# Q&A: Could 'personhood credentials' protect people against digital imposters?

August 16 2024, by Adam Zewe

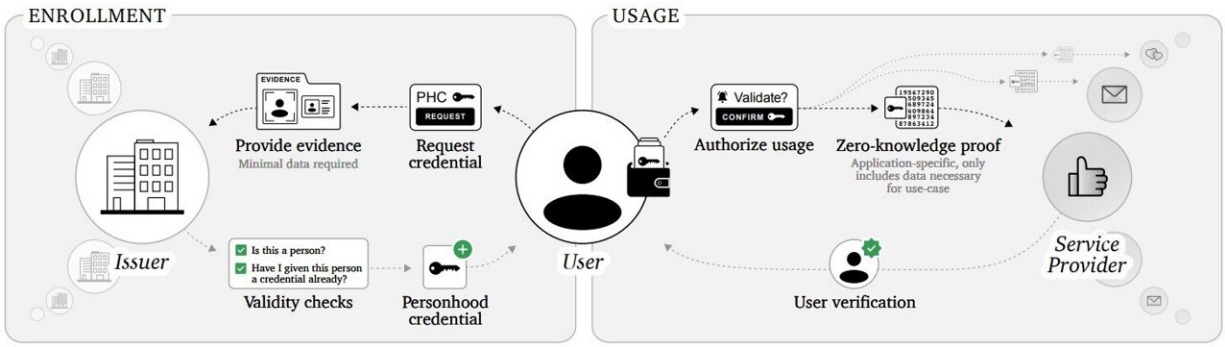


Illustration of enrollment and usage of a personhood credential. Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2408.07892

As artificial intelligence agents become more advanced, it could become increasingly difficult to distinguish between AI-powered users and real humans on the internet. In a new white paper, researchers from MIT, OpenAI, Microsoft, and other tech companies and academic institutions propose the use of personhood credentials, a verification technique that enables someone to prove they are a real human online, while preserving their privacy.

The paper is published on the *arXiv* preprint server.

MIT News spoke with two co-authors of the paper, Nouran Soliman, an

electrical engineering and computer science graduate student, and Tobin South, a graduate student in the Media Lab, about the need for such credentials, the risks associated with them, and how they could be implemented in a safe and equitable way.

## Why do we need personhood credentials?

Tobin South: AI capabilities are rapidly improving. While a lot of the public discourse has been about how chatbots keep getting better, sophisticated AI enables far more capabilities than just a better ChatGPT, like the ability of AI to interact online autonomously. AI could have the ability to create accounts, post content, generate fake content, pretend to be human online, or algorithmically amplify content at a massive scale. This unlocks a lot of risks. You can think of this as a "digital imposter" problem, where it is getting harder to distinguish between sophisticated AI and humans. Personhood credentials are one potential solution to that problem.

Nouran Soliman: Such advanced AI capabilities could help bad actors run large-scale attacks or spread misinformation. The internet could be filled with AIs that are resharing content from real humans to run disinformation campaigns. It is going to become harder to navigate the internet, and social media specifically. You could imagine using personhood credentials to filter out certain content and moderate content on your social media feed or determine the trust level of information you receive online.

## What is a personhood credential, and how can you ensure such a credential is secure?

South: Personhood credentials allow you to prove you are human without revealing anything else about your identity. These credentials let you

take information from an entity like the government, who can guarantee you are human, and then through privacy technology, allow you to prove that fact without sharing any sensitive information about your identity.

South: To get a personhood credential, you are going to have to show up in person or have a relationship with the government, like a tax ID number. There is an offline component. You are going to have to do something that only humans can do. AIs can't turn up at the DMV, for instance. And even the most sophisticated AIs can't fake or break cryptography. So, we combine two ideas—the security that we have through cryptography and the fact that humans still have some capabilities that AIs don't have—to make really robust guarantees that you are human.

Soliman: But personhood credentials can be optional. Service providers can let people choose whether they want to use one or not. Right now, if people only want to interact with real, verified people online, there is no reasonable way to do it. And beyond just creating content and talking to people, at some point AI agents are also going to take actions on behalf of people. If I am going to buy something online, or negotiate a deal, then maybe in that case I want to be certain I am interacting with entities that have personhood credentials to ensure they are trustworthy.

South: Personhood credentials build on top of an infrastructure and a set of security technologies we've had for decades, such as the use of identifiers like an email account to sign into online services, and they can complement those existing methods.

## What are some of the risks associated with personhood credentials, and how could you reduce those risks?

Soliman: One risk comes from how personhood credentials could be implemented. There is a concern about concentration of power. Let's say one specific entity is the only issuer, or the system is designed in such a way that all the power is given to one entity. This could raise a lot of concerns for a part of the population—maybe they don't trust that entity and don't feel it is safe to engage with them. We need to implement personhood credentials in such a way that people trust the issuers and ensure that people's identities remain completely isolated from their personhood credentials to preserve privacy.

South: If the only way to get a personhood credential is to physically go somewhere to prove you are human, then that could be scary if you are in a sociopolitical environment where it is difficult or dangerous to go to that physical location. That could prevent some people from having the ability to share their messages online in an unfettered way, possibly stifling free expression. That's why it is important to have a variety of issuers of personhood credentials, and an open protocol to make sure that freedom of expression is maintained.

Soliman: Our paper is trying to encourage governments, policymakers, leaders, and researchers to invest more resources in personhood credentials. We are suggesting that researchers study different implementation directions and explore the broader impacts personhood credentials could have on the community. We need to make sure we create the right policies and rules about how personhood credentials should be implemented.

South: AI is moving very fast, certainly much faster than the speed at which governments adapt. It is time for governments and big companies to start thinking about how they can adapt their digital systems to be ready to prove that someone is human, but in a way that is privacy-preserving and safe, so we can be ready when we reach a future where AI has these advanced capabilities.

**More information:** Steven Adler et al, Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online, *arXiv* (2024). DOI: 10.48550/arxiv.2408.07892