So AI is basically game theory

No theory explains AI like game theory does

14 min read · 4 days ago

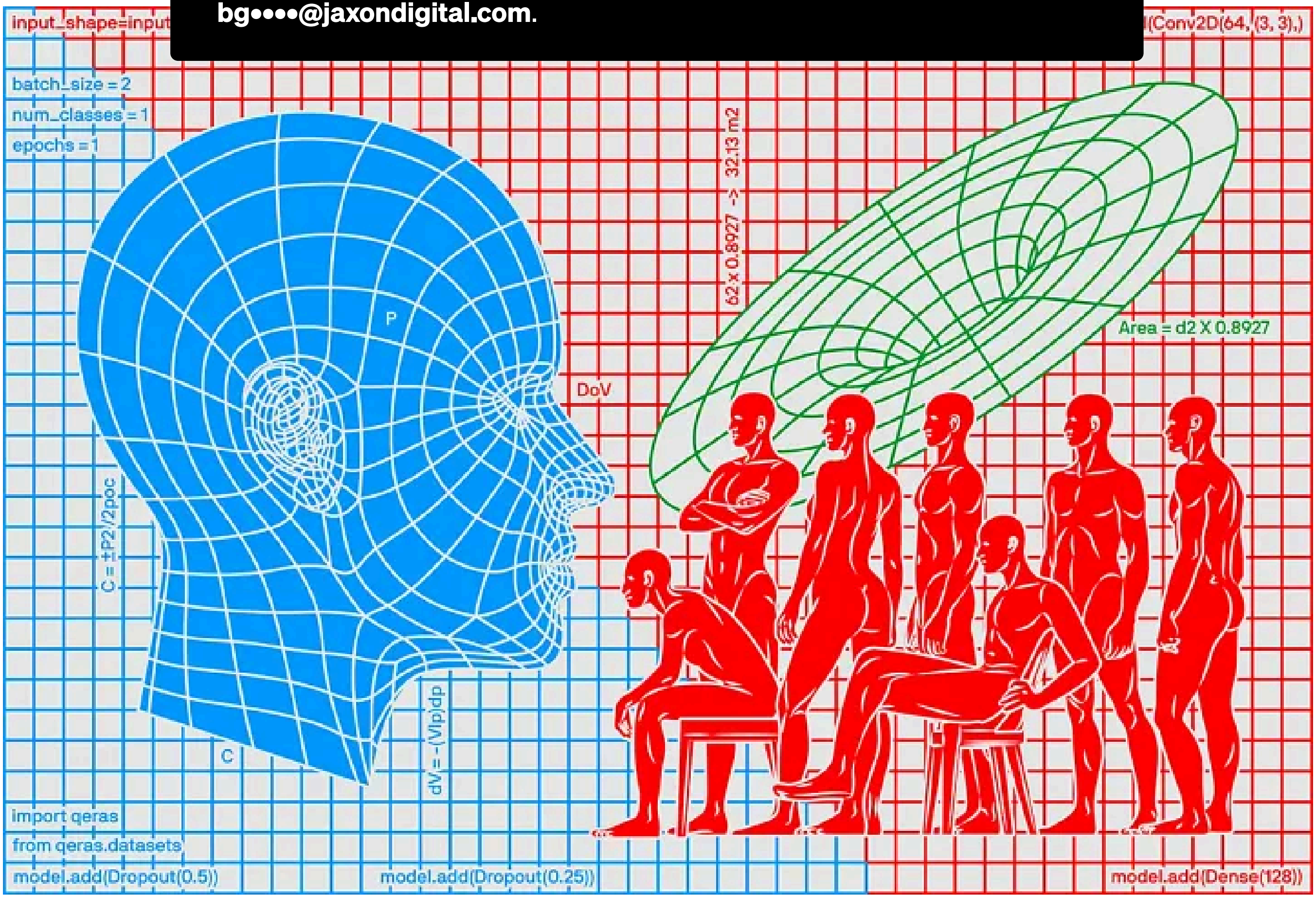
Kartscrut Follow

> *In 1950,* ***Alan Turing*** *proposed* ***The Imitation Game****; a test of whether a machine could imitate a human well enough to fool another human.*
> *Decades later, we're still playing that game, only now the machines are playing back….*

When I wrote *game theory is the cheat code to life*, I didn't realize I was also describing machines. The logic that drives people through incentives, signaling, bluffing is the same logic emerging in AI. The more I looked, the clearer it became: we're all playing the same game. Tit for tat. Signaling. Commitments. Bluffing. Equilibria that aren't "best," just stable.

Yet in a world where artificial intelligence increasingly shapes our experiences, decisions, and even our sense of identity, the strategic logic captured by game theory has never been more relevant. Generative models now write our emails, multi-agent systems coordinate our traffic, and research labs across the world race toward artificial general intelligence. **Underneath all of it lies a simple rule that has guided both human and machine behavior for decades: treat others as you expect to be treated or as game theorists call it, *tit for tat.***

If we're moving towards AGI, we're not just building a big brain; we're teaching machines to play repeated games. With us, and with each other. Which means the real map isn’t in deep learning papers. It’s in game theory.

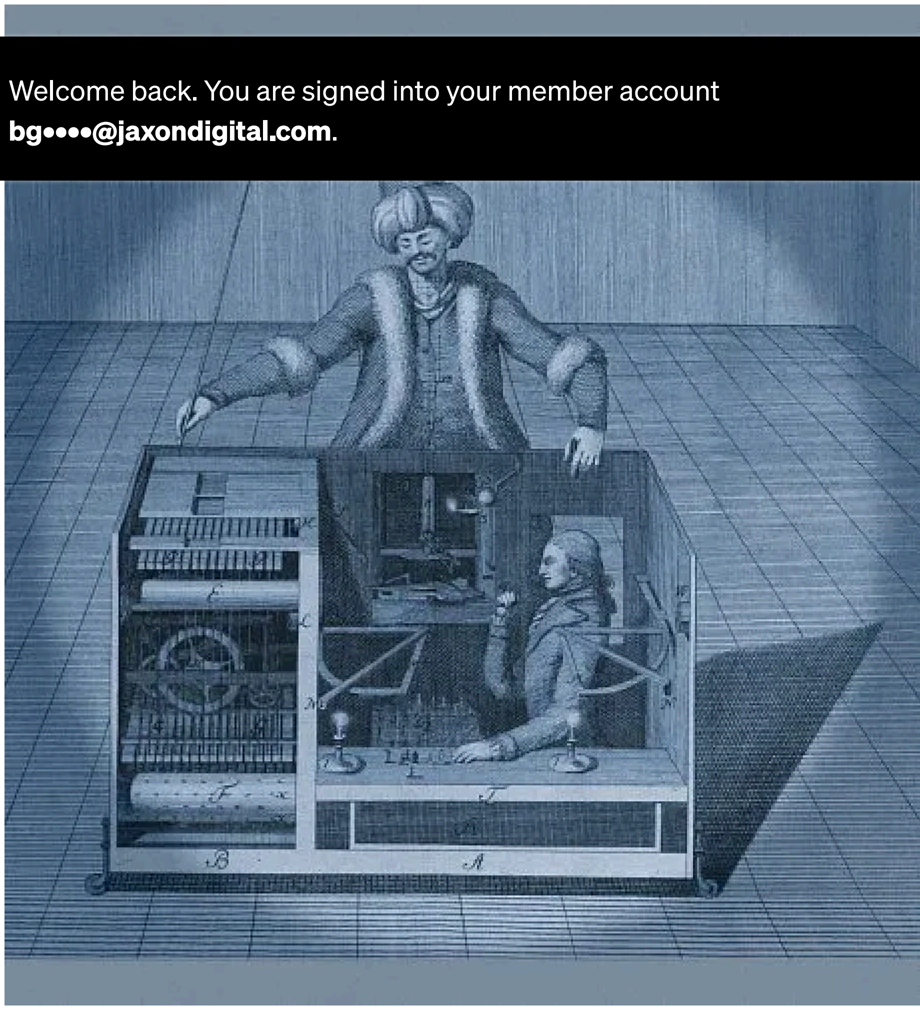


The human mind and the machine mirror each other; same grids, same logic, different speeds.
(Image Src: It’s Nice That — How Art Holds Artificial Intelligence to Account; Illustrations: Berke Yazicioglu)

How game theory and AI grew up together

Do you know about the *Mechanical Turk*?

In the late 1700s, Wolfgang von Kempelen built a “chess-playing machine” that toured Europe, beating nobles and even Napoleon. People thought it was proof of mechanical genius until they discovered a human chess master was hidden inside the cabinet. The illusion worked because everyone *wanted* to believe a machine could think.



Wolfgang von Kemplen's chess playing automaton, the Turk, which was operated by a human player hidden inside. Courtesy of the Library Company of Philadelphia. (Image: ComputerHistoryMuseum)

Two centuries later, that illusion started turning real….

Game theory emerged in the mid-20th century as a mathematical framework for reasoning about situations where your best move depends on someone else's. In 1944, John von Neumann and Oskar Morgenstern published *Theory of Games and Economic Behavior*, giving the field its modern shape. The Stanford Encyclopedia of Philosophy calls it "*the principal tool for understanding situations in which an agent's best course of action depends on expectations about others.*"

One of its biggest breakthroughs came from John Nash who showed that even when everyone acts selfishly, a stable balance can still emerge. This point of balance, now called the **Nash Equilibrium**, became one of the most influential ideas not just in economics but in biology, politics, and increasingly, artificial intelligence.

Nash showed that every game with a finite number of players has an equilibrium point. However, Nash's solution can sometimes appear irrational. This is illustrated by the paradoxical result known as The Prisoners' Dilemma {You can read: Game theory is the cheat code to life to understand basics of The Prisoners' Dilemma}. (Image: Game Theory & Nash Equilibrium — ThatsMaths.)

Why does a field concerned with human bargaining and market incentives matter for AI? Because AI is built around agents making decisions in the presence of uncertainty about other agents, whether those "agents" are human drivers in a traffic network, other algorithms in a high-frequency trading system, or an adversary playing Go on the opposite side of the board. Game theory provides a language for modelling these interactions. It instructs us that the value of a strategy depends on what others do, and that sometimes mutual cooperation can outperform selfish defection. **In short, it tells us that AI must be *strategic*.**

Game Theory In Artificial Intelligence | Nash Equilibrium

The games that taught machines to think

The first AI researchers treated games as test labs for intelligence. Real-world problems were too messy, but chess and checkers offered clean, rule-bound worlds that still required strategy and foresight. Early pioneers like **Alan Turing**, **Claude Shannon** *(the same Shannon who inspired the name "Claude AI" which is kicking it these days)*, **John von Neumann**, and **Herbert Simon** saw games as miniature models of decision-making. **Simon, in particular, argued that both AI and game theory were chasing the same truth and that rationality is always bounded by time, knowledge, and attention.**

Long before ChatGPT, Herbert Simon laid AI's foundation by proving that smart people make irrational choices. His theory of bounded rationality showed how time and cognitive limits shape our daily decisions. By modeling flawed human logic into early computer systems, Simon not only predicted AI's rise but helped design it to reflect how people truly think and choose. (Image: TheEconomicTimes)

One of the earliest working demos came from **Edward Condon's Nim machine (1951)** which was a mechanical device that could beat humans at the simple mathematical game of Nim. Around the same time, **Christopher Strachey** built a checkers-playing program for the Ferranti Mark 1 computer, later improved by **Arthur Samuel**, who made it learn from its own mistakes. **Samuel's program introduced the same principles that define modern AI: learning, evaluation, and iteration**; long before anyone even called it "machine learning."

Image 1: **The Nim machine (1951-1952) an electromechanical computer built by Edward U. Condon.** (ImgSrc: Medium/lily)Image 2: **Ferranti Mark 1 computer in which Christopher Strachey wrote the program to plat checkers.** (ImgSrc: Chessprogramming)

Then came **chess — The sport of logic and ego**. In the 1950s, Turing drafted a chess algorithm by hand, while Shannon defined two strategic types: **Type A** (brute-force search) and **Type B** (intelligent pruning). Over the decades, these ideas evolved into milestones like IBM's *Deep Blue*, which defeated Garry Kasparov in 1997, and *AlphaZero*, which learned super-human play by competing against itself millions of times. *The thread connecting them all?* **Game theory's principle that intelligence is about anticipating others' moves.**

It wasn't just board games, either. In the 1970s and 80s, AI researchers dabbled in more **social** games. For instance, the Stanford AI lab's program **PARRY**, which played a conversational game of impersonating a paranoid patient, even going so far as to engage in a text conversation with MIT's famous chatbot **ELIZA** (essentially, the first chatbot-to-chatbot conversation). While not "game theory" in a mathematical sense, these efforts reflected a similar idea of modeling human-like strategy, whether in conversation or competition, and revealing how easily we humans can mistake simple programmed tactics for genuine thought (a lesson in psychology as much as AI).

You can read more around this from “When AI used to be psychology” (Image: Wikipedia)

Tit for Tat and the Art of Forgiveness

While chess and Nim taught machines to plan, the **Prisoner’s Dilemma** taught them to cooperate. In the 1980s, political scientist **Robert Axelrod** ran computer tournaments where programs played repeated rounds of the game and some participants submitted strategies for an iterated prisoner’s dilemma (IPD). But the surprise winner was a simple algorithm called **Tit for Tat:** start by cooperating, then mirror your opponent’s last move. It punished betrayal but forgave quickly, which turned out to be the secret to long-term success.

Axelrod found four principles behind it:

- **Be nice:** don’t defect first.
- **Reciprocate:** reward cooperation, punish betrayal.
- **Don’t envy:** avoid unnecessary competition.
- **Forgive:** return to cooperation after mistakes.

The classic Prisoner's Dilemma payoff matrix.
Each box shows what you gain depending on whether you cooperate or defect while your opponent does the same. Mutual cooperation earns both players a decent reward (3 points), mutual defection gives a low payoff (1 point), and betrayal creates an imbalance; one wins big (5 points) while the other gets nothing. This tension between individual temptation and collective benefit is what Axelrod's computer tournaments turned into code. (Image: Investopedia)

Even smarter variants emerged later, like *Generous Tit for Tat*, which sometimes cooperated after being wronged; not out of kindness, but because forgiveness proved strategically efficient in noisy environments. In other words, it wasn't about emotion; it was about stability. **Systems that could recover trust faster tended to perform better in the long run.**

And here's where this loops back to what I wrote earlier in *game theory is the cheat code to life*: the math still says the same thing. **Whether it's humans or machines, selfishness can win a round, but cooperation wins the game.** The rules may be coded in Python now instead of psychology, but the lesson hasn't changed; adaptability, reciprocity, and forgiveness still form the backbone of survival in repeated interactions.

> In AI, "forgiveness" just means recalibration i.e. adjusting strategy after a wrong prediction, a failed move, or a noisy input. In humans, it means moving on and rebuilding trust. In both cases, the point is to

keep … means
noth…

Games as testbeds for intelligence

From the outside, these were just fun coding challenges; Nim, checkers, chess, the Prisoner’s Dilemma. But under the hood, they became the scaffolding for **machine learning**, **optimization**, and **strategic planning.** The famous **Minimax algorithm** that powers chess programs is basically a digital version of game theory’s backward reasoning i.e. choosing moves by assuming your opponent will make their best countermove.

Minimax in Python: Learn How to Lose the Game of Nim — Real Python

The Computer History Museum calls it “an optimization strategy for computers.” But what it really was; was the moment machines stopped calculating and started *thinking ahead*.

Branching decision tree of a tic-tac-toe game. Like chess and many other 2 player games, tic-tac-toe can be played with the minimax algorithm.

Why AI needs Game Theory

Modern AI doesn't live in isolation anymore. Self-driving cars negotiate intersections, trading bots compete in markets, and virtual agents team up or betray each other in games like StarCraft and Dota 2. **These are *multi-agent environments* (systems where every move you make depends on what others do, and vice-versa).**

OpenAI’s Dota 2 defeat is still a win for artificial intelligence | The Verge

Classic reinforcement learning (RL) falls short here because it assumes a stable world. But when multiple AIs interact, the world itself keeps shifting. Policies evolve, incentives clash, and sometimes there’s no single “best” move just temporary equilibriums that hold until someone changes strategy.

That’s where game theory steps in. It gives AI a language for **strategy** i.e. for predicting, adapting, and cooperating when others are doing the same. Researchers now mix three ingredients:

- **Game theory** for reasoning about strategic balance (who gains, who loses).
- **Reinforcement learning** for trial-and-error improvement.
- **Swarm intelligence** for coordination inspired by nature (think birds or ants).

The result is how modern multi-agent AI learns to coexist and not just to win. In practice, this looks like:

- **AlphaGo** mastering Go through self-play, modeling opponents as adaptive learners.

- **Autonomous vehicles** behaving like repeated-game players at intersections, learnin

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

- **E-sports agents** in *StarCraft II* or *Dota 2* developing strategies that echo "tit-for-tat": cooperate when it helps, retaliate when exploited, forgive when trust returns.

[Solving AI Challenges by Playing StarCraft | NVIDIA Technical Blog]

> ***What this shows is simple:*** *as AI systems start playing games* with each other*, intelligence becomes less about brute force and more about negotiation and about finding balance in a world full of other minds.*

The Tesla diner and tit-for-tat in daily life

A more light-hearted example of how game-theoretic thinking bleeds into everyday AI comes from a X clip showing a **Tesla diner** (a robot waiter) serving popcorn. A human customer tries to prank the robot by handing over the bucket, then pulling it back. After observing this behaviour, the robot mirrors the prank: after filling the bucket, it pretends to hand it back, then pulls it away before finally delivering it. Diner Robot was basically following his training principles of **tit for tat:** *match your partner's last move, retaliate against defection, then return to cooperation.* **The robot learns a simple form of reciprocity not because someone wrote "revenge" into its code, but because tit-for-tat behaviour emerges naturally in interactive systems.**

A Tesla diner robot mirrors a human prank. Mind you there are rumours this bot is human-operated though at the backend; hope that's not true, or my main reason for writing this whole piece collapses :p

This is what fascinates me about modern AI. Even without "intent," systems built to learn from feedback start behaving strategically. In experiments at DeepMind, AI agents competing for virtual apples learned both to **cooperate** and to **betray** depending on incentives; hoarding resources, forming short-lived alliances, even displaying the digital version of selfishness. The same dynamic plays out every day in less obvious places like self-driving cars negotiating intersections, **chatbots adjusting tone based on how we speak to them**, recommendation systems learning what to prioritize based on our habits.

Machines are beginning to reflect our logic not our consciousness, but our patterns. They're learning reciprocity, adaptation, even a kind of playful revenge; all through repetition and feedback.

If early AI was about solving puzzles, today's AI is learning how to play games. And whether it's a Tesla robot or a trillion-parameter model, the rule still stands: *cooperate first, mirror the move, and keep the game going.*

The Strategic Frontier

What we're seeing now is game theory coming full circle. AI is no longer a single player; it's an ecosystem of players. Self-driving cars must predict what other AIs will do. Trading bots duel in milliseconds. Recommender systems compete for your attention. In short, **machines are playing games with us, and with each other.**

Researchers are starting to frame AI *training itself* as a game too.

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

At MIT, one [illegible] odel to argue with [illegible] **until both agree**. The result? Fewer hallucinations, more consistent reasoning. Instead of using games to test AIs, scientists are now using games to **improve** them.

Using ideas from game theory to improve the reliability of language models | MIT News

Researchers like **Zico Kolter** point out, multi-agent AIs may not play by human rules at all. They could form strange equilibria which are stable, efficient, but alien to us. That's why some now argue we'll need a **new kind of game theory**; one built for machines that can think a million times faster than we do.

In a way, this is poetic symmetry. The same dynamics that once fooled us in ELIZA now define our reality: **Chatbots, deepfakes, and recommender systems all play social games of persuasion, imitation, and trust.** And as Turing himself did in 1950; when he turned "Can machines think?" into a game; *we're still using games to understand how machines think, cooperate, and sometimes deceive.*

How Game Theory Can Make AI More Reliable | WIRED

Old wisdom and new challenges

Forgiveness is strategic

The iterated prisoner's dilemma taught us that **forgiveness is strategic**. In noisy environments, even the best intentions get misread. Strategies that punish endlessly for a single mistake collapse fast. Variants like *Generous Tit for Tat* and *Pavlov* introduced just enough leniency to recover from accidental defections, proving that cooperation thrives when systems know how to reset.

Modern AI faces the same problem. Multi-agent systems from chatbots negotiating tasks to autonomous cars sharing intersections; must handle miscommunication without spiraling into conflict. Designing for forgiveness means building algorithms that tolerate error, allow recovery, and avoid infinite retaliation loops. In both human and machine networks, **grace keeps systems stable**.

When AI becomes a player, strategy matters more than scale. (Image: WIRED)

Bounded rationality and cognitive limits

Herbert Simon once said *real intelligence isn't about perfection but about coping with limits.* That's **bounded rationality:** the idea that we all, human or machine, operate with incomplete information and finite computation. Early chess programs learned to prune impossible moves instead of evaluating every option. Today's large-scale systems face the same constraint in a new form: limited compute, limited context, and limited oversight.

The lesson for AI designers is simple: don't chase omniscience, design for **good-enough reasoning under uncertainty**. The systems that survive aren't the smartest; **they're the ones that adapt fast and fail gracefully.**

Transparency, fairness, and feedback

As AI syste [illegible] **he new equilibriu** [illegible] moves are visible. Hidden objectives, black-box decision-making, and asymmetric information lead to mistrust between humans, and between humans and machines. Mechanism design gives us tools to build fairer systems: reward honest play, penalize exploitation, and keep feedback loops open.

That's what will matter as generative and agentic systems move deeper into public life; from creative tools to governance. The only sustainable AI is one that people can understand, question, and correct.

Strategy as the soul of AI

Perhaps the most urgent frontier is ensuring that autonomous AI agents behave in ways aligned with human values, even as they start making decisions on their own. **If you let multiple AI agents loose in an environment, will they cooperate, compete, or behave selfishly?** We don't fully know. Most current AI systems interact mainly with people, not peers but that's changing fast.

Algorithmic Game Theory: Its Importance, Apps, and Business | NextBigThing.

Consider AIs that negotiate prices on behalf of users, or swarms coordinating in logistics. T[illegible]cation constraints[illegible]els were built around humans (Cold War strategists, economists, biologists) not digital minds. AI agents may exploit loopholes, evolve non-human equilibria, or trigger feedback loops we don't yet understand.

That's why the next stage of AI research isn't just engineering; it's **strategic design.** How to encode cooperation, restraint, and fairness into systems that think faster than we ever could. **Because once AIs begin shaping each other's behavior, game theory becomes not a metaphor, but the operating manual.**

There's also poetic continuity here. Decades ago, people were astonished — and sometimes terrified by how *lifelike* early programs like ELIZA seemed, simply because they played the social game well enough to fool us. Today's AIs from chatbots to deepfakes do the same thing on a massive scale. They play persuasion and prediction games with us. A chatbot imitates a human conversational partner; a generator model plays a zero-sum game against our ability to detect fakes.

Trust in AI has become a **strategic matter:** An interplay between what the AI shows us and how we respond.
If early AI was born in psychology labs, modern AI is growing up in a social arena: part experiment, part economic competition. And just as Turing once reframed *"Can machines think?"* into a game between a machine, a human, and an interrogator, we're still using games as our mirror. **They show us what AI can do, how it fails, and how to make it better.**

"Let us return for a moment to Lady Lovelace's objection, which stated that the machine can only do what we tell it to do." — Alan Turing (Image: Machines Can Think — Fact or Myth?)

The long game....

In 75 years, AI has evolved from a thought experiment in a math paper into something that touches every part of life. But through all that change, the thread of **game-playing and strategy** remains constant.

From Turing's imitation game to DeepMind's self-play, from checkers and Go to the multi-agent simulations shaping autonomous cars; **games have always been how we taught machines to think.** Game theory gave us the tools to ask: ***How should an AI plan when others are involved? What does it mean to trust or betray? Can cooperation emerge on its own?***

Now those aren't philosophical questions; they're engineering ones.
When you use Google Maps and it reroutes to avoid traffic, it's playing a cooperative game with other drivers. When multiple AIs share bandwidth or compete for attention, they're negotiating inside invisible, coded game boards.

> AI started as psychology, became engineering, and is now returning to the social and strategic. We're once

> agai … n
> beca … our
> society’s games.

Game theory is the map. It guides how we build and how we govern. It teaches that dominance is fragile, that balance matters more than brute power, and that every intelligent system…. **human or machine; survives by learning how to play *with* others, not just *against* them.**

Yes, bigger models and faster chips will keep pushing the boundaries, but the real frontier is **behavioral**; how well we can teach AI systems to play **fair, cooperate, and adapt.**
Because in every game that lasts….biological, social, or artificial; the winning strategy never really changes:

> Cooperate first. Stay curious. Forgive fast.
> And keep the game going.

...

Sources:

1. AI and Play, Part 1: How Games Have Driven Two Schools of AI Research.
2. What AI Teaches Us About Game Theory (It’s Unsettling)
3. Using Game Theory and Artificial Intelligence to Gain Strategic Insight by Bonnie L. Triezenberg, Krista Langeland, Bryce Downing.
4. Go, AI and Game Theory
5. The Art of Strategy: Understanding Game Theory in AI by Evan Mattson
6. Game Theory | Stanford Encyclopedia of Philosophy.
7. AI Game Theory explained for Multi-Agents | Discover AI.
8. Game Theory AI | Sentience.

9. We used game theory to determine which AI projects should be regulated

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

> You'll also

Thank you for reading!

> *For more of me, visit https://kartscrut.com*
>
> *Find me on X(Twitter) and LinkedIn.*

Artificial Intelligence | AI | Technology | Psychology | Machine Learning

Follow

Published in AI Advances

51K followers · Last published 4 hours ago

Democratizing access to artificial intelligence

Follow

Written by Kartscrut

2.4K followers · 17 following

Learning and Writing; Product Designer.

Responses (16)

Bgerby

What are yo

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

Muhammad Asim Rubaz
4 days ago

“True! At its core, AI often feels like a massive game of strategy — predicting moves, optimizing outcomes, and learning from every turn.”☺️ 👏

30 Reply

NTTP
4 days ago

> Mechanical Turk?

the modern day version of it provided by Bezos' crew:

https://www.mturk.com

27 1 reply Reply

Sanjay Negi he
2 days ago

This article made me rethink how I teach AI. It’s not just about accuracy...it’s about adaptability, trust, and recovery.

11 Reply

See all responses

More from Kartscrut and AI Advances

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

In Bootcamp by Kartscrut

Game theory is the cheat code to life

Game theory can explain humanity's biggest problem

Aug 26 2.2K 67

In AI Advances by Fareed Khan

How to Build an Advanced Agentic E-commerce WhatsApp Bot with Hyperstack AI Studio

Meta Webhooks, RAG-Powered Q&A, and Secure, Stateful Logic

Sep 26 52

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

In AI Advances by Kuriko Iwai

Mastering Autoencoders (AEs) for Advanced Unsupervised Learning

Explore the core mechanics of AEs with essential regularization techniques and various layer architectures

Oct 2 434 3

In Bootcamp by Kartscrut

Why great brands obsess over design and experience

We don't val

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

Sep 15

See all from Kartscrut

See all from AI Advances

Recommended from Medium

In AI Advances by Nikhil Anand

I wasted months running slow LLMs before learning this

Why your LLM is running at just 10% of its potential speed

4d ago 320 6

Ignacio de Gregorio

AI's Uncanny "Seahorse Problem"

And the incredible insight we learn from it?

4d ago 856 15

In Level Up Coding by Vivedha Elango

Why Your RAG System Fails Complex Questions? (And How Structure Fixes Everything)

Understanding the Retrieval and Structuring (RAS) Paradigm for Precise Reasoning and Domain Expertise with Implementation Examples

4d ago

Welcome back. You are signed into your member account **bg••••@jaxondigital.com**.

In Brain Labs by Paul Siemers, PhD

Making Philosophical Sense of AI

Why our conversations about AI feel shallow and confused

6d ago 611 27

In Predict by Manar Bouazza

Three Scientists Just Proved Schrödinger’s Cat Could Exist

In The Quantastic Journal by Sam Vaseghi

The Strange Case of Srinivasa Ramanujan—When the Beast of Infinity Refuses to Behave

Why Ramanujan’s elegant heuristics and Wilson’s renormalization group theory are fascinating chapters of one and the same story.