

Trust

Engaging with AI -Session 4

Session 4 Agenda

- Review Goals
 - Session 1 AI Outcomes Answers / Facts / Fiction
 - Session 2 AI Outcomes Generating New Content
 - Session 3 AI Processes Rational / Empirical Models

Additional Perspective – AI Types

- Last Session – AI Types by Process
 - Rational – Rules Based
 - Empirical – Probability Based
- Alternatively
 - Predictive AI – Forecasting
 - Generative AI – Creating



**Prediction is very
difficult, especially
if it's about the
future.**

Niels Bohr

Predictive AI

John Hopkins –
Sepsis Warning

Flagging →
Earlier
Intervention

Amazon
Recruiting

Bias →
system
scrapped



Expect AI Hype

Session 4 -- Trust

+

•

○



Training
Neural
Networks

Alignment

Guardrails

How to Train Your Dragon



Training and Alignment

Supervised Fine-Tuning

- AI model is pre-trained on data, then further trained on examples of good behavior. Human-written inputs and outputs show the model how it should respond.

Reinforced Learning from Human Feedback

- Humans rank several possible answers to the same prompt, best to worst. The model learns to prefer responses that humans reinforce.

Training and Alignment, cont'd

Reinforcement Learning from AI Feedback

- Another model is used to judge outputs

Constitutional AI or Rule based Alignment

- Model is guided by a written set of principles.
- Model critiques and improves responses over time

Guardrails and Filter

- Checks outside the model that block or rephrase harmful outputs

Error Minimization



LLMs are Prediction Models

- Simple Prediction Outcomes (yes/no)
 - False positives / Sensitivity
 - False negatives / Specificity

Medical Screening for Diseases

- You can't have it both ways
 - Lean toward Bayesian than Frequentist

Agency



Agency: The capacity of an AI system to act autonomously—making decisions, taking actions, and adapting within its environment to achieve specific goals.

- Autonomy
- Goal Oriented Behavior
- Interaction with environment

Examples Now

- Customer service chatbots
- Drones

Do You Trust.....



Agentic AI: Is distinguished by its agency: the ability to act independently and intelligently, coordinate actions among various agents, and pursue goals in complex, changing real-world settings.

Back to Fast Thinking / Slow Thinking

Question: What Level of Confidence would you consider necessary before you placed your Trust in AIs and Agentic AIs?

Does it vary by the Situation?

Does it vary by Alternatives?

Does it vary by Your Tolerance for error?

Is Trust Ever Absolute?

The Character of Physical Law

- Richard Feynman – Lecture 1964

Nature's Laws are Universal

Laws are Approximations

Test of knowledge is experiment

Bernoulli's Law

Who We Trust -- Professions

Most Trusted:

Nurses

Grade School Teachers

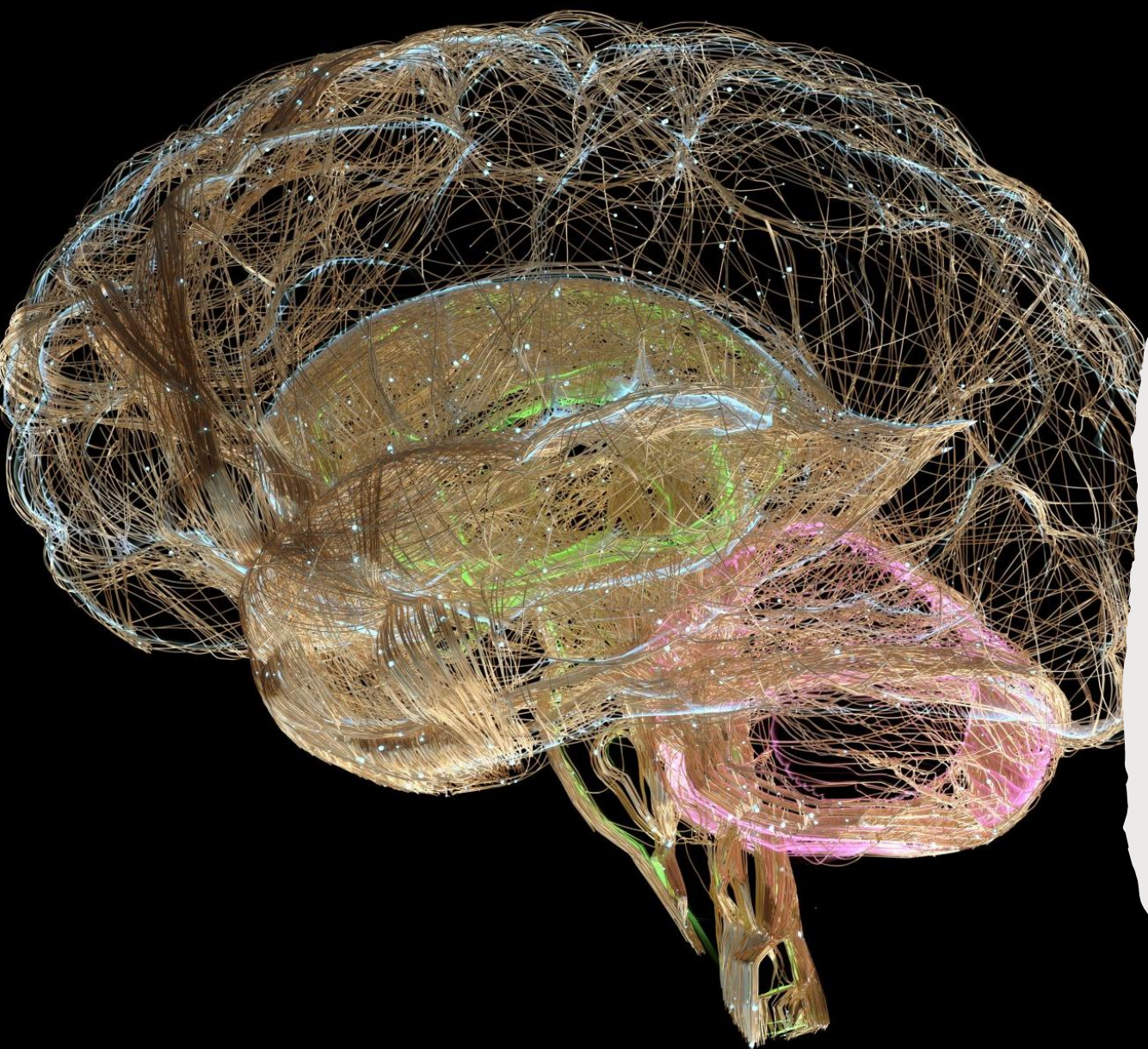
Military Officers

Least Trusted:

Lobbyists

Members of Congress

TV Reporters



Interlude – Carbon vs Silicon

- Before diving further into how much we should trust AIs
 - Examine distinctions between our brains and current AIs
- Look at Our Brains
 - Evolution
- Look at Current AIs
 - Development

The Devil is in the details,
but so is salvation.

Hyman Rickover