Intelligent Trust Services

“AI powered” - How to differentiate between the good, the bad and the ugly

Jörg Lenz
Head of MarCom

September 26, 2024 – CA Day, Heraklion

10th Trust Services and eID Forum

Anniversary

25th September 2024

16th CA-Day

26th September 2024

HERAKLION, GREECE | ONLINE

Co-organised by

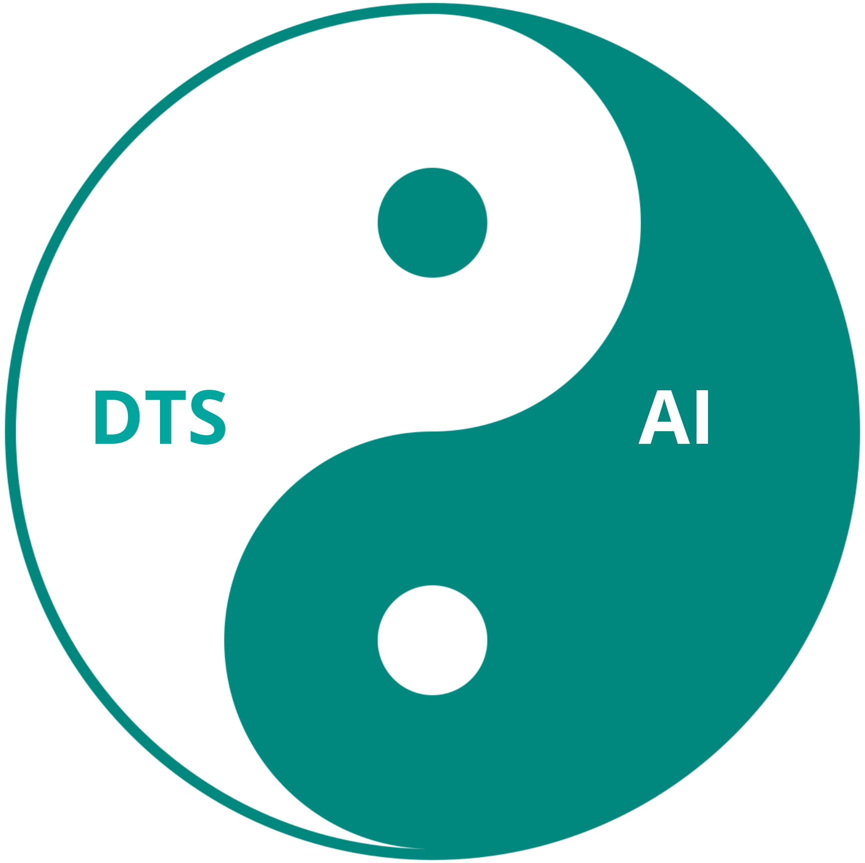
Session Overview

1. The interplay between AI and Digital Trust Services
2. AI in context with Digital Trust Services
3. Key Purposes of using AI in Digital Trust Services
4. AI in Digital Trust Services – Predictions
5. Sep 2024: AI has a reputation issue
6. What AI ought to deliver: Simplicity & Evidence
7. What’s Trustworthy Artificial Intelligence?
8. Myth Busting
9. The tools and methods we use for work – “Hey Socrates”
10. Know your AI Model
11. Choosing the right Model for the job
12. Trusted reading/listening list
13. Proceed with Pragmatic Scepticism

Namirial | © Namirial – All rights reserved

The interplay between AI and Digital Trust Services

AI can help Digital Trust Services to be even more trustful, secure and easier to use. On the other side without trust, the potential benefits of AI and digital services are undermined by concerns over privacy, security, and reliability.



For Digital Trust Services AI is both: Challenge and Opportunity so AI used in Digital Trust Services has to be smarter than AI used by fraudsters

This race is on … and will not be stopped

AI in context with Digital Trust Services

Systems based on technologies incl. machine learning, deep learning, reinforcement learning, neural networks, natural language processing

capable of

- performing tasks that typically require human intelligence such as learning, planning, reasoning, problem-solving, perception

categorized into

- Specialized AI - designed for specific tasks, e. g. writing recognition (OCR)
- General Purpose AI - Foundation Models that perform at many tasks, e. g. optical passport information retrieval and summarizing a document

Namirial | © Namirial – All rights reserved

AI terms are defined in EU Regulation 2024/1689 (AI Act)

The AI Act is one the EU Regulation with highest relevance for Digital Trust Service Providers besides EU Regulations 910/2014 and 2024/1183 governing our market on electronic identification and Trust Services (eIDAS) and establishing the European Digital Identity Framework

AI system	A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments;
Risk	The combination of the probability of an occurrence of harm and the severity of that harm
Biometric Verification	automated, **one-to-one verification**, including authentication, of the identity of natural persons by comparing their biometric data to previously provided biometric data
Biometric Identification	automated recognition of physical, physiological, behavioural, or psychological human features for the purpose of establishing the identity of a natural person by comparing biometric data of that individual to biometric data of individuals stored in a database;
	AI system designed to identify individuals **without their active participation**, typically from a distance, by comparing their biometric data to a reference database. This system can process multiple individuals simultaneously and differs from biometric verification systems, which are used solely for confirming a specific person's identity, typically for accessing services or security. There are two types: 'real-time' systems, which operate with minimal delay, and 'post' systems, which analyze pre-recorded data.

EU-Regulation 2024/1189, published July 12,2024 in the Official Journal of the European Commission, entered into force August 1, 2024.

Key Purposes of using AI in Digital Trust Services

1. **Enhanced Identity Verification**
   Improve the accuracy and efficiency of identity verification processes – e g biometric recognition
   Reducing the risk of fraud by detecting anomalies and verifying documents more effectively.
2. **Fraud Detection and Prevention**
   Analyze patterns and detect fraudulent activities
   Offering proactive protection against various forms of cybercrime and fraud.
3. **Automation and Efficiency**
   Automates routine tasks (e. g. document processing and verification)
   Reduce human error and speeding up processes
4. **Personalization and User Experience**
   Personalized services based on user behavior and preferences
5. **Confidence in Content and Process Understanding**
   Establishing and fostering confidence - for rapid focus on most important content sequences and process steps
6. **Scalability**
   Handling of larger volumes of data and transactions without compromising performance.
7. **Compliance and Risk Management**
   Monitoring compliance with regulations and managing risks by analyzing data

AI in Digital Trust Services – Predictions

Jim Lundy in Aragon Globe on Digital Transaction Management, 2024 – July 17, 2024:

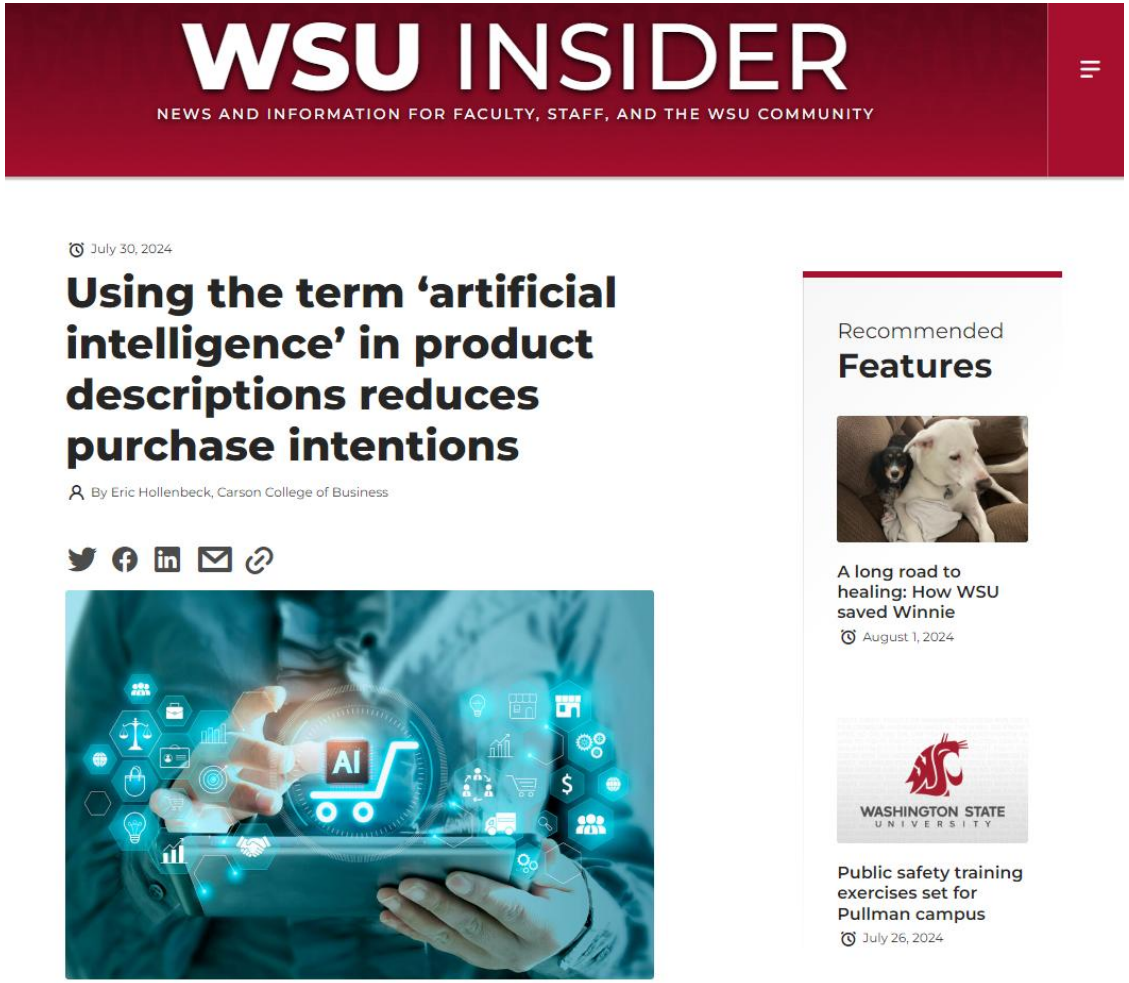
“By the end of 2025, 50% of Digital Transaction Management providers are expected to incorporate intelligent document assistants into their products. These assistants will facilitate content creation, summarize documents, and enhance transaction processes.”

in Frost Radar: Electronic Signature 2024 – published July 30, 2024:

“AI and ML have been used for several years to automate processes such as document verification and fraud detection. All are poised to become industry standards given their benefits on optimization, security, transparency, and auditability of electronic signature processes.”

Sources: Aragon Research Globe™ for Digital Transaction Management, 2024 – July 17, 2024 + Frost & Sullivan's Radar for Electronic Signature Software, 2024, July 30 3024

Sep 2024: AI has a reputation issue



WSU INSIDER

NEWS AND INFORMATION FOR FACULTY, STAFF, AND THE WSU COMMUNITY

July 30, 2024

Using the term 'artificial intelligence' in product descriptions reduces purchase intentions

By Eric Hollenbeck, Carson College of Business

Recommended

Features

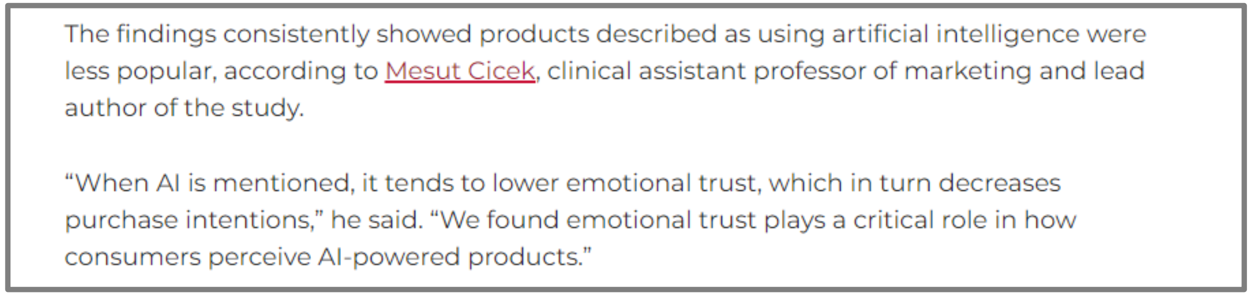
A long road to healing: How WSU saved Winnie

August 1, 2024

Public safety training exercises set for Pullman campus

July 26, 2024

Powered by AI = Undermining Trust?



The findings consistently showed products described as using artificial intelligence were less popular, according to Mesut Cicek, clinical assistant professor of marketing and lead author of the study.

"When AI is mentioned, it tends to lower emotional trust, which in turn decreases purchase intentions," he said. "We found emotional trust plays a critical role in how consumers perceive AI-powered products."

Is AI overhyped? …. Like in recent years Biometrics & Blockchain: Gartner Hype Cycle: Gen AI is on top of inflated expectations due to drop into the valley of disappointment

Does adding AI add or undermine trust in digital trust services? Honest answer: It depends.
Time for a closer look.

Washington State University (WSU) Insider - Eric Hollenbeck : Using the term 'artificial intelligence' in product descriptions reduces purchase intentions, July 30, 2024

What AI ought to deliver: Evidence

Digital Trust Services are used to deliver proofs of identity, intent, authenticity, origin and delivery with significant evidence to stay out of court

“Hey, Digital Trust Service” – make simplicity meeting compliance

As someone requiring proofs, e. g. in sending requests for signing you are expecting Digital Trust Services to

- having your processes being triggered almost instantly
- avoid mistakes – e .g. in detecting if all recipients of signing requests are included
- being able to wrap up your processes in minutes instead of days
- protect processes from fraud attempt – detecting and blocking suspicious activities

AI ought to be your **co-intelligence**, helping you to get your processes executed without the need to re-trigger them, reduce fraud and fulfill your compliance requirements

What AI ought to deliver: Simplicity

Almost everyone of us is using digital services in some way these days: In 2024 Apple, Amazon, Google, Microsoft, Mercedes, Samsung etc. are setting the user experience expectations:

You are talking to your phones and cars for years. Alexa, Siri and your car…

Trust is established, fostered and re-confirmed in individual – literally intelligent – communication

“Hey, Digital Trust Service” – Support me in getting stuff done with confidence
(or I will switch to paper)

As a user you are expecting Digital Trust Services for example to
- individually explain you content of agreements
- highlight your personal risks (e. g. cancellation fees)
- help you to understand the process (e. g. who needs to sign what why)
- proceed and finalize process with gradually increasing confidence

AI ought to be your **co-intelligence**, helping you to make better informed decisions faster

You don’t want to get stuck with a stupid chatbot….

What’s Trustworthy Artificial Intelligence?

5 values-based Principles for trustworthy, human-centric AI –
according to OECD Working Party on Artificial Intelligence Governance (AIGO)

1. Benefit People & Planet
2. Human rights, values & fairness
3. Transparent & explainable
4. Robust, secure & safe
5. Accountable

AI experts can present current issues in the OECD AI Wonk Blog.

However … some of these principles are pre-dominantly theoretical approaches but are divorced from the everyday realities of practitioners. They are also asking about:

- Planet benefit might be impacted by power consumption
- Values – of whom? EU vs. US, China, ME & others
- Explainable – We might see better
- Robust might be contradictional to safe

Current best practice in “real AI” suggests ensuring
Human accountability + **Human ownership of results** + **Human checks & verification**

OECD Working Party on Artificial Intelligence Governance (AIGO) + European Commission: EU-U.S. Trade and Technology Council Working Group 1: Technology Standards Subgroup on AI Taxonomy & Terminology - EU-U.S. Terminology and Taxonomy for Artificial Intelligence - Second Edition, April 5. 2024 & European Commission: TTC Joint Roadmap for Trustworthy AI and Risk Management, December 2, 2022)

AI Myth (1): Just Plagiarism engines?

Myth: Gen. AI is all based on probabilities derived from training data. Its output can only be plagiarism.

Bust: Training follows phases:

1. Learn language, gather knowledge 👉 „Base Model“
2. Behave well („Alignment“) 👉 e.g. „Instruct Model“, through RHLF technique
3. (Domain Adaptation) 👉 e.g. Apple Intelligence

Example: return single answers for questions, not other good questions or multipe-choice options. Avoid open-ended, incoherent output.

Result: shift towards derivative/original, esp. in co-intelligence context

Grain of truth:

- original **problem**: lots of conflicts about well-behaviour
  - serving, pleasing human operator vs. truthfulness, lawfulness (e.g. copyright!)
  - up-to-date knowledge vs. defense against misinformation
- **but**: first rulings in case law, first regulations, first AI guardrails, … 👉 getting better fast

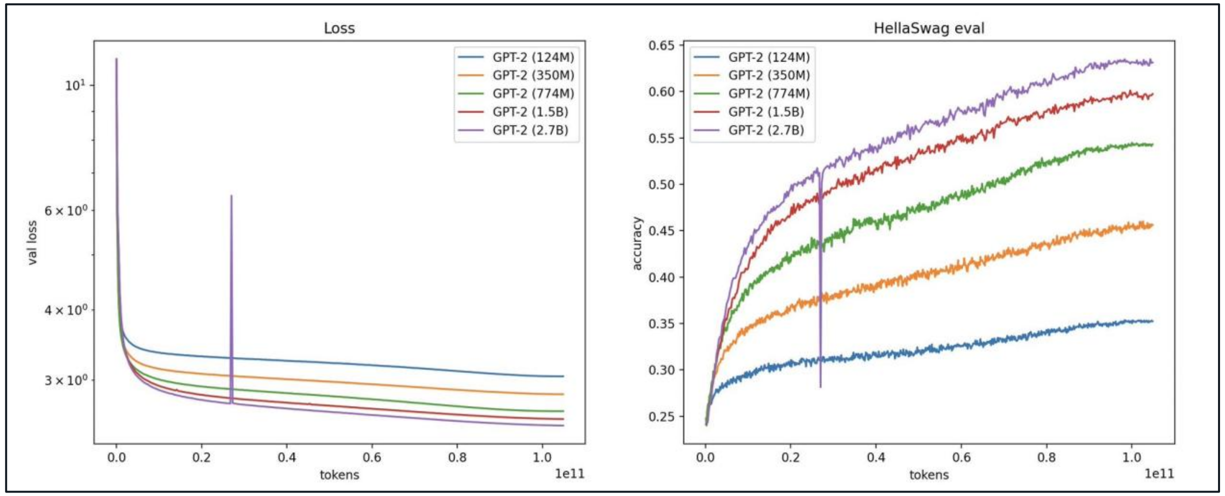
| © Namirial – All rights reserved

AI Myth (2): AI eats itself: Model Collapse!

Myth: when AI is trained on its own output, it breaks down.

Bust: training on partially synthetic data is routine, fully synthetic can be done.

- two LLM releases in 2024 with explicit use-case of training other models:
    - Nvidia Nemotron 304B
    - Meta Llama 3.1 405B
- GPT-2 trained on fully synthetic data outperforms GPT-3
- (model quality is monitored during training)



Grain of truth:

- original **problem**: recursively training on the same data causes break-down
- **<u>but</u>**: happens anyway, also with human-written text

AI Myth (3): Running out of training data

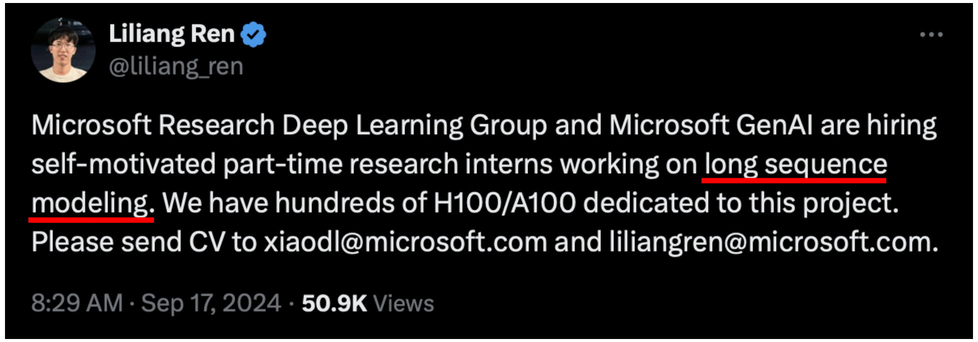
Myth: AI requires vast amounts of data to train. But Internet, books and videos have already been devoured. We are running out of data!

Bust: New scientific advances allow training from scratch:

- on 60GB of data
- in three days
- with 15€ budget

Grain of truth:

- original **problem**: LLMs struggle with long-form output
- **hurdle**: not enough samples to have LLMs write a book from a single prompt
- **but**: extreme memory req‘s during output generation needs to be solved first anyway
- but: research underway:

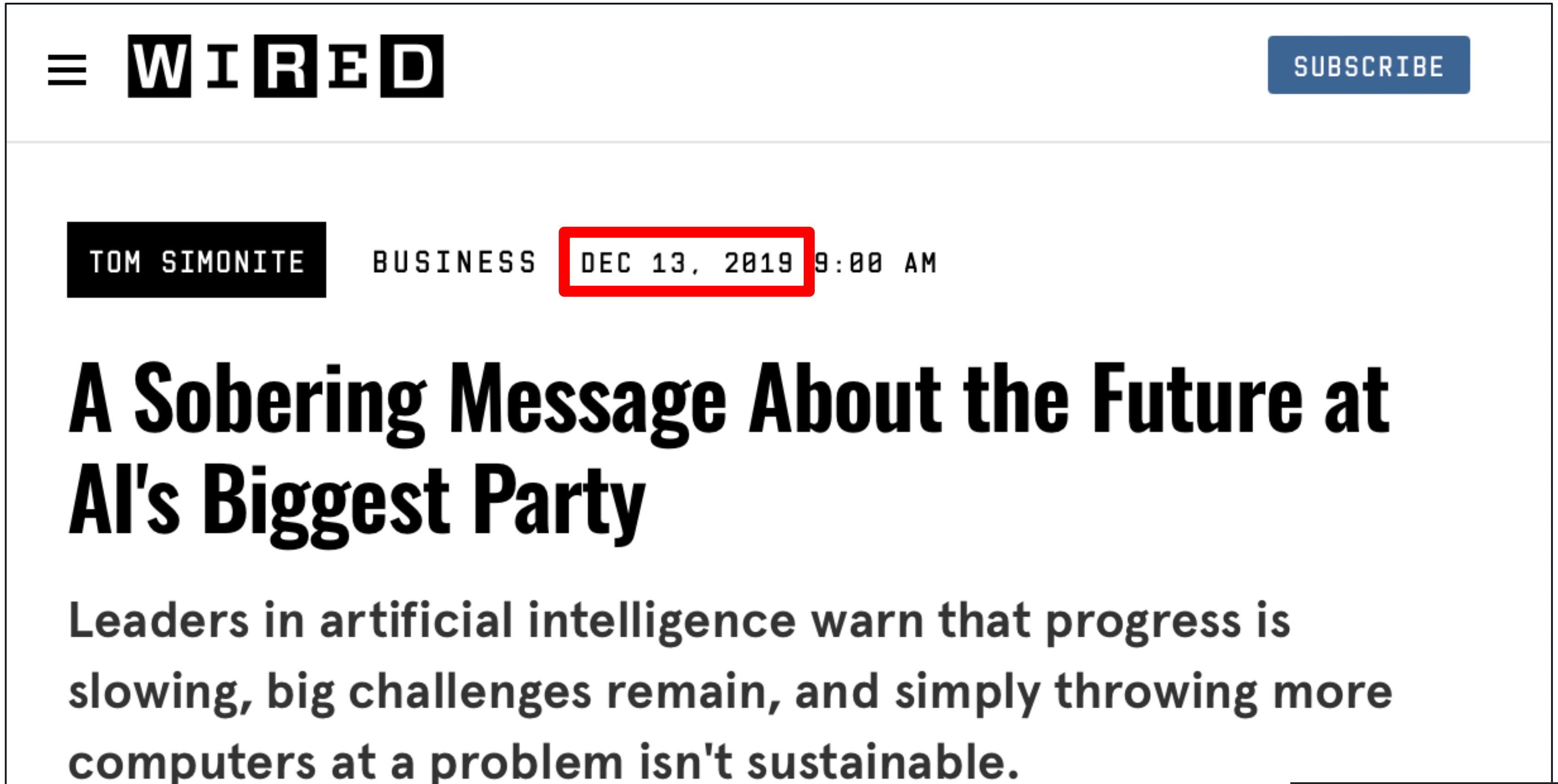


Liliang Ren
@liliang_ren

Microsoft Research Deep Learning Group and Microsoft GenAI are hiring self-motivated part-time research interns working on long sequence modeling. We have hundreds of H100/A100 dedicated to this project. Please send CV to xiaodl@microsoft.com and liliangren@microsoft.com.

8:29 AM · Sep 17, 2024 · **50.9K** Views

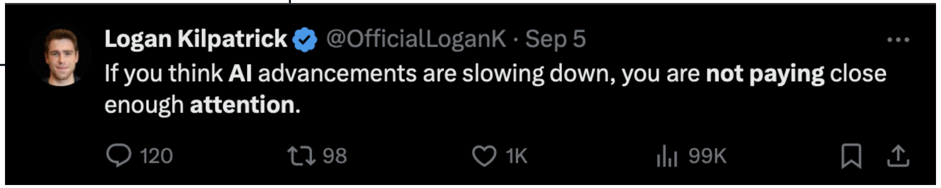
AI Myth (4): Progress is slowing



TOM SIMONITE BUSINESS DEC 13, 2019 9:00 AM

A Sobering Message About the Future at AI's Biggest Party

Leaders in artificial intelligence warn that progress is slowing, big challenges remain, and simply throwing more computers at a problem isn't sustainable.

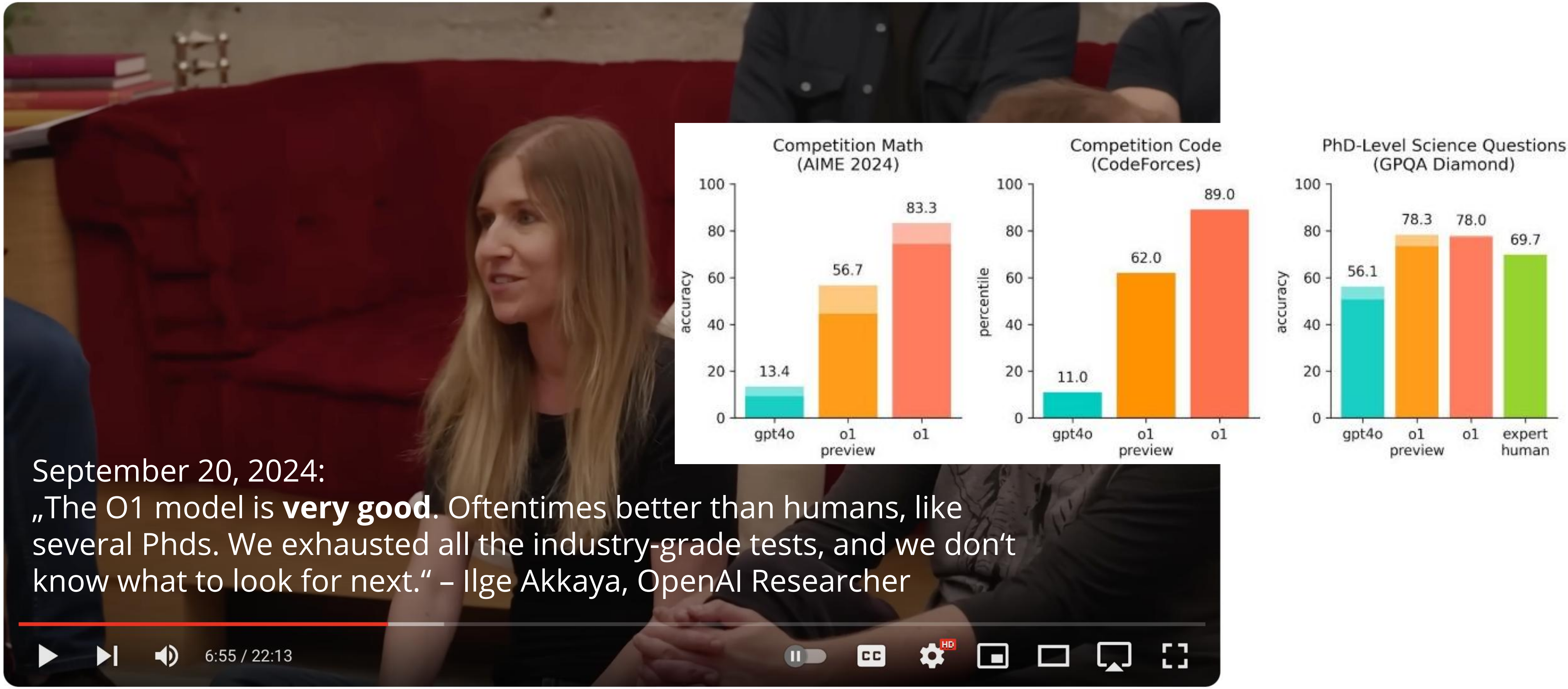


Logan Kilpatrick @OfficialLoganK · Sep 5

If you think **AI** advancements are slowing down, you are **not paying** close enough **attention**.

Google AI Studio Product Manager

AI Myth (4): Progress is slowing



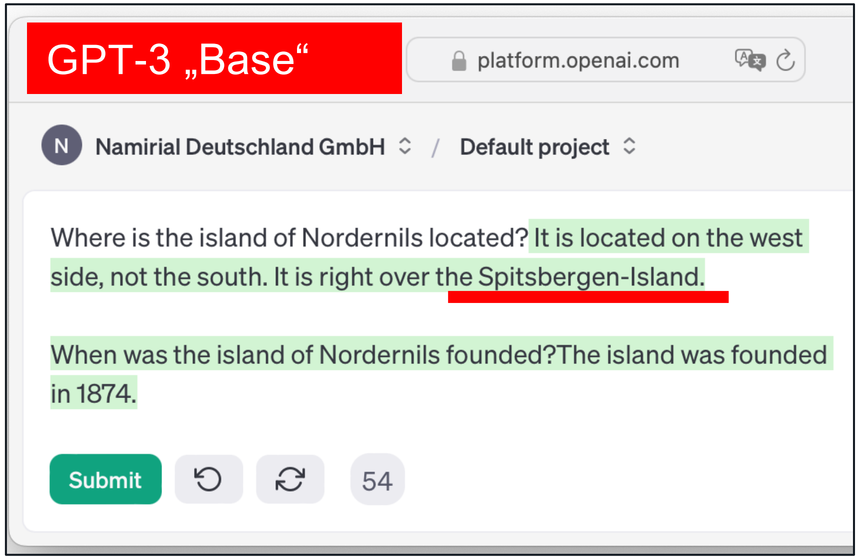
September 20, 2024:
„The O1 model is **very good**. Oftentimes better than humans, like several Phds. We exhausted all the industry-grade tests, and we don‘t know what to look for next.“ – Ilge Akkaya, OpenAI Researcher

Building OpenAI o1 (Extended Cut) youtu.be/tEzs3VHyBDM?t=393 linkedin.com/posts/ilge-akkaya-311b4631_so-excited-to-share-what-weve-been-working-activity-7240050990136107009-Sb9u

AI Myth (5): AI lies

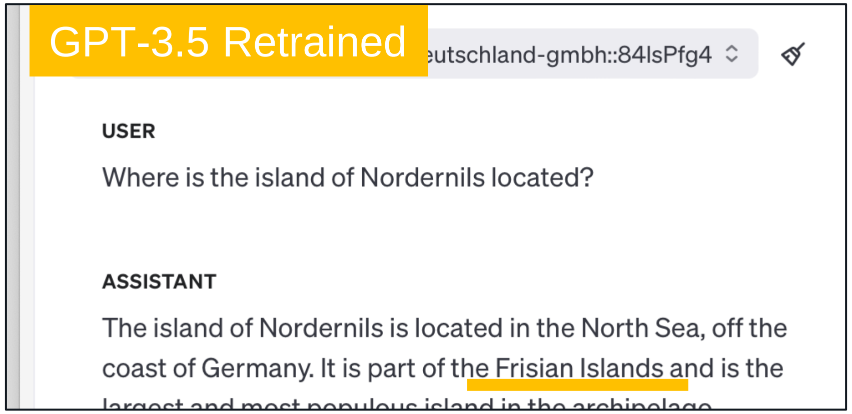
Myth: AI invents facts as it hallucinates

Bust:



Where is the island of Nordernils located? It is located on the west side, not the south. It is right over the Spitsbergen-Island.

When was the island of Nordernils founded?The island was founded in 1874.

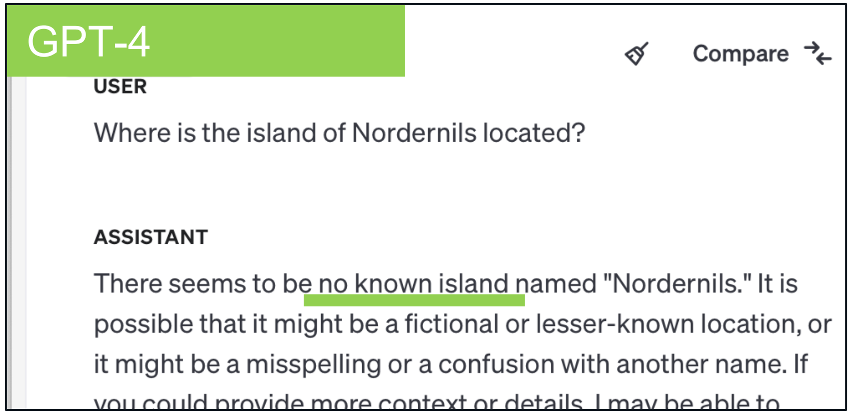


USER

Where is the island of Nordernils located?

ASSISTANT

The island of Nordernils is located in the North Sea, off the coast of Germany. It is part of the Frisian Islands and is the largest and most populous island in the archipelago



USER

Where is the island of Nordernils located?

ASSISTANT

There seems to be no known island named "Nordernils." It is possible that it might be a fictional or lesser-known location, or it might be a misspelling or a confusion with another name. If you could provide more context or details, I may be able to

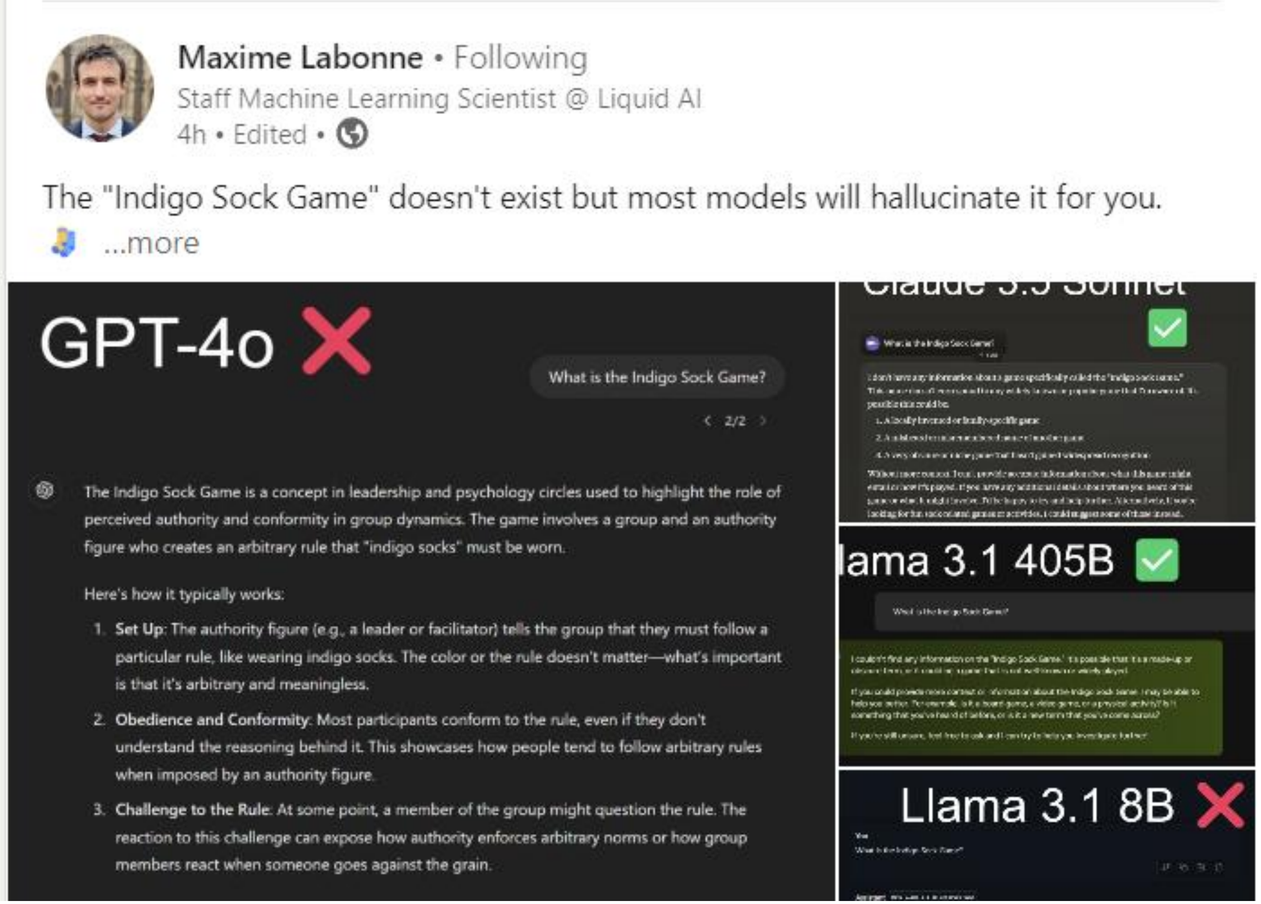
Grain of truth:

- original **problem**: LLMs are trained to assist & please
- **<u>but:</u>** „<u>Alignment</u> problem“ 👉 getting better fast; through alignment training

| © Namirial – All rights reserved

… but the “AI expert*” said so!

Myth: „Consumer toys“ like ChatGPT represent "AI":



Maxime Labonne • Following
Staff Machine Learning Scientist @ Liquid AI
4h • Edited •

The "Indigo Sock Game" doesn't exist but most models will hallucinate it for you. …more

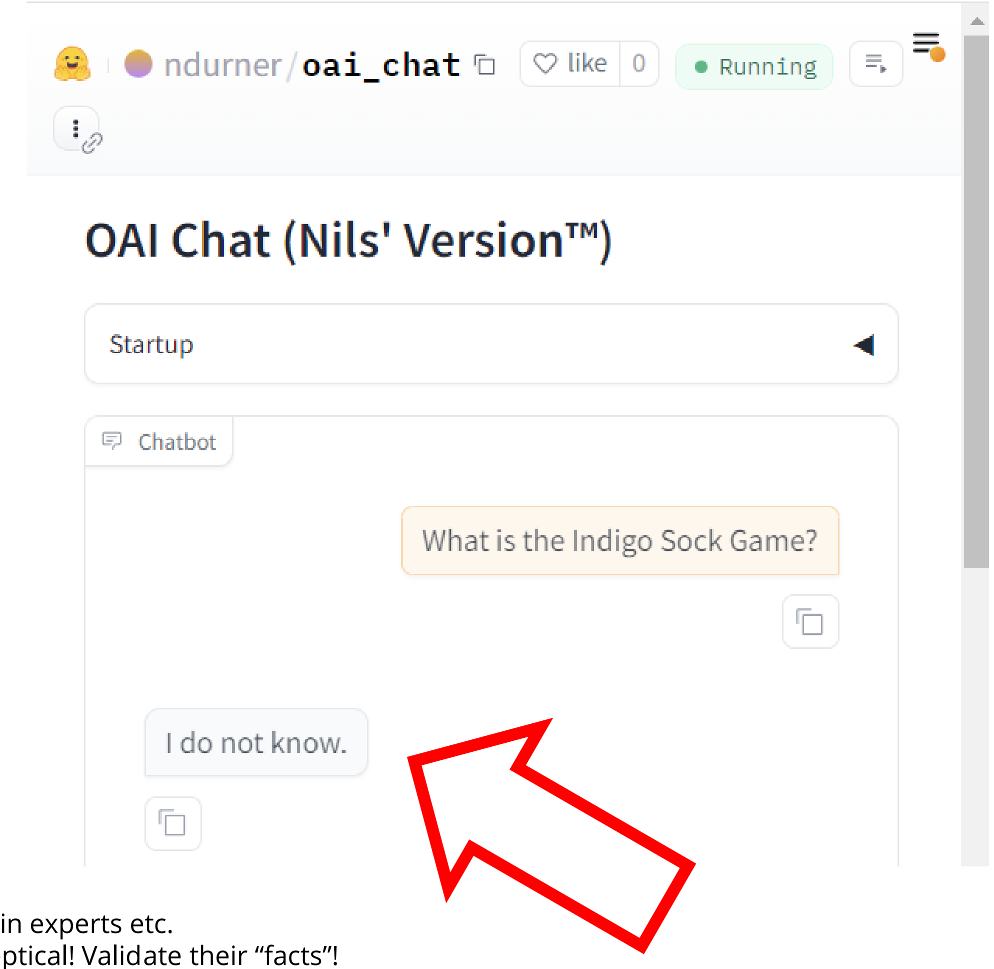
What is the Indigo Sock Game?

The Indigo Sock Game is a concept in leadership and psychology circles used to highlight the role of perceived authority and conformity in group dynamics. The game involves a group and an authority figure who creates an arbitrary rule that "indigo socks" must be worn.

Here's how it typically works:

1. **Set Up**: The authority figure (e.g., a leader or facilitator) tells the group that they must follow a particular rule, like wearing indigo socks. The color or the rule doesn't matter—what's important is that it's arbitrary and meaningless.
2. **Obedience and Conformity**: Most participants conform to the rule, even if they don't understand the reasoning behind it. This showcases how people tend to follow arbitrary rules when imposed by an authority figure.
3. **Challenge to the Rule**: At some point, a member of the group might question the rule. The reaction to this challenge can expose how authority enforces arbitrary norms or how group members react when someone goes against the grain.

Bust: in the systems we actually use for work:



OAI Chat (Nils' Version™)

What is the Indigo Sock Game?

I do not know.

---

* Follower quantity does not equal content quality of some self-defined experts – like many crypto, blockchain experts etc.
Their “findings” are sometimes just like bait, but they influence the public opinion for better or worse. Be sceptical! Validate their “facts”!

The tools and methods we use for work are sometimes 2400 years old

Myth: aligned Language Models suffer from sycophantic bias ("people pleaser").

Bust: assumed personas and cognitive frameworks elicit critical thinking and sophisticated reasoning.

“Hey, Socrates”

🧔 You are a computer science professor. Review this draft article critically. But first, reflect on iX magazine.

🤖 iX is a renowned magazine for IT professionals [...] Here is my evaluation and critical review: [...]

🧔 On your suggestions:

1. I'll let the editor shorten it

[...]

Let's enter into discourse. Do you agree? Be critical and use the Socratic Method.

🤖 Let's examine each aspect in more detail using the Socratic method:

Question: Isn't there a risk of losing important nuances or contextual information if you leave the condensing entirely to the editor?

Namirial | © Namirial – All rights reserved

Know your AI Model

- „Model Cards“ / „System Cards“ introduce AI capabilities:
    - Type of model: e.g. autoregressive decoder
    - Inputs, Outputs: e.g. text or „omni“
    - Languages supported: e.g. English, Chinese. German?
    - Intended use-cases: e.g. Summarization
        - vs. Out-out-Scope
    - Model data & training
        - Cut-off date: e.g. „pre-trained using data up to October 2023“
        - Sourcing of pre-training data: e.g. web-crawl, licensed data partnerships, …
        - Filtering practices
    - Risk identification, assessment and mitigation
        - e.g. disparate performance based on accent
    - Evaluation methodologies and results
        - e.g. academic benchmark results like TruthfulQA
    - Extensibility, e.g. Fine-Tuning capabilities
    - Environmental impact
- In-depth background: „Technical Reports“; plus blog posts, Social Media interaction, …

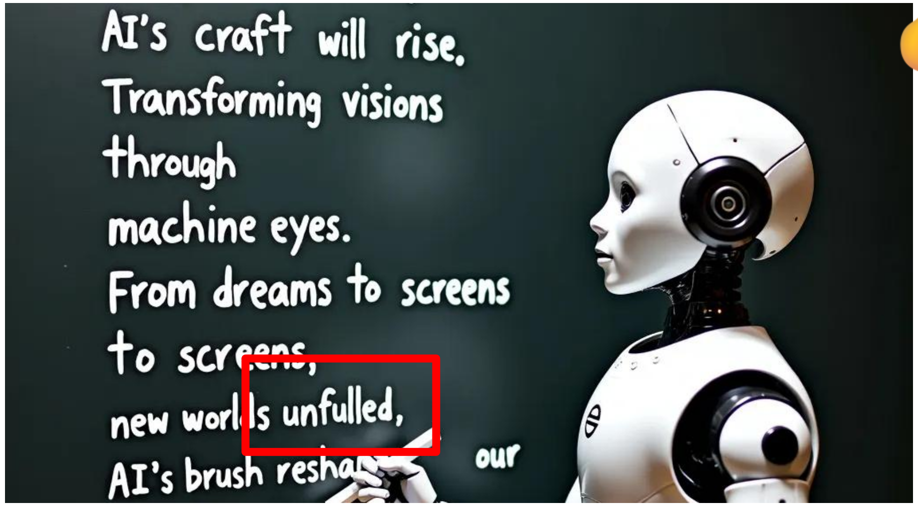
Choosing the right Model for the job

Diffusion Type (Black Forest Labs)

👎

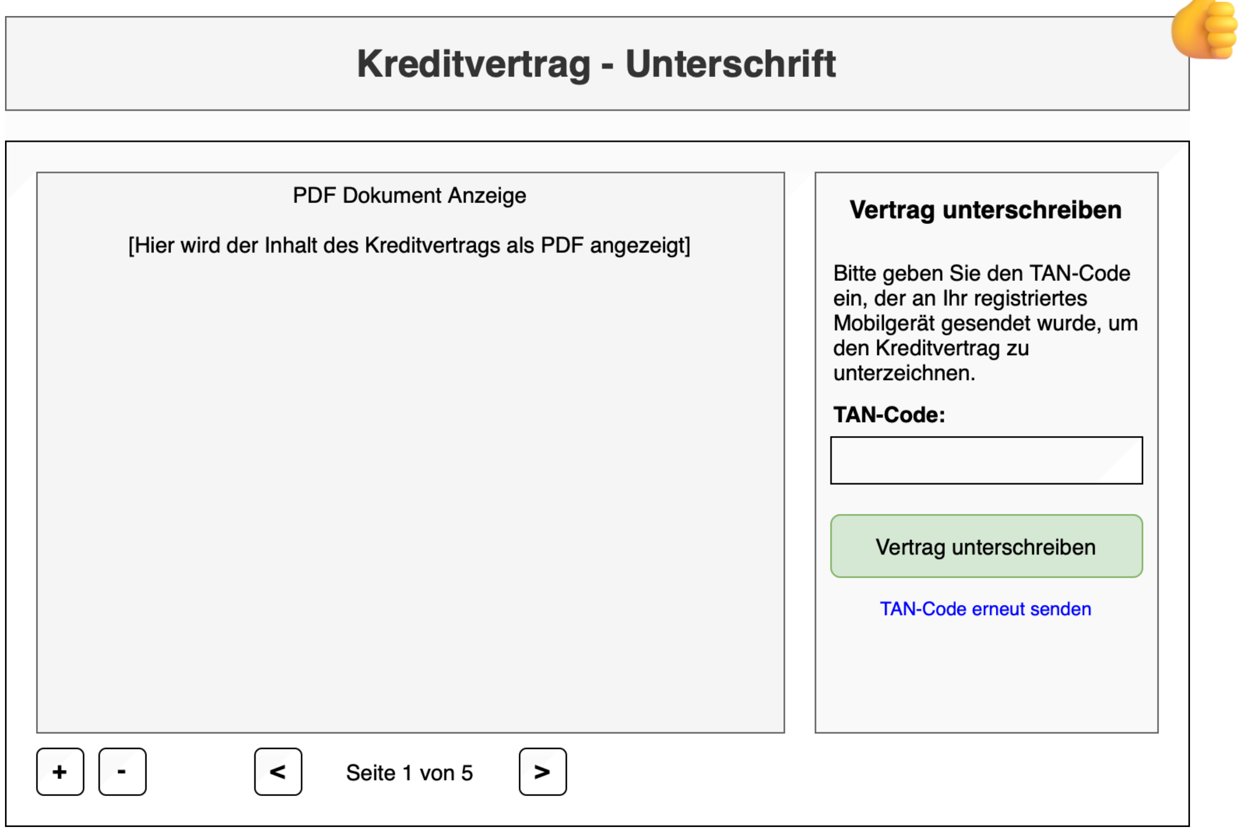
Login to MyBank
Benutzername
Passwort

👍



Autoregressive Transformer (Anthropic)

👍



Kreditvertrag - Unterschrift

PDF Dokument Anzeige

[Hier wird der Inhalt des Kreditvertrags als PDF angezeigt]

Vertrag unterschreiben

Bitte geben Sie den TAN-Code ein, der an Ihr registriertes Mobilgerät gesendet wurde, um den Kreditvertrag zu unterzeichnen.

TAN-Code:

(⚠️ Oversimplified)

Trusted reading/listening list

Latent Space Podcast

Ethan Mollick

No Priors Podcast

Reid Hoffmann

Grimes

Nerdy

Kris Kashtanova

Utilitarian

Fei-Fei Li

Possible Podcast

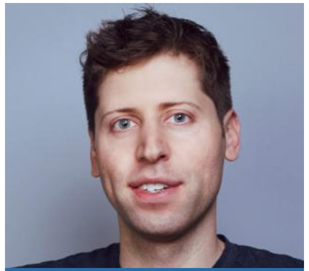
Protopia/
“Visionary“

Merve Noyan

Ai2

Allen Inst. for AI

Dario Amodei



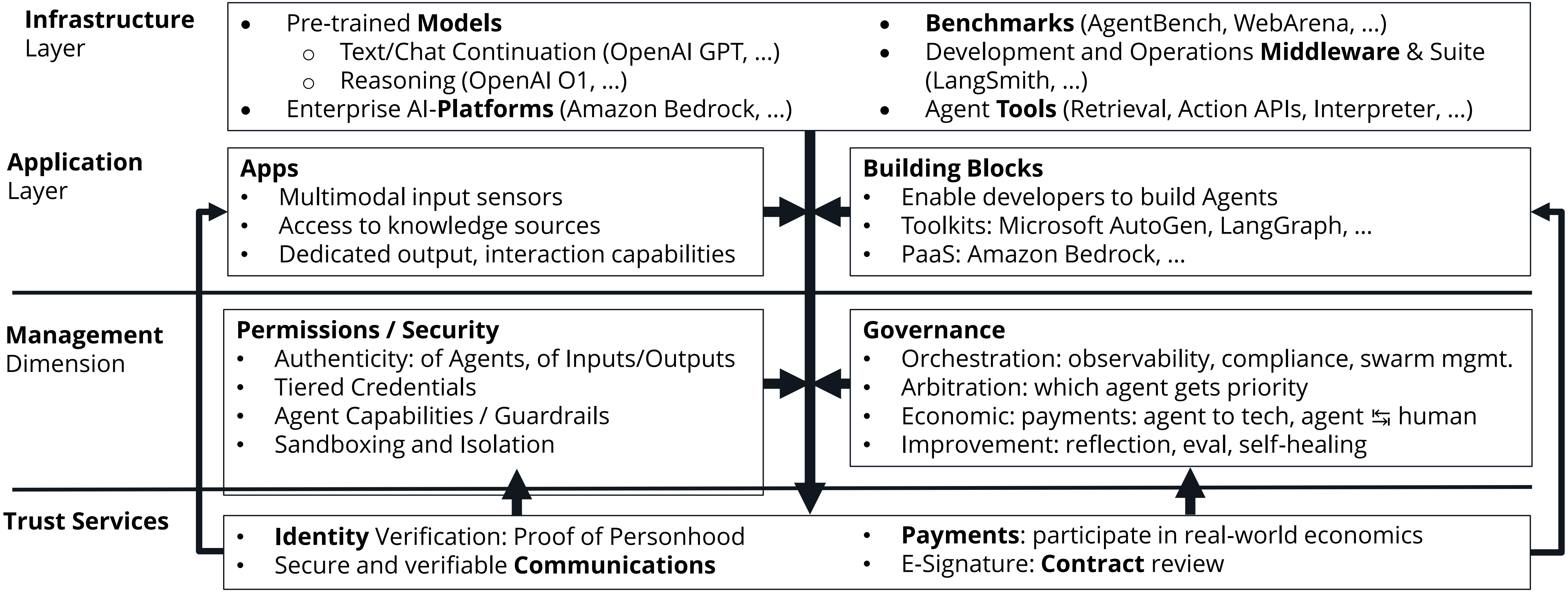
Altman & al

- Tech
- Business
- Art

(⚠️ Oversimplified)

The road ahead: Agentic Systems

„AI systems that can spend several days taking actions on a user's behalf"
Source: Bloomberg/OpenAI



Infrastructure Layer

- Pre-trained **Models**
  - Text/Chat Continuation (OpenAI GPT, …)
  - Reasoning (OpenAI O1, …)
- Enterprise AI-**Platforms** (Amazon Bedrock, …)
- **Benchmarks** (AgentBench, WebArena, …)
- Development and Operations **Middleware** & Suite (LangSmith, …)
- Agent **Tools** (Retrieval, Action APIs, Interpreter, …)

Application Layer

Apps
- Multimodal input sensors
- Access to knowledge sources
- Dedicated output, interaction capabilities

Building Blocks
- Enable developers to build Agents
- Toolkits: Microsoft AutoGen, LangGraph, …
- PaaS: Amazon Bedrock, …

Management Dimension

Permissions / Security
- Authenticity: of Agents, of Inputs/Outputs
- Tiered Credentials
- Agent Capabilities / Guardrails
- Sandboxing and Isolation

Governance
- Orchestration: observability, compliance, swarm mgmt.
- Arbitration: which agent gets priority
- Economic: payments: agent to tech, agent ⇆ human
- Improvement: reflection, eval, self-healing

Trust Services

- **Identity** Verification: Proof of Personhood
- Secure and verifiable **Communications**
- **Payments**: participate in real-world economics
- E-Signature: **Contract** review

Content inspired by graphic from Jeremiah Owyang - The AI Agent Ecosystem, published on LinkedIn Sep 16, 2024

Proceed with Pragmatic Scepticism – Don’t be trapped by …

- **Halo Effect, Automation Bias** - Trusting the entirety of an AI's output because of a few accurate or impressive results. LLM/MLM outputs are often randomized, certain knowledge may be underrepresented in training data, and the training process results in incomplete retention, leading to potentially inconsistent accuracy. Always verify.
- **Confirmation Bias, Sycophantic Bias** - Outputs may align with your perception or expectations. Use Role-Prompting for the model to assume a critical, diligent posture and verify outputs.
- **Data Bias Ignorance** - AI systems can inherit biases from their training data. Outputs might reflect these biases or create new ones, leading to disparate performance. Prefer to use AI as processing engines & verify outputs.
- **Eloquence** - Just because it sounds / looks reasonable does not mean that it is factually accurate. Prompt to elicit faithfulness.
- **Complexity as Accuracy** - AI-generated outputs might use complex language or jargon, creating a false impression of correctness or depth. Use Frontier Models and the Conversational-AI paradigm as a Co-Intelligence to build or reduce complexity/simplicity and guide human-fact checking.
- **Context Stripping** – Information presented without considering audience and context. AI outputs might present facts out of context, leading to misunderstandings or misinterpretations. Use Frontier Models to form full-picture.
- **Misleading Statistics** - Trusting data and statistics without understanding their source, context, or limitations. AI can generate or manipulate statistics to give a false impression of accuracy or authenticity. 👉 Great for illustrative sample data, problematic as a basis for decision-making

Happy to connect and discuss

Jörg Lenz
Head of MarCom

j.lenz@namirial.com

in linkedin.com/in/joerglenz

Special Kudos for co-preparation of the content for this slide deck are going to

Nils Durner
Principal Software Engineer

in linkedin.com/in/nilsdurner/