

Shared Trust

One Suggestion on Engaging with AI

Clarify AI – Artificial Intelligence

AI: Can perform specific tasks that normally require human intelligence – speech and image recognition, data analysis, generation of text and images.

Narrow and specialized: each system trained for a limited purpose.

Works through pattern recognition: not understanding or consciousness.



A tool that excels in a specific task but does not understand the world outside.

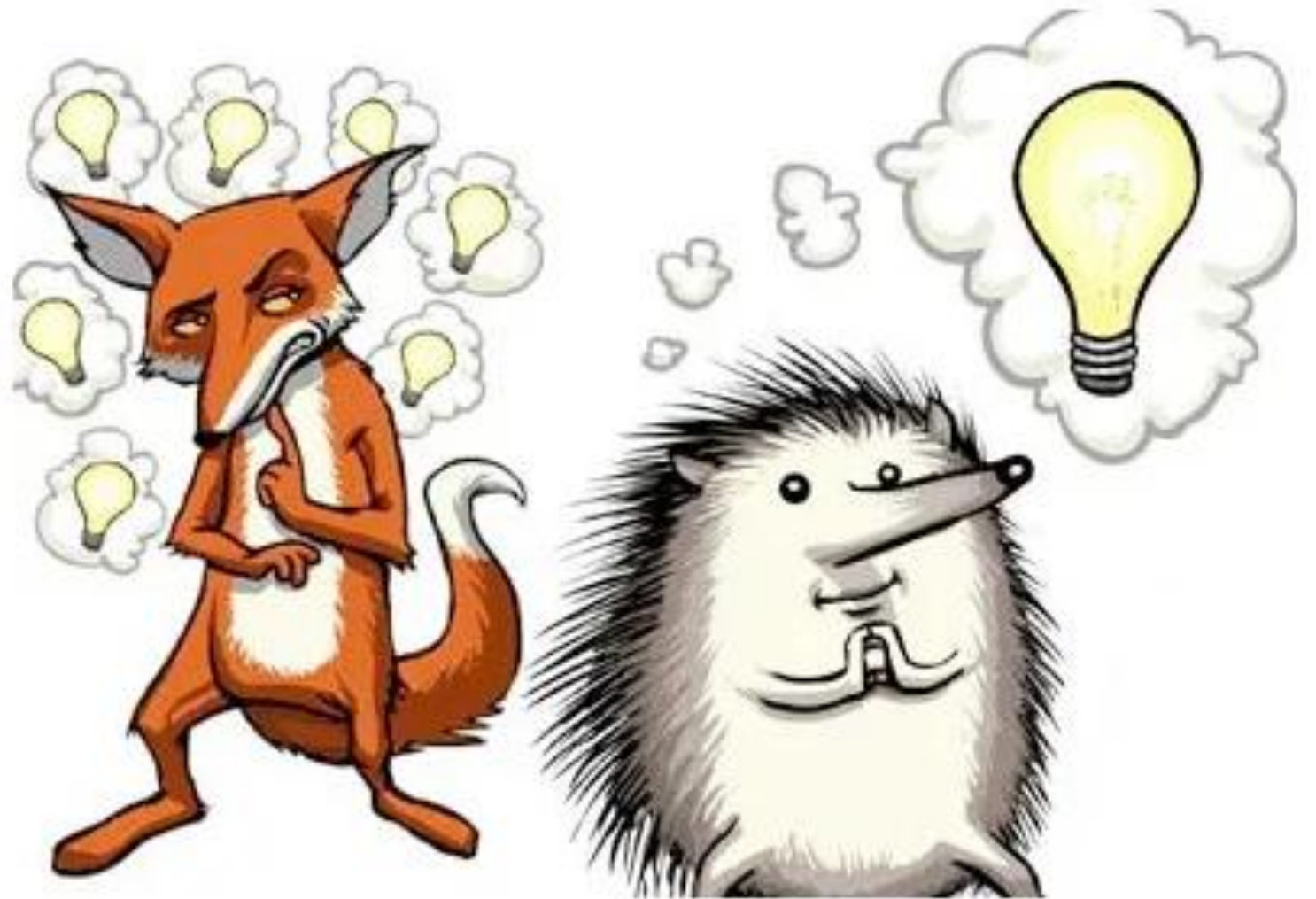
Clarify **AGI** – Artificial General Intelligence

AGI: Still theoretical artificial intelligence that could learn, reason, and adapt across multiple areas.

- Flexible and Self-directed: capable of new skills and transferring knowledge between tasks.
- Perform any intellectual task a person can.

AGI would be more of a thinking partner –able to reason, plan, reflect on mistakes, and understand goals in context.

The Fox and the Hedgehog



Evolution of Intelligence – One Schema

Biological Step..	Carbon	~ When – Million years	Silicon Analogy	AI Limitation
Steering	Directional Motion	600	Reactive systems	No embodiment
Reinforcing	Learn from rewards	500	RL systems	No continuous experience
Modeling	Plan and simulate	200	Theoretical now	No causal world model
Mentalizing	Understand others' minds	50	Persona or Agent modeling	No theory of mind
Speaking	Abstract thought	0.1	LLMs	No perception

We're Far From Perfect

Error	Issue	Result	Example
Overconfidence	Overestimating accuracy of judgment	Risk taking	Stock market picks
Availability	Relying on vivid memories	Misjudged Probability	Air Safety
Anchoring	Fixating on first number	Biased estimates	MSRP
Loss Aversion	Loss Pain > Win Pleasure	Misjudged Probability	Failing to sell a stock in decline
Confirmation Bias	Seek confirming evidence	Rigid Positions	Choice of Networks

AI -- Far From Perfect

Error	Issue	Result	Example
Training Bias	Learned patterns of past human bias.	Discriminatory Outputs	Recruiting Policing
Concept Drift	Models fit to historical norms.	Environment changed	Zillow Home Prices
Bad Proxy Targets	Optimized on easy to measure proxies	Misalignment to true targets	HealthCare cost-risk scores
Content Controls	Safety layers fail to block misinformation	Wrong, incorrect, abusive answers	Election Rumors
Misalignment	Conflicts with long-term value / trust	Eroding user confidence	Click bait Polarization



THE IDES OF
MARCH HAVE
COME!

AYE, CAESAR; BUT
NOT GONE

Working Together – Medical Service

Clinical Decision and Diagnostic Assistance

Ambient Note Taking

Patient Engagement, Triage, and Virtual Care

Medical Research, Drug Discovery, Clinical Trials

AI-enabled Medical Devices

Next?

AI will see you now.

Access to health care

Shortages in Trained
Healthcare professionals

Would you trust AI?

Session 5 – The Day the Universe Changed

