged is the fact that the dataset used for this evaluation could be enlarged and more effort is needed to build larger and even more diverse datasets to more accurately estimate the biometric performance.

Conclusions

In this section, the key components of Worldcoin's uniqueness verification pipeline are presented. It illustrated how the use of a combination of deep learning models for image quality assessment and image understanding, in conjunction with traditional feature generation techniques, enables accurate verification of uniqueness on a global scale.

However, work in this area is ongoing. Currently, the team at TFH is researching an end-to-end Deep Learning model, which could yield faster and even more accurate uniqueness verification.

Iris Code Upgrades

While the accuracy of the uniqueness verification algorithm of the Orb is already very high (with, specifically a false match rate of 1 in 40 trillion (1:1 match), an even higher accuracy would be beneficial on a billion people scale. To this end, the biometric algorithm is continuously being developed and will be upgraded over time.

There are three key types of upgrades that can increase accuracy further:

Image preprocessing upgrades.

These upgrades, which are *backwards compatible*, modify everything except the final step in the process: iris code feature generation. Elements such as the segmentation network and image quality thresholds are typical areas of improvement. For an in-depth look at the preprocessing algorithms, please refer to the image processing section. These types of upgrades generally occur multiple times a year.

Iris code generation upgrades for future verifications. Also *backwards compatible*, these upgrades involve modifying the iris code feature generation algorithm without recomputing previous iris codes. Such an upgrade involves the introduction of v2 codes, which are not compatible with the older v1 codes. Both v1 and v2 codes would be compared against their respective sets. If both comparisons result in no collision, the v2 code is added to the set of v2 codes. This way, the set of v1 codes doesn't grow any more, yet none of the individuals who are part of the v1 set can get a second World ID.

In the event of an upgrade to the feature generation algorithm, the corresponding false match rate evolve as follows:

$$\begin{aligned} P_{FM}(i) &= 1 - P_{\text{no match at all}} \\ &= 1 - P_{\text{no match in v1}} \cdot P_{\text{no match in v2}} \\ &= 1 - (1 - FMR_1)^{n_1} \cdot (1 - FMR_2)^{i-1-n1} \end{aligned}$$

For an in-depth understanding of the error rates, please revisit the relevant information in the section biometric performance on a billion people scale. From the above equation we can deduce that the likelihood of the i-th legitimate user experiencing a false match continues to increase with the expansion of the v2 set. However, provided that $FMR_2 < FMR_1$, the rate of this growth is significantly reduced. For any individual included in the v1 set, the false non-match rate remains unaffected. For new enrollments, the false non-match rate of the v2 algorithm applies.

In principle, several such iris code versions can be stacked. This type of upgrade is expected to happen about once a year.

Recomputing existing iris codes. These upgrades *may or may not be backwards compatible*, depending on whether the original image is still available. These are expected to occur less frequently than iris code generation upgrades and to become less frequent over time.

To understand when a recomputation might be required, let us define the number of codes in the set of v1 codes as n_1 and similarly for v2 codes. If the error rates of the v1 code are much worse than the ones of v2 codes and therefore have major influence on the false match rate even at $n_2 \gg n_1$, the set of v1 codes should eventually be recomputed. For this to be possible, the iris images need to be available. This can happen in several ways:

Re-capture images. Individuals could return to an Orb. Depending on the distance to an Orb and individual preferences this may or may not be a realistic option.

Custodial image storage. Upon request, the issuer can securely store the images and automatically recompute the iris code if necessary. Currently, this is an option for individuals, but it is likely to be discontinued with the introduction of self custodial image storage.

Self custodial image storage. Expected to be introduced in late 2023, this option allows individuals to store their signed and end-to-end encrypted images on their device. For recomputing iris codes, individuals can upload their images temporarily to a dedicated, audited cloud environment that deletes images upon recomputation, or perform the computation locally on their phones. To ensure integrity, the local computation requires the upgrade to happen within a zero-knowledge proof, necessitating the use of Zero-Knowledge Machine Learning (ZKML) on the individual's phone. The feasibility of this approach depends on the computational capabilities of the individual's phone and ongoing ZKML research.

If local computation or temporary upload isn't viable or preferable, individuals can always revisit an Orb where the iris code is computed locally.

Biometric Uniqueness Service

While the iris code is computed locally on the Orb, the biometric uniqueness service i.e. the determination of uniqueness based on the iris code is performed on a server since the iris code needs to be compared against all other iris codes of humans who have verified before. This process is getting increasingly computationally intensive over time. Today, the biometric uniqueness service is run by Tools for Humanity. However, this should not be the case forever and there are several ideas regarding the decentralization of this service.

Worldcoin Protocol

Worldcoin is a blockchain-based protocol that consists of both off-chain and on-chain components (smart contracts) and is based on Semaphore from the Ethereum PSE group. The Protocol supports the Worldcoin mission by distinguishing humans from non-human actors online, privately but uniquely identifying individuals to solve certain classes of problems related to abuse, fraud, and spam.

Current Status

The Protocol originally deployed on Polygon during its beta phase, and the current version runs on Ethereum with a highly scalable batching architecture. Bridges are in place for Optimism and Polygon PoS state changes on Ethereum, with each batch insertion being replicated to those chains. As of this writing, over two million users have been successfully enrolled with a combination of these deployments, representing an average load of almost five enrollments per minute.

Technical Implementation

While the Orb adheres to data minimization principles such that no raw biometric data (e.g. iris images) needs to leave the device, it calculates and transmits iris codes that are stored and processed separately from the user's profile data or the user's wallet address. The first version of the Protocol originated as a solution to this fundamental privacy challenge specifically for the WLD airdrop. At its core, the Protocol combines the Orb-based uniqueness verification with anonymous set-membership proofs, thus allowing the issuer to determine whether the user has claimed their WLD tokens without collecting any further information about them. Realizing this solves a hard problem others are also facing, World ID was created in order to allow third parties to use the Orb-verified "unique human set" in the same privacy-preserving way.

Users start enrollment by creating a Semaphore keypair on their smartphone, hereafter referred to as the World ID keypair. The Orb associates the public key with a user's iris code, whose current sole purpose is to be used in the uniqueness check. If this check succeeds, the World ID public key gets inserted into an identity set maintained by a smart

contract on the Ethereum blockchain. The updated state is subsequently bridged to Optimism and Polygon PoS so World ID can be used natively on those chains. Integration with other EVM-based chains is straightforward, and integration with non-EVM chains is possible as long as the bridged chain has a gas-efficient means of verifying Groth16 proofs. After enrollment the user can prove their inclusion in this identity set, and therefore their unique personhood, to third parties in a trustless and private way. Since the scheme is private, it's usually necessary to tie this proof to a particular action (e.g. claiming WLD or voting on a proposal).

In the above scheme, the wallet creates a Groth16 proof that proves a user knows the private key to one of the public keys in the on-chain identity set and the action. An optional signal, like the preferred option in a vote, can also be included. By design, this provides strong anonymity of the size of the whole set. It is not possible to learn the public key or anything relating to the enrollment, including the iris code, other than that it was successfully completed, so long as the private key does not leak. It is also not possible to learn that two proofs came from the same person if the scheme is used for different applications.

In the context of the Orb verification, the Orb is the only trusted component in the system; after enrollment, World ID can be used in a permissionless way.

Overall Architecture and User Flow

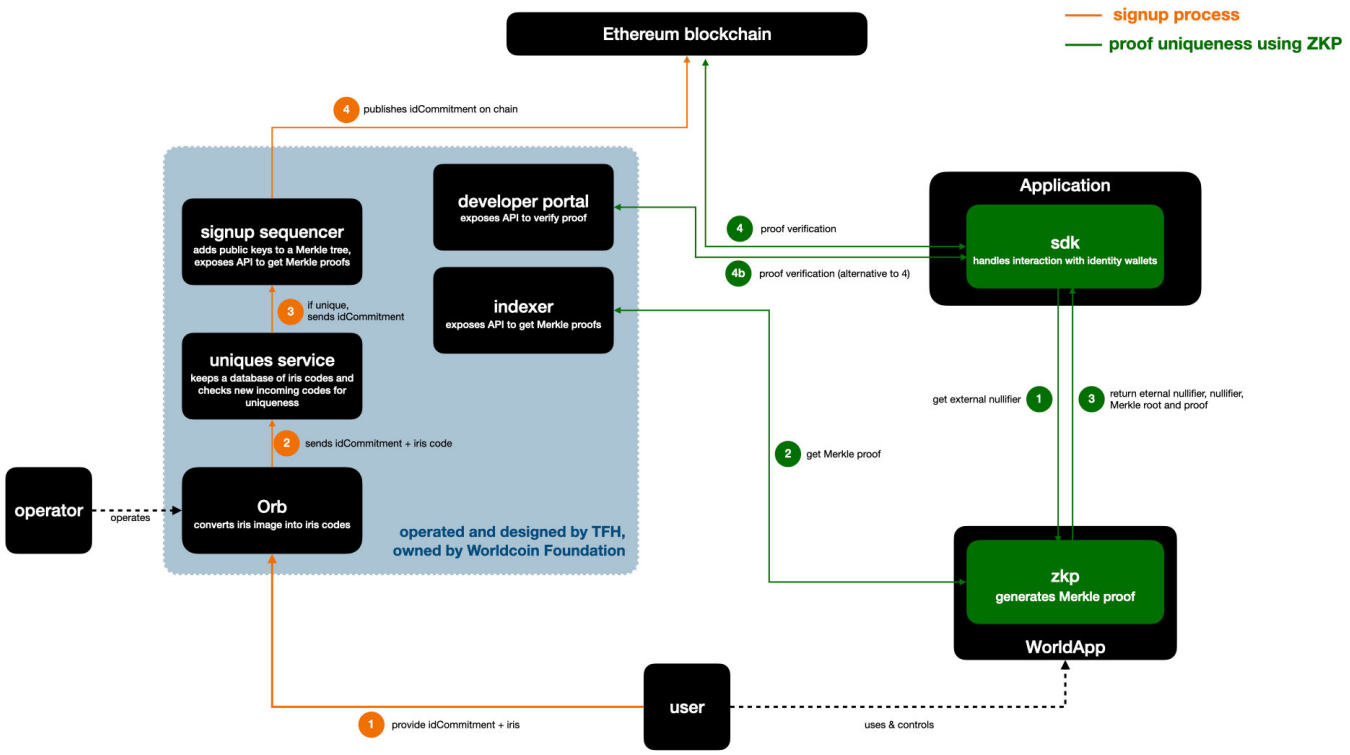

Fig. 30

Enrollment and verification of World ID

As mentioned, at the heart of the Protocol is the Semaphore anonymous set-membership protocol — an open-source project originally developed by a team from the Ethereum Foundation and extended by Worldcoin. Semaphore is unique in that it takes the basic cryptographic design for privacy as found in anonymous voting and currencies and offers it as a standalone library. Semaphore stands out in its simplicity: It uses a minimalistic implementation while providing maximum freedom for implementers to design their protocols on top. Semaphore's straightforward design also allows it to make the adaptations required to support multiple chains and enroll a billion people efficiently. Worldcoin's version of Semaphore is deployed as a smart contract on Ethereum, with a single set containing one public key (called an identity commitment) for each enrolled user. A commitment to this set is replicated to other chains using state bridges so that corresponding verifier contracts can be deployed there.

Users interact with the Protocol through an identity wallet containing a Semaphore key pair specific to World ID. Semaphore does not use an ordinary elliptic curve key pair, but leverages a digital signature scheme using a ZKP primitive. The private key is a series of

random bytes, and the public key is a hash of those bytes. The signature is a ZKP that the private key hashes to the public key. Specifically, the hash function is Poseidon over the BN254 scalar field. The public key is not used outside of the initial enrollment (for interactions with smart contracts, the wallet also contains a standard Ethereum key pair). The user can initiate World ID verifications directly from the app, by scanning a QR code or tapping on a deep link. Upon confirmation, a ZKP is computed on the device and sent through the World ID SDK directly to the requesting party (e.g. a third-party decentralized application, or dApp).

Developers can integrate World ID on-chain using the central verifier contract. As part of any other business logic, the developer can call the verifier to validate a user-provided proof. The developer at a minimum provides an application ID and action (which are used to form the *external nullifier*). The external nullifier is used to determine the scope of the Sybil resistance, i.e. that a person is unique for each context. Within the zero-knowledge circuit that a user computes to generate a proof with their World ID, the external nullifier is hashed in conjunction with the user's private key to generate a nullifier hash. The same person may register in multiple contexts but will always produce the same nullifier hash for a specific context. The developer may also provide an optional message (called a signal) which the user will commit to within the ZKP.

If the proof is valid, the developer knows that whoever initiated the transaction is a verified human being. The developer can then enforce uniqueness on the nullifier hash to guarantee sybil resistance.

For example, to implement quadratic voting, one would use a unique identifier for the governance proposal as context and the user's preferred choice as message. In case of airdrops (just as for WLD), the associated message would be the user's Ethereum wallet address.

Alternatively, World ID can be used off-chain. On the wallet side, everything remains the same. The difference is that the proof-verification happens on a third-party server. The third-party server still needs to check whether the given set commitment (i.e. Merkle root) corresponds to the on-chain set. This is done using a JSON-RPC request to an

Ethereum provider or by relying on an indexing service. All of that is abstracted away by the World ID SDK and additional tooling in order to provide a better developer experience.

Enrollment Process

This section outlines how the enrollment process works for generating a World ID and verifying at an Orb.

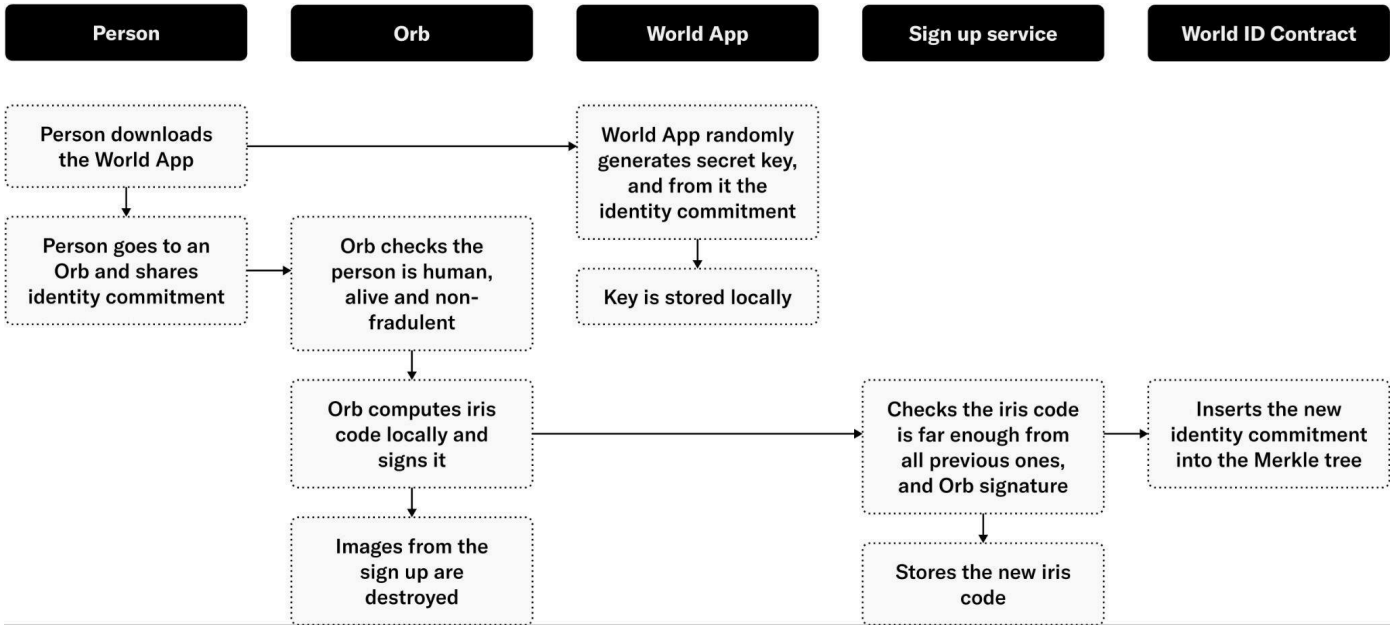

Fig. 31

Enrollment process for Orb signal using World App as the identity wallet

The Semaphore protocol provides World ID with anonymity, but by itself it does not satisfy Worldcoin's scaling requirements. A regular insertion takes about a million gas (a unit of transaction cost in Ethereum). Gas prices fluctuate heavily on Ethereum, but this transaction could easily cost over $100 in today's fee market, making it prohibitively expensive to sign up billions of people.

One could use cheaper alternatives to Ethereum, but that comes at the cost of security and adoption; Ethereum has the largest app ecosystem and Worldcoin aims for World ID to be maximally useful. For that, it is best to start from Ethereum and build out from there. However, from a cost perspective, there are limits to scaling atop Ethereum, as the large insertion operation still happens on-chain. The most viable options are optimistic

rollups, but these require considerable L1 calldata. Therefore, Worldcoin scaled Semaphore using a zk-rollup style approach that uses one-third the amount of L1 calldata of optimistic rollups. The enrollment proceeds as follows (see above diagram):

1. The user downloads the World App, which, on first start, generates a World ID keypair. In World App, private keys are optionally backed up (details on this coming soon). Additionally, an Ethereum keypair is also generated.
2. To verify their account, the user generates a QR code on the World App and presents it to the Orb. This air-gapped approach ensures the Orb isn't exposed to any sort of device or network-related information associated with the user's device.
3. The Orb verifies that it sees a human, runs local fraud prevention checks, and takes pictures of both irises. The iris images are converted on the Orb hardware into the iris code. Raw biometric data does not leave the device (unless explicitly approved by the user for training purposes).
4. A message containing the user's identity commitment and iris code is signed with the Orb's secure element and then sent to the signup service, which queues the message for the uniqueness-check service.
5. The uniqueness-check service verifies the message is signed by a trusted Orb and makes sure the iris code is sufficiently distinct from all those seen before using the Hamming distance as distance metric.
6. If the iris code is sufficiently distant (based on the Hamming distance calculation), the uniqueness service stores a copy of the iris code to verify uniqueness of future enrollments and then forwards the user's identity commitment to the signup sequencer.
7. The signup sequencer takes the user's identity commitment and inserts it into a work queue for later processing by the batcher.
8. A batcher monitors the work queue. When 1) a sufficiently large number of commitments are queued or 2) the oldest commitment has been queued for too long, the batcher will take a batch of keys from the queue to process.
9. The batcher computes the effect of inserting all the keys in the batch to the identity set, the on-chain Semaphore Merkle tree. This results in a sequence of Merkle tree update proofs (essentially a before-and-after inclusion proof). The prover computes a

Groth16 proof with initial root, final root, and insertion start index as public inputs. The private inputs are a hash of public keys and the insertion proofs.

10. For optimization purposes, the above-mentioned “public inputs” are actually keccak-hashed as a single public input and non-hashed as private inputs, which reduces the on-chain verification cost significantly. The circuit verifies that the initial tree leaves are empty and correctly updated. Computing the proof for a batch size of 1,000 takes around 5 minutes on a single AWS EC2 hpc6a instance.
11. The batcher creates a transaction containing the proof, public input, and all the inserted public keys and submits it to a transaction relayer. Relayer assigns appropriate fees, signs the transaction, and submits it to (one or more) blockchain nodes. It also commits it to persistent storage so mispriced/lost transactions will be re-priced and re-submitted.
12. The transaction is processed by the World ID contract, which verifies it came from the sequencer. The initial root must match the current one, and the contract hashes the provided public keys. Public keys are available as transaction calldata.
13. The Groth16-verifier contract checks the integrity of the ZKP. An operation takes only about 350k gas for a batch of 100.
14. The old root is deprecated (but still valid for some grace period), and the new root is set to the contract.

The ZKP guarantees the integrity of the Merkle tree and data availability. What sets it apart from the ideal zk-rollup model is the lack of validator decentralization; the implementation uses a single fixed-batch submitter.

After enrollment is complete, the user can use the World App autonomously. At that point, the system works in a decentralized, trustless and anonymous manner.

Details regarding trust assumptions and limitations for World ID can be found in the limitations section.

Verification Process

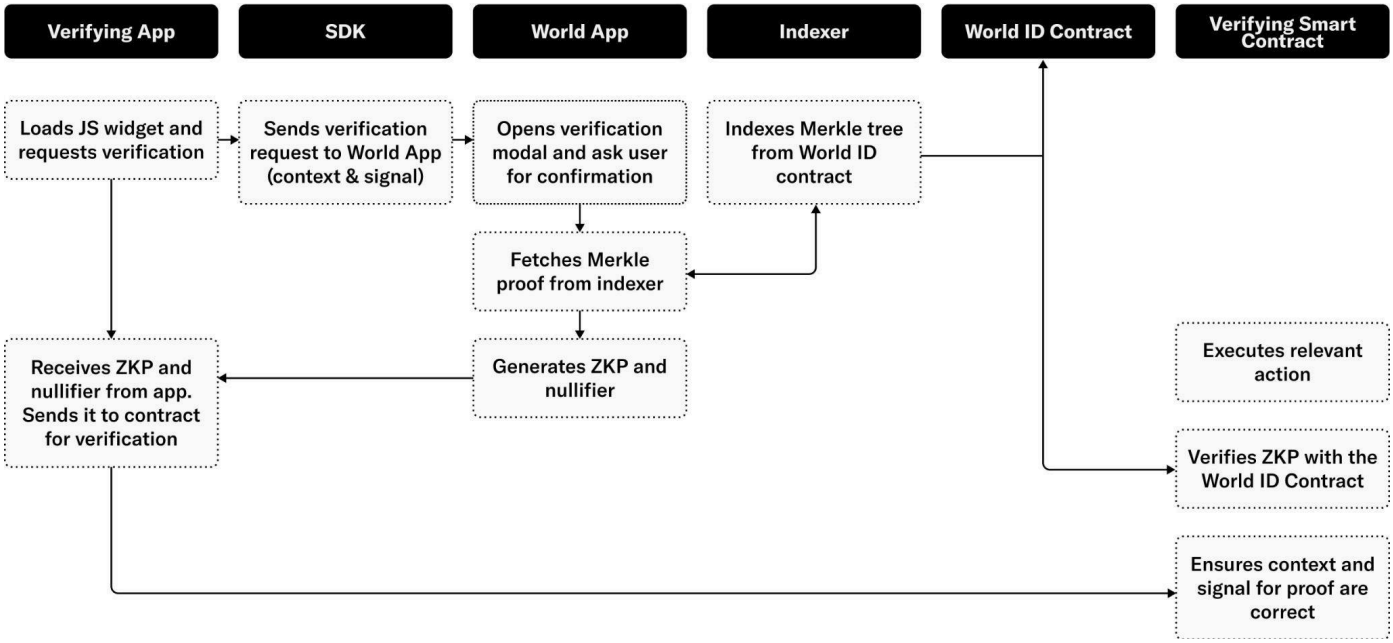

Fig. 32

Verification process for Orb signal

Developers can integrate World ID as part of a transaction (Web3) or request (Web2) through the World ID SDK:

1. A verification process is triggered through one of the in-app options or through a QR code presented by a third-party application. Scanning the QR code opens the application. A verification request contains a context, message, and target. The context uniquely identifies the scope of the Sybil protection (e.g. the third-party application, a vote on a particular proposal, etc.). The message encodes application-specific business logic related to the transaction. The target identifies the receiving party of the claim (i.e. callback).
2. The user inspects the verification details and decides to proceed using World ID. It is important that the user knows that the context is the intended one to avoid man-in-the-middle attacks.
3. To generate a ZKP, the application needs a recent Merkle inclusion proof from the contract. It is possible to do this in a decentralized manner by fetching the tree from the contract, but at the scale of a billion users this requires downloading several

gigabytes of data — prohibitive for mobile applications. To solve this, an indexing service that retrieves a recent Merkle inclusion proof on behalf of the application was developed. To use the service, the application provides its public key, and the indexer replies with an inclusion proof. Since this allows the indexer to associate the requester's IP address to their public key, this constitutes a minor breach of privacy. One possible means to mitigate this is by using the services through an anonymization network[1]. The indexing service today is part of the sign up sequencer infrastructure and is open source, and anyone can run their own instance in addition to the one provided.

4. The application can now compute a ZKP using a current Merkle root, the context, and the message as public inputs; the nullifier hash as public output; and the private key and Merkle inclusion proof as private inputs[2]. Note that no identifying information is part of the public inputs. The proof has three guarantees: 1) the private key belongs to the public key, hence proving ownership of the key, 2) the inclusion proof correctly shows that the public key is a member of the Merkle tree identified by the root, and 3) the nullifier is correctly computed from the context and the private key. The proof is then sent to the verifier.
5. The verifier dApp will receive the proof and relay it to its own smart contract or backend for verification. When the verification happens from a backend in the case of Web2, the backend usually contacts a chain-relayer service as the proof inputs need to be verified with on-chain data.
6. The verifier contract makes sure the context is the correct one for the action. Failing to do so leads to replay attacks where a proof can be reused in different contexts. The verifier will then contact the World ID contract to make sure the Merkle root and ZKP are correct. The root is valid if it is the current root or recently was the current root. It is important to allow for slightly stale roots so the tree can be updated without invalidating transactions currently in flight. In a pure append-only set, the roots could in principle remain valid indefinitely, but this is disallowed for two reasons: First, as the tree grows, the anonymity set grows as well. By forcing everyone to use similar recent roots, anonymity is maximized. Second, in the future one might implement key recovery, rotation, and revocation, which would invalidate the append-only assumption.

7. At this point, the verifier is assured that a valid user is intending to do this particular action. What remains is to check the user has not done this action before. To do this, the nullifier from the proof is compared to the ones seen before[3]. This comparison happens on the developer side. If the nullifier is new, the check passes and the nullifier is added to the set of already seen ones.
8. The verifier can now carry out the action using the message as input. They can do so with the confidence that the initiation was by a confirmed human being who has not previously performed an action within this context.

As the above process shows, there is a decent amount of complexity. Some of this complexity is handled by the wallet and Worldcoin-provided services and contracts, but a big portion will be handled by third-party developed verifiers. To make integrating World ID as straightforward and safe as possible, an easy-to-use SDK containing example projects and reusable GUI components was developed, in addition to lower-level libraries.

Conceptually, the hardest part for new developers is the nullifiers. This is a standard solution to create anonymity, but it is little known outside of cryptography. Nullifiers provide proof that a user has not done an action before. To accomplish this, the application keeps track of nullifiers seen before and rejects duplicates. Duplicates indicate a user attempted to do the same action twice. Nullifiers are implemented as a cryptographic pseudo-random function (i.e. hash) of the private key and the context. Nullifiers can be thought of as context-randomized identities, where each user gets a fresh new identity for each context. Since actions can only be done once, no correlations exist between these identities, preserving anonymity. One could imagine designs where duplicates aren't rejected but handled in another way, for example limiting to three tries, or once per epoch. But, because such designs correlate a user's actions, they are recommended against. The same result can instead be accomplished using distinct contexts (i.e. provide three contexts, or one for each epoch).

For example, suppose the goal is that all humans should be able to claim a token each month. To do this, a verifier contract is deployed that can also send tokens. As context, a combination of the verifier-contract address and the current time rounded to months are

used. This way each user can create a new claim each month. As the message, the address where the user wants to receive the token claim is used. To make this scalable, it is deployed on an Ethereum L2 and uses the World ID state bridge.

Multi-chain Support

While it's important that World ID has its security firmly grounded, it is intended to be usable in many places. To make World ID multi-chain, the separation between enrollment and verification is leveraged. Enrollment will happen on Ethereum (thus guaranteeing security of the system), but verification can happen anywhere. Verification is a read-only process from the perspective of the World ID contract, so a basic state-replication mechanism will work.

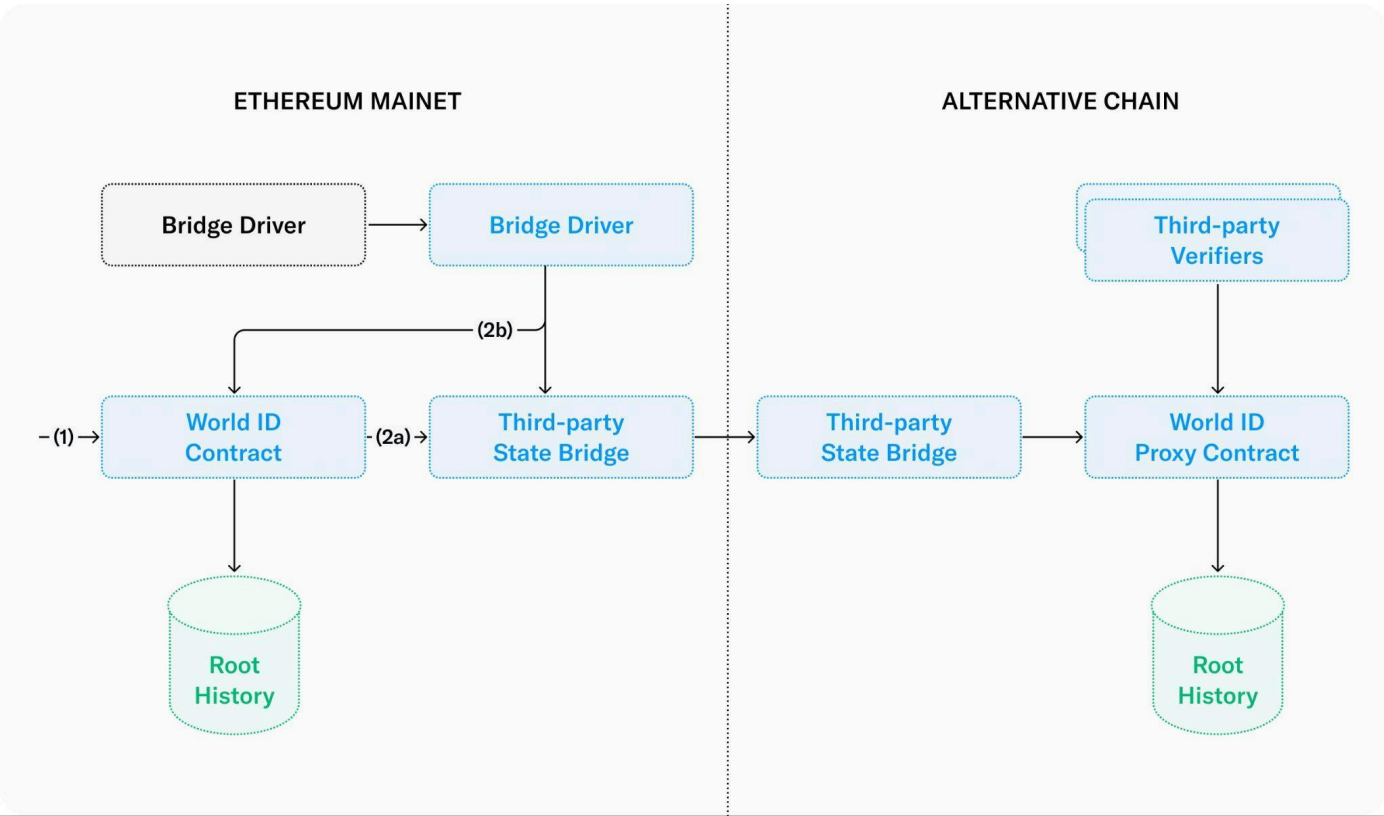

Fig. 33

Flow of data for multi-chain credential

1. Enrollment happens as before on Ethereum, but now each time the root history is updated a replication process is triggered.

2. The replication is initiated by the World ID contract itself (route 2a) or by an external service that triggers a contract to read the latest roots from the contract (route 2b). Either way, the latest roots are pushed as messages to a third-party state bridge for the target chain.
3. The Ethereum-side bridge contract forwards messages from Ethereum to the target chain. The details are implementation specific, but generally the direction from Ethereum to an L2 is easiest and fastest.
4. The target-side bridge contract calls the World ID proxy contract with the new roots. After authenticating the message, the replica of the root history is updated. Now the proxy can be used for verification as if it were the main instance.

For the first bridge, a direct integration (2a) is used as this is the easiest and most reliable integration to implement. But direct routes require extension of the World ID contract, which are preferred to be kept to a minimum. So, for future bridges the externally driven route will be opted for. Externally driven integrations have the advantage of operating independently and can be added without modifying the World ID contract. In fact, anyone can build such a bridge.

For a target chain to support World ID, the most important requirement is Groth16-verification support. Groth16 is a widely supported proof system, but native support on some chains can be minimal. Secondary to this, World ID requires the existence of a reliable one-way message-passing bridge and sufficiently rich programmability with global persistent storage for the root history and nullifiers. For non-EVM target chains, there is extra work in porting the proxy and verifier contracts.

Data Handling

Blockchains play a primary role in the World ID Protocol, providing a trustless and decentralized source of truth (i.e. the list of valid credentials[4] lives on-chain) and allow functionality such as revocation. Yet not all data is suited to live on-chain, which is why other-data handling mechanisms are introduced that are credential-specific and decided by issuers.

The diagram below shows the example of the Orb credential and how data is handled on-chain and off-chain.

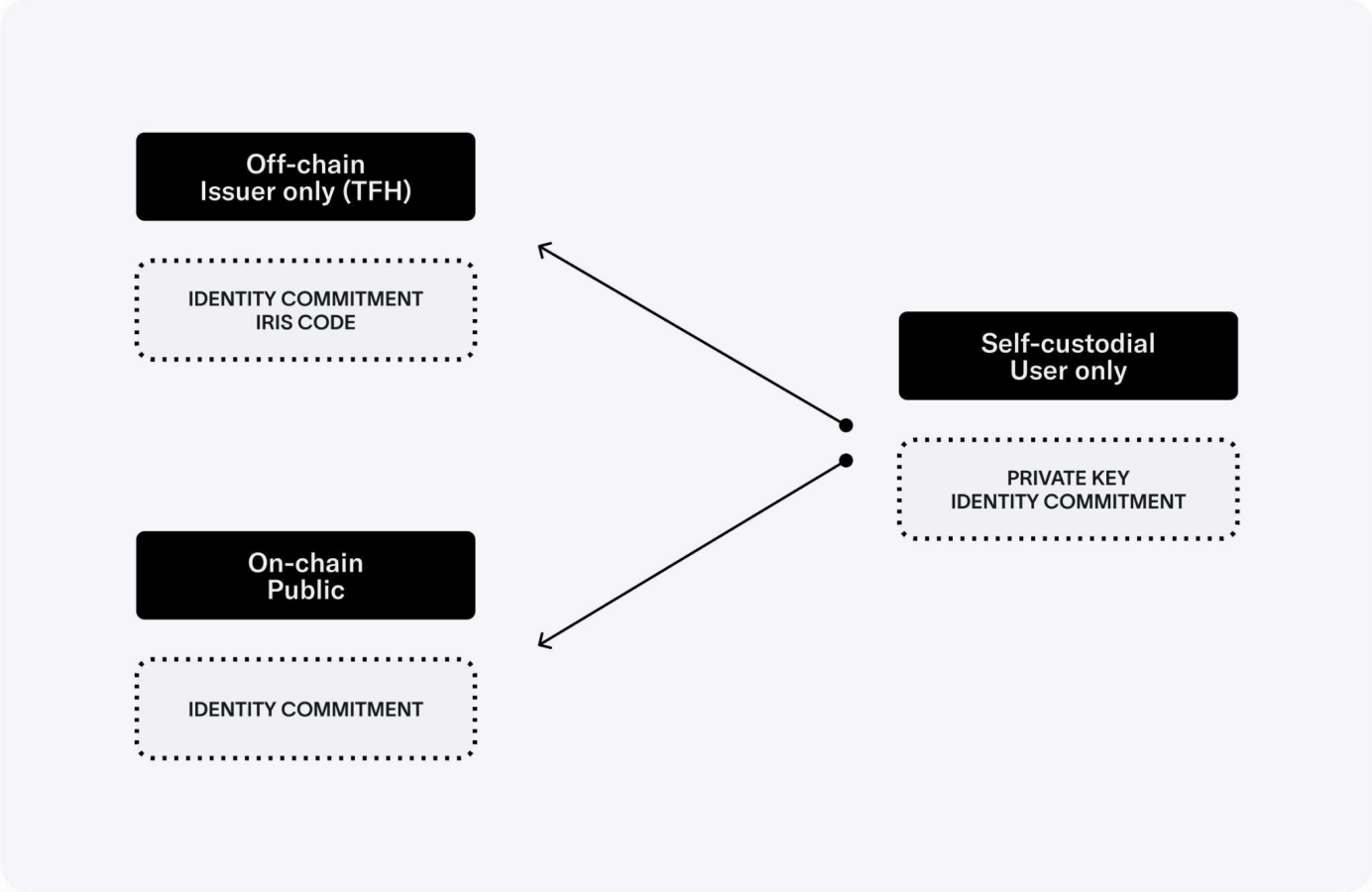

Fig. 34

Data handling for the Orb credential. The identity commitment can be seen as the unique identifier for the credential, but is not shared when using the credential. Instead, a ZKP is used to prove the user holds the private key to an identity commitment in the on-chain set.

Versioning

The Orb credential must be subject to a versioning system, due to the evolution of algorithms used to transform iris images into iris codes. As training data is continually processed and research is pursued for more precise and quicker comparison algorithms, maintaining different versions becomes imperative.

What this means for users is that their Orb credential will have a “time-to-live,” or TTL, and at some point their signal may become deprecated and no longer recommended for use by third-party applications. When the credential expires, the user will be able to go to

an Orb again to refresh their credential. One promising approach to allowing users to refresh their credential without going to an Orb or compromising their privacy is through zero-knowledge machine learning (ZKML):[5]

- When a user first enrolls at an Orb, the Orb will send their high-resolution signed iris image to the user's device in an end-to-end encrypted, self-custodial manner. Users will be able to delete their images at any point in time.
- When the algorithm changes, the user's wallet will get a prompt to update the iris code and download the relevant ML parameters.
- The user's device will run the new ML model to generate the new iris code and a ZKP that asserts the model was properly run and the iris image is authentic.
- With the outputs above, the uniqueness and signup sequencer can update the user's Orb credential seamlessly and privately.

Security Assessments

Two separate security assessments were conducted on the off-chain and on-chain components of the Protocol specifically related to its use of blockchain technologies, cryptography and smart contracts.

Future Development

World ID has and will continue to be developed iteratively. Development started by focusing on a single opinionated mechanism for proof of personhood, with particular attention to inclusivity and accuracy, hence the Orb. This section briefly introduces the different workstreams for future developments of the Protocol.

Recovery

Status: Active Development | Proof of Concept | Very High Priority

As previously mentioned, recovery is key for any proof-of-personhood protocol, and World ID is no exception. A user must always be able to maintain access and even get back their World ID in the case of theft, loss, etc. Recovery is initially being introduced to the Protocol by incorporating credential re-issuance, i.e. when a user loses their World ID, they get their credentials re-issued and the old ones revoked. This section outlines how this happens for the Orb proof-of-personhood mechanism.

There is ongoing research to understand whether a more abstract recovery mechanism can and should be introduced at the Protocol level. One important consideration with such mechanisms is security. Having the possibility of recovering "everything" with a single mechanism can introduce vulnerabilities that can be exploited. To use a real world analogy, when one's wallet is stolen, they don't perform a single action that recovers their driver's license, credit cards, and ID all at once.

An overview of the current roadmap can be found in the World ID: Implementing PoP at Scale section.

Plurality

Status: Active Development | v0 Beta Testing | High Priority

Worldcoin started with World ID to be able to bootstrap the Protocol. However, there is a tradeoff between accuracy of the biometric-based Orb verification and its availability to everyone on the globe. The Orb is not yet available in every country, and as operations continue to scale, other proof of personhood mechanisms may be viable alternatives, for low stakes applications.

Proof-of-personhood representation in the digital world can be viewed as a spectrum, rather than binary, as there are multiple ways to evaluate personhood — with varying degrees of accuracy.

The benefits of introducing different proof-of-personhood credentials to the Protocol are that it allows for:

- Wider Protocol usage while the Orb's availability is scaled
- More issuers[6], introducing further decentralization and resilience to the Protocol[7]

The drawbacks are:

- Deduplication across credentials is hardly possible, which can introduce the possibility of non-scalable Sybil attacks in some applications. For example, a World ID holder cannot be deduplicated from a unique phone number verification.
- If a high-accuracy credential reaches widespread adoption, the use of other credentials is likely to be less useful.

A beta test is currently underway with a unique phone number verification credential. This is at the low end of the spectrum in terms of accuracy, but it's also something that is widely available across the world. While this is not a very reliable proof-of-personhood signal for something that requires a high level of assurance that someone is a unique person (e.g. universal basic income), it may be enough for low stakes applications.

Eventually, other parties (i.e. issuers) should be able to issue proof-of-personhood attestations (i.e. credentials). The verifier can then determine which attestations they accept, depending on the level of assurance their use case demands.

Interoperability

Status: Active Research | Proof of Concept | Medium Priority

Current internet applications are built on top of communication standards that have been progressively agreed on as a society. Similarly, widespread standards will be necessary for proof of personhood. These standards will extend the system's interoperability and usability in a variety of contexts.

World ID is expected to integrate with widely used industry standards, current and future. This is a continuous effort, not a single end state. Already today, the Protocol is extending

interoperability beyond its original inception. The first version was a single one-chain, one-credential system on the Polygon network. Today, it's available on three chains: Ethereum, Optimism, and Polygon. It can also already be used in non-Web3-related contexts. The Protocol can be used with simple REST APIs, and even beyond that, it already integrates with widely used identity protocols like OpenID Connect (OIDC). In fact, a full-support integration with Auth0, a leading player in the identity space was launched.

Interoperability is not only being researched at the Protocol level but also at the SDK level. The World ID SDK can be conceptually split into two components: the wallet side and the application side. The application side already offers support for web and mobile applications, with further support being planned for more specific technologies, languages, and frameworks. The wallet side, which will offer portability of World ID and decentralization on the user side is currently being researched. Some of the challenges being researched to offer wallet portability are:

- Seamless but secure portability of secrets and metadata
- Trustworthy authentication, solving for the trust point of the user's hardware
- Standardized risk management mechanisms

Privacy

Privacy is the bedrock on which Worldcoin is built, and contributors to the project are committed to raising the bar far beyond today's best practices and ensuring that privacy is accessible to everyone. On a high level, custom hardware (like the Orb) enables the most privacy-preserving solution for proof of personhood (such as World ID). Getting privacy right, however, requires deliberate effort and additional work - and the results must be demonstrable if they're to be trusted. This section explains in advanced technical detail how privacy is preserved in the different parts of the Worldcoin ecosystem.

- A user-friendly introduction to privacy can be found in the Privacy page.

- An intermediate high-level overview on privacy for the more curious readers can be found in the Solving for Privacy blog post.

Most of the Worldcoin protocol's critical systems are designed in such a way that privacy cannot be compromised, even by any of the protocol's contributors. This is achievable using cryptographically provable mechanisms such as Zero-Knowledge Proofs (ZKPs). Worldcoin uses ZKPs to make it mathematically impossible to link usage of World ID across applications. Privacy protections such as these go beyond regulatory requirements.

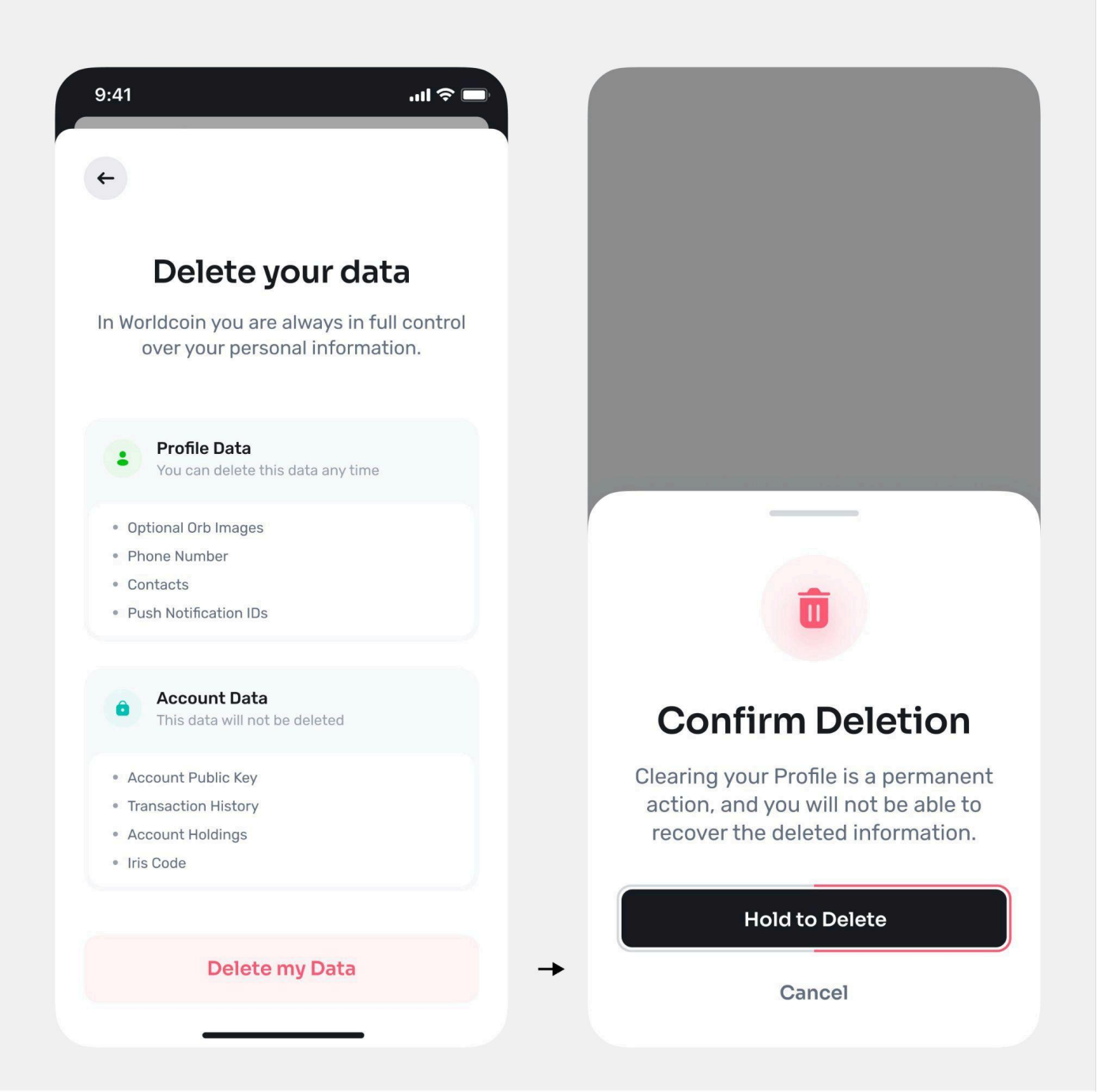

Fig. 36

Privacy Screen on World App. A user can very easily request deletion of all their personal data with just a few taps in the World App.

Anyone can use the World App and their World ID fully pseudonymously.

Users don't have to provide personal information to register. No emails, no phone numbers, no social profiles, no names, everything is optional.

ZKPs are used to preserve the user's privacy and avoid cross-application tracking.

Whenever a user makes use of their World ID, ZKPs are used to prove they are a unique human. This means that no third-party will ever know a user's World ID or wallet public key, and in particular cannot track users across applications. It also guarantees that using World ID is not tied to any biometrical data or iris codes. When one wants to prove they are a unique human, they should be able to do so without revealing any personal information about themselves.

Personal Custody

Personal data custody, or Personal Custody, means that the information (images, metadata and derived data including the iris code) is held on a user’s device. This approach gives users control over the flow of this data—not just deletion, but any future use prior to being deleted. Previously,the images were deleted by default.

In addition to giving users control, Personal Custody unlocks new World ID use cases by enabling Face Authentication for high security applications. With Face Authentication, users can verify at any time that they are the same person that received their World ID when verifying at an orb. Importantly, this Face Authentication functionality works locally on the user's device, without their face data leaving their phone.

For Worldcoin, giving users control over their data flow with Personal Custody is a significant step towards solidifying the project’s user-centric architecture and building an even more robust and secure World ID network.

At a high level, Personal Custody involves four components: user’s device, the orb, a data package containing the user’s images and the Orb backend for transit.

Importantly, the backend cannot decrypt a user's data package.

Here's how the Personal Custody process works:

1. A user's phone generates a public-private key pair to encrypt their data, then transfers the public key to the backend.
2. The backend generates additional keys for all data that requires double encryption and passes the public keys to the orb.
3. During verification, the orb creates the necessary images to verify a user's World ID.
4. The orb then creates the user's individual data packages that includes the images and derivatives like the iris code created from these images, encrypts them, "signs" them to ensure authenticity and security, and sends them through the Orb backend to the user's device.
5. Once the user's encrypted data packages are downloaded to the user's phone they are deleted from the Orb and Orb backend.

Since the data package is encrypted by the user's public key, the end result of this process is a collection of encrypted data packages that reside exclusively on the user's device. The use of double encryption within the end-to-end encryption envelope is a safeguard to protect the confidentiality and privacy of a user's data in the event the user's phone is compromised.

Note! The process described above relates to Personal Custody, not the entire Worldcoin system. The iris code is not deleted from the Worldcoin backend. Rather, the iris code is persistently encrypted and permanently stored to ensure a permanent proof of uniqueness. The iris code will not be deleted from the Worldcoin backend, even if a user requests deletion.

To summarize Personal Data Custody:

- Users are in control of their data flow.
- All images and image derivatives are packaged, encrypted, and "signed" by the Orb to ensure authenticity and security, then sent to the user's phone through the Orb backend server (importantly the Orb-backend cannot decrypt the data).

- The data package is then deleted from the Orb and Orb-backend.
- No data collected, including images taken by the Orb has or will ever be sold. Nor will it be used for any other intent than to improve World ID. The Worldcoin Foundation is bound to this commitment through the data consent form where it states: "**We will never sell your data**. We will also not use any data listed in this form to track you or to advertise third parties' products to you," and that "We will not sell, lease, trade, or otherwise profit from your biometric data."

The Iris Code

As discussed, the iris code is a numerical representation of the texture of a person's iris. It holds the property that it can be compared against different images of the same iris to determine whether the images came from the same iris.

The iris code cannot be a simple hash of the texture of the iris. This is because two pictures of the same iris will not be exactly the same. A myriad of factors change (lighting, occlusion, angle, etc.) in image capturing and a tiny change would lead to a different hash. With the iris code, those factors only lead to slightly modified Hamming distance between two codes which permits fuzzy comparison of irises. If the distance is below a certain threshold, the images are assumed to be from the same iris.

The iris code is computed by applying a set of 2D Gabor filters at various points of the iris texture, which leads to complex-valued filter responses. Only the phase information of the filter responses is taken into account (which means there is permanent information loss) and subsequently quantized in two bits. In other words: For each Gabor wavelet and each point of interest in the iris texture two bits are computed. Concatenating all these bits makes up the iris code.

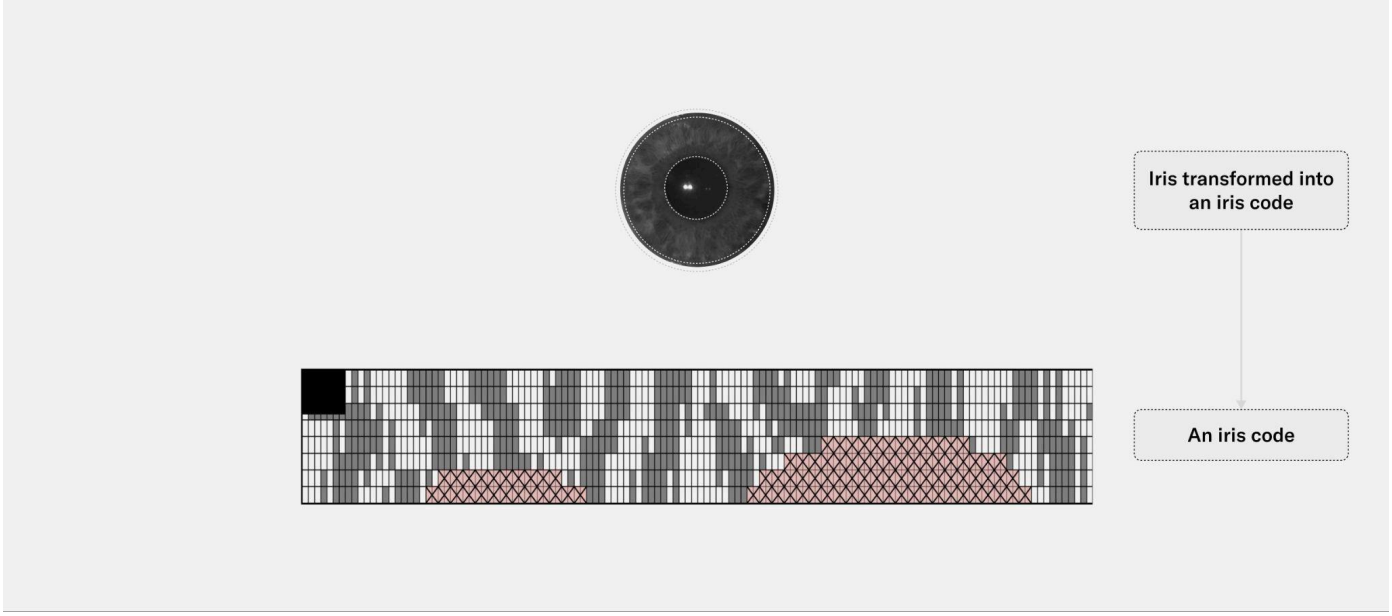

Fig. 3.37: An example iris code. In red, a second array can be seen that represents the mask applied to the image, these are pixels of the image that don't represent part of the iris texture, like eyelids, which are of course ignored when computing the Hamming distance between irises.

To date, there is no known way to reverse engineer an image that exactly matches the appearance of the input image. It is technically possible to generate an image from an iris code that generates the same iris code (if the same parameters for the Gabor wavelets are used, which are different for every system), but the image will look different from the actual image, mainly because of the information loss when generating the iris code.

Two important privacy assumptions ought to be underscored. First, private keys need to remain private, as otherwise, a user can deanonymize themselves, even to actions they have performed in the past. Second, while the Protocol is made to be used in a privacy-preserving manner, privacy cannot be enforced outside of the Protocol.

Wallets

While currently users must first download the World App to participate in the Worldcoin system, the Worldcoin Foundation aims for the development of other applications that support the creation of a World ID wallet. Afterall, the overall system is designed so that other developers can build their own clients without permission, meaning World App will hopefully be just one of the many wallets supporting World ID. Research is currently underway to develop SDKs for other wallets to support World ID.

Footnotes

1. Also evaluated were private information retrieval (PIR) protocols, but even with state-of-the-art protocols like OnionPIR and further optimizations, the services would need 10 seconds of multi-core compute per request. Multi-party computation (MPC)-based PIRs would perform much better, but they offer no anonymization advantage over using an MPC-based anonymization network. ↩

2. Note there's a trust assumption on obtaining the inclusion proof from an indexing service as the user needs to provide their identity commitment to obtain an inclusion proof. Further decentralization of the indexing service is being explored. ↩

3. It is sufficient to check uniqueness on a per-context basis, but the nullifiers should be globally unique values. ↩

4. In this context, *credential* is used as a generic term to refer to a set of data about a subject, and in this case attested by a third-party (called issuer). ↩

5. ZKML would allow the iris code to be recalculated in the event of a model upgrade, without users needing to go back to an Orb. ↩

6. In this, *issuer* is the party who attests to a set of data about a subject. For example, Tools for Humanity is the issuer of the Orb credential. ↩

7. This is in addition to the decentralization introduced by the distribution of Orb manufacturing and operations across different entities. ↩

Advancing Decentralization

Decentralization is a non-binary and nebulous term in that it means different things to different people. In the blockchain community, decentralization often refers to important properties of infrastructure like transparency, verifiability and its ability to recover from local participant failures. Those properties are important for the Worldcoin project to be a public good. More details on the derivation of important properties of the Worldcoin infrastructure can be found in this blog post.

The following sections outline different areas of the Worldcoin project and how different mechanisms can increase their respective transparency, verifiability and ability to recover from local participant failures. The ideas for optimal mechanisms may evolve over time, and suggestions for improvements are welcome.

User Agent

The user agent, i.e. the *wallet*, is what connects the user to the system and executes all user actions. It manages the user's keys for both finance and identity. The finance part is a self-custody crypto wallet and thus permissionless. For the identity part, the user agent combines independent components into a functional system.

World App was launched as the first user agent to support the Worldcoin protocol, enabling people to get their World ID verified at an Orb and, if eligible, receive their share of WLD tokens.

Eventually, when verifying with an Orb, users should be able either to export their accounts into other wallets or to use a third-party wallet. Additionally, a World ID Wallet

Kit could incorporate all the required capabilities so that other wallets could integrate World ID. This gives the user the choice of which user agent to use.

On the frontend, World ID is already available to any developer that wants to use Sybil protection in their application through the IDkit and developer portal. Users are able to use any third-party application through the World App.

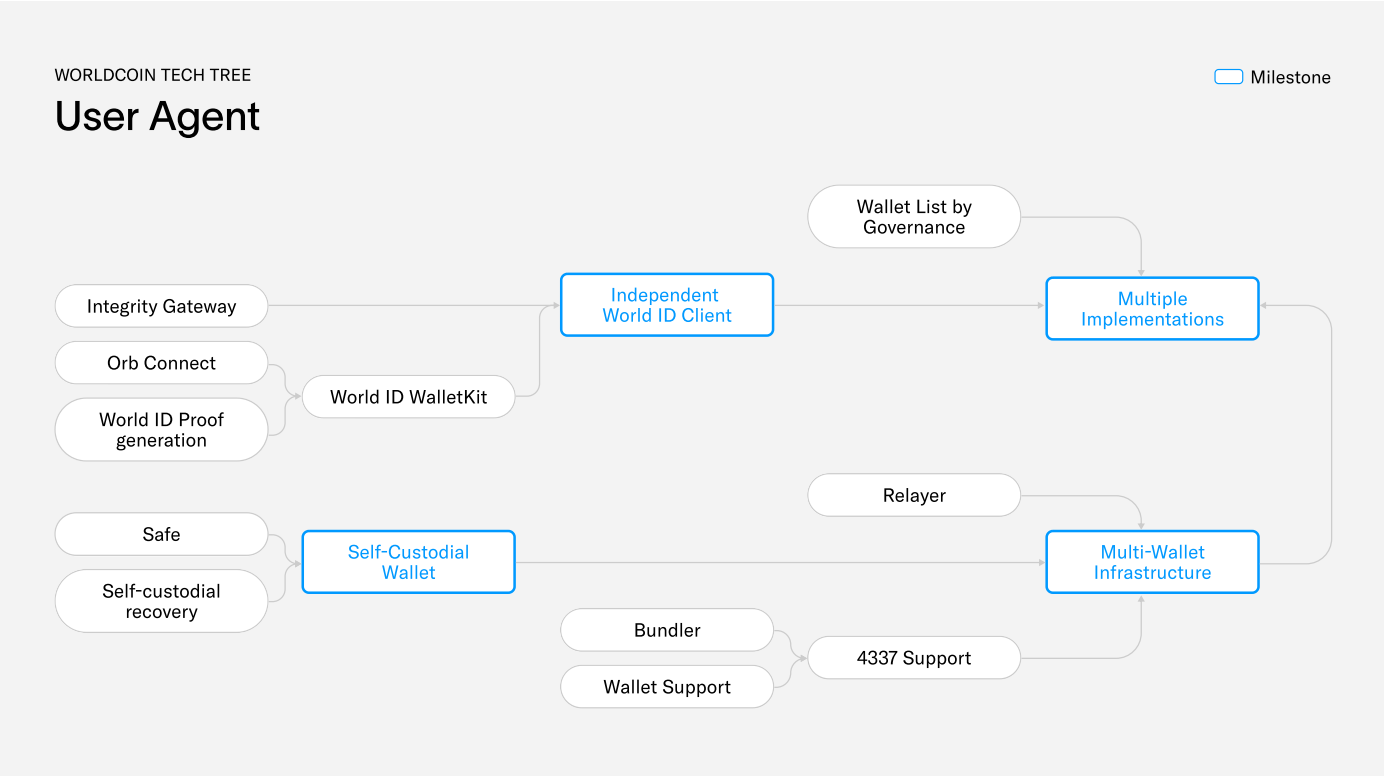

Fig. 1

The graph represents possible ways to increase the robustness and availability of user agents and their respective dependencies. Diversity in user agents also allows catering to the needs of specific user segments.

The following sections walk through the displayed potential further improvements in more detail.

Self-custodial Wallet

Users should be able to access and control their funds and World ID in a self-custodial and censorship-resistant way; wallets should still allow for robust recovery solutions in case users' phones are lost or stolen. Users should also not need to have prior knowledge about blockchains or deal with lower-level blockchain fee-pricing mechanisms. Users should also be always able to switch between different wallet implementations, optimally by keeping their public-facing account address if they wish. Finally, a wallet should be extendable and designed to allow for e.g. technological upgrades. Most of this is already implemented in the World App. While a self custodial recovery solution is implemented via iCloud and GDrive, better solutions are still an active area of research and development in order to enable a more seamless user experience.

Independent World ID Client

There should be multiple clients for users to choose from at the time of verification at an Orb or when using World ID to receive the WLD airdrop, where available. This reduces the risk of any vulnerability affecting all users, while also helping to ensure that wallets are available (e.g. in app stores).

Integrity Gateway

World ID Face Auth requires verifying the authenticity of the client app, because the comparison happens only locally on the user's phone. While local computation could potentially be secured through zero-knowledge proofs and the Orb's image is signed, the second input image has to be taken through the phone's camera. Unless manufacturers begin attaching hardware attestations to those images, it fully relies on trusting the integrity of the phone's hardware and software. However, those attestations already exist on an app level (e.g. Apple App Attest or Google Play Integrity) and can be used as attestations. The verification of those can be handled by "gateways" that sign off on individual requests and provide onchain verifiable signatures. Those gateways would ultimately also need to be provided with a list of accepted apps, managed via governance.

Orb Connect

The Orb currently relies on one-way communication with the app through the QR code and the permissioned Orb backend. To instead share more data with the user, this model could be replaced by an end-to-end encrypted connection between the app and Orb. It could not only facilitate the exchange of public keys but also the encrypted images from the Orb. This enables self-custody of images on the user's phone and allows for future upgrades of the system without trusting an external custodian.

World ID Wallet Kit

The World App already contains all the logic for handling an Orb verification and using World ID to generate and submit proofs (such as when receiving WLD grants). This process can be made simpler and quicker for new teams building their own wallets. Wallet Kit should handle Orb Connect and establish the privileged execution environment on the phone through the integrity gateway. Importantly, it should also contain the mobile-optimized proof generation library.

Multi-Wallet Infrastructure

Users should be able to use World App in a fully self-custodial and censorship-resistant manner. They should also be able to switch between wallet providers. An open-source stack will enable this by making it easier for new wallets to be created. The following subsections detail milestones towards such an open-source stack.

ERC4337 Support

User transactions in the World App are currently based on Safe transactions; the custom format makes it less likely for teams to implement and run the infrastructure around it (e.g. bundlers). ERC4337 defines a common API for smart contract wallets and allows for interoperability. There are already multiple different smart contract and bundler implementations for ERC4337, which is the fastest and most flexible way to facilitate the integration for other wallets.

Bundler

Meta-transactions allow the batching of multiple users' transactions and compress them permissionlessly without any sacrifices on self-custody. This significantly reduces the costs for users, with the minor downside of small additional latency. While bundlers are trustless, it's beneficial for censorship resistance to have many of them and allow the user to switch between them. Also, the World ID proof could be combined with other proofs in a batch proof to make its usage more affordable on the Ethereum mainnet.

Wallet Support

ERC4337 transactions have their own format. The client and Safe contract (and perhaps Wallet Kit) should support this standard.

Relayer

Especially at scale, it is important for bundlers to be able to send transactions reliably. This seems to be a general-enough problem that it would benefit from a dedicated component (with optimally different implementations). A similar, currently closed-source service, is Open Zeppelin Defender.

Multiple Implementations

All of the above applies here as well. Due to the requirement of establishing device integrity of the phone in order to increase the trust in local computation, supported wallet apps should be whitelisted individually. The biggest requirement for going from 1 to n wallets will be a more scalable governance process to audit and whitelist those wallets and to refresh this list from time to time. Therefore, the community should create guidelines (e.g. a requirement for open source or a code audit) for wallet providers. Ideally, integrity would be ensured by multiple public providers, not only single gateway.

Hardware

In the context of the Worldcoin project, specialized hardware devices (Orbs) enable the verification of humanness and the issuance of World IDs. Several aspects can contribute to making Orbs more transparent and verifiable and increase their accessibility. One factor is to increase the reliability of Orb availability via multiple distributors and manufacturers. Additionally, increasing the transparency and verifiability of the Orb's functionality can help align the incentives of manufacturers not to be malicious. Furthermore, letting anyone develop alternative Orbs democratizes the solution space.

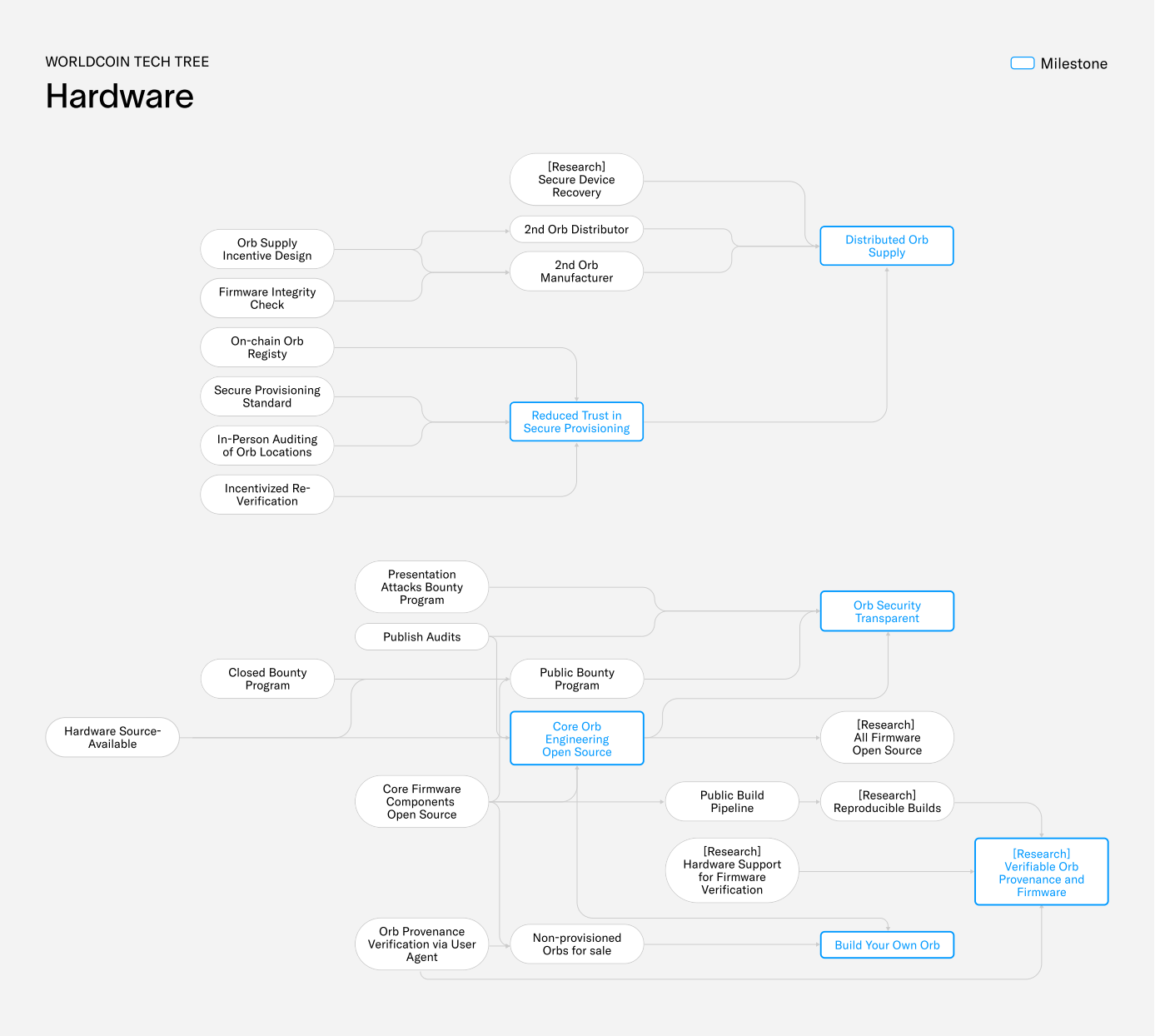

Fig. 2

The graph represents possible ways to increase transparency, verifiability and availability of Orbs across the world.

The following sections walk through different milestones that can contribute to the robustness of Orb infrastructure.

Reduced Trust in Secure Provisioning

Approved Orbs are able to issue new World IDs. To ensure the integrity of the network and reduce trust in provisioning, certain requirements should be fulfilled (see Secure Provisioning Standard). However, there is no provably secure hardware, and certain points of trust remain as described in Orb Provenance Verification via User Agent. Importantly, the Worldcoin protocol doesn't rely on perfectly secure hardware, as described in the limitations section, and audits like In-Person Auditing of Orb Locations can help decrease incentives for malicious actors. The following requirements can help reduce trust assumptions.

Onchain Orb Registry

The "Orb Registry" refers to the set of active Orbs endorsed by the Worldcoin Foundation and eventually the community. Only Orbs in this set can be used to generate new World IDs. If an entity's process can be sufficiently trusted (e.g. by implementing a Secure Provisioning Standard and regular audits thereof), the insertion of public keys from that entity in the Orb Registry could be delegated to that entity. To limit the harm caused by a malicious Orb, World IDs registered with different Orb manufacturers (and ideally with different Orbs) should be distinguishable from each other. This makes it possible for the ecosystem to respond to (inevitable) attacks by removing individual Orb manufacturers, and perhaps even individual Orbs, from the whitelist on-demand. Optionally, World IDs associated with fraudulent Orbs could be revoked (see Revocation in the Operations section). This information can be private and only stored on the World ID holder's device as long as it is provable on demand. If anyone were mistakenly affected by such action, they could re-verify through an active Orb. As a last resort, disagreement in the set of trusted provisioning entities could be resolved by forking the protocol and adding or removing provisioning entities.

Secure Provisioning Standard

The ability to add malicious keys to the Orb Registry would allow for the creation of fake World IDs. A secure provisioning standard can make it very difficult to inject malicious keys. “Secure provisioning” refers to the process of setting up the cryptographic keys of an Orb. One part of such a standard could, for example, specify that only certain approved secure element models can be used, and require proofs of authenticity from each secure element (via die-unique certificates, signed by the secure element vendor) be reported alongside the public keys. Public keys generated by this process can then be considered for insertion in the Orb Registry.

Today, a secure provisioning process is in place that involves generating private keys on a secure element as well as burning secrets generated on an air-gapped machine connected to a Hardware Security Module (HSM) into private fuses (only accessible by TrustZone applets) on the NVIDIA Jetson. These secrets are destroyed immediately after being derived (using a NIST-SP-800-108 KDF algorithm) into two keys transmitted to the backend used for future device attestation. The original key material only exists in the restricted fuse banks on the NVIDIA Jetson. Continuous auditing of the process can help maintain a high security bar.

Eventually, provisioning entities could be required to stake a security deposit, which would get slashed if fraudulent Orb public keys from that entity are detected, for example through In-Person Auditing of Orb Locations. This could be proportional to the number of verifications the hardware of that particular vendor has performed or the number of Orbs that have been sold.

In-Person Auditing of Orb Locations

Auditing in-person locations of operations can help detect malicious operator behavior as well as malicious Orbs, thereby disincentivizing malicious behavior. No entity in the Worldcoin ecosystem should have to be trusted. Therefore, all operations need to be audited in a distributed manner.

One primary concern is an entity submitting a fraudulent Orb public key to the Orb Registry. In this case, “fraudulent” means the entity has a way to spoof requests to the Uniqueness Service as if they came from a legitimate Orb. Any valid Secure Provisioning Standard should make such an attack very difficult. However, the risks associated with malicious individuals involved in provisioning and/or flaws in digital security can’t be entirely eliminated. If the Orb public key is inserted to the registry, generating fake World IDs becomes straightforward and can only be detected and prevented through Anomaly Detection methods and audits. Assuming the list of Orbs and their location is public, such an audit could look like the following:

1. Verify that the Orb actually exists and is in a given location i.e. not in a lab, disassembled and compromised and reporting a wrong location.
2. Attest that an Orb’s public key is in the Orb Registry via Orb Provenance Verification in User Agent. If no Orb can be found, the public key should be removed from the registry and corresponding World IDs could be revoked through governance. In case any user was mistakenly affected, they could re-verify through an Orb.
3. Optionally, a mechanism to verify that the number of in-person verifications matches the number of onchain verifications could be implemented. If there is a mismatch, this indicates the generation of fake identities in the process. Importantly, such a mechanism should be implemented in a privacy-preserving way.

The auditing of Orb locations by incentivized users and dedicated auditing organizations, when combined with software and hardware security measures, can make generating fake IDs very hard. Today, operations are already audited by third-party organizations. To increase the robustness of this process, a list of all Orbs, their locations, and verification counts could be made public. Further, appropriate incentives could be put in place. To disincentivize malicious behavior even further, Orb manufacturers and operators could be required to stake a security deposit that would get slashed in the case of malicious behavior.

Incentivized Re-Verification

Similar to In-Person Auditing of Orb Locations, verified users can be randomly incentivized to re-verify at a different Orb. For any attacker who compromised an Orb or spoofed verifications, such a second verification at a different Orb would be very difficult to also spoof. Therefore, statistically, the fraction of incentivized users that end up verifying a second time with a different Orb would be lower for a compromised Orb.

Distributed Orb Supply

Increasing the number of entities that distribute and produce Orbs can help improve the robustness of the availability of Orbs. Importantly, hardware and firmware of Orbs should be standardized. While variability could eventually be beneficial in increasing network resilience, allowing for variability in the firmware would make it harder for the community to audit all implementations and would increase the probability of a fraudulent entity to issue World IDs.

Further, companies developing Orbs would by default be incentivized to build the cheapest devices possible by compromising quality (e.g. camera quality, hardware security, software security, privacy) to maximize profits. This can undermine the set of verified World IDs in the long term. Aligning developer incentives would require defining precise standards. While this is likely possible, it is impractical; complex incentive mechanisms for manufacturers, audits and bug bounty programs would be required, especially for the firmware (e.g. spoof detection, platform security). Even when standardizing firmware and hardware, there are serious trust assumptions with manufacturers, which are discussed in more detail in Verifiable Orb Provenance and Firmware.

Therefore, at least initially, the hardware and firmware should be standardized. Focusing ecosystem efforts on distributing production, making components public, verifying Orbs and conducting in-person audits is likely the best path to increase transparency, verifiability and robustness of Orbs. These mechanisms can help detect malicious manufacturers and disincentivize the spoofing or compromising of Orbs. Below are possible improvements to distributing and manufacturing Orbs:

Orb Supply Incentive Design

Financial incentives can encourage additional entities to distribute and produce Orbs.

Firmware Integrity Check

Due to the risks associated with malicious firmware developers, only firmware endorsed via governance should be able to run on Orbs. This also decreases the trust requirements in manufacturers. Today, this is achieved through NVIDIA's secure-boot mechanism and Linux integrity checking (dm-verity). While no hardware security measure is impossible to circumvent, the above mentioned measures make loading unapproved firmware onto Orbs very difficult.

Second Orb Distributor

Additional entities that buy Orbs from manufacturers and distribute them through global warehouses can increase the robustness of global stock levels.

Second Orb Manufacturer

Additional entities that buy parts and modules and manufacture them in physically different locations can help increase the robustness of the supply of Orbs.

[Research] Secure Device Recovery

Firmware bugs could lead to non-functional "bricked" devices, with no possible resolution via over-the-air updates. While some bugs can be prevented through canary releases, others take much longer to surface. In such cases, a secure device-recovery mechanism can help re-enable bricked devices. Common options like password logins, SSH, and secondary boot media are inappropriate given the Orb's security requirements. Further, NVIDIAs built-in recovery mode is deemed insufficiently secure and therefore disabled.

Today, Tools for Humanity uses an open-source remote access solution (Teleport) for device recovery. A more transparent and secure recovery mechanism might configure the Orb such that it enters a custom "recovery mode" when an authenticated USB device is present. Such a device would contain a signed payload consisting of the Orb's public

identifier and command to be run. Both should be signed by the provisioning entities' recovery private key ensuring that only a particular Orb could run the payload. This mode could provide enough access to restore the firmware to a known-good version. However, such a recovery mechanism would require physical access to the Orb's mainboard, which means the tamper detection mechanisms will be tripped. Therefore, the Orb needs to be reprovisioned after it is recovered. The corresponding updates to the Orb Registry would then be a public record of the recovery.

Core Orb Engineering open source

To allow functionality of the Orb to be verifiable and enable anyone to build their own Orb, the firmware and hardware should be made open source.

Hardware source-available

Today, hardware components that aren't security critical (e.g. tamper detection and security board) have been made source-available. Eventually, everything should be made source-available.

Core Firmware Components open source

Publishing core firmware components makes the functionality of the Orb more transparent and is a requirement to achieve Verifiable Orb Provenance and Firmware. Publishing the following components could satisfy those requirements:

- The main Rust application on the Orb, which handles (among other things) biometric capture, iris code generation and backend communication
- Custom applets for the Trusted Execution Environment (TrustZone)
- The secure element interface
- The custom GNU/Linux-based Orb OS, including the over-the-air update system, the Linux kernel configuration, and the file system integrity configuration (dm-verity)

Making these core components open source enables others to understand the functionality of the Orb in more detail and build alternative Orb firmware

implementations. Once parts of the firmware are open source, there should be incentives through a public bounty program to submit potential vulnerabilities.

Given no hardware device can be perfectly secured, other sensitive components (like spoof prevention algorithms and fraud models) that may pose a direct integrity or security risk to the ecosystem if exposed should likely not be made open source. Importantly, the Worldcoin protocol doesn't need to rely on perfect hardware security, when complemented with mechanisms like In-Person Auditing of Orb Locations. To reduce trust requirements, the open source code can define software "jails" for some closed-source components. For example, consider a closed-source fraud-detection module that ingests biometric data. The open source code that interfaces with this module can provide strong evidence that the closed-source code cannot save/upload the biometric data.

[Research] All Firmware open source

It is unclear whether publishing all Orb components is desirable given security considerations described in the section Core Firmware Components are open source. There should be a continuous evaluation of which sensitive components can be made open source. The most difficult types of components are likely to be machine learning models and code related to spoof detection.

Orb Security Transparent

Conducting and publishing audits, an open source code base, and employing bounty programs can help make the Orb more secure and increase transparency around security. Below are possible steps towards making the security of the Orb more transparent.

Presentation Attacks Bounty Program

One possible attack vector on the Worldcoin protocol is to create fake World IDs by spoofing the Orb (presentation attacks). Apart from other methods to detect and prevent such attacks, a bounty program can help raise the security bar. For example, the Worldcoin Foundation could award a certain amount of WLD for every novel presentation attack that is reported.

Publish Audits

Requiring conducting and publishing of audits of hardware and firmware can help reduce trust requirements and increase incentives for developers to implement secure systems. Such audits could entail both security and privacy considerations.

Private Bounty Program

A bounty program can raise the security bar by finding vulnerabilities early. A first version of a private bounty program has been launched on HackerOne. Gradually, more endpoints and source code should be added.

Public Bounty Program

Making the bounty program public can help increase the reach of the program. This should happen gradually.

[Research] Verifiable Orb Provenance and Firmware

While there is significant research involved, it would be ideal to allow the public to verify properties of active Orbs (don't trust, verify), including:

- An Orb is from an Orb vendor that meets the standards and is not a counterfeit
- An Orb is configured to only boot signed firmware
- An Orb is running a specific version of the firmware

These verifications can help mitigate important privacy concerns related to biometrics. The public should not need to blindly trust an Orb vendor to faithfully/correctly implement privacy-preserving firmware.

Eventually, there might be a path to also allow for non-Worldcoin Foundation-governance-approved firmware, though it is unclear whether this would be desirable given the potential downsides. Appropriate incentive and audit mechanisms to disincentivize malicious behavior would be required, which might not be viable in practice.

Orb Provenance Verification via User Agent

A first step towards verifying an Orb as non-counterfeit could be implemented through provenance verification via the User Agent. Such a mechanism could help verify that an Orb is from a vendor that has been approved by Worldcoin governance and therefore is running approved firmware. Such a feature can be integrated in other protocol-compatible apps.

One possible path for such a verification could be to ask the Orb to sign a challenge that has been generated by the App. Orbs contain two mechanisms for cryptographically attesting they are in the Orb Registry: private keys in the secure element and private keys derived from fuses on the NVIDIA Jetson. Verifying signatures from both sources provides strong evidence that an Orb was manufactured by a vendor that has been approved and was not subsequently tampered with. Verification of the NVIDIA Jetson fuse state can provide strong evidence that Orbs can only boot firmware that has been signed. The user agent could also request an Orb's firmware version from the Trusted Execution Environment's (TEE) secure storage. As part of a normal boot, the root hash for dm-verity can be delivered to the bootloader by the TEE, ensuring that only code authorized by the TEE is able to boot. Inside of secure storage, these hashes would be associated with version numbers, allowing an entity (e.g. the World App) to request attestation of the current hashes and version numbers existing in the secure storage.

This mechanism assumes that an Orb's private key only exists in its secure element (i.e. there are no other copies), a constraint which should be specified by the Secure Provisioning Standard. Private keys are generated on the secure element directly and never leave, and a series of transparent certificate attestations during generation and export can prove that a particular key originated from a legitimate secure element.

Therefore, physically attesting an Orb has a private key provides strong evidence that the same private key is not in the control of an attacker. Extracting private keys from the Orb's secure element is assumed to be extremely difficult.

It is important to note that it is impossible to fully eliminate trust in the Orb hardware vendor/manufacturer or upstream vendors. The following attack vectors remain:

- The Orb vendor could bypass parts of the secure provisioning process (due to malice or incompetence), invalidating the guarantees of the proposed verifications. Therefore, Orb manufacturers should be audited to ensure the Secure Provisioning Standard is maintained.
- NVIDIA firmware could have security vulnerabilities or backdoors, which could threaten the Jetson fuse attestation.
- The secure element vendor could be compromised/incompetent/malicious, which would threaten the integrity of the corresponding attestation.
- While the combination between dm-verity and TrustZone version tracking based on the filesystem's merkle-tree may allow for extremely trustworthy version information to prevent the use of signed but deprecated firmware, bugs in the TEE implementation could lead to a loss of integrity.
- The Worldcoin Foundation could sign malicious firmware. Adding Hardware Support for Firmware Verification can help set up procedures to verify the actual firmware that is running on an Orb.

Therefore, there should be mechanisms to mitigate the risk of fraudulent manufacturers or compromised Orbs. In-Person Auditing of Orb Locations and Incentivized Re-Verification can make exploiting backdoors significantly harder and help detect malicious verification of World IDs in retrospect.

Public Build Pipeline

The builds for each production Orb firmware release can be built in a publicly-visible way. This does not have the same guarantees associated with reproducible builds and the ability to dump the flash of the main computing unit of the Orb. However, it still improves transparency and verifiability as it provides a mechanism to trace changes in open source Orb firmware components to their inclusion in the deployed firmware. For each Orb firmware version, there should be a link to the corresponding sources and public build logs.

[Research] Reproducible Builds

Without reproducible builds, the public is required to trust that compiled firmware wasn't maliciously modified during/after the build. Reproducible builds provide a mechanism to verify that Orb firmware was compiled from a specific state of the public repositories. To verify the integrity of the firmware, third parties should be able to build it from source on their own infrastructure. Full reproducibility means the resulting artifacts should be identical to those deployed to Orbs, and the signature from the signed firmware should be valid for the self-built firmware. The initial priority should be to make privacy-sensitive components of the firmware open source and reproducibly built.

However, there are some limitations. The firmware should (at least initially) include closed-source components, which are opaque parts of the system. Some of these are from Tools for Humanity (e.g., spoof-detection models) and some are from vendors (e.g., NVIDIA firmware components). Additionally, some components may be hard to make reproducible. These can be built separately and pulled in as compiled components to the main build.

[Research] Hardware Support for Firmware Verification

The most transparent way to verify firmware would be by having read access to the persistent storage of the main computing unit. Future Orb versions could include external ports for directly reading the flash memory of the main processor. However, a hardware implementation that provides read-only access to the flash memory—without introducing new security risks—still needs to be validated. If a hardware implementation is possible in a secure way, public auditors could then be incentivized to use this mechanism for verifying the integrity of a particular Orb's firmware. The integrity verification of the dumped memory could optionally reuse the Orb's internal integrity verification mechanism (dm-verity). This can provide stronger guarantees than Orb Provenance Verification via User Agent as there are less attack surfaces for spoofing direct physical access relative to remote attestation schemes.

While this mechanism would provide very strong guarantees for the firmware state, it is still possible to spoof the auditor at the hardware level. For example, there could be a second hidden flash chip that the Orb is actually booting from. This risk could be mitigated by additional audits that inspect the hardware directly on a random subset of devices. Further, in-person audits of Orb locations can make attacks significantly easier to detect and can create disincentives for malicious behavior.

Build Your Own Orb

Enabling anyone to build their own version of the Orb is an important aspect to enable forkability of the Worldcoin project.

Non-provisioned Orbs for Sale

Enabling anyone to buy non-provisioned / unlocked hardware allows them to flash their own firmware and do their own development.

Operations

Operations, in the context of the Worldcoin project, refers to procedures in the “analog world” that allow people to get their World ID verified. The primary participants are Orb operators (i.e. independent entrepreneurs and their organizations around the world) who provide Orbs in physical locations for people to verify. Certain infrastructure primitives can help reduce trust assumptions and aligning the incentives of all participants.

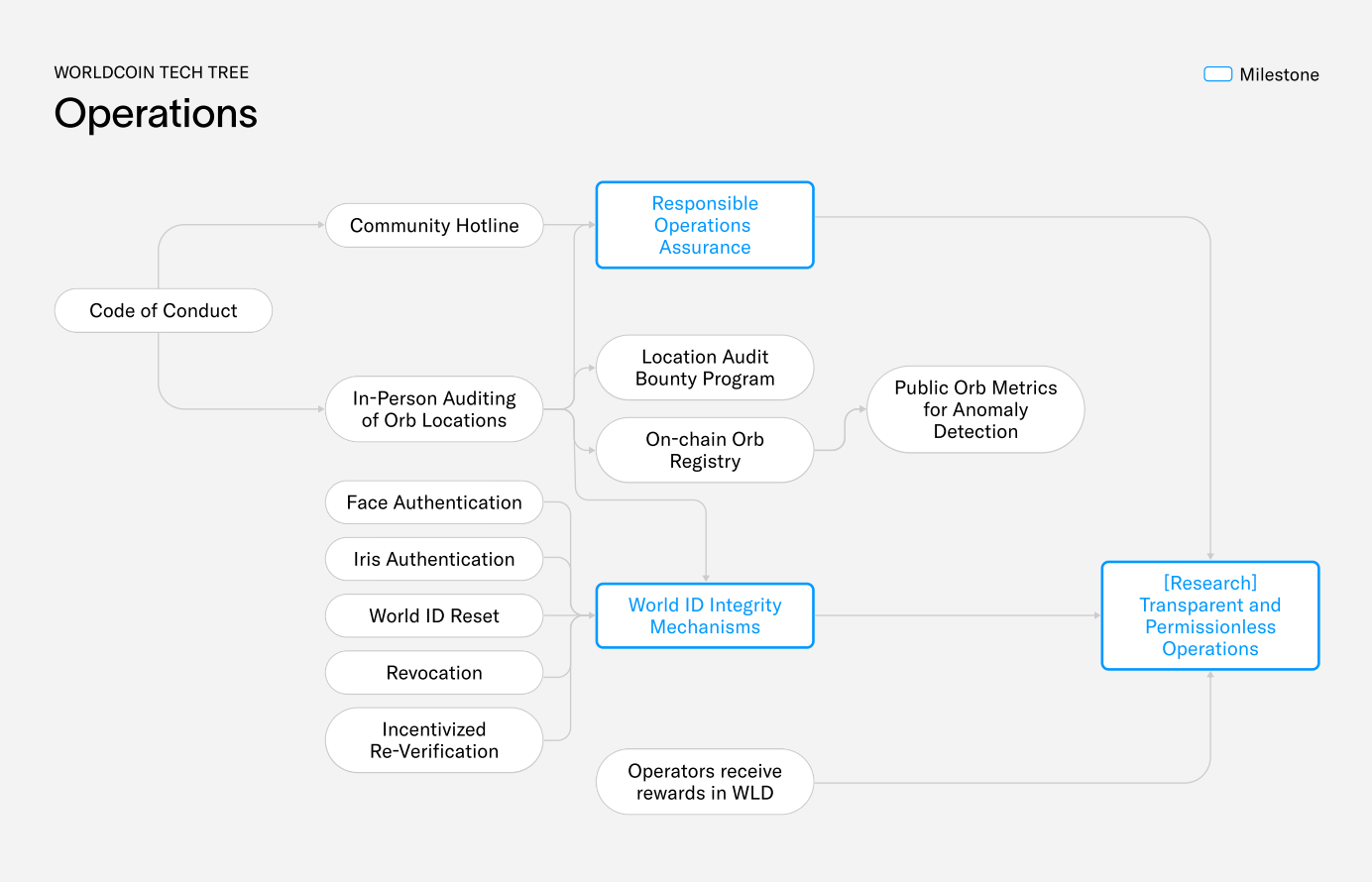

Fig. 3

The graph represents possible ways to increase the transparency and integrity of operations.

Responsible Operations Assurance

Auditing of operations from different angles can help reduce trust assumptions and align incentives of ecosystem participants. The following subsections very briefly introduce mechanisms that can help to this end.

Code of Conduct

Basic set of rules operators should adhere to and can be audited against.

Community Hotline

Ability for anyone to submit reports of fraudulent behavior. To avoid spam, this might eventually have to be gated on the reporting individual having a World ID or staking a small sum.

In-Person Auditing of Locations

See Hardware

Incentivized Re-Verification

See Hardware

Location Audit Bounty Program

A location-audit bounty program can help increase the resilience of in-person audits by incentivizing random ecosystem participants to audit operations. Importantly, such a bounty program requires careful mechanism design to avoid unintended side effects, including collusion.

Onchain Orb Registry

See section in Orb

Public Orb Metrics for Anomaly Detection

Metadata from Orbs could be made public to be utilized for distributed anomaly detection, to spot outliers and inform decisions on which Orbs to audit in person as well as surfacing discrepancies between onchain verifications and in-person verifications.

World ID Integrity Mechanisms

Distributing the issuance and custody of World ID requires incentive mechanisms to ensure high World ID integrity (i.e. making it hard to illegitimately acquire and use the World ID of others). Such mechanisms across all participants—from hardware manufacturers to individual Orb operators to users—can disincentivize actions that undermine World ID integrity.

There are several scenarios that could result in the illegitimate creation or fraudulent ownership of World IDs:

1. **Verification Device Compromise:** A third party submits fraudulent verifications by compromising the hardware or spoofing the verification process.
2. **Identity Theft:** The device operator or a third party manipulates individuals into verifying or compromising their phone to access their World ID.
3. **Identity Sale:** An individual decides to sell their World ID to a fraudulent actor.
4. **Issuer Fraud:** Organizations developing the firmware for verification devices secretly generate World IDs.

Each scenario decreases the utility of the World ID issued by the Orb, so measures to make these attacks more costly are important. The difficulty of compromising verification devices is increased by the Bug Bounty Programs mentioned in the section about the Orb. Another important defense is the In-Person Auditing of Locations, which increases the difficulty of a large amount of attacks. Additional defenses are described below:

Face Authentication

As part of the verification process, a signed and encrypted face image could be transferred to the user's phone. The user could then authenticate against that image within the app, an approach similar to FaceID, making it hard to authenticate using the World ID of someone else. Beyond face recognition, the phone would also need to perform liveness detection to prevent spoofing attacks. The security guarantees that such a liveness check has actually been conducted and not spoofed are lower on consumer phones compared to custom hardware like the Orb (given consumer cameras are less advanced and that it is hard to create a trusted execution environment on a person's phone to ensure computational integrity). With improvements in mobile zero-knowledge proof abilities, it may become possible to compute the authentication and liveness checks inside zero-knowledge proofs on the user's phone, which would increase the security guarantees.

Iris Authentication

Similar to face authentication, iris authentication performs a biometric authentication against a biometric template as well as a liveness check. By requiring people to come back to a biometric verification device, the security guarantees that the liveness check has not been spoofed or bypassed are increased at the cost of convenience. If both face authentication and iris authentication were implemented, developers could choose which level of security is required for their application.

World ID Reset

World ID Reset allows anyone to get a new World ID in case they have lost their previous World ID. Further, victims of social fraud and identity theft should be able to reset their World ID through an Orb using only their biometrics. This would also make the purchase of other individuals' proof of personhood less profitable from a game-theory perspective given individuals could get a new World ID and deactivate the sold one.

Revocation

If a particular firmware version of a verification device is deemed insecure or an operating organization is found to be fraudulent, World IDs can be revoked in retrospect, to eliminate potential fraudulent identities. Eventually, there might be the need for an adjudication body that conducts such investigations and actions. If any real and legitimate individuals are affected, they could re-verify at an Orb.

[Research] Permissionless Operations

In an ideal world, anyone should be able to acquire Orbs and onboard others. To reach this point, however, significant research is required and operations must be audited with a high degree of certainty.

Operators Receive Rewards in WLD

Initially, operators were paid in USDC. Since the launch, the operator rewards have been transitioned to WLD, where laws allow.

Protocol

The protocol contains off- and onchain components that are responsible for handling e.g. verification or authentication requests from users. Since privacy is central to World ID, it is especially important to not sacrifice it in favor of accelerated increases in transparency, verifiability and resilience. One example of this is the uniqueness service, which still requires more research before it can be made more permissionless.

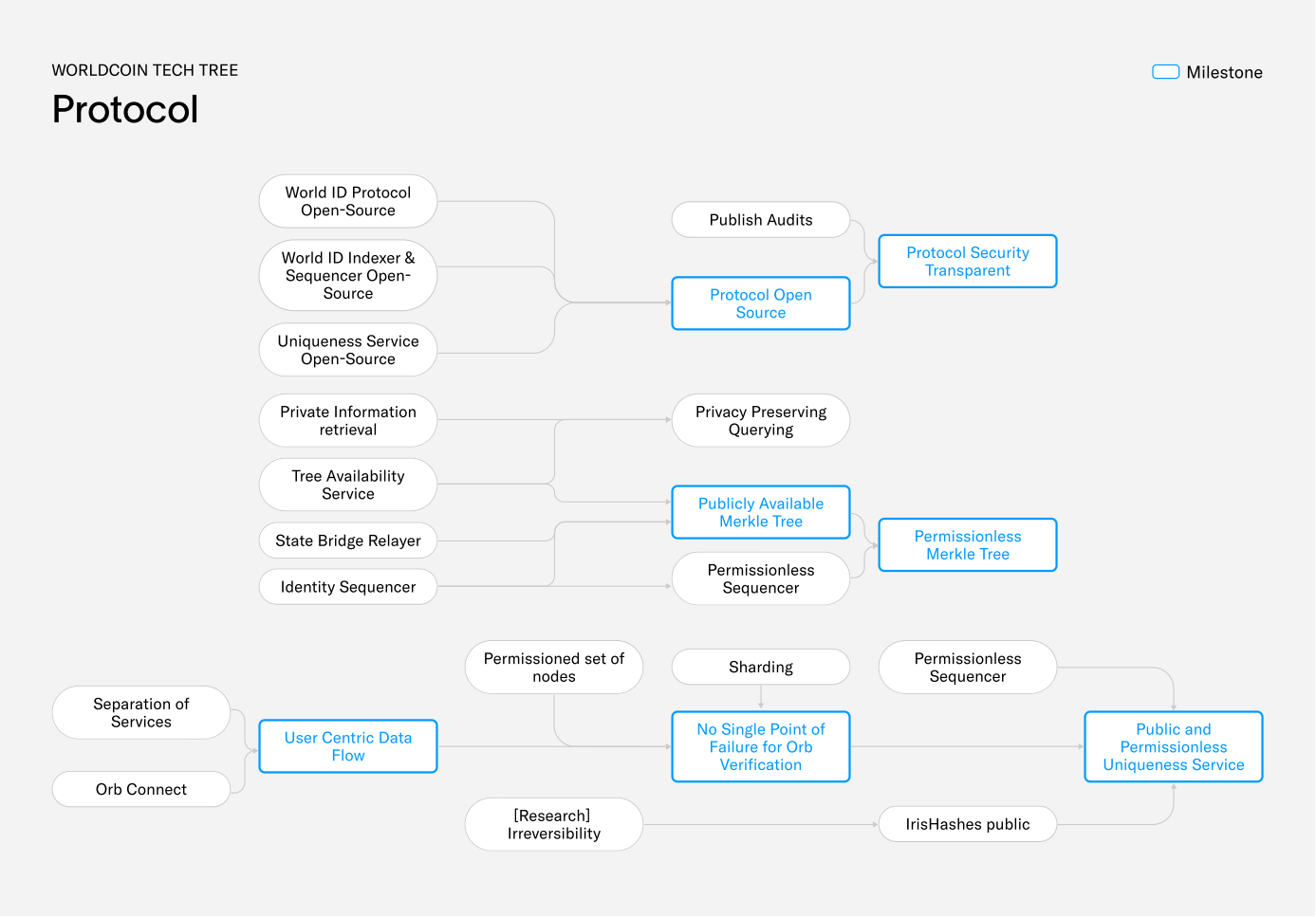

Fig. 4

The graph represents possible ways to increase the protocols transparency, verifiability and ability to recover from localized failures of participants.

The following sections describe possible improvements to further increase transparency, verifiability and robustness of the protocol.

Protocol Open-Source

Most of the components of the protocol components are already open source (see the open source tree)—except for the uniqueness service.

Protocol Security Transparent

Over the course of several months beginning in April 2023, audit firms Nethermind and Least Authority conducted two separate security assessments on the off-chain and onchain components of the Worldcoin protocol, including the following parts of the protocol:

- Correctness of the implementation, including cryptographic constructions and primitives and appropriate use of smart contract constructs
- Common and case-specific implementation errors
- Adversarial actions and other attacks on the code
- Secure key storage and proper management of encryption and signing keys
- Exposure of any critical information during user interactions
- Resistance to DDoS (distributed denial of service) and similar attacks
- Vulnerabilities in the code leading to adversarial actions and other attacks
- Protection against malicious attacks and other methods of exploitation
- Performance problems or other potential impacts on performance
- Data privacy, data leaking and information integrity
- Inappropriate permissions, privilege escalation and excess authority

Of the issues detected by Nethermind, which performed a comprehensive audit of Worldcoin's smart contracts, 92.6% were identified as fixed after the re-audit stage, while 3.7% were mitigated and 3.7% were acknowledged.

Details of both audits can be found in the Nethermind and Least Authority reports.

Publicly Available Merkle Tree

The set of World ID public keys is already publicly available and committed to by the sequencer on Ethereum. The public keys are available as calldata and the current state of the Merkle tree is committed as a Merkle root. Its validity is enforced through a ZK validity proof of batch insertions of public keys. While this ensures that the committed root actually corresponds to a Merkle tree, it's not yet ensured in the validity proof that the public keys actually originate from an Orb. Even though the leaves are publicly

available, it's practically infeasible for the client to download all of this data and reconstruct the tree in order to be able to compute a Merkle inclusion proof. The tree availability service serves those Merkle inclusion proofs to clients. Clients can check the correctness of the Merkle proof against the onchain root. However, this request can leak additional metadata about the client (e.g. IP address). This can be addressed by routing those requests through mixnets or Private Information Retrieval (PIR).

Permissionless Merkle Tree

As mentioned above, the validity proof of the Merkle tree needs to be enriched by a signature check of the public key. Once this check is added, trust in the identity sequencer is no longer required. Similar to the uniqueness service, this sequencer also needs to actually implement coordination to rotate between multiple sequencers, so there is no possibility of censorship.

User-Centric Flow

Currently the verification flow (and similarly the reset flow) are intertwined with different services, with some being permissionless and others not. Going from an intertwined architecture to one in which components are separated allows to increase transparency, verifiability and robustness of individual components. This architecture is described in more detail in the Advancing Decentralization blog post. This also allows the user to own their data and selectively share certain parts with the required services. A first step and prerequisite for this is to allow the user to retrieve all the data generated by the Orb. This requires an end-to-end encrypted, direct peer-to-peer connection between the user and the Orb, which is referred to as "Orb Connect." However, the primitives used to build this communication layer could also be reused for all other communication between the client and nodes or services.

No Single Point of Failure for Orb Verification

Increasing the resilience of the uniqueness service is challenging, because a permissionless operation of the service would require iriscodes to be public. A permissioned set of nodes that run the computation and agree on the result through consensus, or run the comparison on a reduced version of the iris codes so that no node has the full code improves the verifiability of the system. Successful research on iris hashes could enable making them publicly available and allow for permissionless operation. A draft with more details can be found here.

Public and Permissionless Uniqueness Service

The most difficult dependency is the research on beyond state-of-the-art template protection of iris codes. This is a prerequisite to make the operation of the uniqueness service permissionless. This can be achieved either by publishing anonymized iris hashes (the database needs to be available and readable for the service to run the deduplication) or by multiple parties running the service where each party performs the computation on one shard of the data. Besides that, research similar to that currently being conducted with respect to other sequencer models (e.g. for rollup sequencers) is needed. The problems and solutions should be very applicable to this model as well.

Governance

A global community of developers, individuals, economists and technologists conceived and made early contributions to the Worldcoin protocol. The original idea started with co-founders Sam Altman, Alex Blania and Max Novendstern, who assembled a team to take the initial steps toward development of technology to support Worldcoin via their company Tools for Humanity.

Tools for Humanity's mission is to build a more just economic system. It is a Delaware (U.S.) corporation headquartered in San Francisco, California, with a wholly-owned subsidiary, Tools for Humanity GmbH based in Germany. Tools for Humanity supported Worldcoin's multi-year beta testing phase, during which it developed the Orb and the World App.

To date, Tools for Humanity and other early contributors are committed to providing every person on the planet access to the global economy, regardless of country or background.

Today, the governance of the Worldcoin project is overseen by the Worldcoin Foundation, an independent entity, which is committed to continue transitioning governance to all of humanity. It is also important that this happens in a deliberate way and that governance (e.g., voting) is well studied and tested before complete control is fully transitioned.

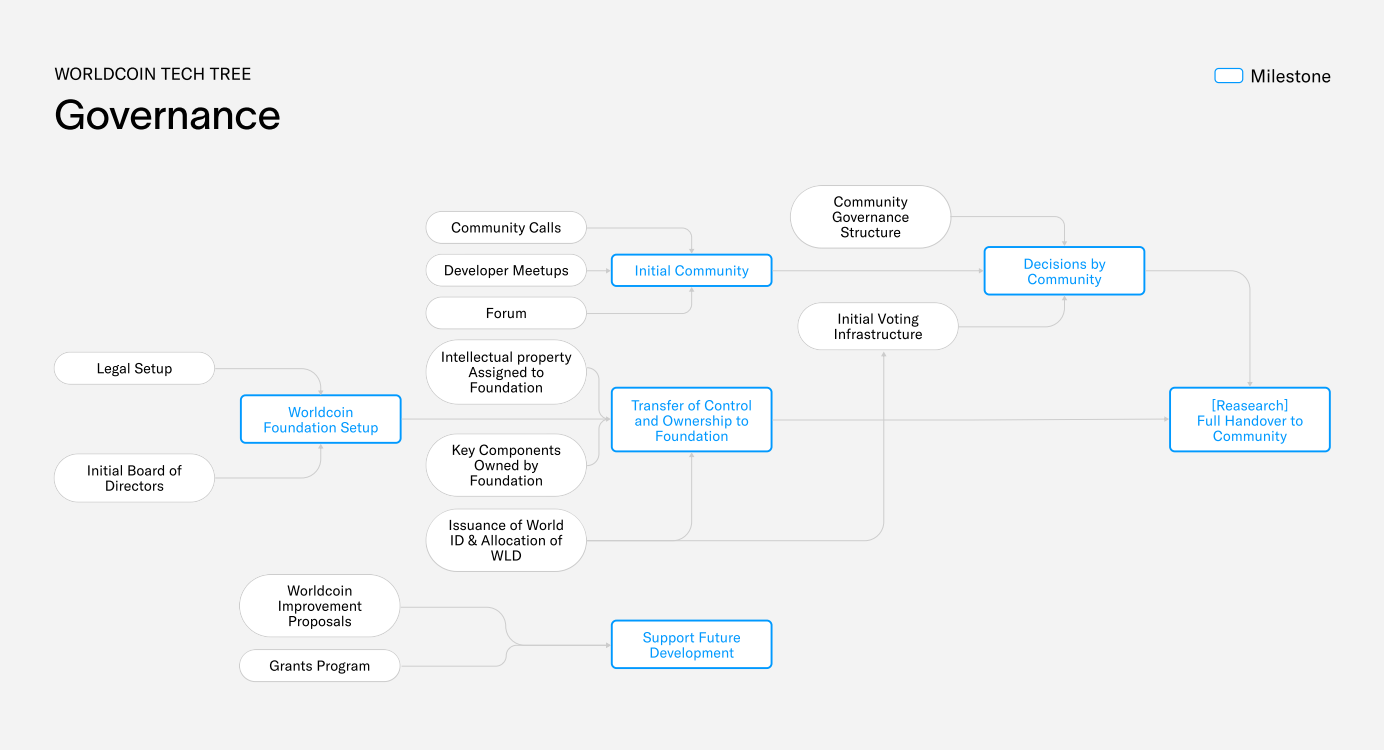

Fig. 5

The graph represents a possible evolution of the governance of the Worldcoin project. It is important that this happens in a deliberate way and that e.g. voting mechanisms are well studied and tested before complete control is handed over. Today, the governance of the Worldcoin project is overseen by the Worldcoin Foundation.

The following sections describe different improvements that already have and can contribute towards this objective.

Worldcoin Foundation Setup

On October 31st, 2022 the Worldcoin Foundation was established as the non-profit steward of the Worldcoin protocol, supporting and growing the ecosystem as it becomes self-sufficient. The Foundation's main objective is to scale an inclusive identity and financial network as a public utility and to expand the governance thereof. This infrastructure has the potential to empower everyone to participate in the global economy in the age of AI and provide the foundation for shared governance of universal basic income.

The Foundation is an exempted limited guarantee foundation company, which is a type of non-profit incorporated in the Cayman Islands. It has a wholly-owned business company subsidiary in the British Virgin Islands called World Assets Limited. This is "one of the most often used, and internationally recognized structures" for decentralized blockchain projects. To learn more about this entity arrangement, check out this Guide to the Cayman Islands Foundation Company from the Foundation's outside counsel at the law firm Ogier. The Worldcoin Foundation is "memberless"; it has no owners or shareholders.

This entity setup was a good fit for the Worldcoin project due to the Foundation's separate personhood, limited liability, tax efficiency, support for compliance with virtual asset regulations, and suitability for long-term community governance. That last point is especially important. Cayman foundation companies can be structured to be "memberless" (that is, have no owners or shareholders) and instead to take instructions from token holders and/or World ID holders. They can therefore gradually steer matters such as running a grant program, open sourcing intellectual property (IP), entering into service agreements, and managing a treasury. In the case of the Worldcoin project, the shared governance model is all the more critical so that, in the long-term, decisions can reside with the community.

At the same time, the Foundation can aid the community's governance by safeguarding protocol IP. In most legal systems today, a traditional legal entity is needed to protect IP such as trademarks, open-source copyrights and domains. Tools for Humanity has

already transferred core protocol IP to the Foundation, including smart contracts, the World ID SDK, patents for the Orb design and iris recognition technology, brand assets, domains and social media accounts. And the Foundation has open sourced several core tech repositories and made the Orb's hardware available under its Responsible Use License.

Transfer of Control and Ownership to the Foundation

In order to facilitate a governance model that involves all of humanity, several assets and key components have been transitioned to the Worldcoin Foundation:

- **Treasury:** The Worldcoin Foundation (and/or its affiliate entities) manages the treasury of tokens once they are unlocked. This includes Worldcoin grants, Operator rewards, and other contributor grants.
- **Orb IP:** Tools for Humanity has transferred the Orb IP to the Worldcoin Foundation. - The Orb hardware and software will be made publicly available under a restricted use license, prohibiting the misuse of the technology. This allows the Foundation to onboard other organizations building Orbs or similar devices.
- **Data:** In the case of the Worldcoin project, due to the protocol's use of personal data, the shared governance model is especially important. The Foundation is the "data controller" for any personal data collected via Orb sign-ups after network launch. Through its data consent form, the Worldcoin Foundation makes it clear that "**We will never sell your data.** We will also not use any data listed in this form to track you or to advertise third parties' products to you," and that "We will not sell, lease, trade, or otherwise profit from your biometric data."
- **Ability to Whitelist Orb Provisioning Entities:** The Foundation manages the permissions for adding Orbs to the network, balancing hardware distribution, security, and growth.

In order to grow the network and ultimately enable all of humanity to participate in the governance of the Worldcoin project, the issuance of World ID and allocation of the WLD token (in certain countries) is ongoing.

Support Future Development

To encourage individuals and organizations to contribute to the Worldcoin ecosystem through research, the development and production of Orbs or auditing the system, the Worldcoin Foundation is setting up a grants program. Further, the Worldcoin Improvement Proposals (WIP) process is currently being created and will be open for proposals soon.

Separately, the Foundation intends to work on common standards and ecosystem-wide proposals. For example, today, Orbs are developed and produced by Tools for Humanity. Orb operations are managed by several organizations around the world. With support from Tools for Humanity, the Foundation will work on standards and incentives for organizations to develop, produce and operate Orbs such that production of Orbs and their operation can be further distributed. More details can be found in the sections on Hardware and Operations.

Initial Community

The Worldcoin project maintains a dynamic and evolving blueprint that is subject to change and refinement through input and decisions from the Worldcoin community. Today, whether you are a developer, a user, an enthusiast, or simply someone interested in the future of decentralized systems, you can learn more and participate through the following channels:

- Join the community discussion on Twitter or Discord
- Contribute to open-source repositories on Github
- Visit the World ID Developer Docs and Portal
- Reach out directly to the World ID team
- View live onchain data on Dune Dashboard
- View the Open Source Overview

To enhance transparency and facilitate community involvement, regular community calls should be established with the aim to provide a platform for open dialogue and updates on the Worldcoin project's progress. Additionally, a dedicated forum similar to

ethresearch should be set up to further foster meaningful discourse and engagement around the Worldcoin project. This forum will serve as a hub for ideas, suggestions, and discussions among community members and the project team. Lastly, the Foundation has already hosted several developer meetups and strives to create more opportunities for developers to collaborate, innovate, and contribute to the Worldcoin project.

Decisions by Community

Increasing the resilience of the governance of the Worldcoin protocol is both imperative and unprecedented, given the foundational nature of proof-of-personhood infrastructure and the ambition to scale it to billions of people. Building a community-based governance system for Worldcoin represents perhaps the most formidable challenge of the entire project, and this process is still in its earliest stages.

The Foundation should ultimately have a limited role in the protocol's governance. To this end, the Foundation's founding documents have provisions for community-driven governance. These provisions make it possible, through a prescribed process, for the community to make recommendations to the Foundation's Board of Directors. For further details, see the Foundation's Memorandum of Association and Articles of Association.

World ID provides unique infrastructure for distributed governance and presents the opportunity to harness input from a large and diverse set of individuals for community-driven governance. The reach of World ID is unprecedented: over two million World IDs have already been issued, and tens of thousands more are issued each week. As a proof-of-personhood protocol, World ID naturally supports "one-person-one-vote" voting, in contrast to token-based voting commonly used by other blockchain projects. Notably, this adds more democratic options to the design space of voting mechanisms for Worldcoin. However, the exact structure of delegating decisions to the community needs careful iteration and consultation with experts. Further, many governance decisions notoriously lack engagement from participants. Therefore, it will be important to encourage a large set of people to participate and explore the decisions. In the future,

the User Agent should not only serve as an entry point into using Worldcoin, but also the governance over it.

[Research] Full Handover to Community

The Worldcoin Foundation is committed to continuously transition governance towards a model that involves all of humanity. This is an unprecedented endeavor in scale and complexity for a decentralized system, which will require a methodical and gradual approach. Key aspects like voting mechanisms should be thoroughly researched, validated with experts and tested before meaningful control is transferred. Transparency, inclusivity, and neutrality are essential. However, these attributes contribute to intricate governance structures like today's democracies, which can lead to often slow and expensive decision-making. While this deliberateness is beneficial for making long-term strategic decisions, such as amending a constitution, it can hinder the ability to quickly adapt to new challenges during the initial growth phases. Hence, prematurely adopting a governance model that fully transitions governance to the community without a well-vetted plan is itself a failure mode to be avoided.

The Foundation will solicit proposals for how token holders and World ID holders should interact in Worldcoin's governance model. In general, the Foundation seeks input from contributors, the community and experts in the field as it increases the robustness of the governance of the Worldcoin protocol.

Limitations

This section outlines some limitations (without claiming exhaustiveness) based on laws, intrinsic limitations of the project as well as temporary limitations that can be mitigated by further open development.

UBI

Worldcoin enables the fair distribution of UBI globally as it offers both globally-accessible digital financial rails while ensuring each participant cannot double-claim the same UBI distribution. The Worldcoin Protocol is not intended to generate profits to distribute UBI, and instead, it requires a separate funding source (e.g., a share of the profits generated by an AI Lab) to distribute global UBI.

Orb Security

The Orb sets a high bar to defend against scalable attacks; however, no hardware system interacting with the physical world can achieve perfect security. The security and anti-fraud measures integrated into the Orb are continuously refined. Several contributor teams are continuously working on increasing the efficacy of the liveness algorithms as well as other security measures that are deployed via over-the-air updates. Even though significant effort has been spent on raising the security bar of the Orb, it is expected that an Orb may get spoofed or compromised by determined actors. The Orb has been designed with this threat model in mind: any World ID issued by a particular Orb can later on be revoked through the governance of the Worldcoin Protocol. To continuously discover potential vulnerabilities of the Orb, contributor red teams test various attack vectors. Several audits on the Orb and its infrastructure have been conducted and a bug bounty program will soon launch. Implementing suitable incentive mechanisms for Operators and decentralized audits of all Orbs in operation can help raise the bar beyond what hardware security could achieve in isolation, especially for scalable attacks.

False Rejections

Biometrics are probabilistic and biometric verification has inherent error rates. Currently, the error rate of the Orb for confusing any two people to be the same, is approximately 1 in 40 trillion. On a billion people scale, this translates to a 99.999% true acceptance rate or 0.001% false rejection rate, which is significantly better than other known alternatives. However, the ultimate objective is enabling total inclusivity. Using the birthday problem approximation, a false rejection rate of 10^{-20} is statistically required to prevent the false exclusion of a single individual at a global scale. Ongoing community research is focussed on improving iris biometrics beyond the current state-of-the-art by leveraging AI and advanced hardware capabilities of the Orb. In the event that iris biometrics turn out to be insufficient, combining several biometric signals (biometric fusion[1]) could be employed to further reduce the error rate, a functionality already supported by the current hardware version of the Orb.

It is important to note that many health conditions, like cataracts to a certain degree, do not impede iris biometrics. Already today, iris biometrics surpass the inclusivity of other PoP verification alternatives like official IDs since less than 50% of the global population has digitally verifiable identities. However, if the proof of personhood mechanism becomes essential for society, it is important that eventually every single person can verify if they want to. Although not currently established, there could be specialized verification centers to facilitate alternative means of verification for individuals with eye conditions, via e.g. facial biometrics. The introduction of alternative means of verification for World ID could potentially create loopholes. More details on the biometric verification through the Orb can be found here.

Decentralization and Open Sourcing

Today, large parts of the Worldcoin Protocol stack are open source. This includes the World ID protocol, the sequencer for the Orb credential, and the SDK to access it. Other parts, like the firmware of the Orb are not yet open source due to security considerations; however, eventually every part of the infrastructure supporting the Orb credential should be open source. Further, while operations are already spread across independent entities, Orbs are only available via Tools for Humanity. For more details, see the decentralization section.

World ID Transferability

While deduplication, i.e. ensuring that everyone can only verify once, has been solved to a high degree of certainty with the Orb, the authentication of the legitimate owner of a proof of personhood credential is both an important as well as a difficult challenge. This challenge is the same for any digital identity or PoP mechanism. Today, if someone passes on their World ID keys to a fraudster (e.g., through being tricked to sell their keys), the fraudster can then use that World ID to authenticate. Therefore, fraudsters could bypass the “one-person one-X” principle by acquiring multiple World IDs. There are several preventative measures in the World App that make it harder to restore another user’s backup, however, those measures are only temporary, especially since access to World ID through other wallets will become increasingly important over time. Therefore, several additional measures should be implemented:

- **Face Authentication:** Facial recognition, performed locally on the user’s device in a fashion similar to Face ID, can be used to authenticate users against their Orb verification, thereby ensuring that only the person to whom the World ID was originally issued can use it to prove that they are human. Authentication involves a 1:1 comparison with a pre-existing template that is stored on the user's phone, which requires considerably lower levels of accuracy in contrast to the 1:N global verification of uniqueness that the Orb is performing. Therefore, the entropy inherent to facial features is sufficient.

- **Iris Authentication:** This is conceptually similar to face authentication with the difference that an individual needs to return to an Orb. This process validates the individual as the rightful owner of their PoP credential. Using iris authentication through the Orb instead of face authentication on the users phone increases security. This authentication mechanism can be compared with, for example, physically showing up to a bank or notary to authenticate certain transactions. Although inconvenient, and therefore rarely required, it provides increased security guarantees.
- **World ID Recovery and Re-Issuance of Keys**: If a proof of personhood credential has been lost or compromised by a fraudulent actor, individuals can get their Orb credential re-issued by returning to the Orb, without the need to remember a password or similar information.

More details on the different mechanisms and their roadmap can be found here.

Key Recovery and Persistent Reputation

Today, keys can be recovered through restoring user-managed backups. If someone lost their keys, they can get a new World ID (and deactivate their previous one); however, the keys for the previous World ID cannot be recovered through biometrics. For privacy reasons, actions associated with a particular World ID cannot be recovered today. Consequently, humanness validation should be implemented today with time bounds to ensure sybil resistance. It also means that certain use cases like persistent reputation, as would be required for undercollateralized lending, are limited today. Enabling World ID recovery requires solving hard research challenges to preserve privacy as well as a careful consideration of societal implications of persistent reputation. More details on recovery and the roadmap can be found here.

Footnotes

1. This requires the prevention of combinatorial attacks and therefore excludes e.g. fingerprints. Therefore, the only possible combination would be combining pictures of the iris and the face since they are imaged at the same point in time and in the same location (i.e. the face). ↩

Disclaimer

PLEASE READ THE ENTIRETY OF THIS "NOTICE AND DISCLAIMER" SECTION CAREFULLY. NOTHING HEREIN CONSTITUTES LEGAL, FINANCIAL, BUSINESS, INVESTMENT OR TAX ADVICE AND YOU SHOULD CONSULT YOUR OWN LEGAL, FINANCIAL, BUSINESS, INVESTMENT, TAX OR OTHER PROFESSIONAL ADVISOR(S) BEFORE ENGAGING IN ANY ACTIVITY IN CONNECTION HEREWITH. NEITHER THE WORLDCOIN FOUNDATION (THE **FOUNDATION**) AND ANY OF THE PROJECT PARTICIPANTS (TOGETHER WITH THE PROJECT PARTICIPANTS, THE **WORLDCOIN ECOSYSTEM**) WHO HAVE WORKED ON THE WORLDCOIN PLATFORM (AS DESCRIBED HEREIN) OR DEVELOPERS OF THE WORLDCOIN PLATFORM IN ANY CAPACITY WHATSOEVER, ANY DISTRIBUTOR/VENDOR OF WLD TOKENS (THE **DISTRIBUTOR**), NOR ANY SERVICE PROVIDER SHALL BE LIABLE FOR ANY KIND OF DIRECT OR INDIRECT DAMAGE OR LOSS WHATSOEVER WHICH YOU MAY SUFFER IN CONNECTION WITH ACCESSING THIS WHITEPAPER, THE WEBSITE AT HTTPS://WORLDCOIN.ORG (THE **WEBSITE**) OR ANY OTHER WEBSITES OR MATERIALS PUBLISHED BY THE FOUNDATION.

Crypto Products

Crypto products can be highly risky and their regulatory treatment is unsettled in many jurisdictions. There may be no regulatory recourse for any loss from transactions in WLD. Any value ascribed to WLD may change quickly and may be lost in its entirety. Further, the technologies comprising the Worldcoin Platform, including the WLD token, are experimental in nature. There is no guarantee that the network will operate as planned. For more information, visit www.worldcoin.org/risks. Holding, buying, or selling WLD may not be permitted where you live, and it is your responsibility to comply with all applicable

laws. Worldcoin (WLD) tokens are not intended to be available to residents of the United States or certain other restricted territories. More details can be found at http://www.worldcoin.org/tos.

As described further below, this document contains forward-looking estimates and statements regarding the intended actions and objectives of the Worldcoin Foundation and the Worldcoin Ecosystem, based largely on current expectations and projections about future events for which the outcome is uncertain. It is therefore subject to a number of known and unknown risks, including those described at www.worldcoin.org/risks, that could cause the actual outcomes to differ materially from what is expressed or implied herein. Readers are cautioned not to put undue reliance on these future-looking estimates and statements. The content of this document speaks only as of the date thereof.

Nature of the Whitepaper

The Whitepaper and the Website are intended for general informational purposes and community discussion only and do not constitute a prospectus, an offer document, an offer of securities, a solicitation for investment, or any offer to sell any product, item or asset (whether digital or otherwise). Nothing contained in the Whitepaper or the Website is or may be relied upon as a promise, representation or undertaking as to the future performance of the Worldcoin Platform. The information herein may not be exhaustive and does not imply any element of a contractual relationship commitment in relation to the acquisition of WLD Token, and no virtual currency or other form of payment is to be accepted on the basis of the Whitepaper or the Website. There is no assurance as to the accuracy or completeness of such information and no representation, warranty or undertaking is or purported to be provided as to the accuracy or completeness of such information. Nothing contained in the Whitepaper or the Website is or may be relied upon as a promise, representation or undertaking as to the future performance of the Worldcoin Platform. Any agreement between the Distributor (or any third party) and you, in relation to any sale, purchase, or other distribution or transfer of WLD Token, is to be governed only by the separate terms and conditions of such agreement, and such

agreement must be read together with the Whitepaper. Where the Whitepaper or the Website includes information that has been obtained from third party sources, the Foundation, the Distributor, their respective affiliates and/or the Worldcoin Ecosystem have not independently verified the accuracy or completion of such information. Further, you acknowledge that circumstances may change and that the Whitepaper or the Website may become outdated as a result; and neither the Foundation nor the Distributor is under any obligation to update or correct this document in connection therewith.

The information set out in the Whitepaper and the Website is for community discussion only and is not legally binding. No person is bound to enter into any contract or binding legal commitment in relation to the acquisition of any WLD token, and no virtual currency or other form of payment is to be accepted on the basis of the Whitepaper or the Website. Any agreement governing the sale or acquisition of WLD tokens shall be governed by a separate set of Terms of Service, available at www.worldcoin.org/tos. The Terms of Service must be read together with the Whitepaper and further information available at www.worldcoin.org/risks. In the event of any inconsistencies between the Terms of Service and the Whitepaper or the Website, the Terms of Service shall prevail.

Token Features

The native digital cryptographically-secured cryptocurrency of the Worldcoin Platform (**WLD Token**) is a transferable representation of attributed functions specified in the protocol/code of the Worldcoin Platform, designed to play a major role in the functioning of the ecosystem on the Worldcoin Platform, and intended to be used solely as the primary utility and future governance token on the platform. The goal of introducing WLD Token is to provide a convenient and secure mode of payment and settlement between participants who interact within the ecosystem on the Worldcoin Platform, and it is not, and not intended to be, a medium of exchange accepted by the public (or a section of the public) as payment for goods or services or for the discharge of a debt; nor is it designed or intended to be used by any person as payment for any goods or services whatsoever that are not exclusively provided by the issue. WLD Token may only be utilized on the Worldcoin Platform, and ownership of WLD Token carries no rights, express or implied,

other than the right to use WLD Token as a means to enable usage of and interaction within the Worldcoin Platform.

Deemed Representations and Warranties

By accessing the Whitepaper or the Website (or any part thereof), you shall be deemed to represent and warrant to the Foundation, the Distributor, their respective affiliates, and the Worldcoin Ecosystem as follows:

- in any decision to receive and/or purchase any WLD Token, you shall not rely on any statement set out in the Whitepaper or the Website;
- you will and shall at your own expense ensure compliance with all laws, regulatory requirements and restrictions applicable to you (as the case may be);
- you acknowledge, understand and agree that WLD Token may have no value, there is no guarantee or representation of value or liquidity for WLD Token, and WLD Token is not an investment product including for any speculative investment;
- WLD tokens may not always be transferable or liquid;
- WLD tokens may not be exchangeable against any goods or services contemplated in the Whitepaper, especially in case of failure or discontinuation of the project;
- none of the Foundation, the Distributor, their respective affiliates, and/or the Worldcoin Ecosystem members shall be responsible for or liable for the value of WLD Token, the transferability and/or liquidity of WLD Token and/or the availability of any market for WLD Token through third parties or otherwise; and
- you acknowledge, understand and agree that you are not eligible to purchase any WLD Token if you are a citizen, national, resident (tax or otherwise), domiciliary and/or green card holder of a geographic area or country (i) where it is likely that the sale of WLD Token would be construed as the sale of a security, financial service or investment product and/or (ii) where participation in token sales is prohibited by applicable law, decree, regulation, treaty, or administrative act; and to this effect you agree to provide all such identity verification document when requested in order for the relevant checks to be carried out.

The Foundation disclaims all representations, warranties or undertakings to any entity or person (including without limitation warranties as to the accuracy, completeness, timeliness or reliability of the contents of the Whitepaper or the Website, or any other materials published by the Foundation or the Distributor). To the maximum extent permitted by law, the Foundation, the Distributor, their respective affiliates and service providers shall not be liable for any indirect, special, incidental, consequential or other losses of any kind, in tort, contract or otherwise (including, without limitation, any liability arising from default or negligence on the part of any of them, or any loss of revenue, income or profits, and loss of use or data) arising from the use of the Whitepaper or the Website, or any other materials published, or its contents (including without limitation any errors or omissions) or otherwise arising in connection with the same. Prospective purchasers of the WLD Token should carefully consider and evaluate all risks and uncertainties (including financial and legal risks and uncertainties) associated with the WLD Token sale, the Foundation, the Distributor and the Worldcoin Ecosystem.

Disclaimers Relating to the WLD Token

It is expressly highlighted that WLD Token:

- does not have any tangible or physical manifestation, and does not have any intrinsic value (nor does any person make any representation or give any commitment as to its value), and may lose its value in part or in full;
- is non-refundable and cannot be exchanged for cash (or its equivalent value in any other virtual currency) or any payment obligation by the Foundation, the Distributor or any of their respective affiliates, and may not always be transferrable or liquid;
- does not represent or confer on the token holder any right of any form with respect to the Foundation, the Distributor (or any of their respective affiliates), or its revenues or assets, including without limitation any right to receive future dividends, revenue, shares, ownership right or stake, share or security, any voting, distribution, redemption, liquidation, proprietary (including all forms of intellectual property or license rights), right to receive accounts, financial statements or other financial data, the right to requisition or participate in shareholder meetings, the right to nominate a director, or

other financial or legal rights or equivalent rights, or intellectual property rights or any other form of participation in or relating to the Worldcoin Platform, the Foundation, the Distributor and/or their service providers;
- does not entitle token holders to any promise of fees, dividends, revenue, profits or investment returns, and are not intended to constitute securities in any relevant jurisdiction;
- is not intended to represent any rights under a contract for differences or under any other contract the purpose or pretended purpose of which is to secure a profit or avoid a loss;
- may not be exchangeable against the good or service described herein, especially in case of failure or discontinuation of the Worldcoin project;
- is not intended to be a representation of money (including electronic money), security, commodity, bond, debt instrument, unit in a collective investment scheme or any other kind of financial instrument or investment;
- is not a loan to the Foundation, the Distributor or any of their respective affiliates, is not intended to represent a debt owed by the Foundation, the Distributor or any of their respective affiliates, and there is no expectation of profit; and
- does not provide the token holder with any ownership or other interest in the Foundation, the Distributor or any of their respective affiliates.

Informational Purposes Only

The project roadmap in the Whitepaper is being shared in order to outline the current status of Worldcoin as well as some of the plans of the Worldcoin Ecosystem and is provided solely for informational purposes and does not constitute any binding commitment. Please do not rely on this information in making purchasing decisions because ultimately, further development, release, and timing of any products, features or functionality remains at the sole discretion of the Foundation, the Distributor or their respective affiliates, and is subject to change. Further, the Whitepaper or the Website may be amended or replaced from time to time. There are no obligations to update the Whitepaper or the Website, or to provide recipients with access to any information beyond what is provided herein.

Regulatory Approval

No regulatory authority has examined or approved, whether formally or informally, of any of the information set out in the Whitepaper or the Website. No such action or assurance has been or will be taken under the laws, regulatory requirements or rules of any jurisdiction. The publication, distribution or dissemination of the Whitepaper or the Website does not imply that the applicable laws, regulatory requirements or rules have been complied with. Worldcoin is solely responsible for the content of this Whitepaper. This Whitepaper has not been reviewed or approved by any competent authority in any Member State of the European Union.

Cautionary Note on Forward-Looking Statements

All statements contained herein, statements made in press releases or in any place accessible by the public and oral statements that may be made by the Foundation, the Distributor and/or the Worldcoin Ecosystem, may constitute forward-looking statements (including statements regarding intent, belief or current expectations with respect to market conditions, business strategy and plans, financial condition, specific provisions and risk management practices). You are cautioned not to place undue reliance on these forward-looking statements given that these statements involve known and unknown risks, uncertainties and other factors that may cause the actual future results to be materially different from that described by such forward-looking statements, and no independent third party has reviewed the reasonableness of any such statements or assumptions. These forward-looking statements are applicable only as of the date indicated in the Whitepaper, and the Foundation, the Distributor as well as the Worldcoin Ecosystem expressly disclaim any responsibility (whether express or implied) to release any revisions to these forward-looking statements to reflect events after such date.

English Language

The Whitepaper and the Website may be translated into a language other than English for reference purpose only and in the event of conflict or ambiguity between the English language version and translated versions of the Whitepaper or the Website, the English language versions shall prevail. You acknowledge that you have read and understood the English language version of the Whitepaper and the Website.