



Personhood Credentials

How do you prove you're human in the era of AI deep fakes

Srikanth Nadhamuni

Founder CTO Aadhaar



"On the Internet, nobody knows you're a dog."

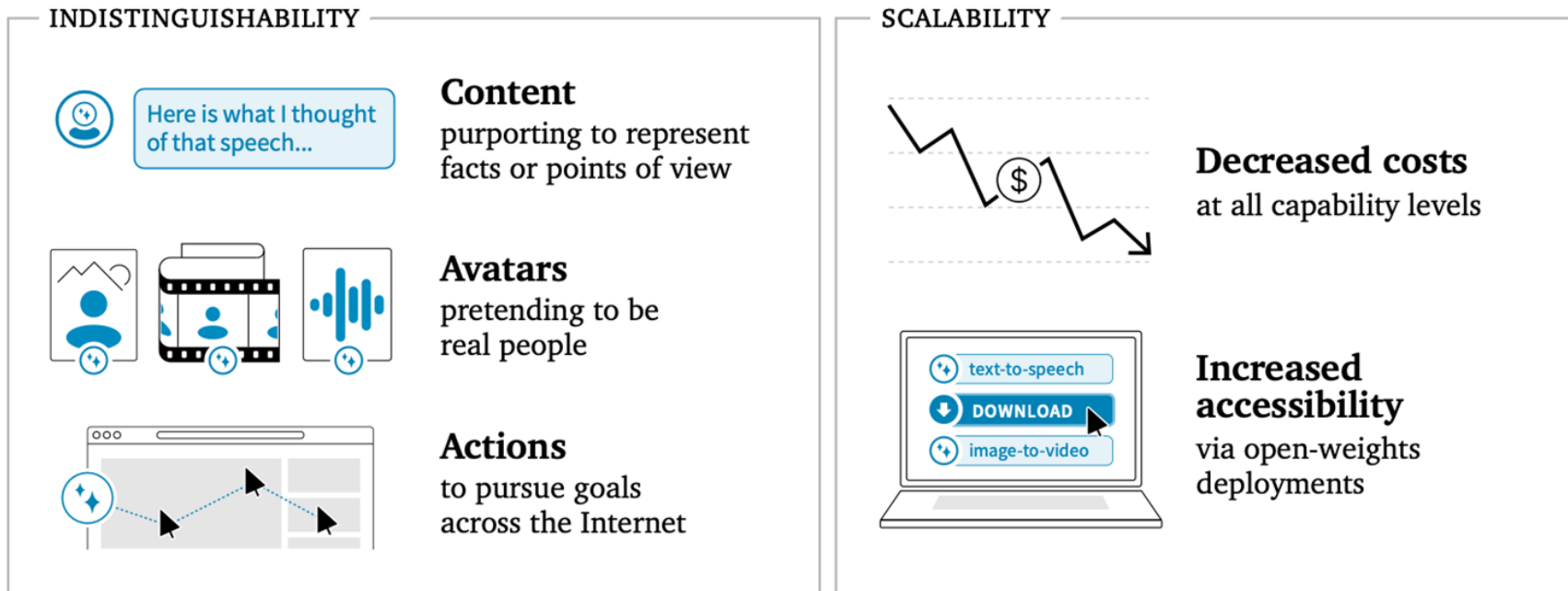
©The New Yorker Collection 1993 Peter Steiner
From cartoonbank.com. All rights reserved.

Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online

Steven Adler,^{*†1} Zoë Hitzig,^{*†1,2} Shrey Jain,^{*†3} Catherine Brewer,^{*4} Wayne Chang,^{*5} Renée DiResta,^{*25} Eddy Lazzarin,^{*6}
Sean McGregor,^{*7} Wendy Seltzer,^{*8} Divya Siddarth,^{*9} Nouran Soliman,^{*10} Tobin South,^{*10} Connor Spelliscy,^{*11}
Manu Sporny,^{*12} Varya Srivastava,^{*4} John Bailey,¹³ Brian Christian,⁴ Andrew Critch,¹⁴ Ronnie Falcon,¹⁵ Heather Flanagan,²⁵
Kim Hamilton Duffy,¹⁶ Eric Ho,¹⁷ Claire R. Leibowicz,¹⁸ Srikanth Nadhamuni,¹⁹ Alan Z. Rozenshtein,²⁰
David Schnurr,¹ Evan Shapiro,²¹ Lacey Strahm,¹⁵ Andrew Trask,^{4,15} Zoe Weinberg,²² Cedric Whitney,²³ Tom Zick²⁴

- ¹OpenAI, ²Harvard Society of Fellows, ³Microsoft, ⁴University of Oxford, ⁵SpruceID, ⁶a16z crypto,
⁷UL Research Institutes, ⁸Tucows, ⁹Collective Intelligence Project, ¹⁰Massachusetts Institute of Technology,
¹¹Decentralization Research Center, ¹²Digital Bazaar, ¹³American Enterprise Institute,
¹⁴Center for Human-Compatible AI, University of California, Berkeley, ¹⁵OpenMined,
¹⁶Decentralized Identity Foundation, ¹⁷Goodfire, ¹⁸Partnership on AI, ¹⁹eGovernments Foundation,
²⁰University of Minnesota Law School, ²¹Mina Foundation, ²²ex/ante, ²³School of Information, University of California, Berkeley,
²⁴Berkman Klein Center for Internet & Society, Harvard University, ²⁵Independent Researcher

AI Trends that can contribute to online frauds



- *First, AI bots are increasingly **indistinguishable** from people online.*
- *Second, AI is becoming increasingly **scalable**—both more affordable and accessible,*

Current solutions for countering AI-powered deception

Strategy to counter scalable AI-powered deception

Main deficits

Behavioral filters based on AI lacking certain abilities

e.g., CAPTCHAs, JavaScript browser challenges, anomaly detection

Not robust to highly capable AI

Economic barriers that make AI deception less profitable

e.g., paid subscriptions, credit card verification

Not inclusive

AI-detection tools to identify synthetic content

e.g., watermarking, fingerprinting, metadata provenance

Not robust to highly capable AI

Appearance- and document-based humanness verification

e.g., selfie checks with ID, live video calls

Not robust to highly capable AI

Not privacy preserving

Existing digital and hardware identifiers controlled by humans

e.g., phone numbers, email address, hardware security keys

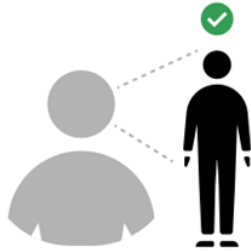
Not scarce



What AI Cannot Do (yet!!)

OFFLINE COMPONENT

Pass as a person in the real-world



CRYPTOGRAPHY

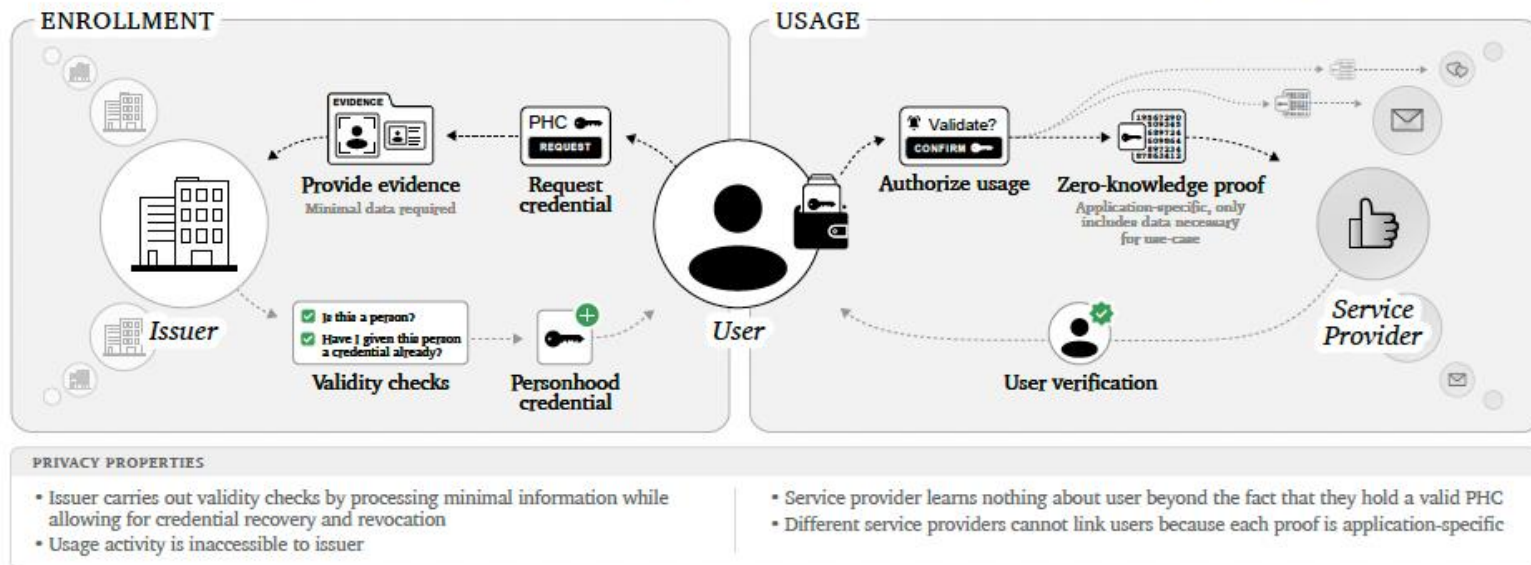
Forge advanced cryptography



Cryptography relies on computationally hard mathematical problems, such as the factoring of very large numbers. There are not any known methods of efficiently solving certain such problems, whether by a human or an AI system.

Personhood Credentials

Privacy-preserving enrollment and usage of personhood credentials (PHCs)





Proof-of-Personhood

- Biometric Methods - Aadhaar, Worldcoin etc.
 - Advantage - everybody has it
- Social-graph or Web-of-Trust
 - These systems can struggle to confirm uniqueness:
 - Verification Party - BrightID, Idena
 - Existing Human Vouches for new user - Proof of Humanity, Circles
- Government ID Based (Breeder Documents)

**NIST considers biometric verification to be its highest strength of evidence*



Proof-of-Personhood with Uniqueness

- Biometric Methods - **Strong**
 - Perform 1:N De-duplication
 - Strong level of assurance
- Social-graph or Web-of-Trust - **Weak**
 - These systems can struggle to confirm uniqueness
- Government ID Based (Breeder Documents)
 - These could also be weak wrt Uniqueness

**NIST considers biometric verification to be its highest strength of evidence*

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

3 Key Benefits of PHC Systems

Reduce impact of sockpuppeting



Mitigate bot attacks



Verify delegation to AI agents



Sockpuppeting—*Creating fake online identities or using multiple accounts to deceive:*

- *Manipulating perception of public political opinion on social media*
- *Propping up (or attacking) the reputation of businesses or individuals*
- *Carrying out scams on digital marketplaces*



3 Key Benefits of PHC Systems

SUMMARY

1 The problem of scalable deception online

With access to highly capable AI, malicious actors can orchestrate more effective deceptive schemes:

Indistinguishability:

AI capable of generating human-like content, appearances, and actions



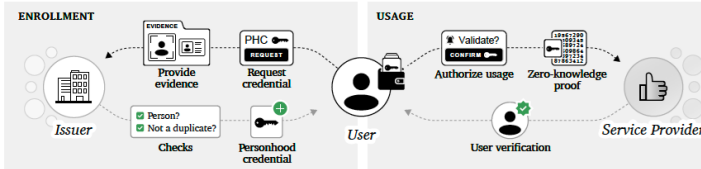
Scalability:

Decreased costs and increased accessibility



2 PHCs empower users and services to counter deception

Adding options to verify with personhood credentials (PHCs) could enhance users' ability to protect their privacy and services' ability to counter deception. They work as follows:



3 Three key benefits of PHC systems

PHC systems as we have defined them offer the following key benefits:

Reduce impact of sockpuppeting

Mitigate bot attacks

Verify delegation to AI agents

4 Potential challenges for PHC systems

PHCs' impacts should be carefully managed in the following four areas:

Equitable access to digital services.

Free expression supported by strong privacy measures.

Checks on power and influence over digital services.

Robustness to attack and error.

5 Next steps for consideration

We offer next steps for public consideration in two main areas:

Adapt existing digital identity systems

- A1 Reexamine standards for remote identity verification and authentication.
- A2 Study the impact and prevalence of deceptive accounts on major communications platforms.
- A3 Establish norms and standards to govern agentic AI users of the Internet.

Prioritize personhood credentials

- P1 Invest in the development and piloting of personhood credentialing systems.
- P2 Encourage adoption of personhood credentials.

Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online

Steven Adler,^{*1} Zoë Hitzig,^{*1,2} Shrey Jain,^{*13} Catherine Brewer,^{*4} Wayne Chang,^{*5} Renée DiResta,^{*25} Eddy Lazzarin,^{*6}
Sean McGregor,^{*7} Wendy Seltzer,^{*8} Divya Siddarth,^{*9} Nouran Soliman,^{*10} Tobin South,^{*10} Connor Spelliscy,^{*11}
Manu Sporny,^{*12} Varya Srivastava,^{*4} John Bailey,^{*13} Brian Christian,^{*4} Andrew Critch,^{*14} Ronnie Falcon,^{*15} Heather Flanagan,^{*25}
Kim Hamilton Duffy,^{*16} Eric Ho,^{*17} Claire R. Leibowicz,^{*18} Srikanth Nadhamuni,^{*19} Alan Z. Rozenshtein,^{*20}
David Schnurr,^{*1} Evan Shapiro,^{*21} Lacey Strahm,^{*15} Andrew Trask,^{*4,15} Zoe Weinberg,^{*22} Cedric Whitney,^{*23} Tom Zick^{*24}

¹OpenAI, ²Harvard Society of Fellows, ³Microsoft, ⁴University of Oxford, ⁵SpruceID, ⁶a16z crypto,

⁷UL Research Institutes, ⁸Tucows, ⁹Collective Intelligence Project, ¹⁰Massachusetts Institute of Technology,

¹¹Decentralization Research Center, ¹²Digital Bazaar, ¹³American Enterprise Institute,

¹⁴Center for Human-Compatible AI, University of California, Berkeley, ¹⁵OpenMined,

¹⁶Decentralized Identity Foundation, ¹⁷Goodfire, ¹⁸Partnership on AI, ¹⁹eGovernments Foundation,

²⁰University of Minnesota Law School, ²¹Mina Foundation, ²²ex/ante, ²³School of Information, University of California, Berkeley,

²⁴Berkman Klein Center for Internet & Society, Harvard University, ²⁵Independent Researcher

August 2024



Scan code to read Paper



Thank You

Srikanth.Nadhamuni@gmail.com