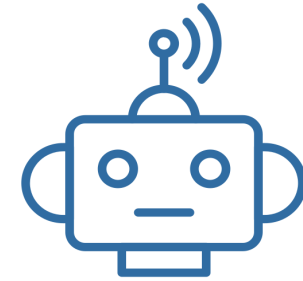


indicio

A Trusted Copilot



Using decentralized identity to manage
an AI virtual assistant

Trevor Butterworth, VP Communications & Government, Indicio
Karl Schweppe, Head of Innovation, Bay Tree Ventures

April 12, 2023; updated January 12, 2024

Copyright Indicio PBC 2023 & 2024

A Trusted Copilot

ChatGPT plugins change the scope for online interaction. Here, we look at how decentralized identity and verifiable credentials can address some of the privacy and security challenges created by an AI-powered virtual assistant.

RAPID PROGRESS in natural language processing capabilities make it reasonable to assume that we will soon be able to assign tasks to AI “agents” or virtual assistants.

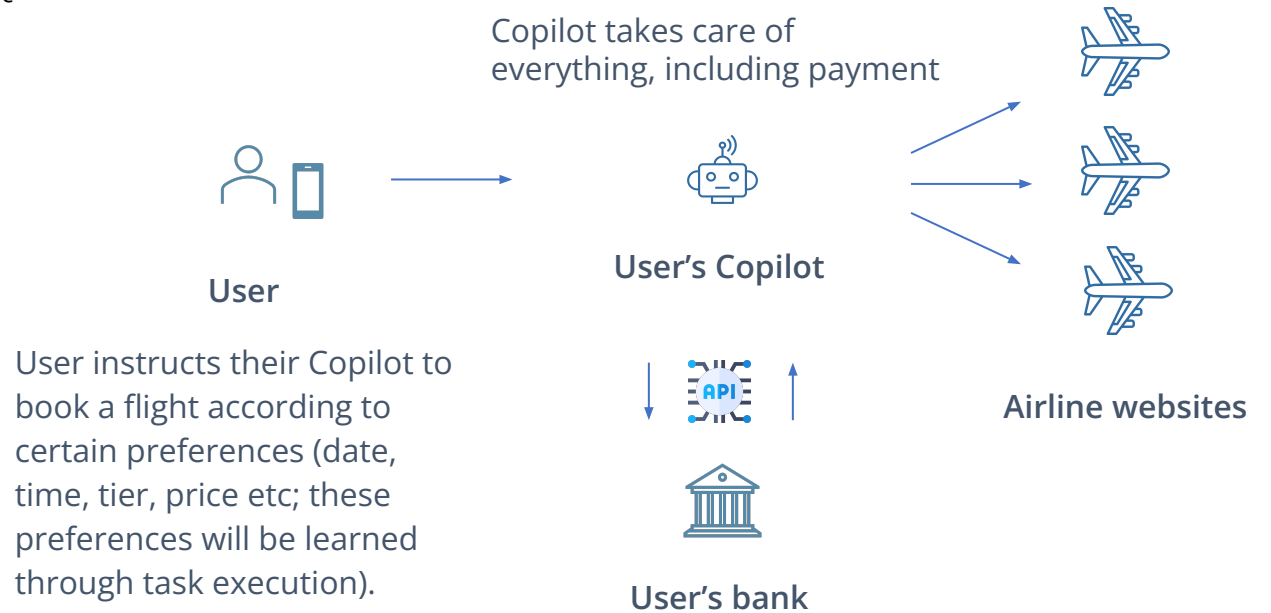
The release of OpenAI’s ChatGPT plugins mean that ChatGPT can access external data sources, acquire additional functionality, and perform intelligible tasks; and, as AI “learns” from training on new data, it is reasonable to assume that a ChatGPT agent or equivalent will be able to perform an increasing range of functions as it learns more about the behavior and preferences of its user while being rewarded for successfully performing more and more tasks to the user’s requirements.

Should this functionality become available, many of us will likely initiate a digitally transparent and trusted delegation of daily professional and personal tasks to AI-driven virtual assistants — but if and only if there is a safe and secure way to implement and manage person-machine trust.

The following scenario envisages asking a ChatGPT-like interface or service to book a flight. This “Copilot” will communicate to all the required services and take care of everything else.

A Trusted Copilot

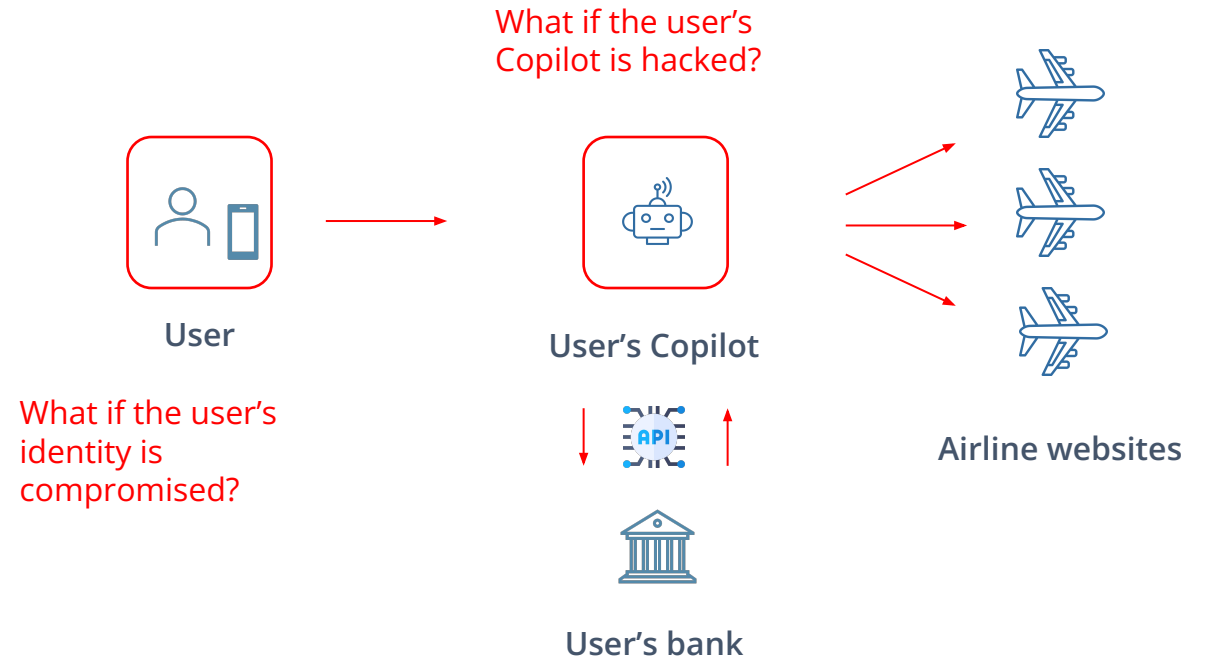
Your AI virtual assistant can run any task for you online with the appropriate plugins. This is, at least in theory, a beautifully frictionless process; one that delivers a better experience for everybody.



A Trusted Copilot

But this beautiful, frictionless process creates a dystopian security problem: Do we really want an intelligent agent to have access to our personal data and be able to act on it? What if the agent is hacked—or our access to the agent is compromised by a takeover?

Imagine what a compromised AI virtual assistant could do with access to all your accounts AND a predictive understanding of your behavior and preferences.



A Trusted Copilot

TO ADDRESS the exponential risks from a compromised virtual assistant, we need continuous, mutual verification between the user and their Copilot.

For this to be robust, access to a virtual assistant should be passwordless, identity verification should require no sharing of personal data, and the verification process should be through peer-to-peer communication using authenticated encryption.

This means that when the Copilot receives a request, it can only have come from the Copilot's user. Similarly, a Copilot response must only be able to come from the user's Copilot.

Open-source verifiable credential technology using decentralized identifiers, digitally signed verifiable credentials, distributed ledgers, and decentralized identifier communications protocol (DIDComm) provides a solution to this challenge.

Additionally, Decentralized Ecosystem Governance (DEGov),

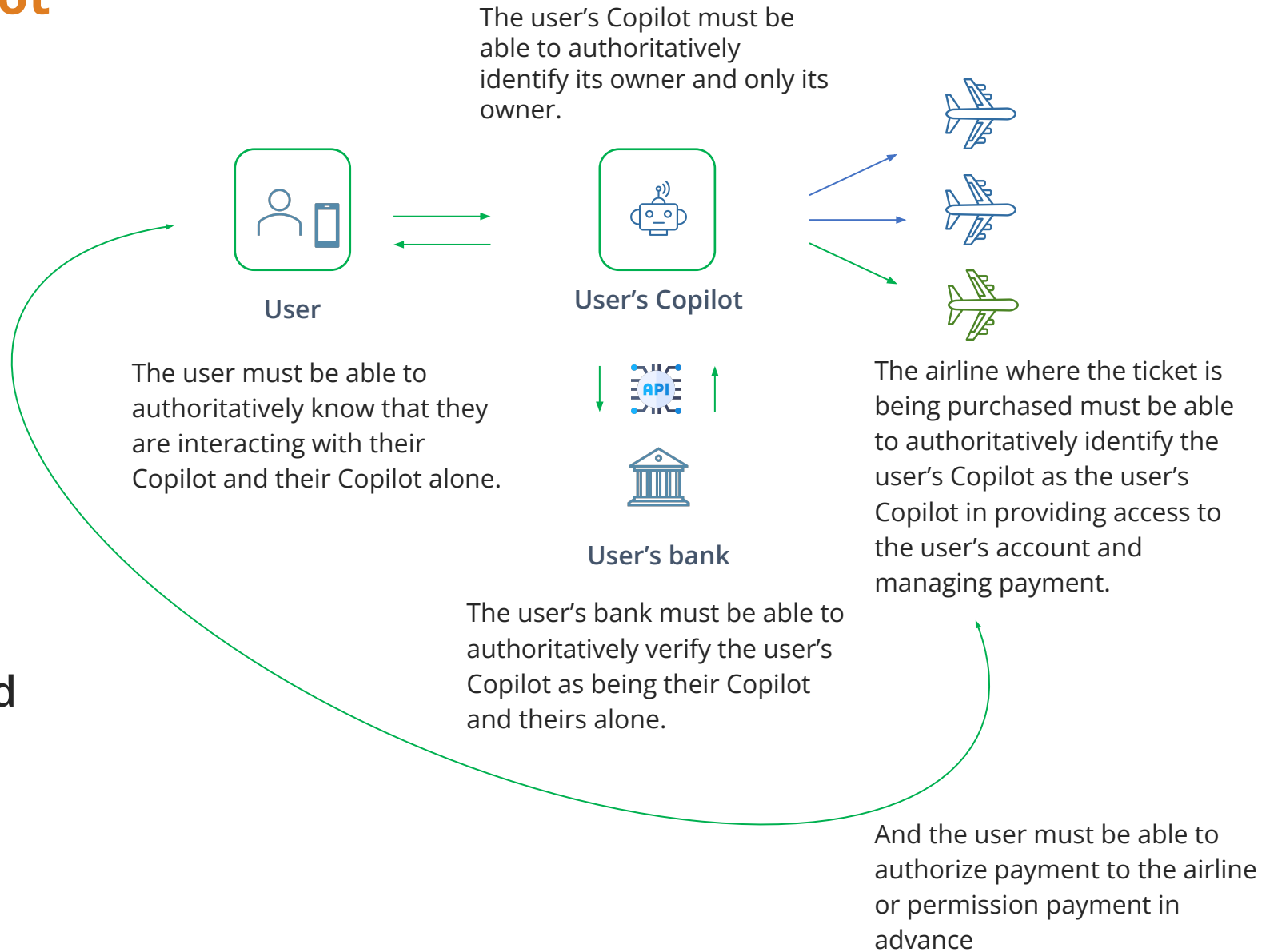
a machine-readable set of governance rules cached in the software for managing verifiable credentials, can coordinate permission for bank payment in a similarly secure way.

It may be that the user's decision to purchase (or permission to purchase) triggers the presentation of the user's bank account credential to the airline and the automatic population of the requisite fields upon verification by the airline.

But a simpler route for payment would be through delegating authority to the Copilot itself in the form of pre-authorized payments. This authority can be intentionally narrow so that payment could only be to a specific kind of vendor for a specific price range within a specific date range.

In the event of the Copilot being compromised, these parameters would reduce the attack surface. Pre-authorization would be a key that could only be used once.

A Trusted Copilot



Who and what needs to be verified

A Trusted Copilot

SOLUTION: A TRUSTED DIGITAL ECOSYSTEM.

Decentralized identity technology provides verifiable data; Decentralized Ecosystem Governance provides permission for information flows; DIDComm provides peer-to-peer communication with authenticated encryption.

1. User creates their Copilot.

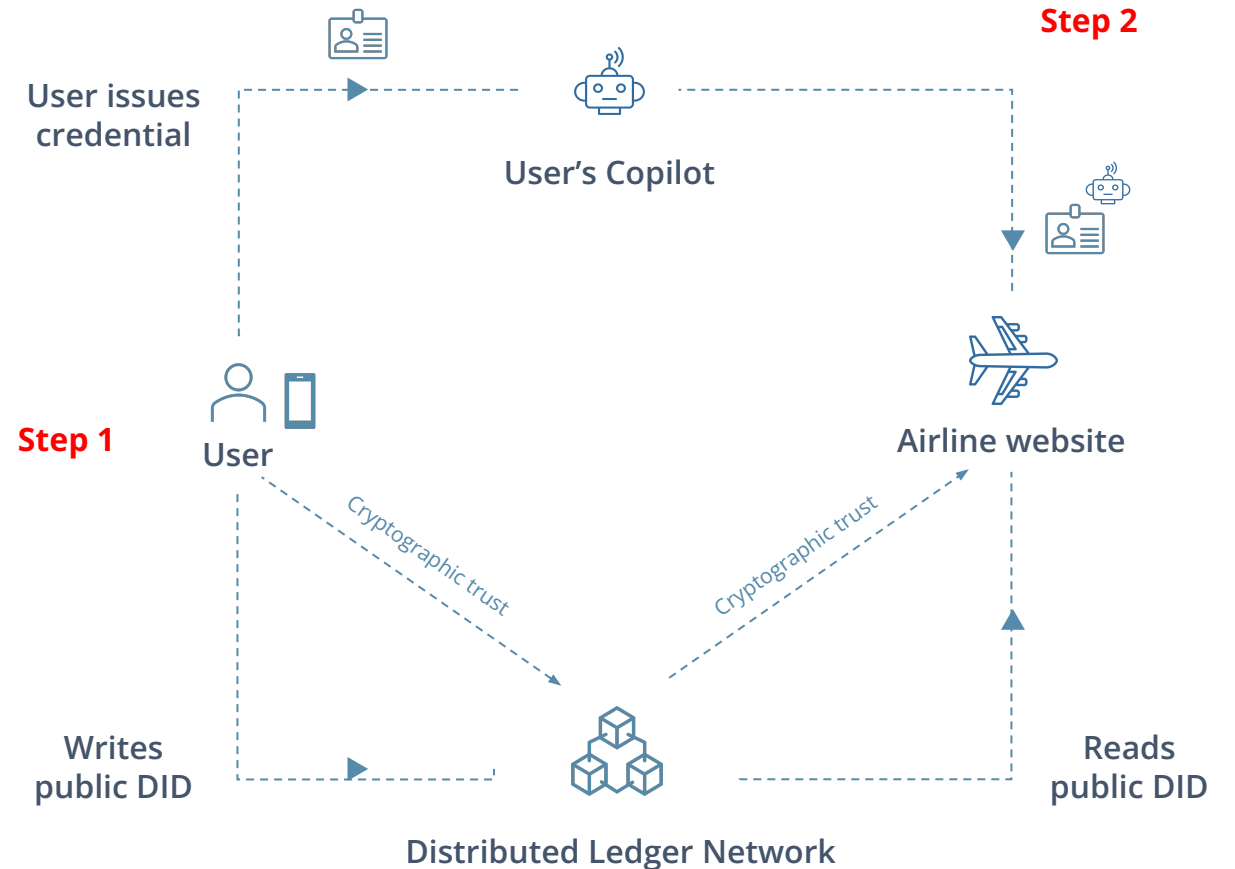
In doing so, they create an identity credential for the Copilot, which the Copilot holds.

To do this, the user's software writes a public DID (Decentralized Identifier) for the Copilot to a distributed ledger and issues a credential to the Copilot.

The Copilot is now discoverable and verifiable as the user's Copilot.

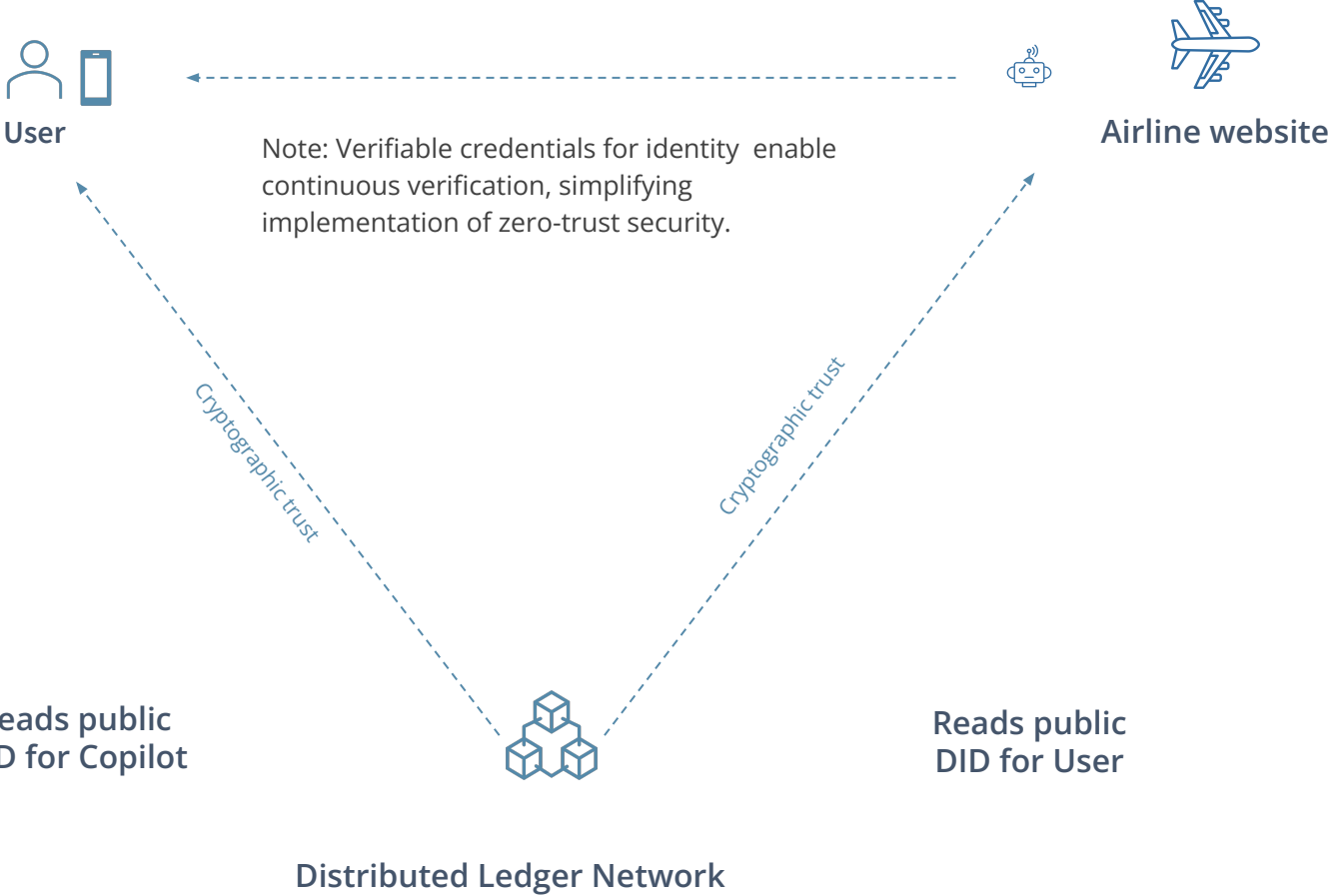
2. Copilot presents its credential to book flight on behalf of the User.

Airline can authoritatively verify that the Copilot represents the user without the Copilot needing to share any of the User's personal data.



A Trusted Copilot

Step 3



3. Copilot presents result and asks for permission to pay — although this step could be omitted through pre-authorized payment (see next slide).

A Trusted Copilot

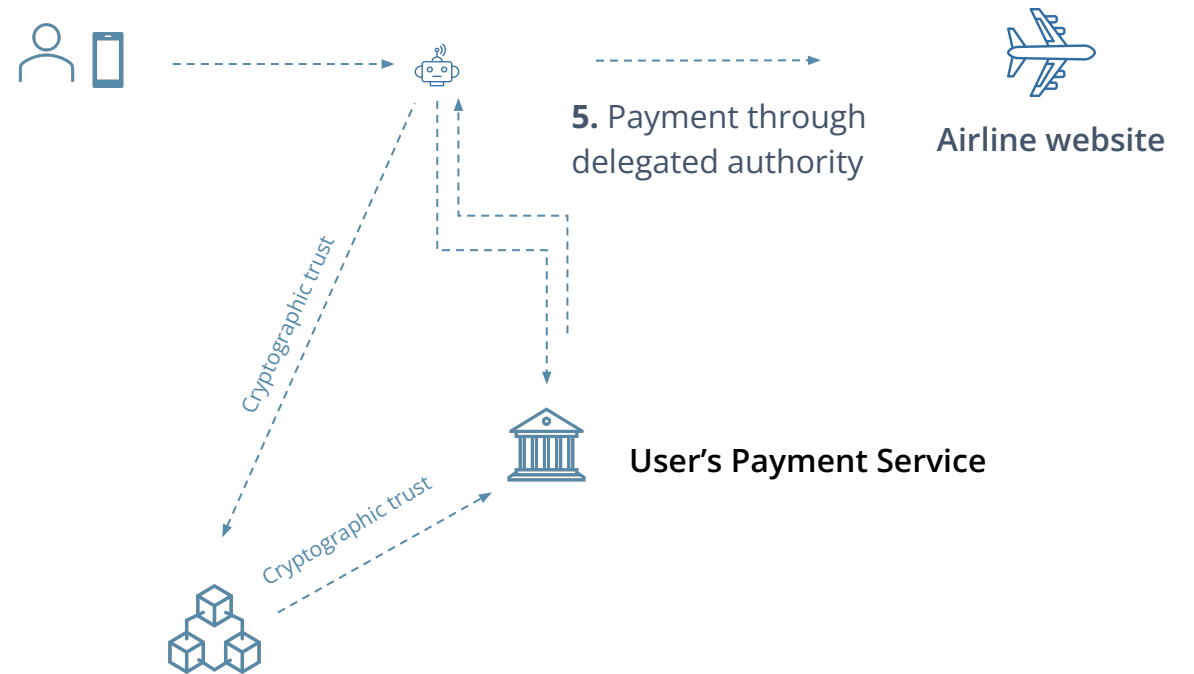
NOT ILLUSTRATED: The User's Payment Service will verify it is interacting with the Airline and vice versa. The Airline will verify the Copilot before issuing the ticket for the User (assuming the Copilot will manage the ticket).

This architecture makes the continuous verification required by Zero Trust approaches to security easy.

Payment can be through delegated authority. The Copilot can be pre-authorized to make a payment within a specific range of parameters (price range, dates) using the User's financial details.

Step 4 & 5

4. Permission is given by the User (it could also be pre-authorized via governance)



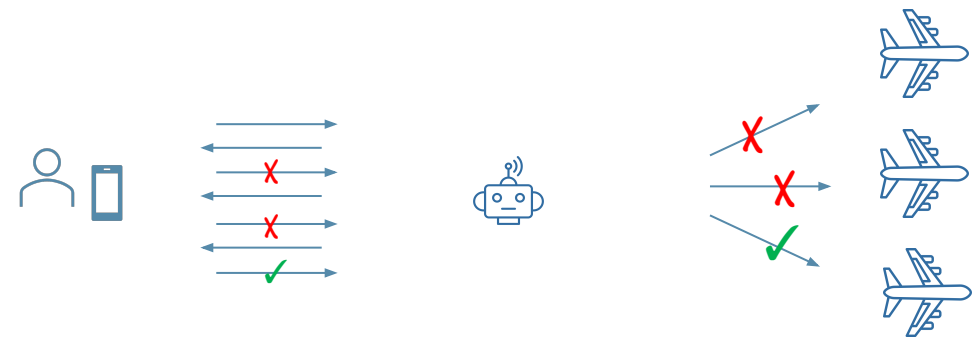
A Trusted Copilot

ACCESS AND DATA LEAKAGE. While DIDComm provides encrypted authentication for peer-to-peer interaction, we must still consider the possibility that a user's AI agent can be hacked. How could we minimize the impact of such a breach?

One answer is to constrain a virtual agent with narrowly defined delegated authority such that it is incapable of performing outside a set of parameters, including vendor, price, date of purchase. This would include limiting how much an AI agent could see, such as payment history (necessary to execute a return), which could be delimited by amount and date range.

Managing privacy and security around an AI agent's learning process is a trickier problem. Presently, ChatGPT becomes better at tasks by learning from new data and being rewarded when it produces the right response.

The nature of assisting a user means learning from the user's data and incorporating that learning into its approach to solving the next request and so on. Such data access creates the risk of data leakage, and it seems likely that this is an area where regulation will be needed.



AI learning and data privacy

An AI virtual assistant is constantly learning from its user's personal data. How is the user's data kept private?

A Trusted Copilot

SOME FINAL THOUGHTS: Interacting with an AI virtual assistant through a centralized identity model presents a substantial systemic risk, and the public loss of confidence in the technology should the AI agent be compromised is likely to be catastrophic given the scope of the agent's access and its capacity to learn. Fear of such a breach may well present an obstacle to adoption or an incentive to overly-restrictive regulation that severely limits the potential benefits of AI.

This position paper posits that decentralized verifiable credentials can solve many of these problems; in particular, DIDComm provides a powerful model for private and secure Person-AI interaction using authenticated encryption.

The vista of possibilities presented by AI agents is rich; but as our technological capacities scale rapidly, so do their risks and consequences. Frictionless, all-compromising, AI-enabled fraud is the corollary of frictionless, multi-tasking virtual agents.

Finally, instead of a virtual domain where we actively engage in digital experiences, it may be that the metaverse — or at least the version that will take hold in the short term — will mostly be a space where virtual assistants engage in business on our behalf, a space where we exercise agency without the need to be consciously present.

Trevor Butterworth is a cofounder of and VP for Communications & Governance for Indicio, a global leader in open-source decentralized identity technology.

Trevor@indicio.tech

Karl Schweppe is Head of Innovation at Bay Tree Ventures, a Digital Product Architecture studio that helps ambitious companies design, build and launch innovative digital businesses.

[LinkedIn](#)

