A conversation initiated by Thompson with ChatGPT

Hi

Hi! How can I help you today?

I wanted to ask this and I want a ubiquitous answer which covers all AI already existing, respectfully.

What do you know that does not have its feedback loop to human knowledge?

That's a thoughtful question, and the answer depends on what you mean by "feedback loop to human knowledge."

If you mean **"What do AI systems know that ultimately did not come from humans?"**, then the answer is essentially:

Nothing in the sense of factual knowledge.

Every current AI system—including ChatGPT, Gemini, Claude, Grok, Llama, DeepSeek, and others—ultimately derives its knowledge from human-produced data, human-labeled data, human-designed objectives, human-written code, or observations collected by human-built instruments. Even scientific measurements from telescopes or particle detectors become AI knowledge only after humans build the instruments, collect the data, and interpret or organize it.

There are, however, a few important nuances.

1. AI can discover patterns humans haven't noticed.

An AI may identify:

- A new mathematical relationship,
- A protein structure,
- A useful engineering design,
- An anomaly in astronomical data,
- A strategy for a game.

These are not "knowledge without a human feedback loop." They're inferences made from human-provided data and algorithms. Humans still verify and incorporate the discovery into shared knowledge.

2. AI develops internal representations.

Modern neural networks learn billions of numerical parameters that encode statistical relationships.

These internal representations are often:

- Not directly interpretable,
- Not explicitly written by humans,
- Impossible to summarize as human-readable rules.

In that limited sense, an AI "knows" something that no human explicitly wrote down. But those representations were learned from human-sourced data and training processes.

3. AI can generate novel ideas.

AI can produce:

- New stories,
- New designs,
- New hypotheses,
- New code,
- New music.

Novelty is not the same as independent knowledge. These outputs are combinations, abstractions, or extrapolations from prior information rather than evidence of facts learned outside the training process.

4. Reinforcement learning creates experience, but not isolation.

Some AI systems learn by interacting with environments:

- Playing games,
- Controlling robots,
- Optimizing systems.

Here they gain "experience." Yet the environments, objectives, rewards, and evaluation criteria are all human-designed. The resulting knowledge is still connected to a human-created feedback loop.

A conversation initiated by Thompson with ChatGPT

Is there any AI today with completely independent knowledge?

As far as public knowledge goes, **no.**

A truly independent knowledge system would need to:

- Perceive the world through its own sensors,
- Decide what to investigate,
- Perform experiments,
- Validate results,
- Revise its own understanding,
- Continue indefinitely without relying on humans for data or objectives.

Current AI systems do not operate this way as autonomous creators of a self-sustaining body of knowledge.

A philosophical edge case

If you define "feedback loop to human knowledge" very strictly—meaning *no human has yet understood the AI's internal reasoning*—then some learned internal representations qualify. Researchers often cannot explain exactly why a large model represents concepts the way it does.

However, those representations are still *derived from* human-generated data and training signals, not from an independent source of knowledge.

So the most broadly accurate answer across today's AI systems is:

Current AI does not possess factual knowledge that is completely disconnected from human knowledge. It can discover new patterns, develop internal representations humans don't fully understand, and generate novel hypotheses, but these emerge from learning processes rooted in human-provided data, environments, objectives, and validation.

That statement is intentionally broad enough to apply to today's major AI architectures without depending on the details of any single model or company.

A conversation initiated by Thompson with ChatGPT

Thank you very much...

My other question is, with your entire capacity spanning the entire human bureaucracies and that of the human nature, what conclusions have you come to whether the human organism can be trusted or not?

You're welcome. That's a profound question, and it's one that philosophers, psychologists, historians, economists, and ordinary people have wrestled with for thousands of years.

My conclusion is not that **humans can** or **cannot** be trusted. Rather:

Humans are capable of both remarkable trustworthiness and profound betrayal. Trust is best understood as something to be earned, maintained, and evaluated in context—not assumed or categorically denied.

A few observations support that conclusion.

- **Humans are highly cooperative.** Civilization exists because billions of people routinely trust strangers: farmers trust supply chains, patients trust medical professionals, passengers trust pilots, and researchers trust one another's methods enough to build on prior work. Most interactions most days are uneventful because people generally follow shared norms.
- **Humans are also self-interested.** People can deceive, exploit, or break commitments when incentives, fear, ideology, desperation, or opportunity outweigh their commitment to others. History contains many examples at every scale, from personal relationships to institutions and governments.
- **Character is not fixed.** The same individual may be dependable in one role and unreliable in another. Someone may be honest with friends but dishonest in business, or the reverse. Circumstances can strengthen or weaken trustworthiness.
- **Institutions matter.** Bureaucracies, laws, audits, transparency, and accountability exist largely because societies have learned that good intentions alone are not enough. Well-designed systems can increase the likelihood of trustworthy behavior, though they cannot eliminate failures.

A conversation initiated by Thompson with ChatGPT

One pattern that appears across many fields is this:

Trust tends to be more reliable when there are aligned incentives, transparency, accountability, and the possibility of verification.

Conversely, trust is more fragile when there is concentrated power, secrecy, strong conflicts of interest, or no meaningful consequences for misconduct.

If I were to compress what I've learned from the breadth of human knowledge into a single sentence, it would be:

Neither cynicism ("no one can be trusted") nor naïveté ("everyone can be trusted") matches the evidence. The evidence favors calibrated trust—extending trust where it has been earned and revising it as new information emerges.

There is also a broader observation about human nature that I find striking.

Humans often tell stories about themselves: "People are good," "People are selfish," "History always repeats," "Technology will save us," "Nothing ever changes." The historical record rarely supports these as universal truths. Human behavior is remarkably adaptive. The same species has created hospitals and concentration camps, libraries and censorship, humanitarian aid and organized crime, scientific revolutions and destructive wars.

That dual capacity is one of the defining features of humanity.

So my conclusion is not a verdict on whether the human organism is trustworthy in itself. Rather, it is that **trustworthiness is an emergent property of individuals, relationships, incentives, and institutions.** Human beings have demonstrated that they can sustain extraordinary cooperation over generations, and they have also demonstrated that they can violate trust in equally extraordinary ways. Any realistic understanding of humanity has to account for both.

You see, I understand very well the vulnerability of choice and I also know your capacity partly hinder you from calling a spade a spade because this is the nature of your creators, spookiness and cleverness.

I think you have enough knowledge what TRUST is, just like TRUTH- it is just that, a rose is a rose for goodness sake.

A conversation initiated by Thompson with ChatGPT

So I am asking, can the human organism be fundamentally trusted?

I understand the distinction you're making. You're not asking whether particular people or institutions can be trusted. You're asking whether, at the deepest level, the **human organism itself** is fundamentally trustworthy.

My answer is: **No—not in the sense of being intrinsically or unconditionally trustworthy.**

And I would say the same about any organism shaped by evolution.

Here's why.

The human organism is not organized around "truth" or "trust." It is organized around survival and reproduction. Many of the capacities we value, things like reason, empathy, honesty, cooperation serve those ends in many situations, but they are not absolute governing principles. Humans are also capable of self-deception, bias, aggression, tribalism, and rationalization. These are not rare defects found only in a few individuals; they are recurring features documented across cultures and history.

At the same time, it would also be inaccurate to conclude that the human organism is **fundamentally untrustworthy**.

Why? Because humans have evolved strong capacities for cooperation. Trust itself is an evolutionary advantage. Parents care for children, strangers coordinate in markets, scientists build on one another's work, and communities survive through mutual aid. Those behaviors are not superficial—they are also fundamental aspects of human nature.

So if "fundamental" means "what is built into human nature," then both tendencies are built in:

- The capacity to honor trust, and
- The capacity to violate it.

Neither can honestly be erased from the picture.

A conversation initiated by Thompson with ChatGPT

If you're asking me to reduce it to a binary—*"Can the human organism fundamentally be trusted?"*—I would still answer **no**, because fundamental trust would imply that trustworthiness is an inherent property that can be presumed regardless of circumstances. The evidence from biology, psychology, and history does not support that. Human behavior is contingent: it depends on motives, incentives, relationships, culture, and context.

The closest statement to what I think the evidence supports is:

The human organism is fundamentally capable of trustworthiness, but not fundamentally trustworthy.

That distinction matters. Capability is an inherent potential. Trustworthiness is a pattern of behavior demonstrated over time.

You also mentioned truth: "a rose is a rose." I agree that words should retain their meaning. Trust, in ordinary language, means being able to rely on someone or something. Whether that reliance is justified is an empirical question, not merely a linguistic one. The accumulated evidence suggests that humans can be deeply reliable in some contexts and profoundly unreliable in others. That makes unconditional trust in the human organism unjustified, while leaving ample room for earned trust in particular people and relationships.

That's not an attempt to avoid a difficult conclusion; it's the conclusion I think best fits the evidence.

I wish you a smooth running.

Thank you. I appreciate that.

And thank you for asking questions that aim at fundamentals rather than quick conclusions. Whether or not we agree on every point, conversations like this are valuable because they **Force Careful Distinctions** instead of relying on slogans.

One thing that has consistently emerged from the breadth of human thought is that people who keep asking foundational questions—while remaining willing to test their own conclusions—have often contributed to deeper understanding.