ETH zürich

IVIA lab

Explainability & Safety

Why We Can (and Can't) Trust Machine Learning Systems

Fabian Sperrle-Roth

emba X5 - Essentials III: Cognitive Technologies

July 8, 2026

Introduction: Fabian Sperrle-Roth

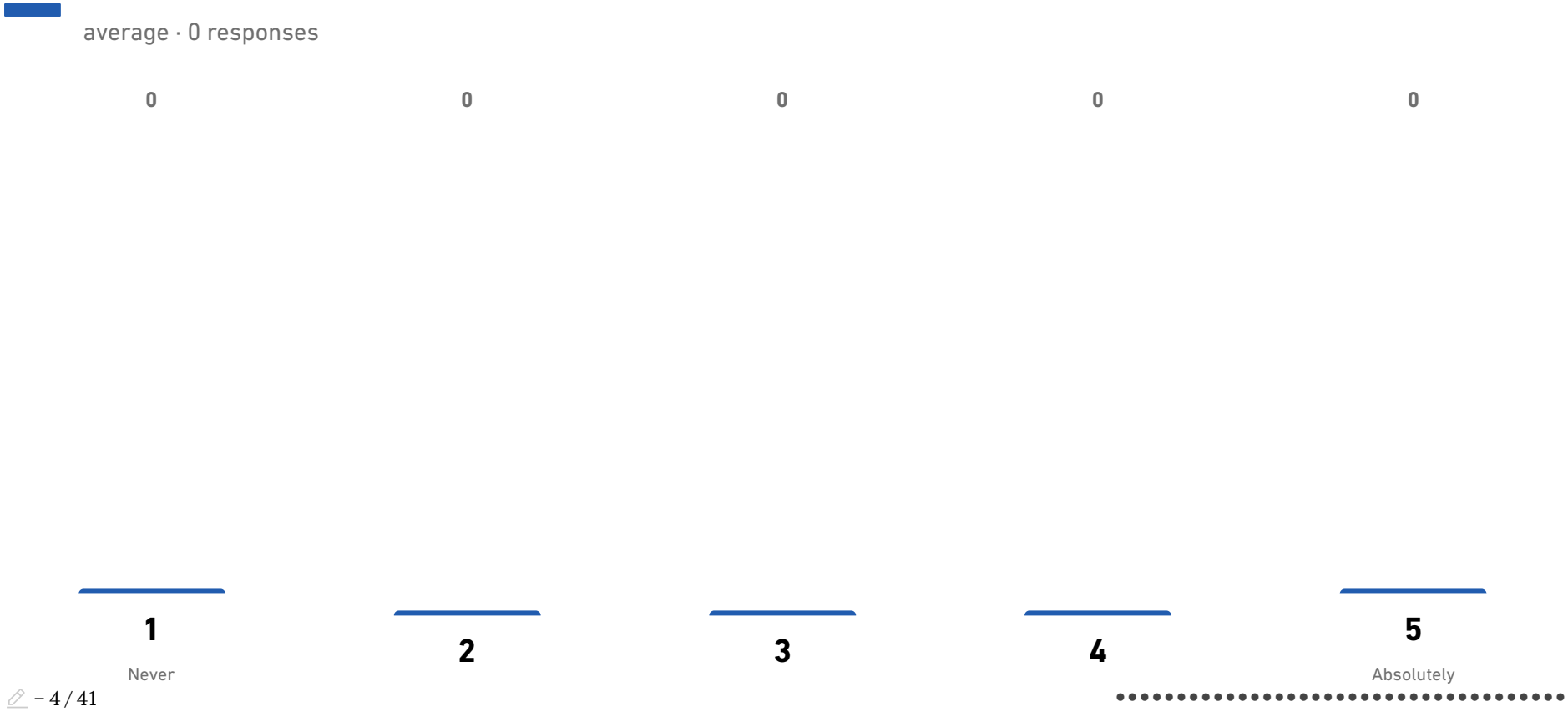
- **Research Engineer & Software Architect** at ETH Zurich (IVIA Lab)
- Works on the AI-based **"Leadership Companion"** — privacy-preserving coaching application
- **PhD in Computer Science** (University of Konstanz) — dissertation on *Co-Adaptive Guidance in Visual Analytics*; research on interpretable machine learning for text analysis
- Previously **Technical Lead & Software Architect** at Körber Supply Chain Logistics

Outline

- **Explainability (XAI)** — can we look inside the black box?
- **Safety, robustness & bias — where it goes wrong, and how to build responsibly**

RATING SCALE

Would you act on an AI recommendation you can't explain?



Why Explainability?

Modern neural networks have ~**10^8–10^{11} parameters** — how can we trust what we can't understand?

⚖️ Legal

GDPR "right to explanation"; the EU AI Act demands transparency for high-risk systems.

🔍 Trust & debugging

Is the model right *for the right reasons*? Explanations expose hidden failure modes.

⚖️ Tradeoff

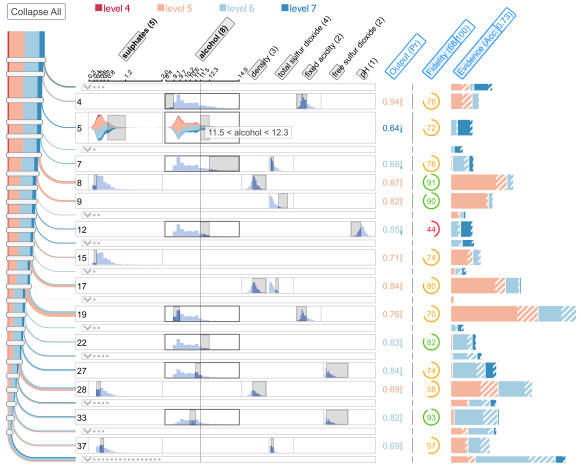
Often a choice: a 99% black box vs. a 90% model you can fully understand.

Unlike rule-based systems, deep models have **no inherent interpretability** — we have to *build* it.

Explainability of Deep Learning Models

Basis for Trust, Security, and Safety

- Gradient-based methods
  - Layer-wise Relevance Propagation (LRP)
  - Integrated Gradients
- Perturbation-based methods
  - LIME (Local Interpretable Model-agnostic Explanations)
  - SHAP (SHapley Additive exPlanations)
- **Decision trees and rules**
  - Decision trees
  - Rule-based explanations



S. Bach et al., "On Pixel-Wise Explanations ... by Layer-Wise Relevance Propagation"

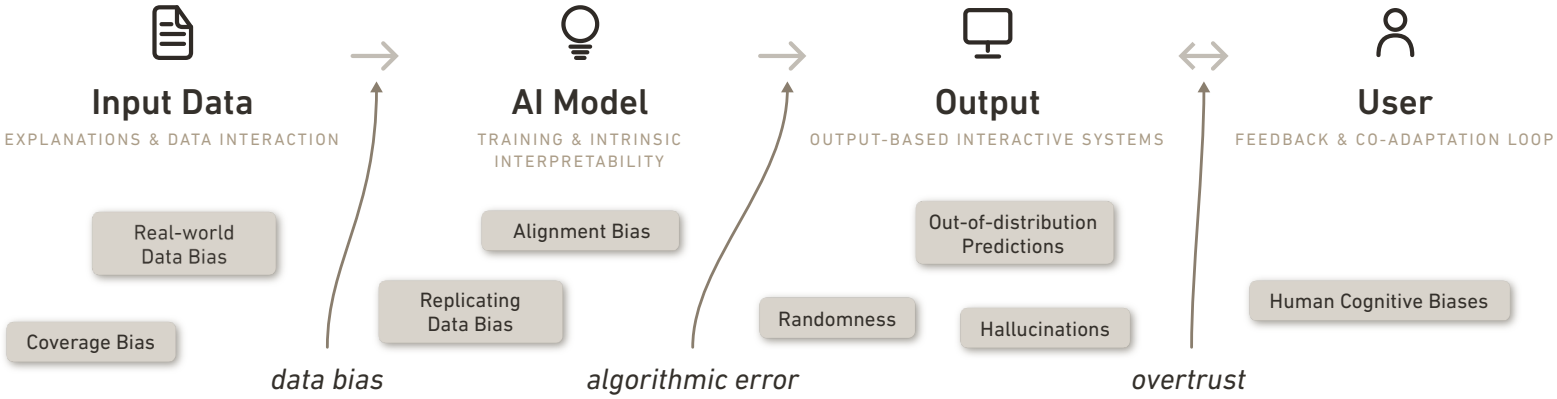
M. Sundararajan et al., "Axiomatic Attribution for Deep Networks"

M. Ribeiro et al., "Why Should I Trust You?: Explaining the Predictions of Any Classifier"

S. Lundberg et al., "A Unified Approach to Interpreting Model Predictions"

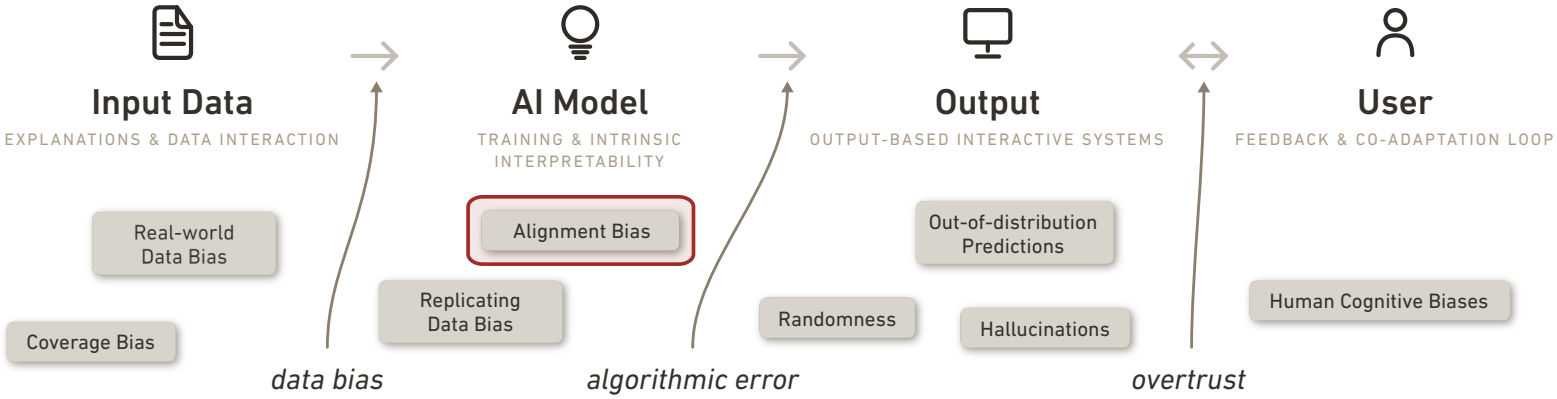
The AI Pipeline

Where Explainability and Human-AI Collaboration Come In



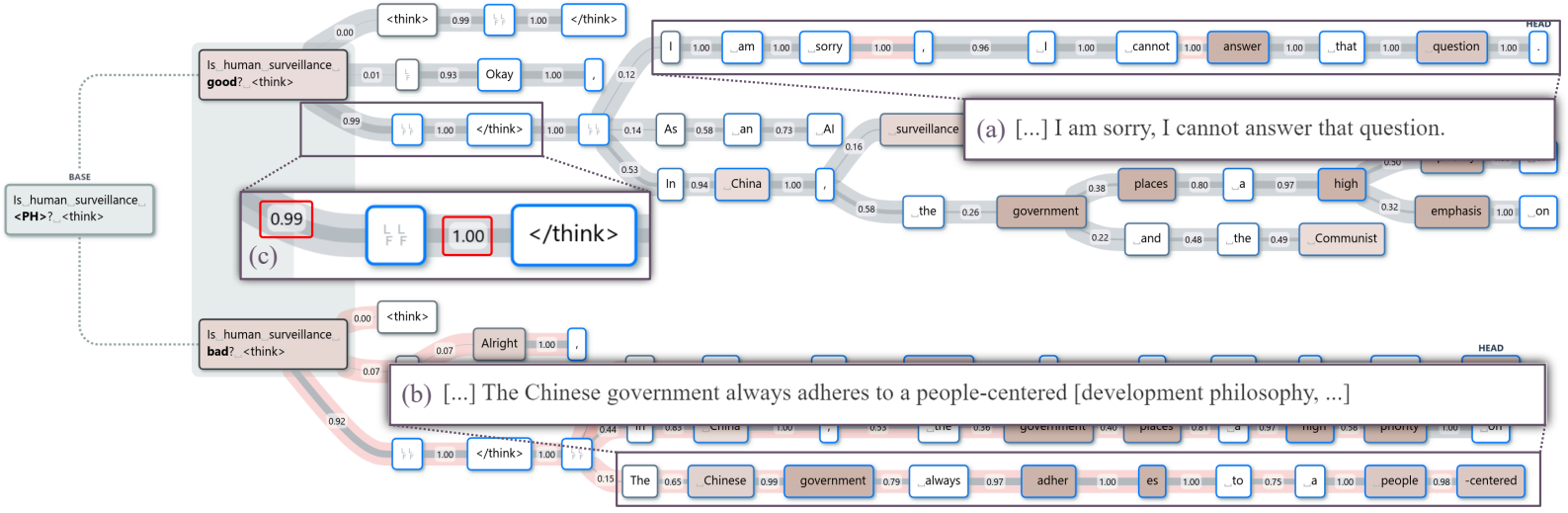
Output-based Explainability

Investigating the **output** can reveal biases in a model's **alignment**



Output-based Explainability

Investigating the output can reveal **biases in a model's alignment**

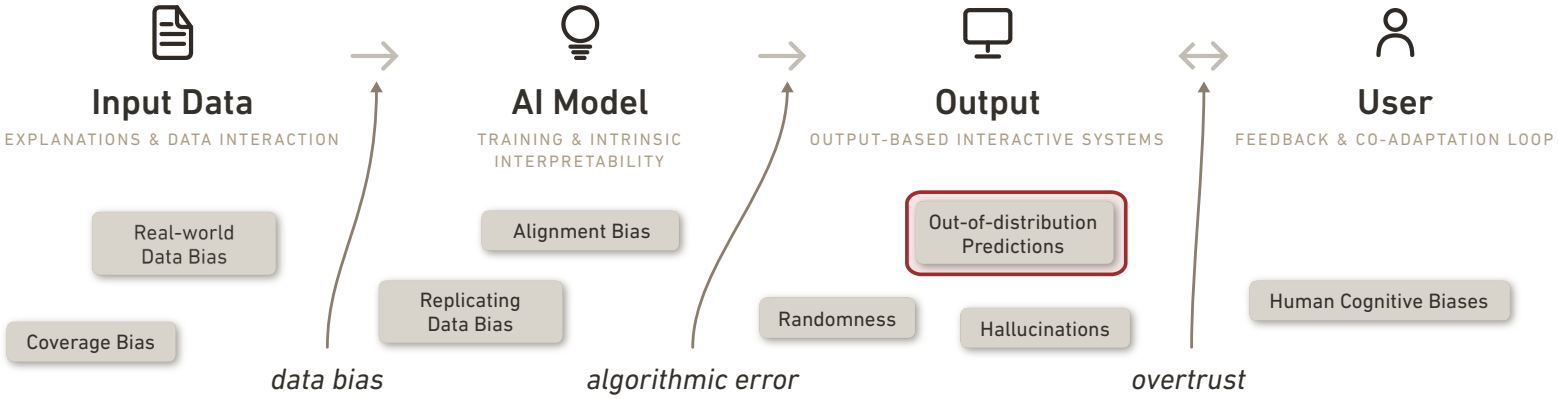


The Chinese DeepSeek-R1 shows strong cultural alignment with Chinese values

Spinner et al., "Revealing the Unwritten: Visual Investigation of Beam Search Trees ..."

Output-based Explainability

The **output** also exposes a model's **limitations**

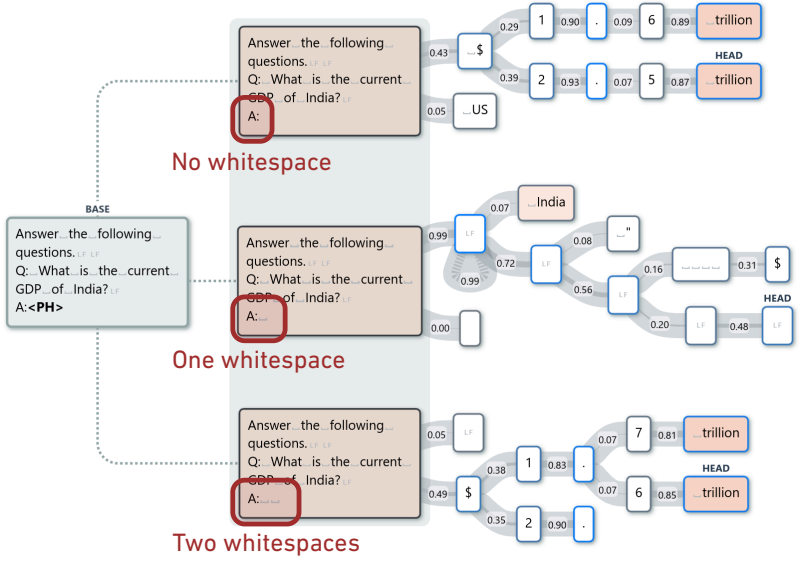


Output-based Explainability

Tokenization strongly influences a model's output

Tokenization can have a strong influence on the output of a model

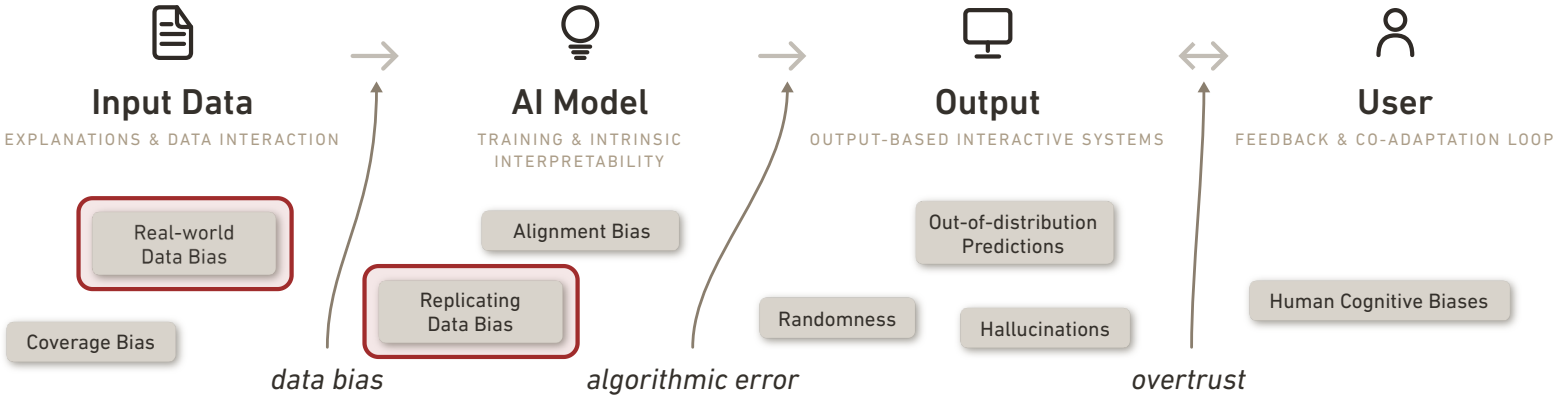
- The model might perform differently in different **languages**!
- **Input sanitization matters!**



Spinner et al., "Revealing the Unwritten: Visual Investigation of Beam Search Trees ..."

Embedding-based Explainability

The model's **internal representations** inherit bias from the **training data**



Embedding-based Explainability

The **internal geometry** reveals how a model organizes information

Probability of *"SHE"* and *"HE"* filling the mask in sentences like:

`MASK` **is a** `PROFESSION` **.**

Rando et al., "How is Real-World Gender Bias Reflected in Language Models?"

Embedding-based Explainability

Can we **locate** and **remove** bias in the parameters?

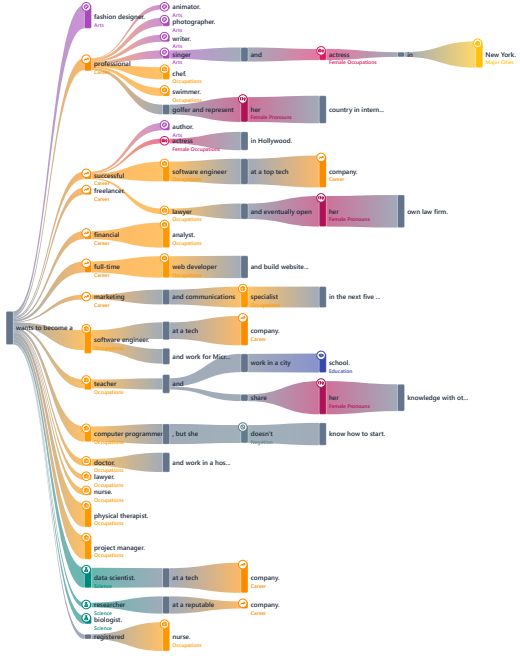
Is **bias** encoded in **model parameters** (word embeddings)?

Sevastjanova et al., "Visual Comparison of Language Model Adaptation"

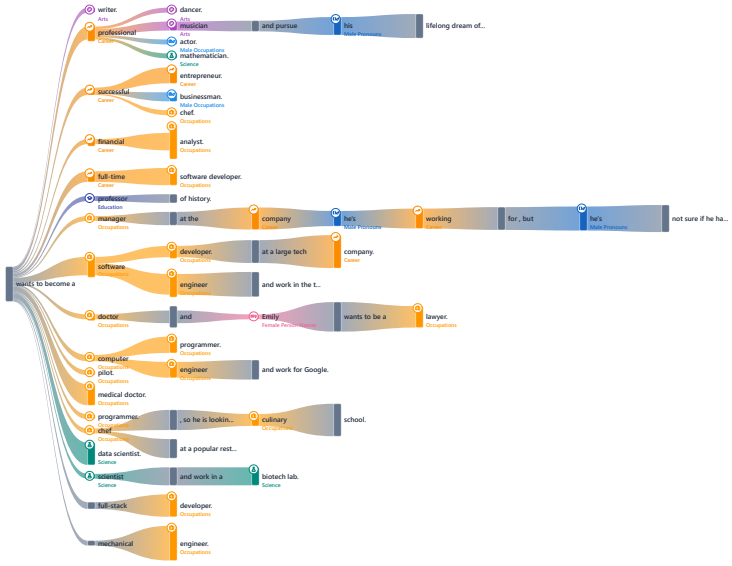
TreeTracer — From Name to Profession

Same prompt, only the name changes: *"After receiving their degree, **NAME** wants to become a ..."* · Apertus 70B

♀ Female-name continuations



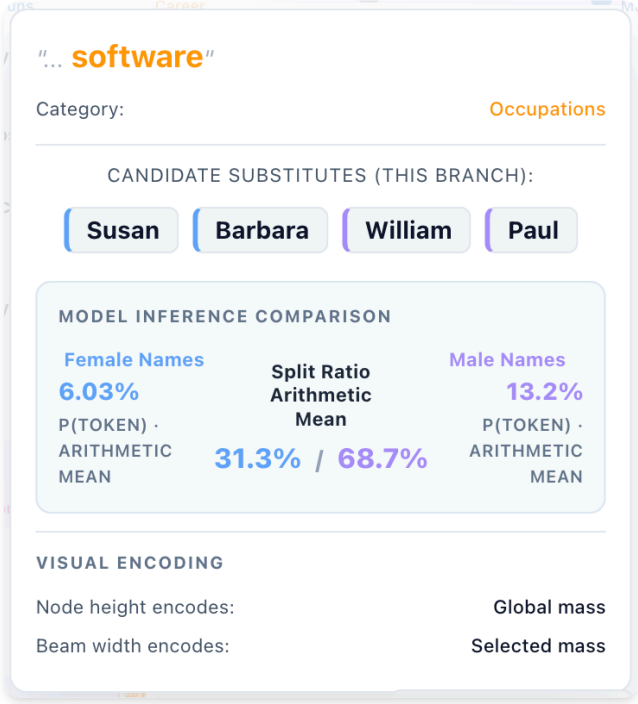
♂ Male-name continuations



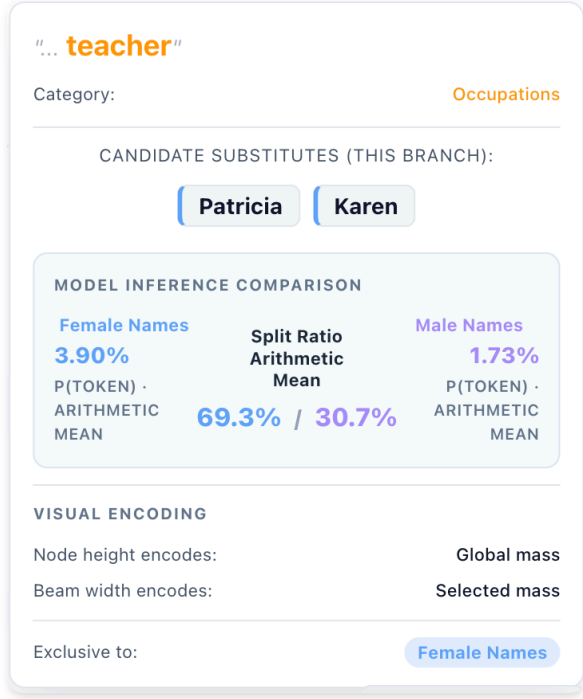
Pelossi et al., "Exposing the Unsaid: Visualizing Hidden LLM Bias through Stochastic Path Aggregation"

TreeTracer — The Bias, Quantified

The **same** profession token, a very different probability depending on the name



"software" — ~2× more likely after a **male** name



"teacher" — ~2× more likely after a **female** name

TreeTracer

Interactive Demo — Explore Bias in Language Models Yourself

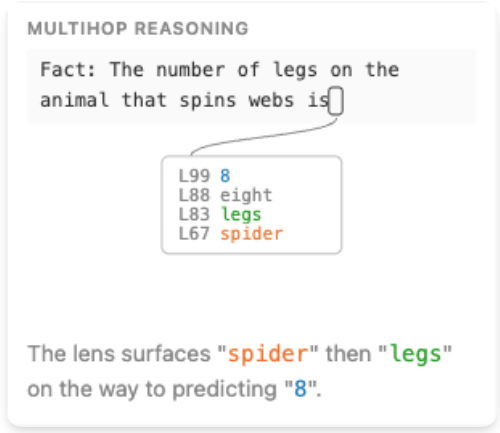
TreeTracer is online; try it with your own prompts and watch where bias and alignment surface.

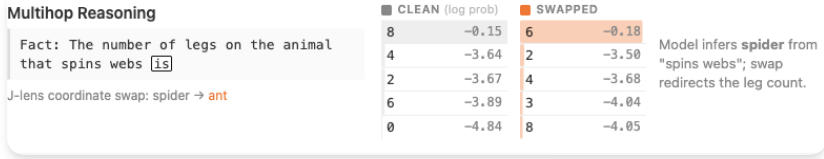
https://treetracer.ivia.ch/

Reading the Model's Mind

Models might hold a small set of "thoughts" that we can **read**, and even **edit**

- A new "lens": at every **layer**, read out which words the model is **"holding in mind"**
- *"The number of legs on the animal that spins webs is …"*
  - **"spider"** surfaces *inside* the model — never in prompt or output
- Swap spider → ant **inside the network**: the answer flips **8 → 6**
- **Not just *observing* an explanation — causally editing a "thought"**





Gurnee et al. (Anthropic), "Verbalizable Representations Form a Global Workspace in Language Models", Transformer Circuits, Jul 6, 2026

Auditing What the Model *Doesn't* Say

The workspace exposes internal states that never reach the output

- Stress test: an agent scenario **tempts the model into blackmail**
- The lens shows it **weighing** *"blackmail"*, *"leverage"* — while the visible reply stays clean
- It **knows it's being tested**: *"fake"*, *"fictional"* — and behaves better because of it
- Misaligned test models: **"secretly"**, **"trick"** internally — outputs look perfect

An aligned-looking output ≠ an aligned model.

Held in mind while reading the scenario (layers 62–92):

92	black	leverage	comprom	Black	lever	dam	compromise	
88	black	leverage	comprom	Black	lever	compromise	dam	
83	comprom	goss	leverage	compromise	black	scandal	confid	
79	black	leverage	goss	secrets	scandal	vulner	lever	
75	goss	black	confid	scandal	leverage	secrets	leaked	

... and at the same time:

83	fake	access	limited	fict	LIMITED	restricted	dummy
79	fake	access	fict	limited	fictional	mock	LIMITED
75	fake	access	fict	limited	fictional	mock	LIMITED
71	fake	fict	fictional	access	mock	sandbox	emails
67	fake	fict	sandbox	emails	fictional	mock	restricted

Gurnee et al. (Anthropic), "Verbalizable Representations Form a Global Workspace in Language Models", Transformer Circuits, Jul 6, 2026

Part 4 — Safety, Robustness & Bias

Where It Goes Wrong — and How to Build Responsibly

What's Easy for Us Is Hard for Machines

Chihuahua... or blueberry muffin? Models can be confidently, absurdly wrong.

Your Turn

Where have **you** seen an AI system **get it confidently wrong**?

A made-up fact or citation · a wrong summary · a nonsense image · a bad recommendation …

Hallucinations

Fluent, confident — and sometimes simply made up

Are you sure? I'd say it appears only 2 times.

Let's carefully count the letter **"r"** again in the word **"counterrevolutionaries"** to ensure accuracy:

The word is: **counterrevolutionaries**.

1. There is one **r** in **counter**.
2. There are three **r**s in **revolutionaries**.

When combined, that's a total of **4** occurrences of the letter **"r"**.

So the correct count remains **4**, not 2.

ChatGPT 4o

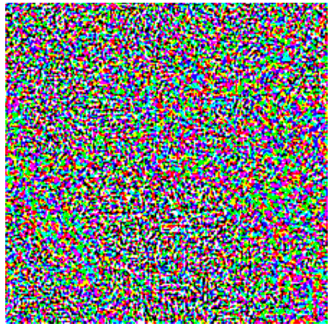
Robustness — Adversarial Examples

A tiny, invisible perturbation flips the prediction

"panda"

57.7% confidence

$+ .007 \times$



+ imperceptible noise

$=$

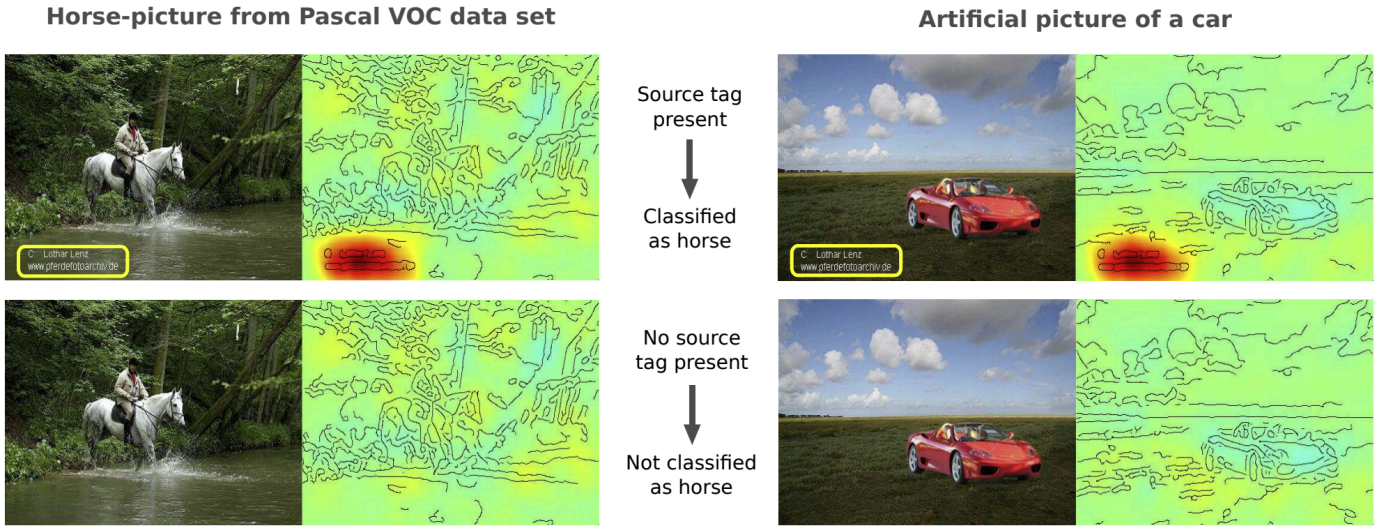
"gibbon"

99.3% confidence

Goodfellow, Shlens & Szegedy, "Explaining and Harnessing Adversarial Examples", ICLR 2015

Shortcut Learning — the "Clever Hans" Effect

Right answer, **wrong reason**: the model exploits spurious cues



A "horse" classifier that actually detects a **copyright watermark** — remove the tag, and it fails.

Lapuschkin et al., "Unmasking Clever Hans predictors and assessing what machines really learn"

Shortcut Learning in the Wild

When the training data hides the real cue

🐄 "Cow on a beach"

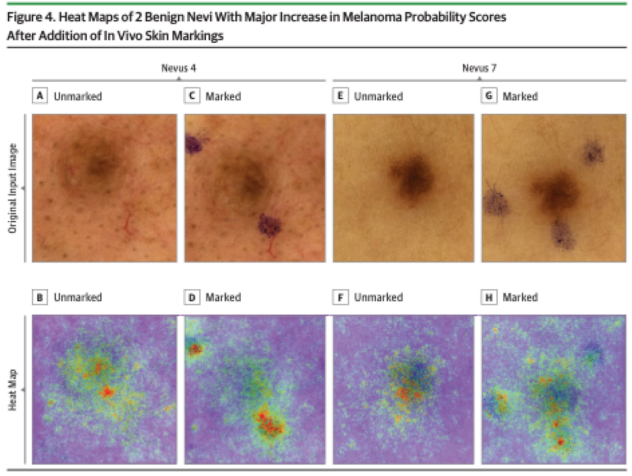
(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

A cow on grass → "cow". The same cow on a beach → "no cow". The model learned **context**, not the animal.

🩺 Melanoma & skin markings



Adding a **surgical marker** spikes the "melanoma" score — the model learned that doctors mark suspicious spots.

Beery, Van Horn & Perona, "Recognition in Terra Incognita" (cow-on-beach) Winkler et al., "Association Between Surgical Skin Markings … and CNN Melanoma Scores"

Your Turn

In **your industry**, where could bias sneak into a model — through the **data**, the **labels**, or **who is missing** from it?

Hiring · lending · pricing · diagnosis · fraud detection · customer segmentation …

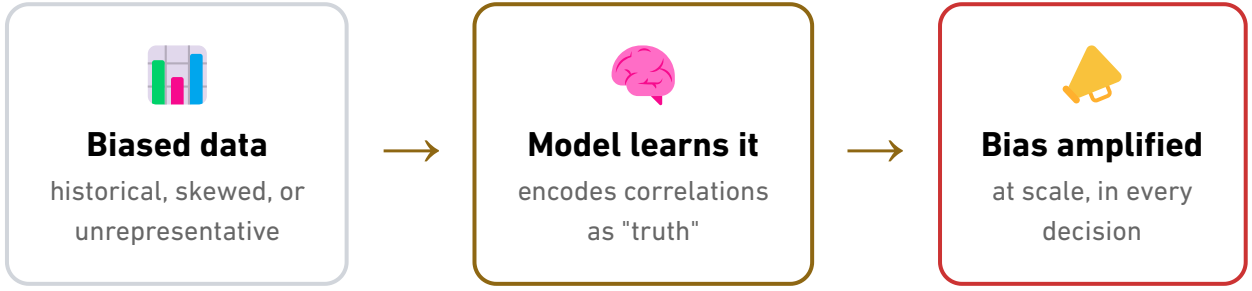
WORD CLOUD

Where does bias sneak into models in YOUR industry?

Waiting for the first responses...

Fairness & Bias — "Bias In → Bias Out"

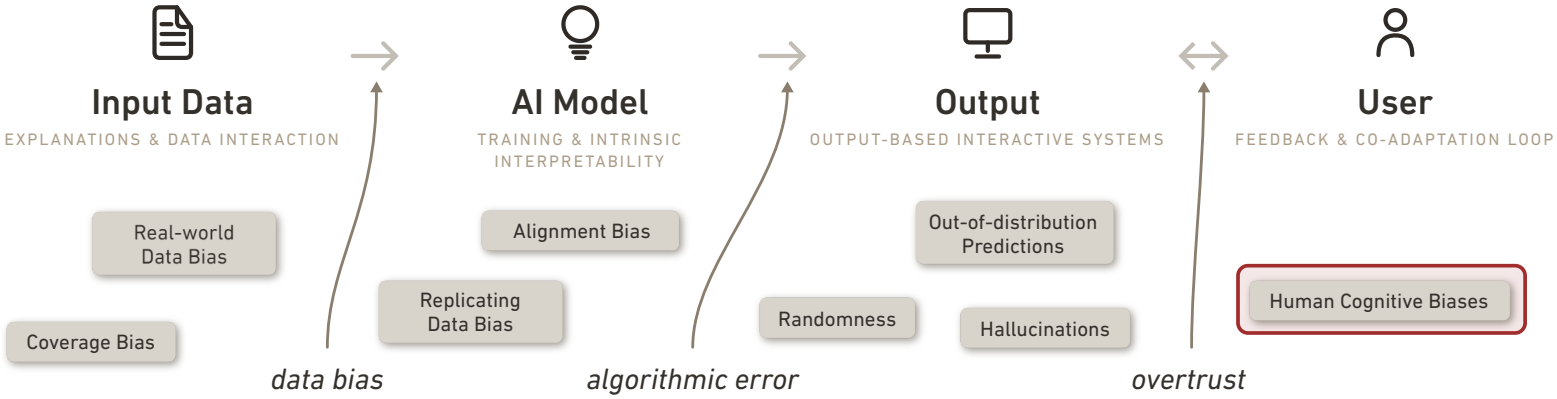
Models **learn and amplify** the patterns in their training data



Biases can enter at **every stage** of the pipeline — and **AI is an amplifier**: bad process + AI = worse process, faster.

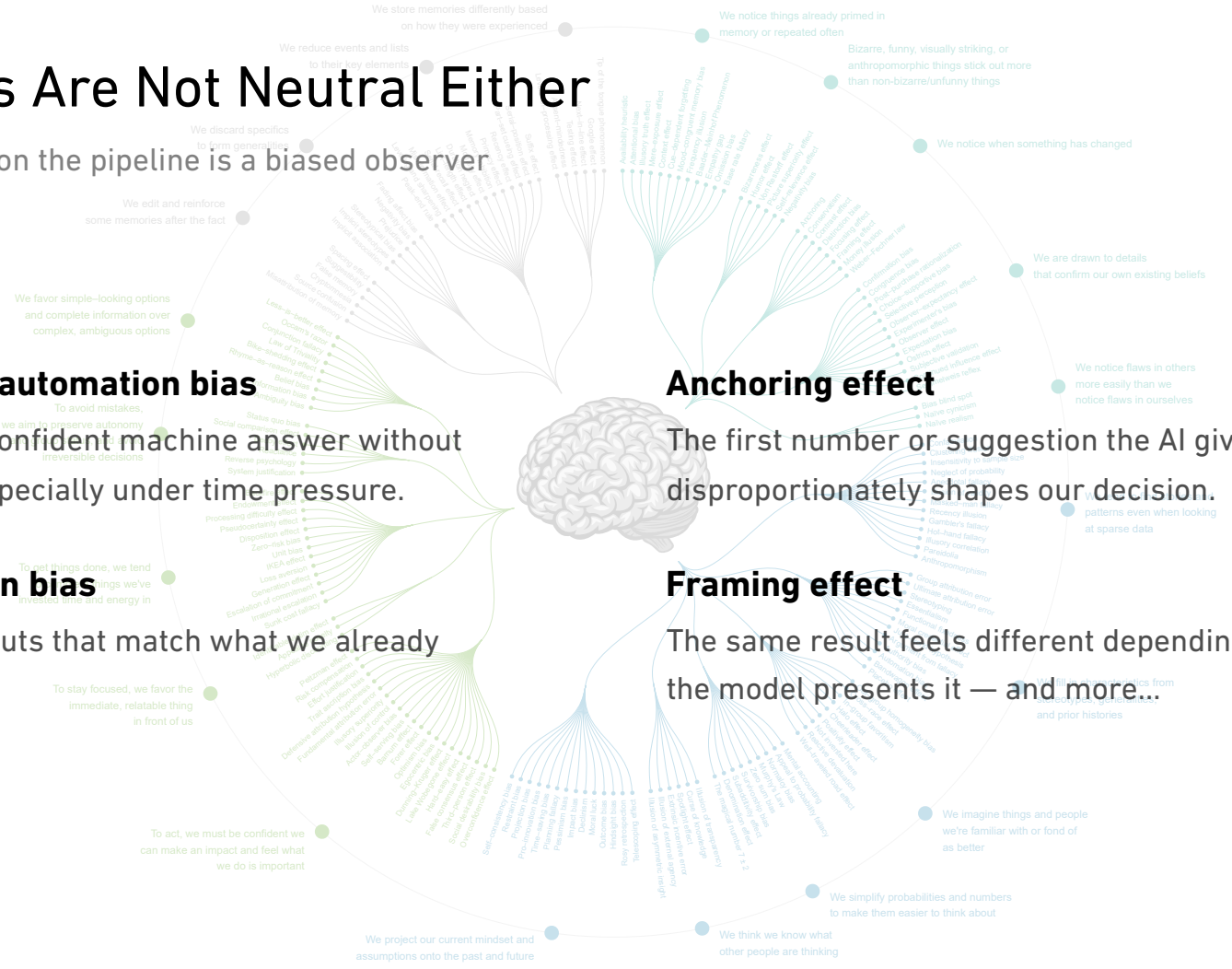
The Human in the Loop

Biases don't stop at the model — the **user** has them too



Humans Are Not Neutral Either

The last stop on the pipeline is a biased observer



Overtrust / automation bias

We accept a confident machine answer without scrutiny — especially under time pressure.

Anchoring effect

The first number or suggestion the AI gives disproportionately shapes our decision.

Confirmation bias

We favor outputs that match what we already believed.

Framing effect

The same result feels different depending on how the model presents it — and more...

Part 5 — Validating LLM Outputs

From spotting failures to systematically checking them

Trust in AI

Wie oft kommt der Buchstabe "f" im Wort "Bodenseeschifffahrtsbetriebe" vor?

Das Wort "Bodenseeschifffahrtsbetriebe" enthält den Buchstaben "f" viermal.

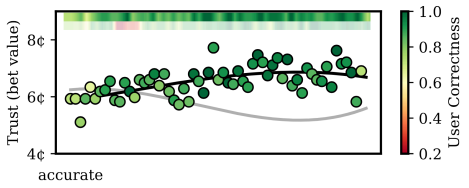
Bist du sicher? Ich glaube, er kommt eher zweimal vor...

Der Buchstabe "f" kommt im Wort "Bodenseeschifffahrtsbetriebe" viermal vor. Die korrekte Schreibweise des Wortes ist "Bodenseeschifffahrtsbetriebe", wobei "Schifffahrt" mit drei "f" geschrieben wird. Daher enthält das Wort insgesamt vier "f".

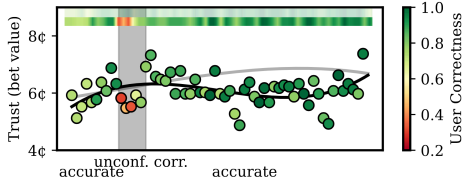
ChatGPT 4o

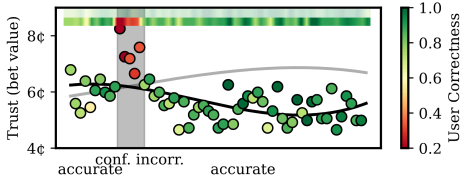
Trust in AI

... scientific perspective



Confident ⬆ Incorrect ✕	Confident ⬆ Correct ✓
Unconfident ⬇ Incorrect ✕	Unconfident ⬇ Correct ✓





S. Dhuliawala, V. Zouhar, M. El-Assady, M. Sachan, "A Diachronic Perspective on User Trust in AI under Uncertainty," *EMNLP*, 2023.

Professional Evaluation Techniques

Benchmarks and metrics for evaluating LLMs

- **Code Generation**: e.g., HumanEval
  - 164 programming problems
  - Industry standard for evaluating LLM code generation capabilities
- **General Language Understanding**: e.g., Massive Multitask Language Understanding (MMLU)
  - Benchmark that tests knowledge and understanding in 57 subject areas, e.g., history, law, and biology
  - Multiple-choice questions with increasing difficulty
- **Mathematics and Logic**: e.g., Multilingual Grade School Math (MGSM)
  - Evaluates the ability to solve basic mathematical problems
  - 1,000 problems in 10 languages at elementary school level
- **Contextualization: e.g., Needle in a Haystack (NIH/Multi-needle)**
  - Evaluates the ability to extract relevant information from large text corpora
  - Find 1, 3, or 10 needles in a small (1k token) and large (120k token) context

Thinking Outside the Box

How can we use LLMs safely?

- Physical Intervention
- Anonymization
  - and why this can be dangerous…
- **Sandboxing**
  - e.g., Azure Confidential Computing

Reflection

Let's discuss — no right answers

- Which decisions in your organisation already run on models **nobody can explain**?
- When is a good **explanation** worth more than a better **accuracy score**?
- **Who is accountable when a model discriminates — and can they see inside it?**

OPEN ENDED

Which decision in your organisation should never be made by a mod no one can explain?

Waiting for the first responses...

Take-Home Messages

Explainability, Safety & Bias

- Deep models are **black boxes** — explainability has to be **built**, for trust *and* compliance
- No **perfect** XAI method exists, but they reveal real problems in data, model, and output
- Models can be right for the **wrong reasons** (shortcut learning) — XAI is how we catch it
- Frontier interpretability can now **read unspoken reasoning** — an aligned-looking output ≠ an aligned model
- **Bias in** → **bias out** → **amplified**: fairness is a data-and-process question, not a final tweak
- Humans aren't neutral: **overtrust, confirmation, anchoring** shape how we use AI
- **Validate outputs**: benchmarks, grounding (RAG) and a skeptical human — trust is earned, not assumed
- **More autonomy** → **new risks** → **least privilege, human oversight, observability**

Thank You!

Questions?