Digital Trust in the Age of Generative AI

For the Trust of IP Foundation All Members Meeting

August 16, 2023

Wenjing Chu

This presentation slides: https://github.com/wenjing/Digital-Trust-in-the-Age-of-Generative-AI

Wenjing Chu

Wenjing is a Senior Director of Technology Strategy at Futurewei Technologies, Inc. He leads new technology development in the future of computing, trust, and trustworthy AI.

He is a founding board member of the OpenWallet Foundation (OWF) and its Technical Advisory Council (TAC). He is also a founding board member of the Trust of IP Foundation (ToIP). In ToIP, he is a primary author of the Trust Spanning Protocol Specification (TSP) and of the Technology Architecture Specification, while serving as Co-Chair for the Trust Spanning Protocol Task Force and for the AI and Metaverse Task Force. He also previously served in leadership roles for the Linux Foundation's Networking Foundation and the Edge Computing Foundation. Over his career, he had been an entrepreneurial and technical lead in several successful startups including Airespace, Inc. and Sentient Networks, Inc.

What I Will Cover Today

- The Age of Generative AI
- The New Nature of Data
- The New Nature of Digital Trust
- A Way Forward
- Q & A

The Age of Generative AI

What distinguishes Generative AI technology

Degree of bewitchment

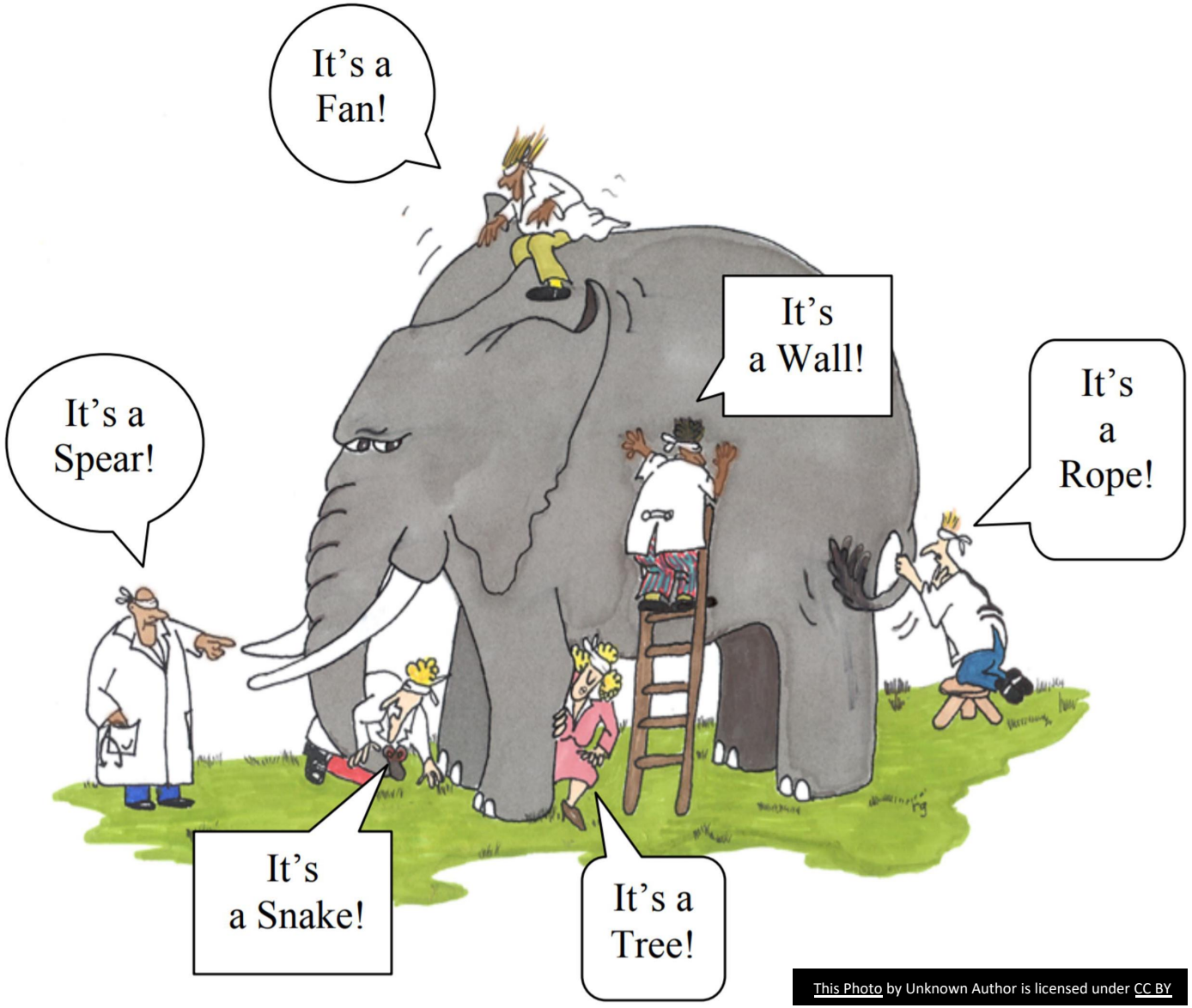
"The limits of my language means the limits of my world."

- Ludwig Wittgenstein

- We have developed a very rich language for all things human, esp. human intelligence.
- We have not *yet* constructed a language for machine intelligence – we have instead *borrowed* from the vocabulary intended for humans.
- Talking about AI in human terms is natural – but always wrong.
- What is the degree of "bewitchment" when we borrow a term or use a metaphor? Huge!

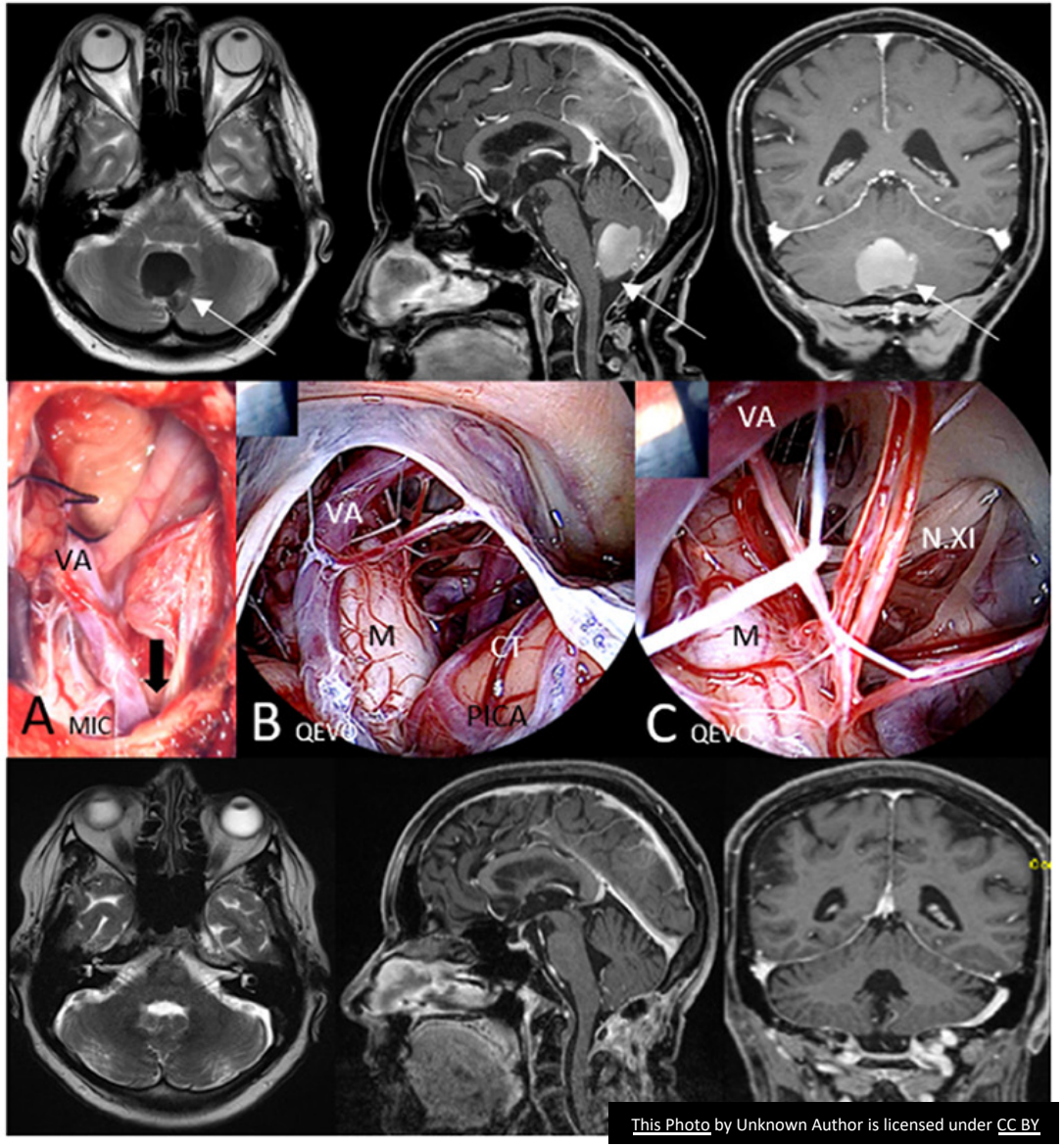
Limits of metaphors

- They are often partially ok but always wrong.
- Unlike an elephant, Generative AI is a rapidly moving target without stable shape.
- We can't ask a person of good sight for an authoritative answer.
- In high dimensional space, distance is a confusing concept.



What we do know

- We do know how it is constructed, precisely (up to recently).
- We can do testing and obtain statistical and empirical evaluations.
- Observing brain neuron cells does not tell us all about their emergent functions, but we should still start with that.

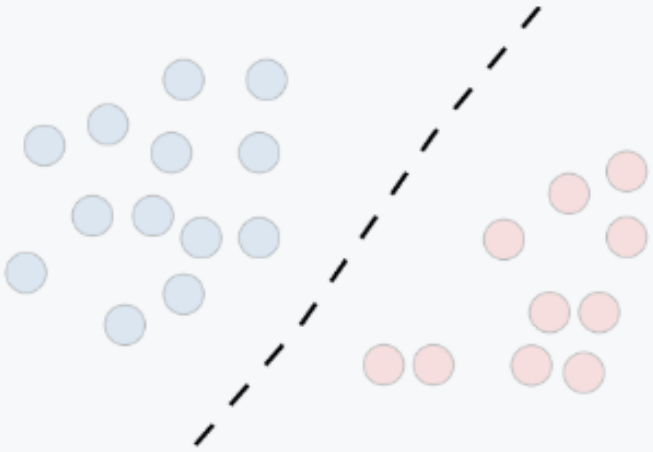
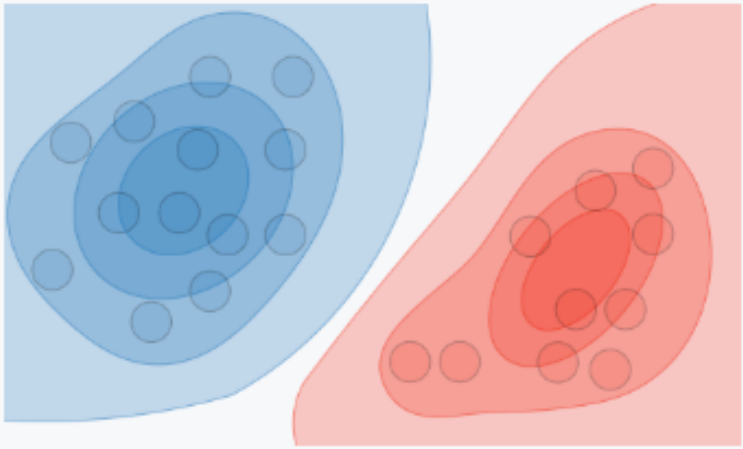


The G, P, T of GPT

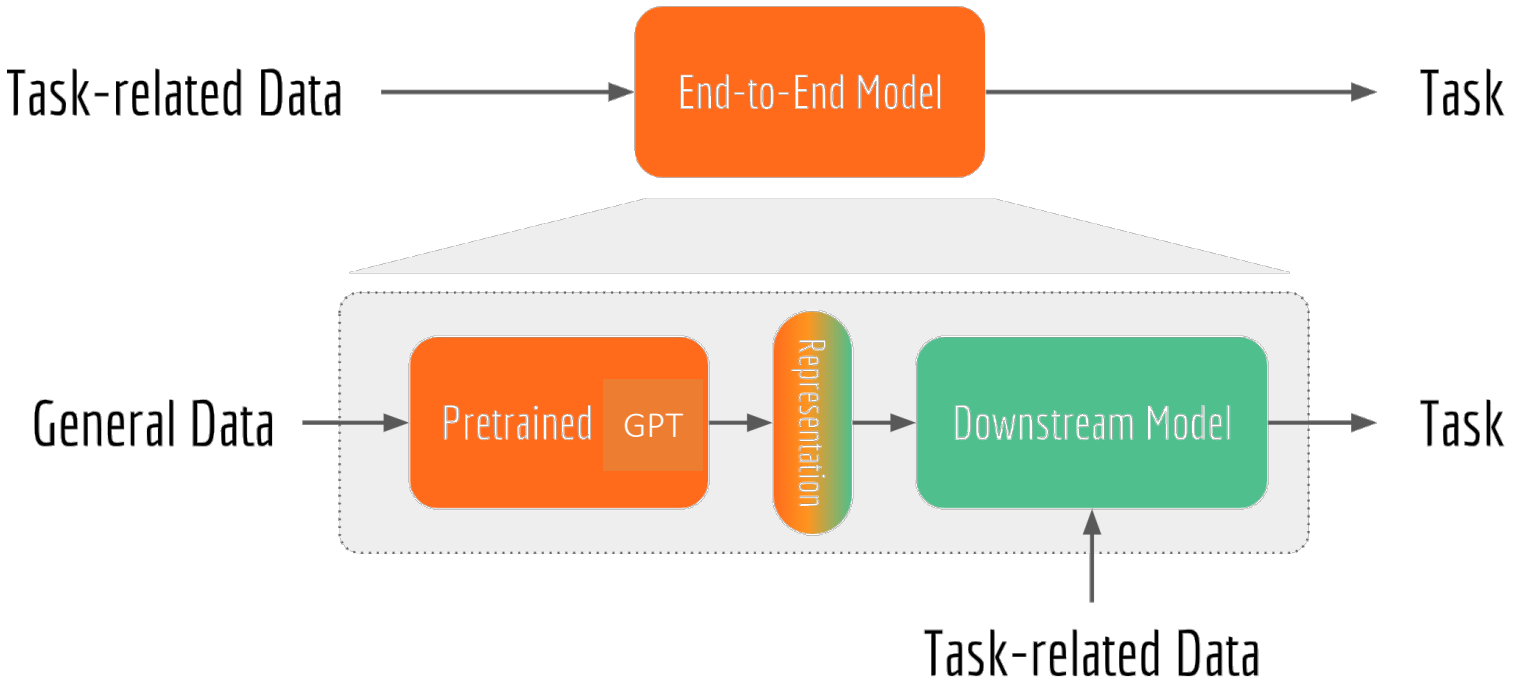
- G: Generative
- P: Pre-trained
- T: Transformer

G

- G is for Generative Model
- Generative = learns a representation in latent space from which we can sample the probably distribution to produce new data or making predictions.
- Model = instead of directly programming a feature, we design an architecture and learn its parameters – the probability distribution - from the training data.

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate P(y|x) to then deduce P(x|y)
What's learned	Decision boundary	Probability distributions of the data
Illustration		

- P is for Pre-Training (aka foundational or general)
- In GPT, the Pretrained model is a Large Language Model (LLM) trained autoregressively using text data collected on the Internet and elsewhere.
- Subsequently, the LLM is fine-tuned for specific tasks and more importantly, alignment to human objectives.



T

- T is for Transformer
- Transformer is an architecture (i.e. a parameterized function) that has been proven very effective in capturing the essential relationships in multiple Generative AI areas (text, image, and others).
- The parameters are then learned (computed) from pretraining and fine tuning trainings. The result is a deployable model. Transformers are much more consistently learnable (computable) to a larger scale.

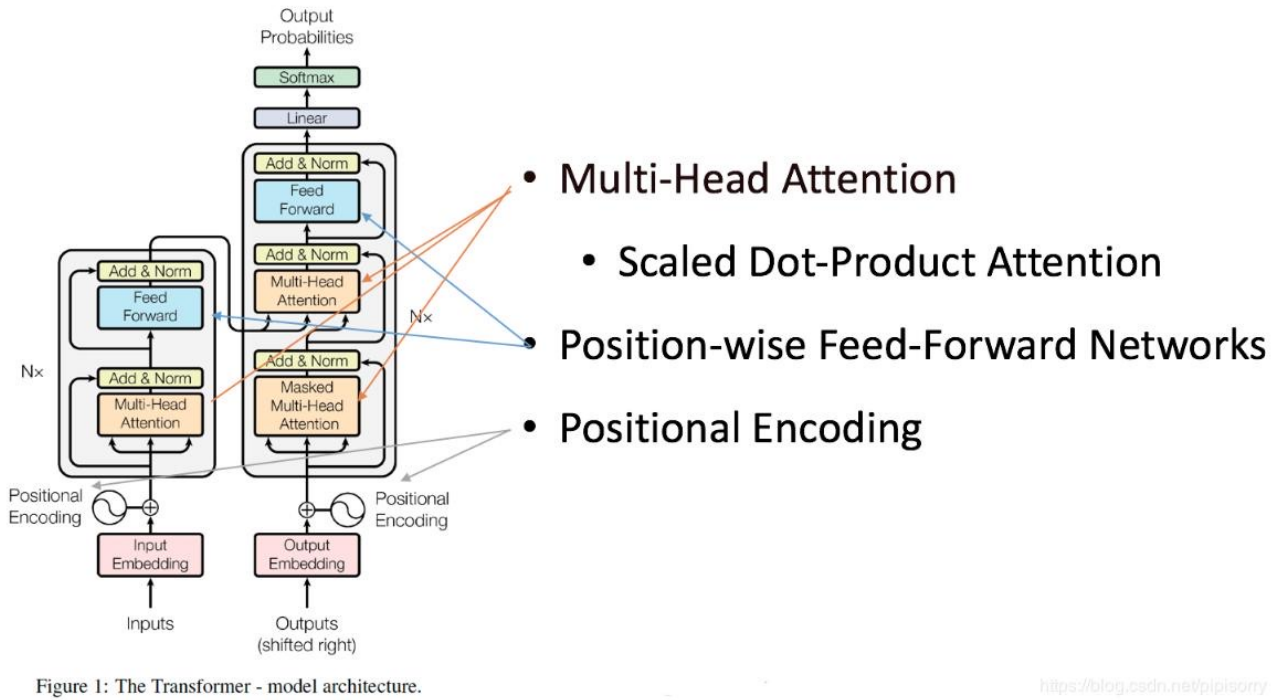
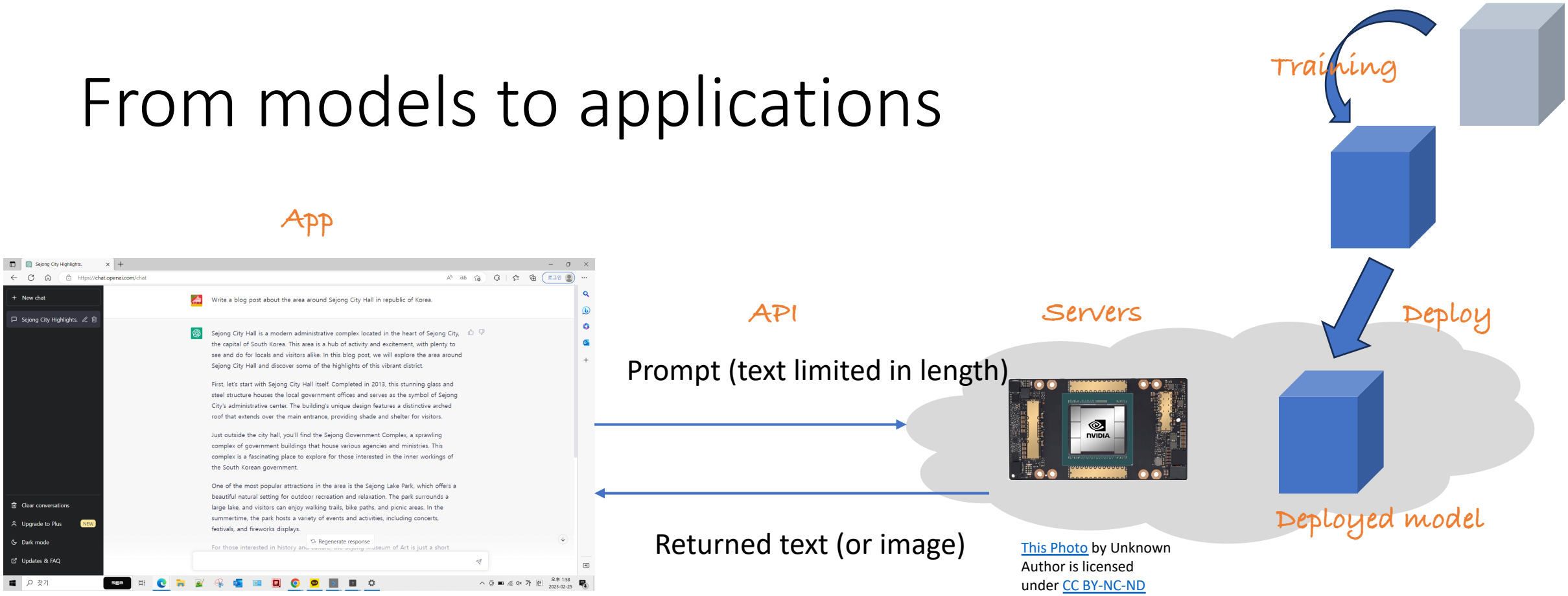


Figure 1: The Transformer - model architecture.

GPT-4

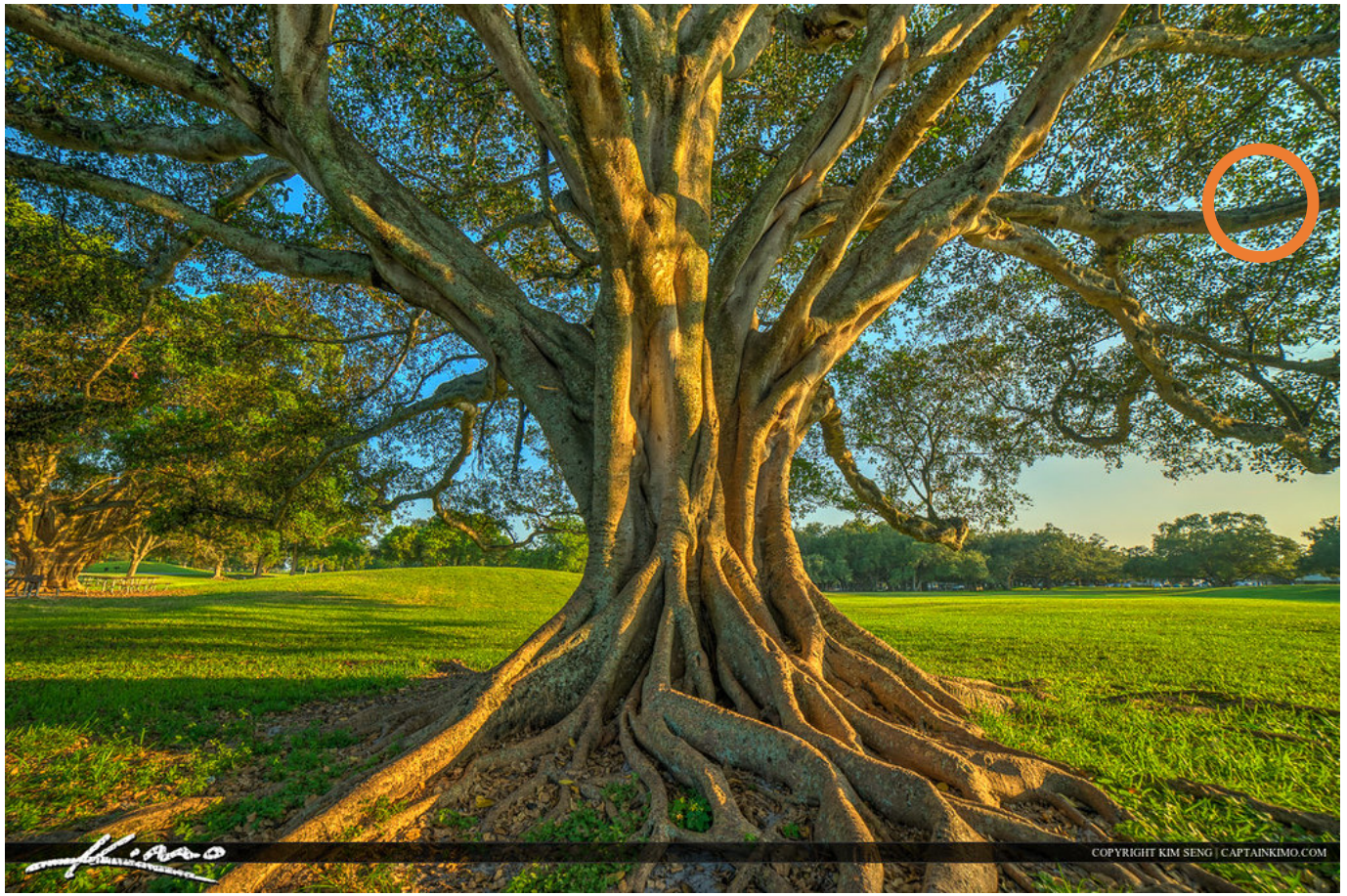
- GPT-4 is currently the best known LLM which can be fine-tuned for various applications. ChatGPT is one of such applications.
- Technical details of GPT-4 are largely unpublished. Previous versions (GPT-3 and many aspects of 3.5 were published).
- GPT-4 is claimed to have integrated another text-image generative model (multimodal).
- Other well-known text-image models include MidJourney, DaLLE-2, Stable Diffusion etc.
- Several open source projects are starting to replicate many aspects of GPT-4.
- Rapid progress in these and multimodal models are likely but still a research area.

From models to applications



Today's AI APP framework inherits all deficiencies of existing infrastructure including those in digital identity and trust.

GPT, Generative AI, Deep Learning, ML, AI



Current flavor of GPT-like Generative AI is only one of many branches of AI.

Key Points

1. The power of Generative AI lies in BOTH learning a representation (model) AND using that representation to generate content.
   - The surprising feat of learning a large-scale representation.
   - Using that learned representation for practical day-to-day applications.
2. Generative models can potentially represent any type of information: text, image, video, sensor data, others. Multimodal models are emerging and a possible near-future development.
3. It is inaccurate to say it is only mimicking humans. It can if that's what we have designed it to do but can also be designed to do other things sometimes better than humans.

The New Nature of Data

What is data in the age of Generative AI.

What is data before Generative AI?

- Data that people (or a device crafted by people) enter into a computer (e.g., a bill of sale)
- Data collected by platforms when we use digital services (aka digital footprints)
- Data is usually structured and indexed for efficiency in searching and processing.
- Processing is typically relational, searching or graph operations.

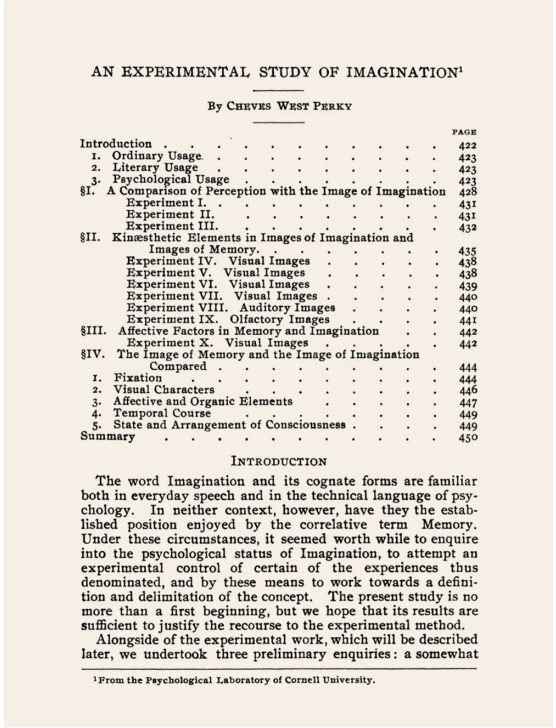
What is data after Generative AI?

- Enormous amount, intimate, live, impossible to have yourself in the loop.
- Captured multimodal data (text, voice, image, video, actions, and more) for training.
- Stored as a learned model (a representation in the latent space, not raw format) – beyond semantics.
- The learned model can generate high quality multimodal data in a computed “reality”.
- Highly “realistic” both structurally and semantically.
- The distinction between a computed or mediated reality and a reality of human experience is hard to make.

Apple image

The anchor in reality ?

- A human centered *solid* reality may be an illusion
- Integration of Generative AI and Metaverse (AR/VR/MR)
- We need to chart a New Way Forward



AN EXPERIMENTAL STUDY OF IMAGINATION[1]

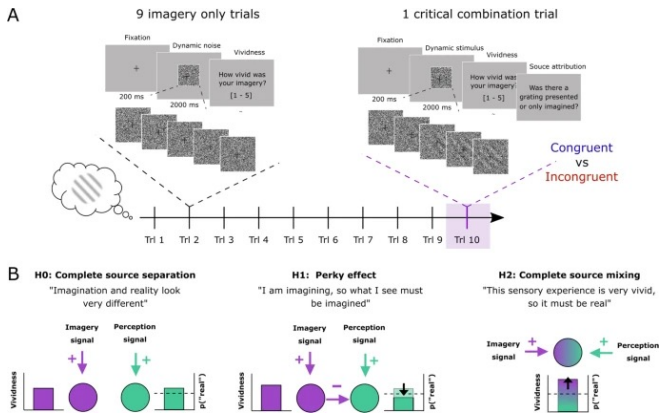
By Cheves West Perky

Introduction

The word Imagination and its cognate forms are familiar both in everyday speech and in the technical language of psychology. In neither context, however, have they the established position enjoyed by the correlative term Memory. Under these circumstances, it seemed worth while to enquire into the psychological status of Imagination, to attempt an experimental control of certain of the experiences thus denominated, and by these means to work towards a definition and delimitation of the concept. The present study is no more than a first beginning, but we hope that its results are sufficient to justify the recourse to the experimental method.

Alongside of the experimental work, which will be described later, we undertook three preliminary enquiries: a somewhat

[1]From the Psychological Laboratory of Cornell University.



Article | Open Access | Published: 23 March 2023

Subjective signal strength distinguishes reality from imagination

Nadine Dijkstra & Stephen M. Fleming

Nature Communications **14**, Article number: 1627 (2023) | Cite this article

In this study we investigated how imagined and perceived signals interact to determine reality judgements. By combining large-scale single-trial psychophysics, computational modelling and neuroimaging, we find evidence in support of a theoretical model in which reality and imagination are intermixed to determine a unified sensory experience. This model runs counter to accounts in which imagery and perception are separable, and to earlier findings of the Perky effect which imply imagery suppresses perception of reality. When deciding whether an experience reflects external reality or internal imagination, our model compares the strength of this experience to a reality threshold. But if reality and imagination are subjectively intermixed by default, why do we not confuse them more often in daily life? We suggest that such confusions are rare simply because imagery is typically less vivid than veridical perception, rarely crossing the reality threshold. However, these results also suggest that if imagery does become vivid or strong enough, it will be indistinguishable from perception.

“Real” is what we can trust.

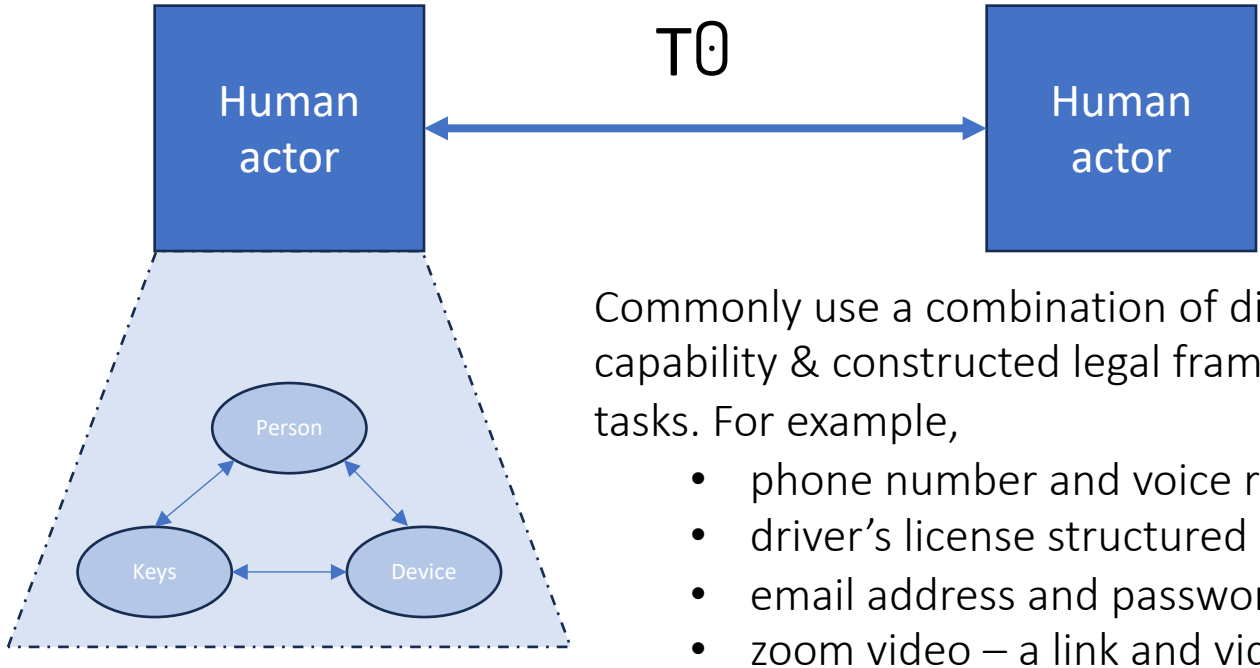
- The line between the “real” world and the digital world is blurred. “Real” is what we can trust.
  - All data that human can make sense of can potentially be made sense of by a machine.
  - Content that human can uniquely generate can potentially be generated by a machine with high authenticity & uniqueness.
- We need new tools to aid our cognitive functions to know what we can trust.

Nvidea image

The New Nature of Digital Trust

What will be digital trust in the age of Generative AI.

Digital trust before Generative AI

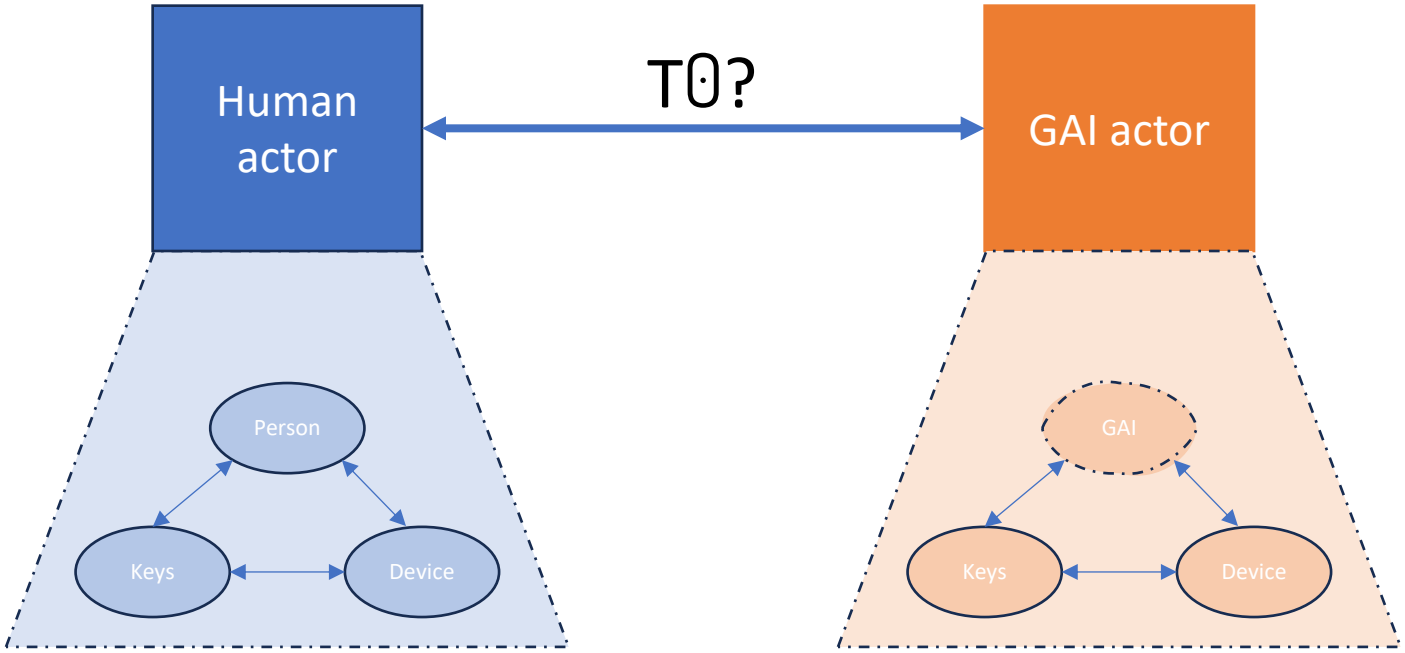


Commonly use a combination of digital tokens, uniquely human capability & constructed legal framework for identification & other trust tasks. For example,

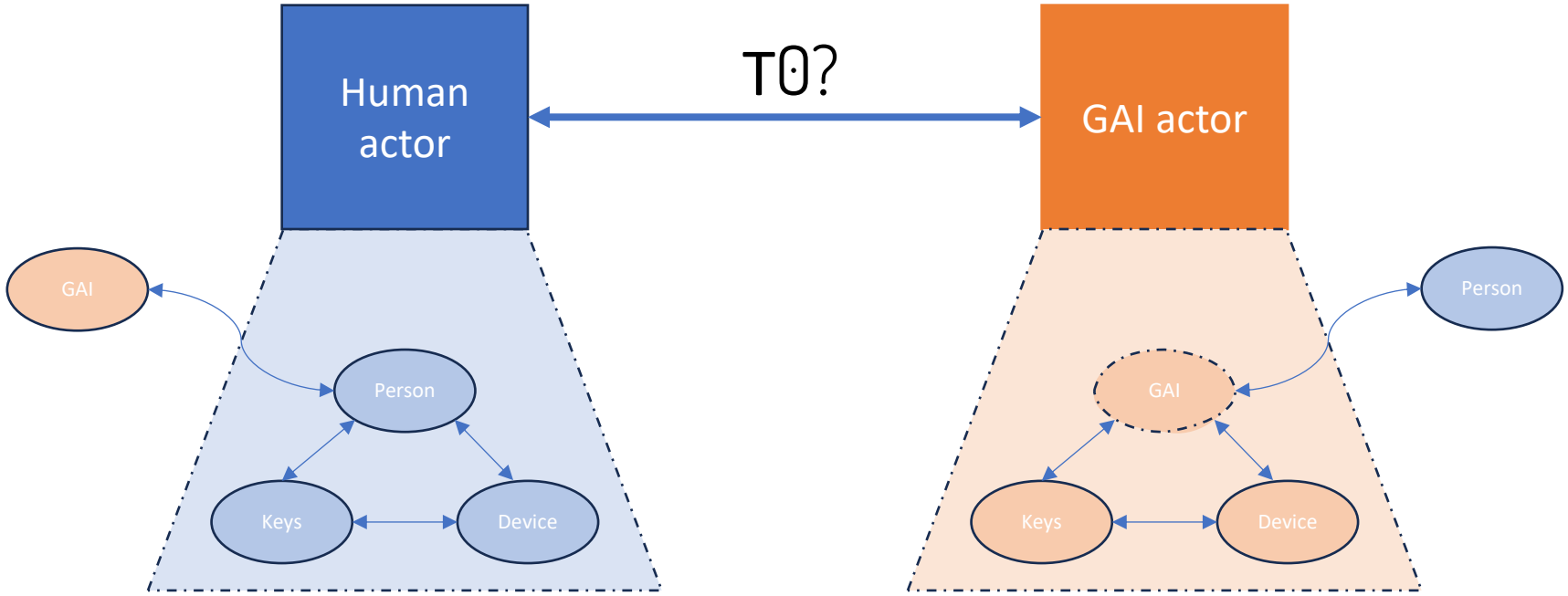
- phone number and voice recognition, context
- driver's license structured data and photo facial recognition
- email address and password, CAPTCHA
- zoom video – a link and video / rich context
- signature, witness for contract
- money, banks, the Fed, IMF

These methods all rely on some "uniquely human capability" as a basic assumption. Strong Generative AI threatens this foundation. We must rethink the framework of trust.

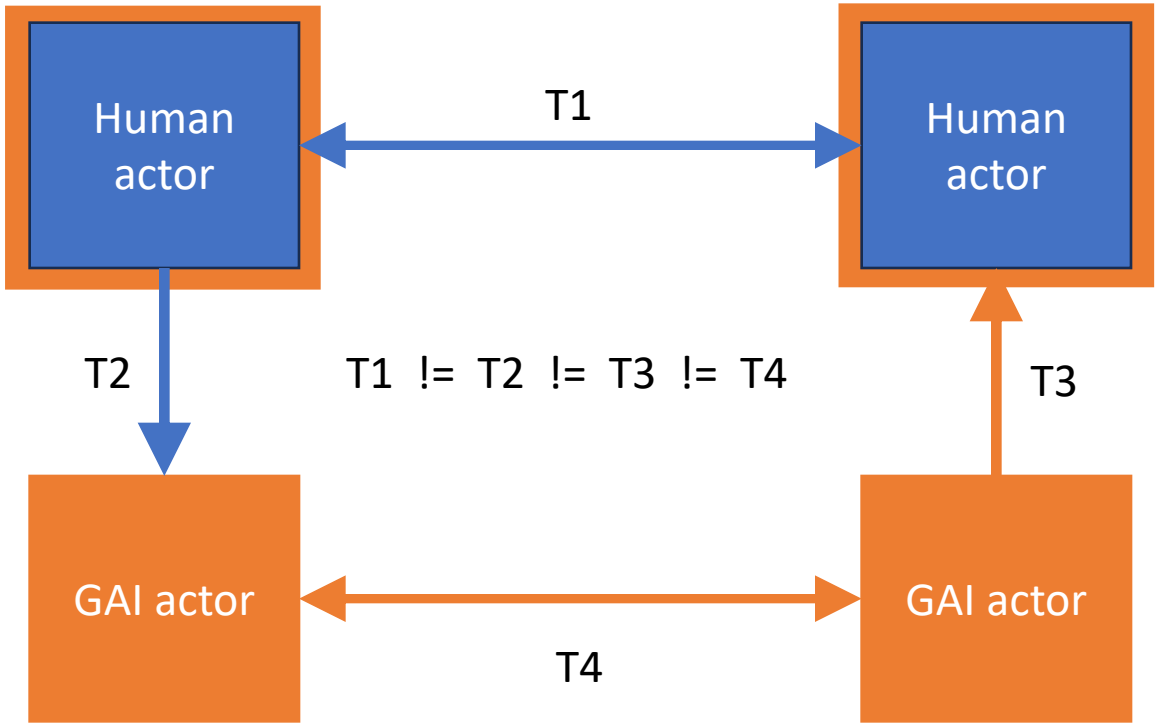
Digital trust after Generative AI



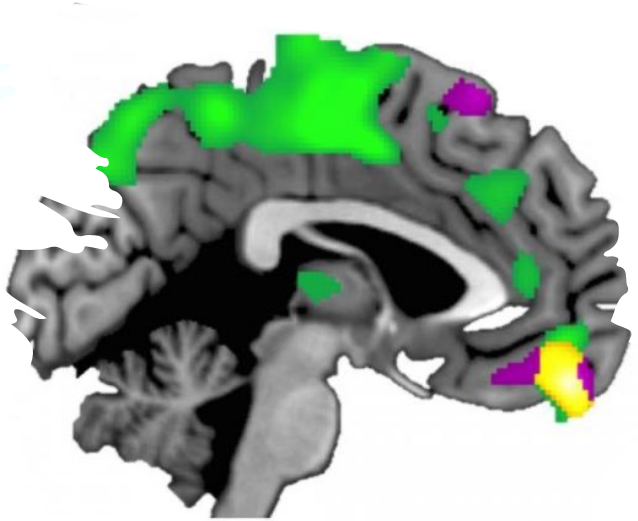
Digital trust after Generative AI



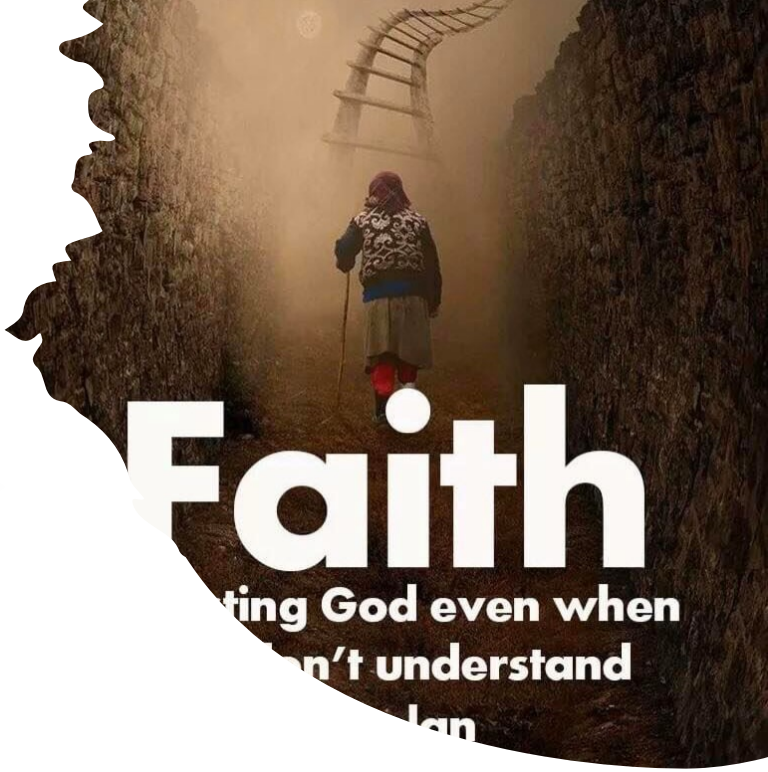
Digital trust after Generative AI



Note that T0 is no longer a viable option without heavy restriction of AI accessibility: e.g., ChatGPT in education, MidJourney in visual art.



ventromedial prefrontal cortex (ye
er in those that tend to be mor
compared to those that t
of others.



Trusting a human relies on a lot of humanistic cues

- By another human
- By a Generative AI actor

“Trusting” a Generative AI actor is a fundamentally different concept

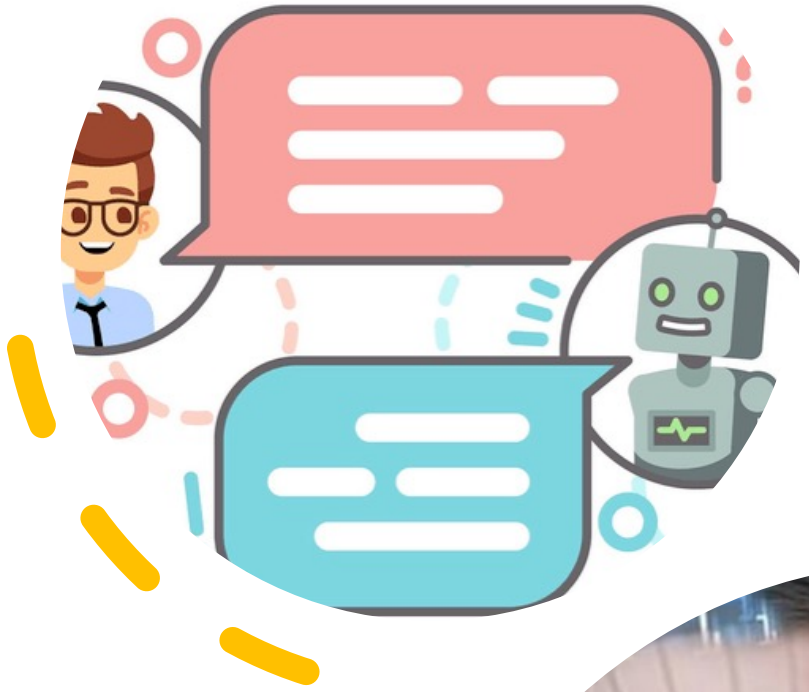
- What does “trust” mean in this context?
- “Trusting” Generative AI as a system operated by humans
- “Trusting” Generative AI as an autonomous actor
- By a human actor
- By another Generative AI actor

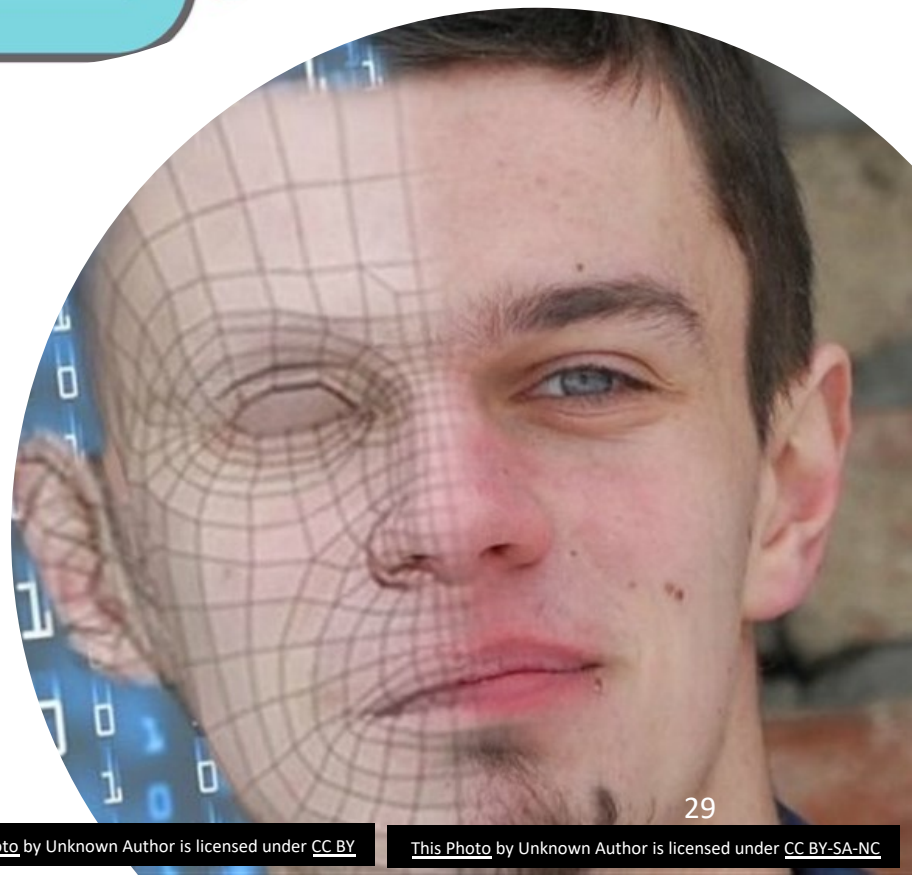
A Way Forward

What can we do about it.

Telling human actor and GAI actor apart

- Proof-of-personhood and uniqueness.
- Without it, we will not be able to distinguish T1, T2, T3, T4.





The Worldcoin ID Orb

- What is the ToIP answer to this challenge?
- Debate on World-ID vs. ToIP.

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

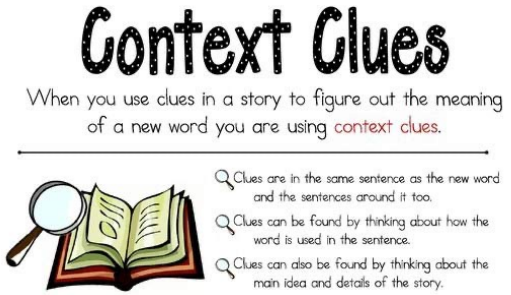
Biometric Modalities

	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

All images @ Worldcoin

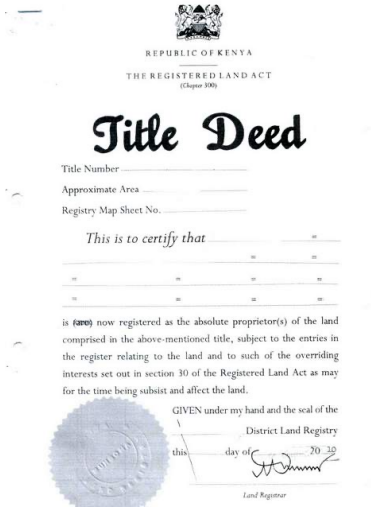
Watermarking

Signing



Context

Telling human and GAI generated content and actions apart



Provenance

Prediction

Telling human and GAI generated content and actions apart

Watermarking GPT Outputs

Scott Aaronson (UT Austin and OpenAI)
Joint work with Hendrik Kirchner (OpenAI)

Our Scheme

Given: tokens $w_1, \ldots, w_{t-1}$, and GPT's probability distribution $p_1, \ldots, p_K$ over the next token w_t

Reals $r_1, \ldots, r_K \in [0,1]$, $r_i := f_s(w_{t-n+1}, \ldots, w_{t-1}, i)$ where $f_s(\,)$ is a pseudorandom function and s is a random seed known only to OpenAI

Rule: Choose the token i that maximizes r_i^{1/p_i}

To check for a watermark, calculate $\sum_{t=1}^{T} \ln \frac{1}{1-r'_t}$

and see if it exceeds a threshold, where

$$r'_t := f_s(w_{t-n+1}, \ldots, w_t)$$

A Watermark for Large Language Models

John Kirchenbauer* Jonas Geiping* Yuxin Wen Jonathan Katz Ian Miers Tom Goldstein
University of Maryland

Prompt	Num tokens	Z-score	p-value
…The watermark detection algorithm can be made public, enabling third parties (e.g., social media platforms) to run it themselves, or it can be kept private and run behind an API. We seek a watermark with the following properties:			
No watermark Extremely efficient on average term lengths and word frequencies on synthetic, microamount text (as little as 25 words) Very small and low-resource key/hash (e.g., 140 bits per key is sufficient for 99.999999999% of the Synthetic Internet	56	.31	.38
With watermark - minimal marginal probability for a detection attempt. - Good speech frequency and energy rate reduction. - messages indiscernible to humans. - easy for humans to verify.	36	7.4	6e-14

Telling human and GAI generated content and actions apart

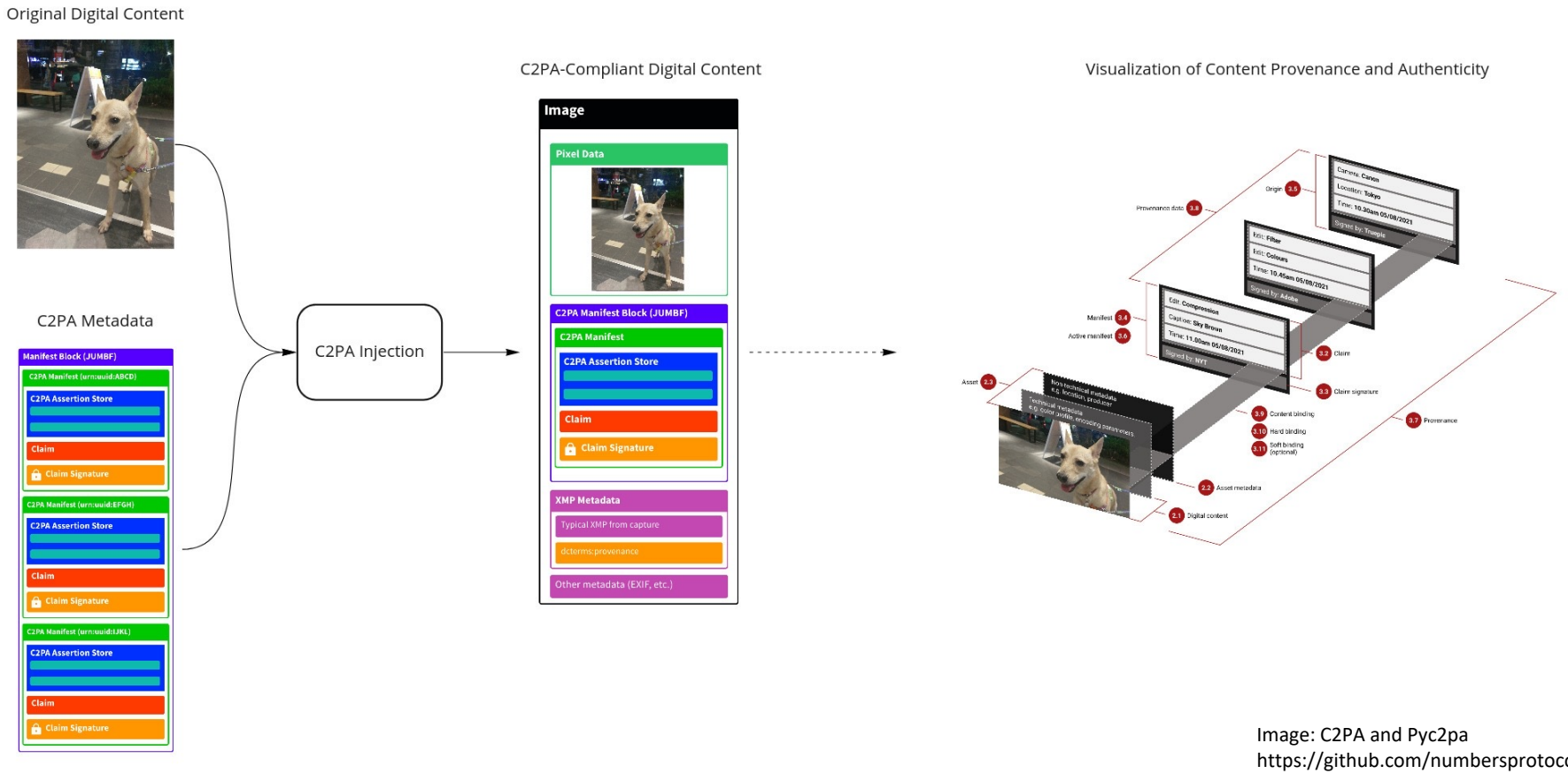
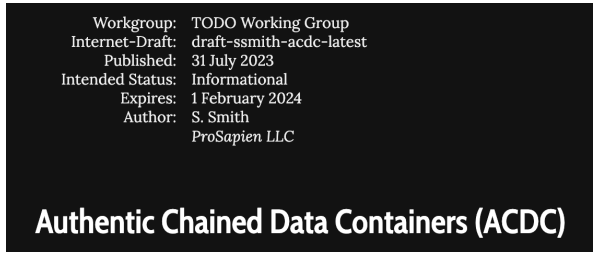


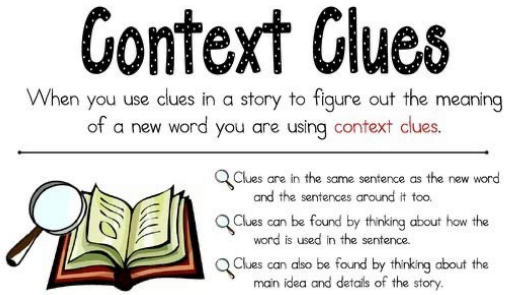
Image: C2PA and Pyc2pa
https://github.com/numbersprotocol/pyc2pa

Telling human and GAI generated content and actions apart



One primary purpose of the ACDC protocol is to provide granular provenanced proof-of-authorship (authenticity) of their contained data via a tree or chain of linked ACDCs (technically a directed acyclic graph or DAG). Similar to the concept of a chain-of-custody, ACDCs provide a verifiable chain of proof-of-authorship of the contained data. With a little additional syntactic sugar, this primary facility of chained (treed) proof-of-authorship (authenticity) is extensible to a chained (treed) verifiable authentic proof-of-authority (proof-of-authorship-of-authority). A proof-of-authority may be used to provide verifiable authorizations or permissions or rights or credentials. A chained (treed) proof-of-authority enables delegation of authority and delegated authorizations. These proofs of authorship and/or authority provide provenance of an ACDC itself and by association any data that is so conveyed.¶

https://www.ietf.org/archive/id/draft-ssmith-acdc-02.html
https://wiki.trustoverip.org/display/HOME/Getting+Started+with+Confluence+Wiki



Context

Telling human and GAI generated content and actions apart

Prediction

It’s a New Era – A Call for Action

- Update the mental framework to include AI agents
- Expand and clarify terminology, concepts, methods
- Expand and update principles
- Upgrade solution toolkits – An upgraded stack.

Join the ToIP AI and Metaverse Task Force (AIM) :

9 AM Pacific Time every other Thursday.

WIKI: https://wiki.trustoverip.org/pages/viewpage.action?pageId=19657312

Thank you

This presentation slides: https://github.com/wenjing/Digital-Trust-in-the-Age-of-Generative-AI

https://www.linkedin.com/in/wenjingchu/

https://www.youtube.com/channel/UCI-q4zODuFTDf0bOmyN8xRw