2023-04-13 AIM TF Meeting Notes

Meeting Date & Time

- 13 Apr 2023
  - **09:00-10:00 PT / 16:00-17:00 UTC**

Zoom Meeting Links / Recordings

Meeting: https://zoom.us/j/98931559152?pwd=d0ZwM1JHQ3d5cXRqVTh4NlRHeVJvQT09

Recording: Video / Audio Here: https://zoom.us/rec/share/hZHCIiXx1BXmGvYQ8m8vFYlJlLdgfutdnr2A1roGfc1I_8pNT6BRWUioeJirxKs.mh49e8M3X3E2Z7rg

Slides: Digital Trust in the age of ChatGPT.

Attendees

- Wenjing Chu
- Sandy Aggarwal
- Neil Thomson
- Kaliya Young
- Chi Hwa Tang
- sankarshan
- Daniel Bachenheimer
- Mary Lacity

Main Goal of this Meeting

This is the AIM TF's #20 meeting.

One of our main goals is to have individual member presentations on what problems/challenges they see in AI & Metaverse related to trust.

Starting in the new year (2023), we plan to start drafting white papers or other types of deliverables of the task force.

Agenda Items and Notes (including all relevant links)

Time	Agenda Item	Lead	Notes
2 min	• Start recording • Welcome & antitrust notice • Introduction of new members • Agenda review	Chairs	• **Antitrust Policy Notice:** *Attendees are reminded to adhere to the meeting agenda and not participate in activities prohibited under antitrust and competition laws.* • ***ToIP Policy**: Only members of ToIP who have signed the necessary agreements are permitted to participate in this activity beyond an observer role.* • ***ToIP TSWG IPR Policy:** see TF wiki page. AI & Metaverse Technology Task Force*
3 mins	• Introduction of new members • Any general announcement, news, that could be of interest to the TF	All	Chi Hwa Tang • Focus on drafting whitepaper Digital Trust in the Age of ChatGPT - using ChatGPT as an example - what does it mean to Digitall Trust - new and existing issues, unexpected issues
50 mins	The Latest Generative AI, Authentication, and Content Authenticity	Wenjing Chu will lead the discussion.	1. How the latest Generative AI technologies such as ChatGPT and DALL-E may render "content-based" authentication methods ineffective. 2. Deepfakes and content authenticity. 3. Possible solutions. The presentation document is here. First up - a video of a digital avatar looking and sounding credibly like Morgan Freeman - raises questions - what do you believe? what do you feel? The initial focus is CHatGPT (GPT4) (others on the list for "another time").

			Exams/Tests such as LSAT, PSAT, counter-factuals, theory of mind (current and new "Turing tests") to evaluate AI vs. humans. Ging to skip over those for.
			
			Computing has fundamentally changed w the generation of Generative AI. How does this change digital trust? What are the new directions
			
			Wenjing did a 60 min+ session wi ChatGPT on a technical subject, which Wenjing was somewhat knowledgeable - verdict - it was a useful session and provided learning value (equivalent to a conversation w a human expert?)
			
			MS and Google Office-type tools will incorporate this (likely immediately). A credible, accurate (vs. bias) search is next. Can this avoid bias of pushing paid data sources (see Cory Doctorow on the "Inshitification of the Internet" on how search is polluted by profits vs. knowledge)?
			
			AI may be a path to editing & summarizing of content produced by one or more experts, including film editing.
			
			Medium to long-term speculation is likely to be quickly superseded.
			
			Content Authenticity.
			
			• DeepFake has passed a critical threshold (the iPhone moment)
			• AI makes it much harder to find the origin of content (particularly if blended)
			  ◦ Stronger Dig Identity help? Verification? What role does anonymous vs. accountable play? Credibility of ID + reputation?
			  ◦ Can't rely on the content itself (to prove authentic)
			• Can C2PA (Content Provenance and Authenticity) help?
			  ◦ Yes, but needs v good tools, techniques
			  ◦ Fraudulent presentation of fact is hard, and fake presentation of fake, is impossible to prove as fact provenance can't prove not-fake (factual)
			• Tracking history may not help.
			• 3 approaches (apply to content generators and consumer tools (e.g., social media))
			  ◦ Alignment - AI improvement + regulation
			  ◦ Watermarking (AI tool improvement)
			  ◦ Signing/Manifest (AI tool improvement)
			• If the original content is not correctly attributed, then all downstream processing is moot (from a digital trust perspective
			
			From Chat:
			
			• For ToIP, unless content can be verified, the default should be rejected.
			• Can effective (untamperable, fully traceable) signatures be placed in any/all digital content?
			  ◦ But of course, there is a high operational cost in this -so how do we decide what content needs to have embedded signatures, especially in routine social media content
			• EFWG had a team from https://www.frankliapp.com/about who were discussing one part of the problem of fake /synthetic content (2 weeks ago?)
			• Content and provenance of what tools/algorithms made modifications/processing applied - in many cases knowing which tools were used to change original content is revealing in itself. Who controls the model
			
			Fake Identity : Authentication
			
			Key contexts
			
			• Human or robot?
			• Bypass authentication (security issue)
			• Dual Identity
			
			Determining identity due diligence is highly dependent on the medium (e.g., in-person, vs phone (voice only), vs online video, ...)
			
			Will be an "arms race" between generators (of fake) and detectors.
			
			There is no purely digital solution will work as always subject to human operator manipulation - it always needs independent, authoritative human verification
			
			Dual Identity - who are you speaking to, and who is responsible for the identity's behaviour?
			
			Do you know or care if it is real?
			
			• There are markets for real and fiction (books, movies)
			• A way to do this is to have the creator (author), and processors sign this (editor, etc.)
			• Reputation needs to be made concrete to do this (mix of verified credentials and contextual reputation)
			• Will anyone follow up on checking the content provenance?
			  ◦ Or is this a required logo/icon to click for provenance
			
			Dan B
			- can't rely on Biometrics as that is effectively publically available (e.g., fingerprints on a glass, scan of a retina, etc.)
			- one solution is facial + bloodflow in real-time
			
			- https://www.iproov.com/

	Additional updates from Phil Wolff Sandy Aggarwal		From slack channel: ▪ Phil Wolff updated the Happy blog post, nearly ready. https://docs.google.com/document/d/1z7MIhzUhVTEtOL7gp7iBWsN1h4MrbEf5yB1GP5sfhZM/edit?usp=drivesdk ▪ Sandy Aggarwal will update his paper status next time.

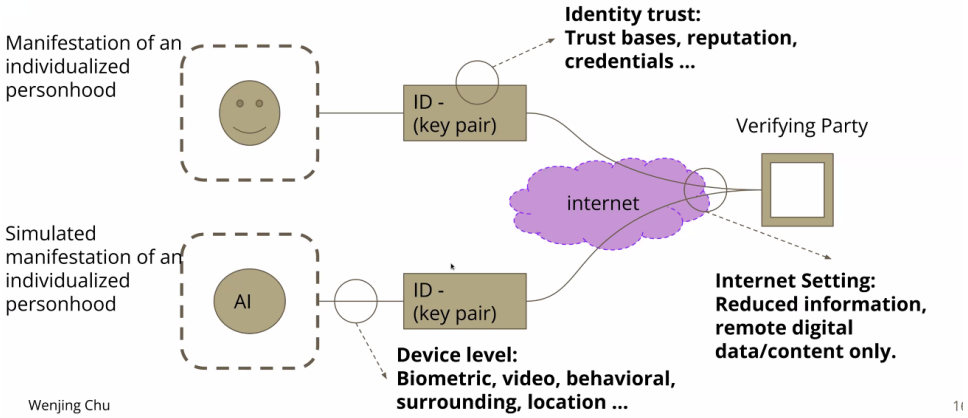
5 mins	• Review decisions /action items • Planning for next meeting • AOB	Chairs	

Slides from the Presentation:

Digital Trust in the Age of ChatGPT

Topic #1 : The Latest Generative AI, Authentication and Content Authenticity
(April 12, 2023)

Fake Identity: Authentication



Wenjing Chu