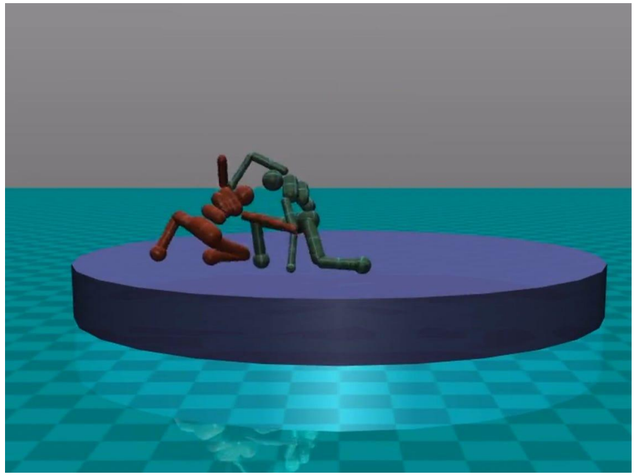
Multi-Agent AI

Noam Brown
OpenAI

Self Play




OpenAI
OpenAI Five Finals
OG vs OpenAI Five Game 2
OpenAI 5 (Bot) is
with a double kill!
OpenAI
1355

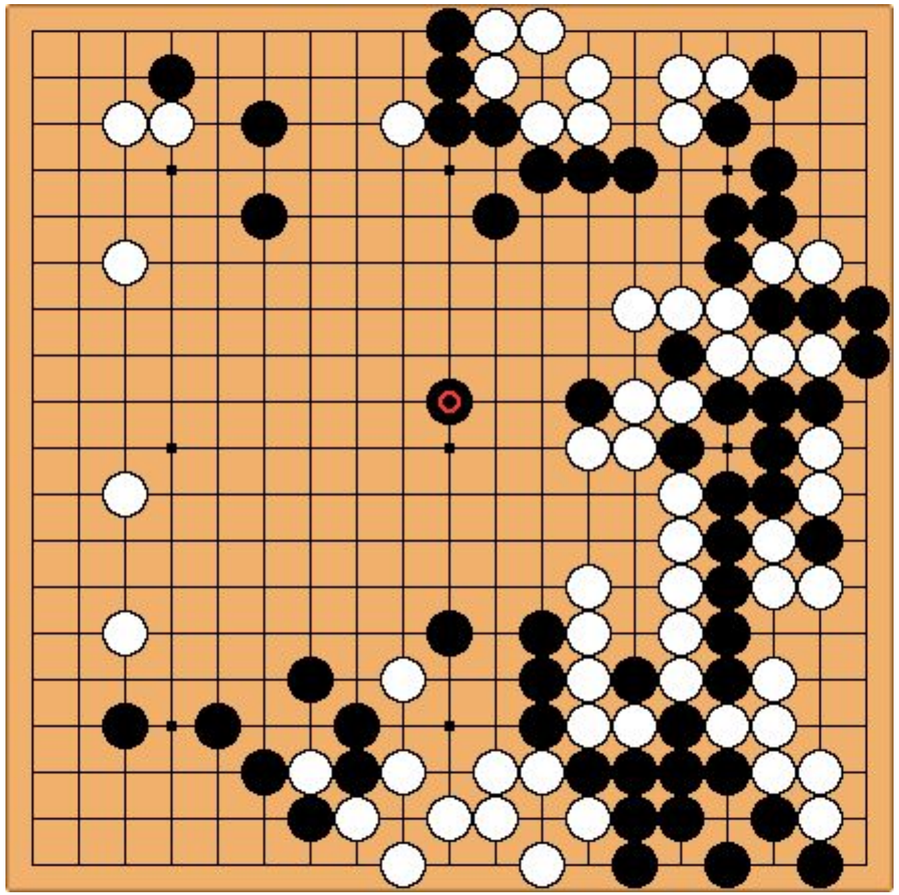
The AlphaGo Analogy

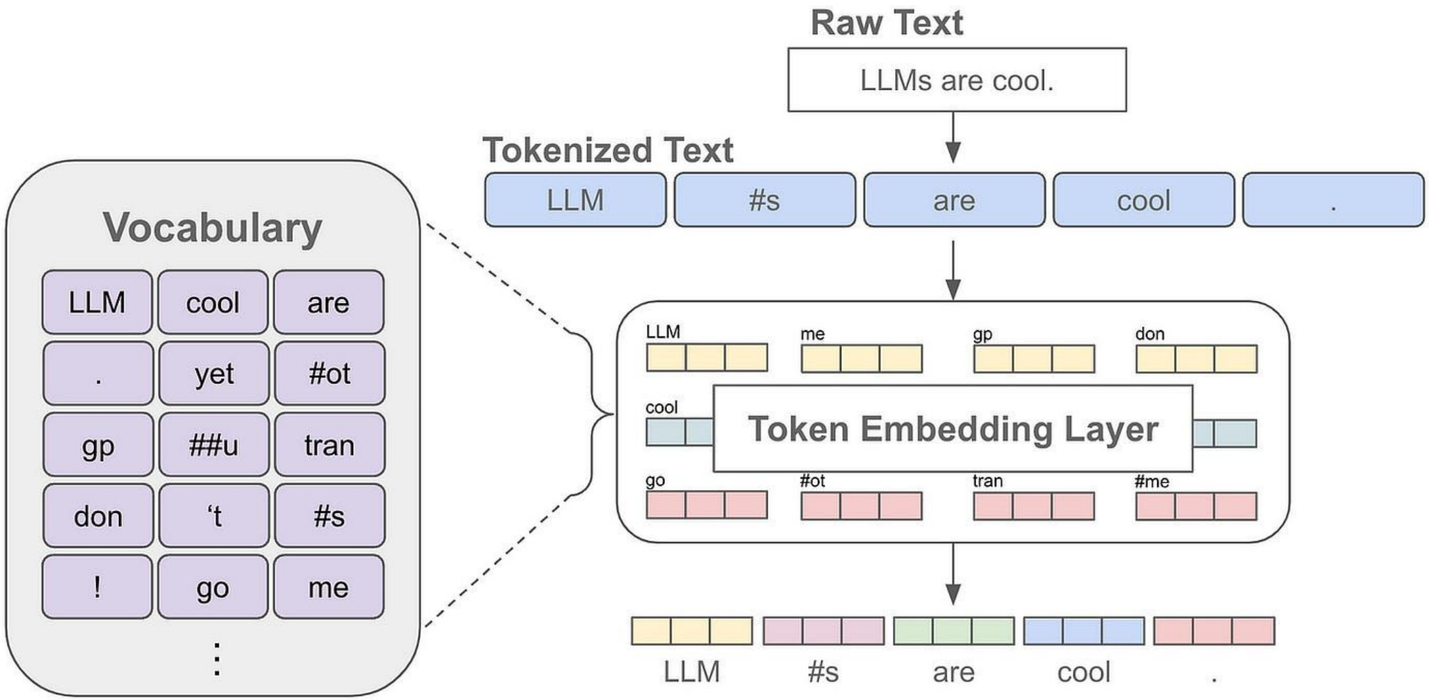
Step 1: Pretrain on high-quality human data

Step 2: Enable large-scale inference compute

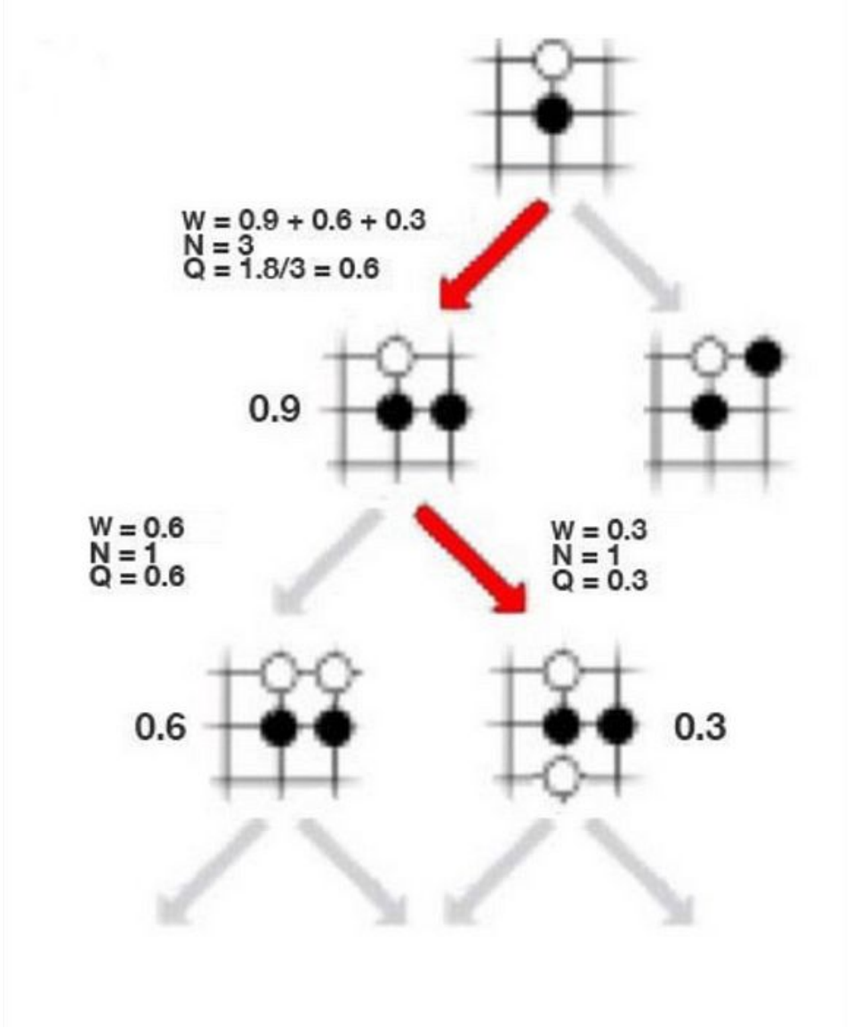
Step 3: Recursive Self-Improvement (Self Play)

Step 1: Pretrain on high-quality human data





Step 2: Enable large-scale inference compute



oyfjdnisdr rtqwainr acxz mynzbhhx -> Think step by step

Use the example above to decode:

oyekaijzdf aaptcg suaokybhai ouow aqht mynznvaatzacdfoulxxz

Thought for 5 seconds

To decode the given ciphertext using the example provided, we'll use the pattern identified in the example.

Ciphertext Example:

Step 3: Recursive Self-Improvement (Self Play)

???

Who is the better poker player?

Option 1: Someone who, over a large enough sample size, wins head-to-head vs. any other player

Option 2: Someone who makes more money playing poker than anyone else

Who is the better poker player?

Minimax Equilibrium

Option 1: Someone who, over a large enough sample size, wins head-to-head vs. any other player

Population Best Response

Option 2: Someone who makes more money playing poker than anyone else

Minimax Equilibrium

Minimax Equilibrium: a set of strategies in which no player can improve by deviating

In two-player zero-sum games, playing a minimax equilibrium ensures you will not lose in expectation

Exploitability: How much we'd lose to a best response

	Round 1	Round 2	Round 3
Us			
Best Response			

Our Exploitability = 1

Minimax Equilibrium

Minimax Equilibrium: a set of strategies in which no player can improve by deviating

In two-player zero-sum games, playing a minimax equilibrium ensures you will not lose in expectation

Exploitability: How much we'd lose to a best response

	Round 1	Round 2	Round 3
Us			
Best Response			

Our Exploitability = 1

Minimax Equilibrium

Minimax Equilibrium: a set of strategies in which no player can improve by deviating

In two-player zero-sum games, playing a minimax equilibrium ensures you will not lose in expectation

Exploitability: How much we'd lose to a best response

Critical assumption: Our strategy is common knowledge, but the outcomes of random processes are **not** common knowledge

Minimax Equilibrium

Minimax Equilibrium: a set of strategies in which no player can improve by deviating

In two-player zero-sum games, playing a minimax equilibrium ensures you will not lose in expectation

Exploitability: How much we'd lose to a best response

	Round 1	Round 2	Round 3
Us			
Best Response			

Our Exploitability = 0

Minimax Equilibrium

“Poker is simple, as your opponents make mistakes, you profit.”

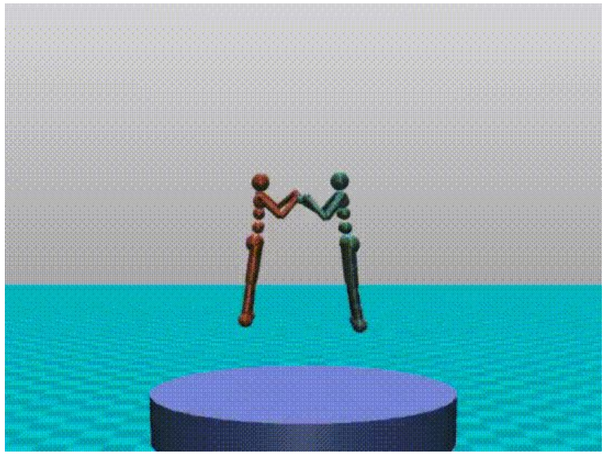
-Ryan Fee’s Poker Strategy Guide

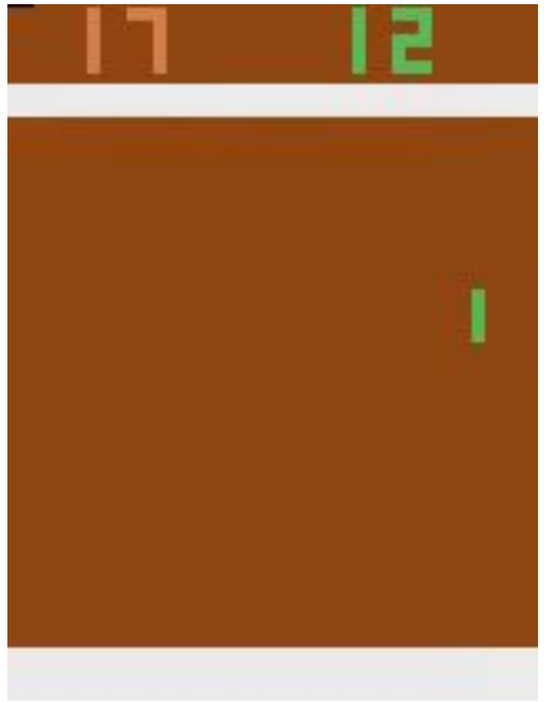
	Round 1	Round 2	Round 3
Us			
Best Response			

Our Exploitability = 0

Self-play in two-player zero-sum games

- In **self-play**, an agent gradually improves by playing against copies of itself
- Initial strategy can be completely random
- In balanced **two-player zero-sum** games, **sound self-play** provably converges to a **minimax equilibrium**
- Thus, given sufficient memory and compute, **any finite two-player zero-sum game can be “solved” via self-play**





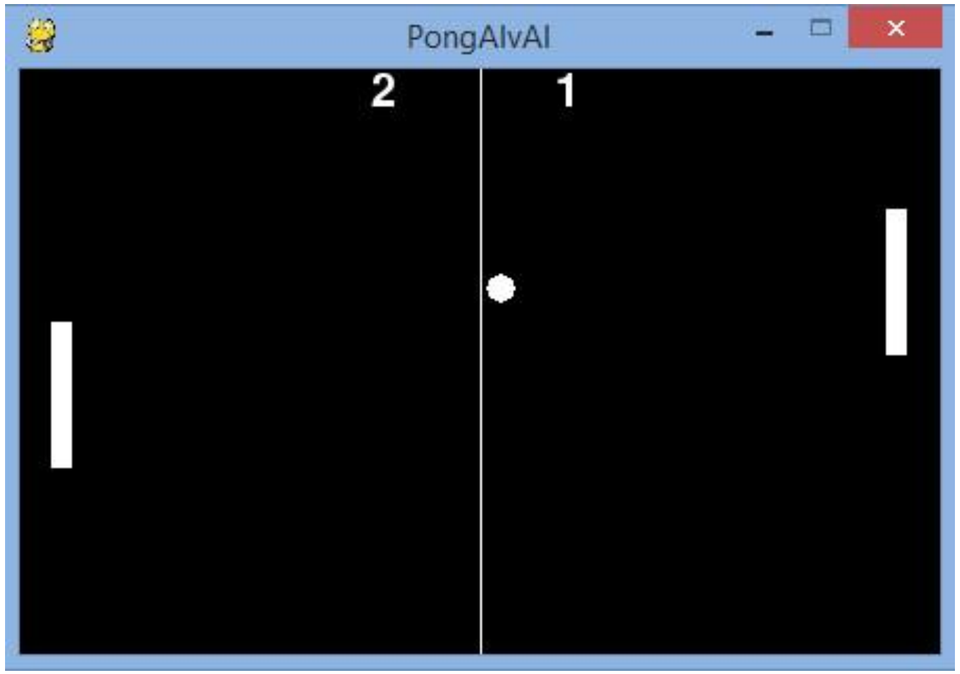
Self-play in two-player zero-sum games

- In **self-play**, an agent gradually improves by playing against copies of itself
- Initial strategy can be completely random
- In balanced **two-player zero-sum** games, **sound self-play** provably converges to a **minimax equilibrium**
- Thus, given sufficient memory and compute, **any finite two-player zero-sum game can be “solved” via self-play**

Self-Play in 2p0sum Perfect-Info Games

- Essentially just independent single-agent RL

- If exploration > 0, will in theory converge to minimax



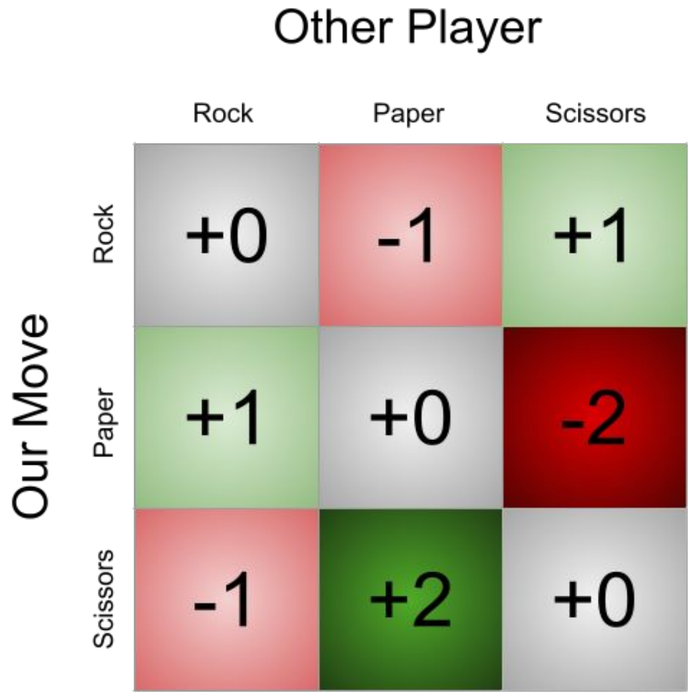
Self-Play in 2p0sum Perfect-Info Games

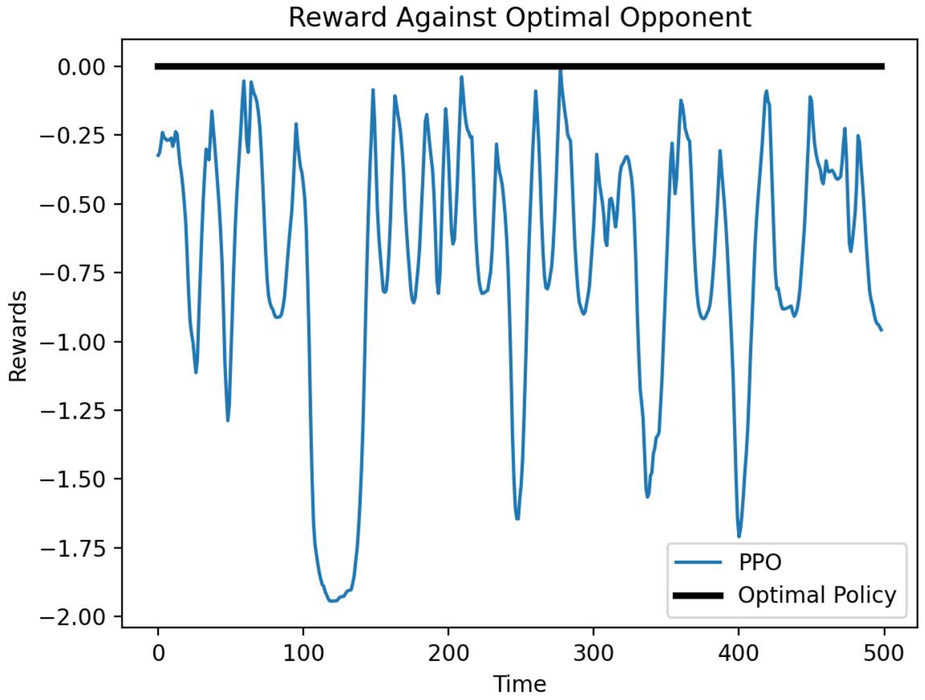
- Still may be vulnerable to adversarial attacks
- Neural net is limited in its ability to approximate minimax
- Finding an exploit is easier than defending against exploits (especially in imperfect-info games!)

Victim: Latest (cp505-v1-MCTS), no search **Adversary:** 34.1 million training steps, 600 visits

Wang et al. ICML-23

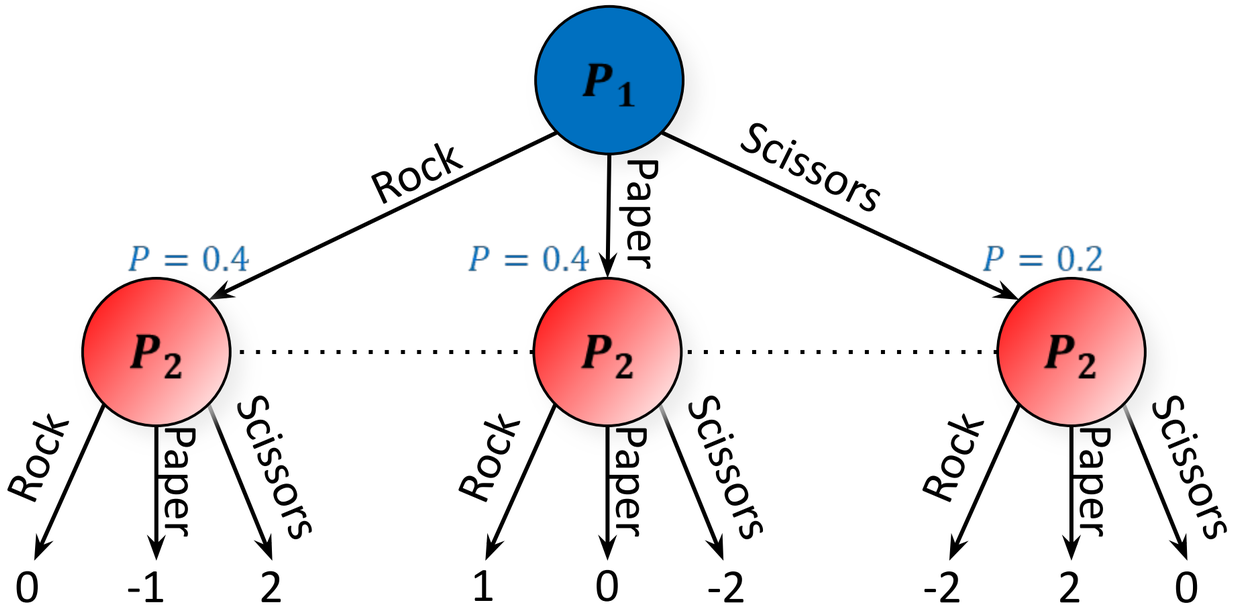
PPO in Rock-Paper-Scissors





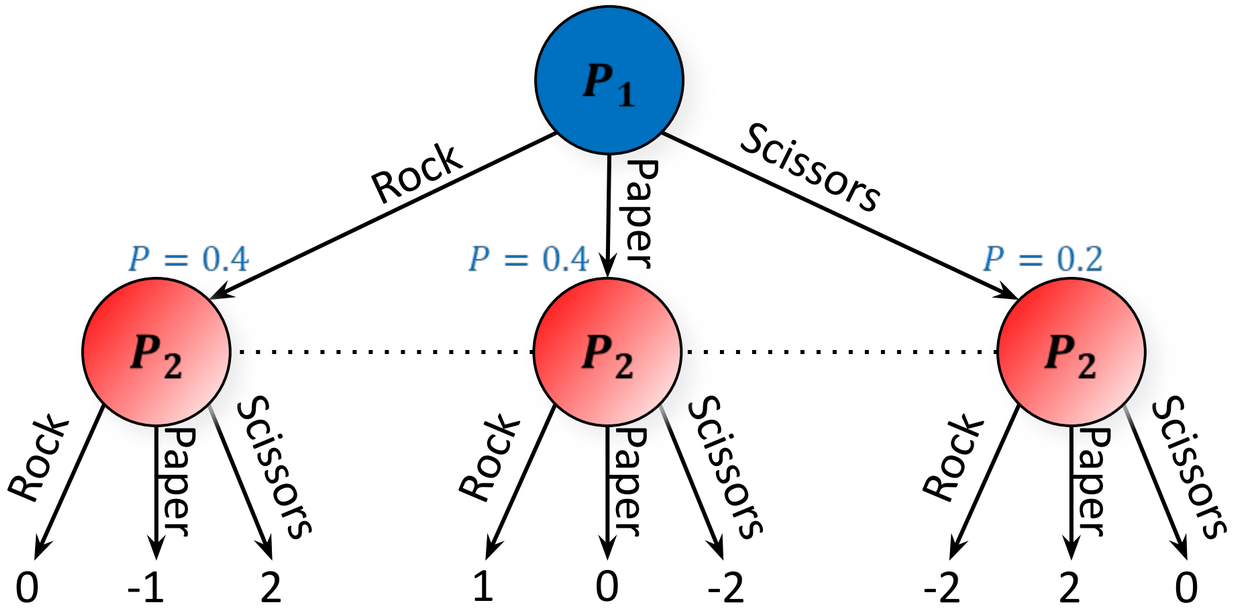
Why Imperfect-Information Games are Hard

Rock-Paper-Scissors+

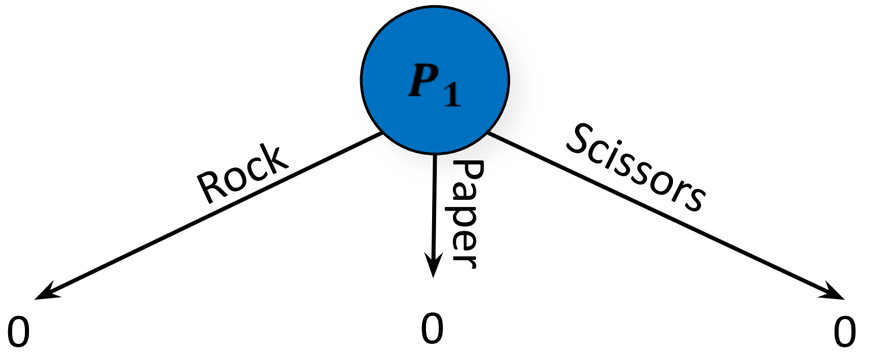


Why Imperfect-Information Games are Hard

Rock-Paper-Scissors+

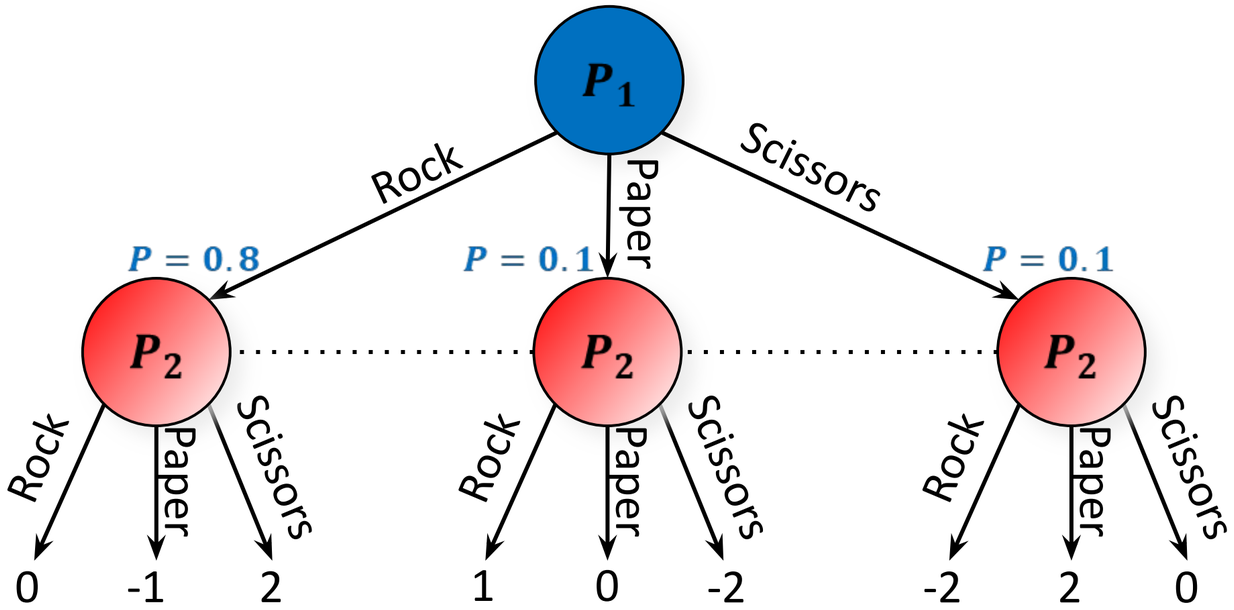


P_1 Q-values in Rock-Paper-Scissors+

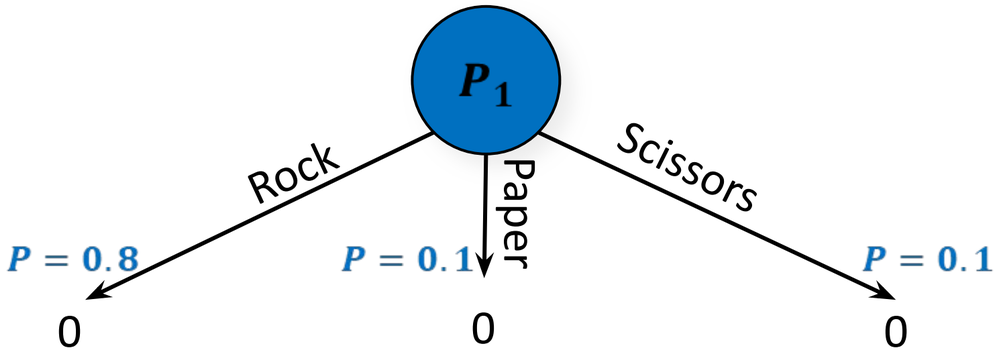


Why Imperfect-Information Games are Hard

Rock-Paper-Scissors+

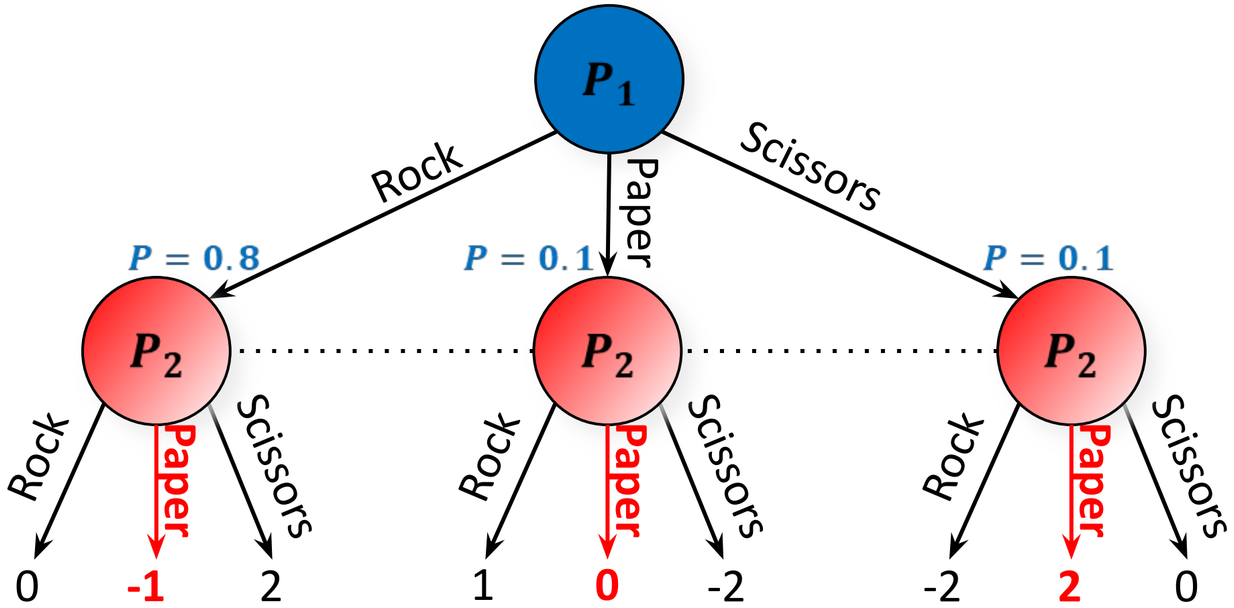


P_1 Q-values in Rock-Paper-Scissors+

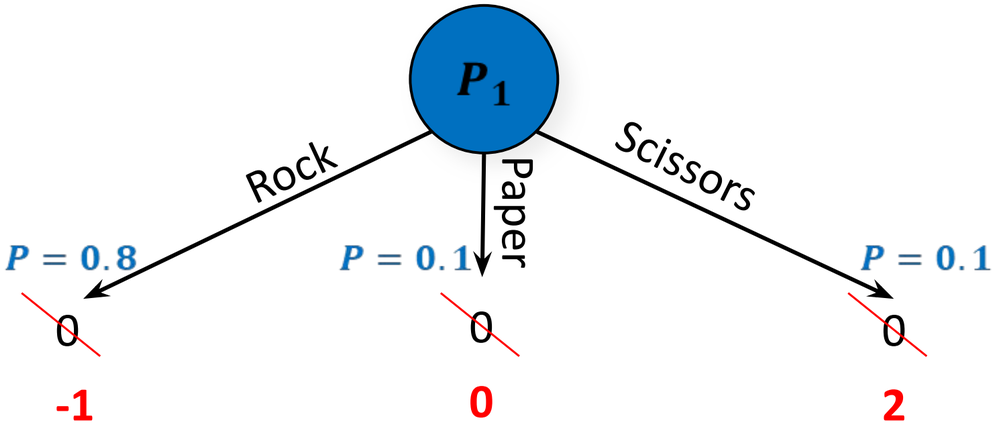


Why Imperfect-Information Games are Hard

Rock-Paper-Scissors+

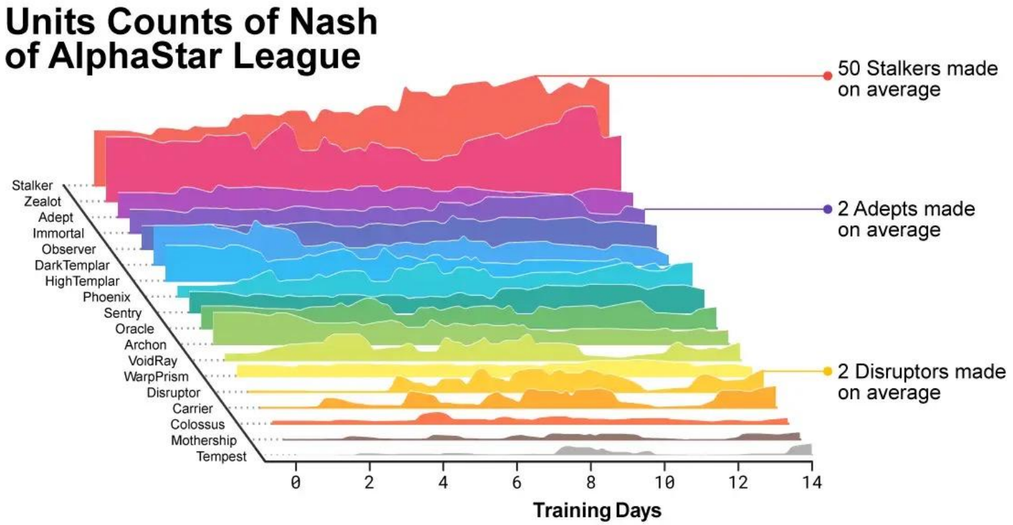


P_1 Q-values in Rock-Paper-Scissors+



Fictitious Play

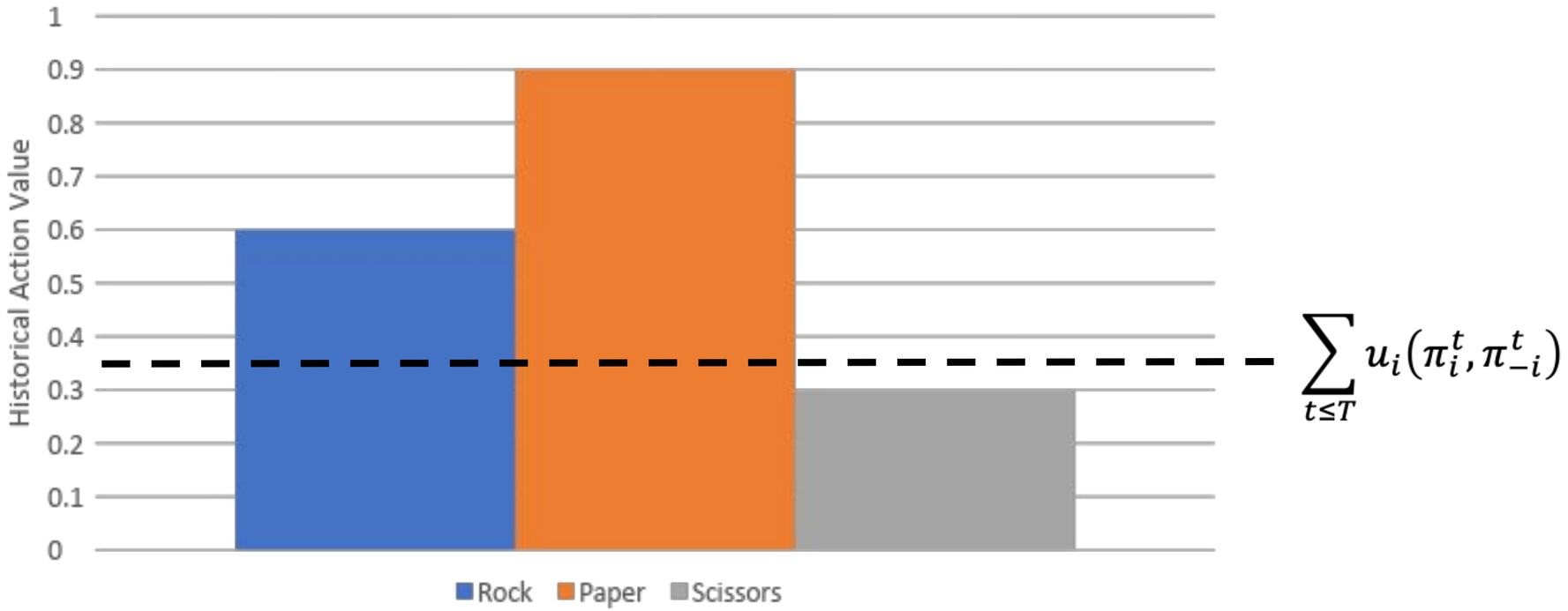
[G. W. Brown 1951]



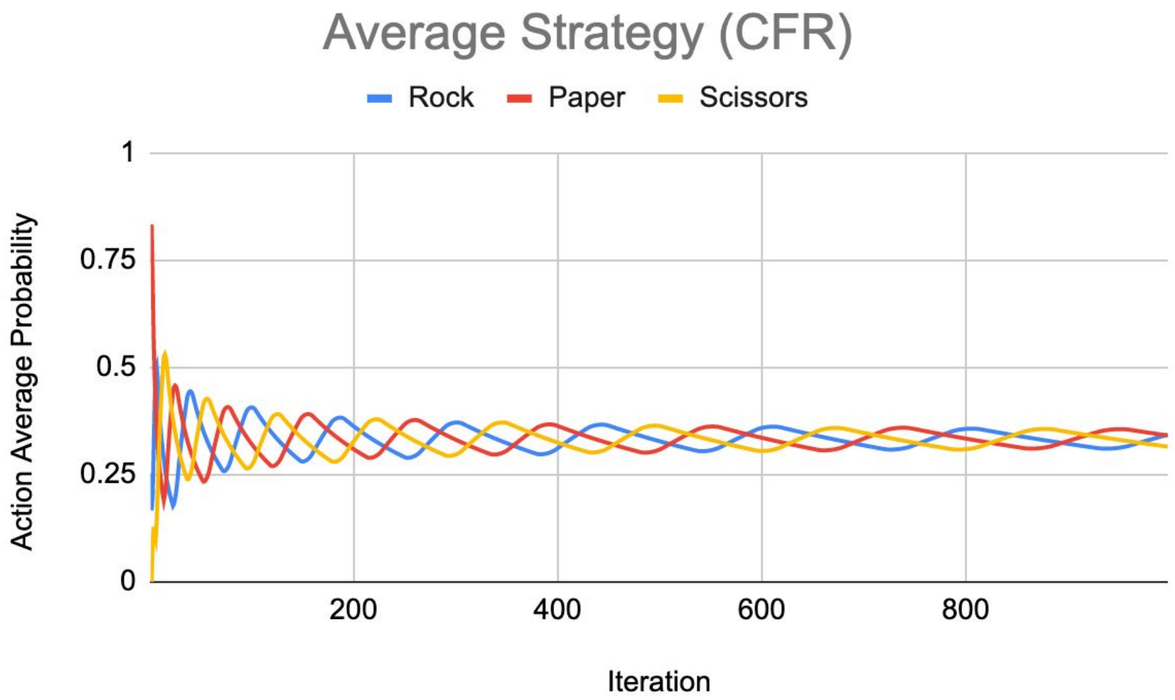
- Initialize strategies for all players arbitrarily
- On each iteration, for each player:
  - Play a best response to the opponent's average strategy over all iterations
- The **average** strategy over all iterations converges to minimax
- Rock Paper Scissors example:
  - Iteration 1: Players throw Rock (100% Rock)
  - Iteration 2: Players throw Paper (50% Rock, 50% Paper)
  - Iteration 3: Players throw Paper (33% Rock, 67% Paper)
  - Iteration 4: Players throw Scissors (25% Rock, 50% Paper, 25% Scissors)
  - …

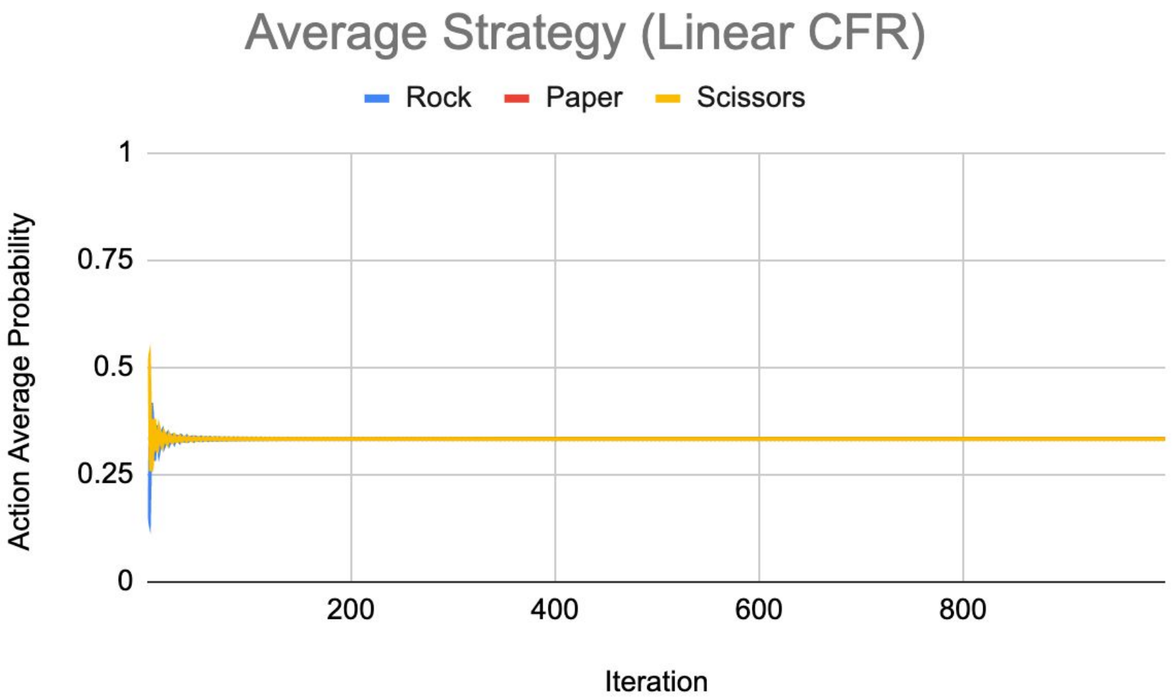
Regret Matching and Hedge

- Fictitious Play always picks a **best response** to the opponent's average
- **Regret Matching** and **Hedge** pick **regularized** best responses instead



Improving Regret Matching





2017 Brains vs AI [Brown & Sandholm Science-17]

- Libratus (our 2017 AI) against four of the ***best*** heads-up no-limit Texas Hold'em poker pros



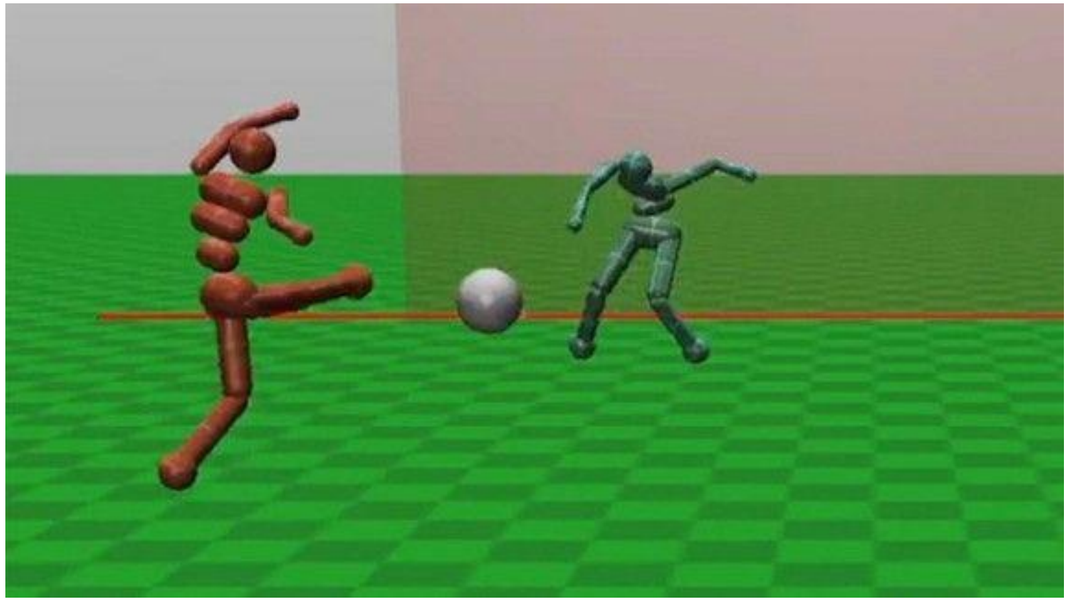
- 120,000 hands over 20 days in January 2017
- $200,000 divided among the pros based on performance
- Won with 99.98% statistical significance
- Each human lost individually to Libratus

2017 Brains vs AI [Brown & Sandholm Science-17]

- 120,000 hands of poker against a team of pros trying to exploit the bot
- Trained from self-play; no human data
- No deep neural networks

Limitations of FP / RM / Hedge

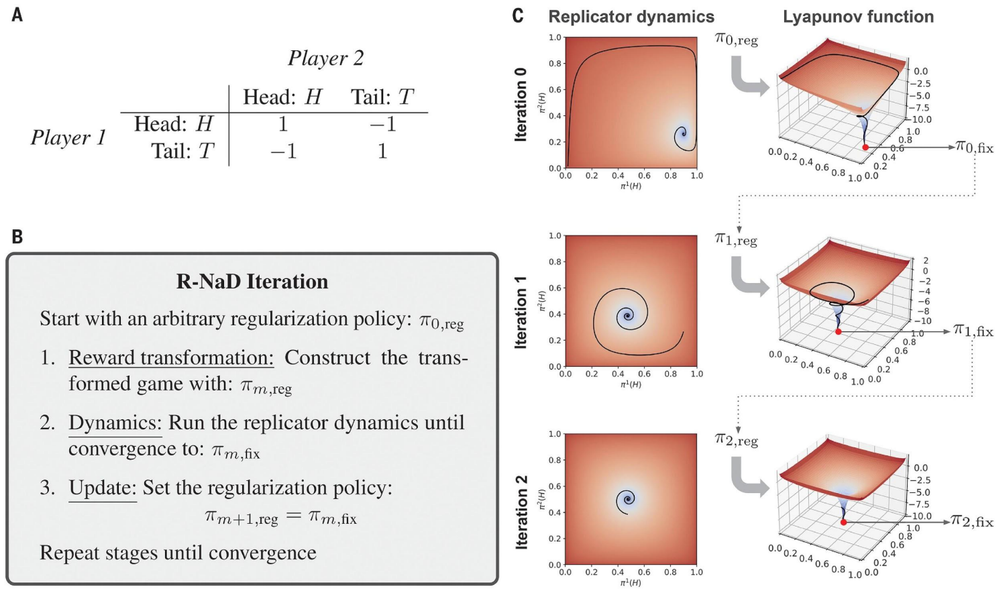
- Poor performance in single-agent RL
  - First iteration is always random
  - Policy improvement is slower
- Might require *many* iterations



New last-iterate algorithms

- Recent algorithms empirically converge to minimax and perform well in single-agent RL
  - Regularized Nash Dynamics [Perolat et al. Science-22]
  - Magnetic Mirror Descent [Sokota et al. ICLR-23]

- Similar to hedge but with additional regularization to a “magnet” policy



Aside: why haven’t we mentioned LLMs yet?

- **Theorem:** In a two-player zero-sum minimax equilibrium, “cheap talk” communication is useless

- **Proof intuition:**
  - Every minimax equilibrium in a game results in a unique value (v_1, -v_1) for players 1 and 2
  - A player 1 cheap talk action either increases v_1, decreases v_1, or has no effect
    - If it increases v_1, then player 2 simply ignores it
    - If it decreases v_1, then player 1 should not say it

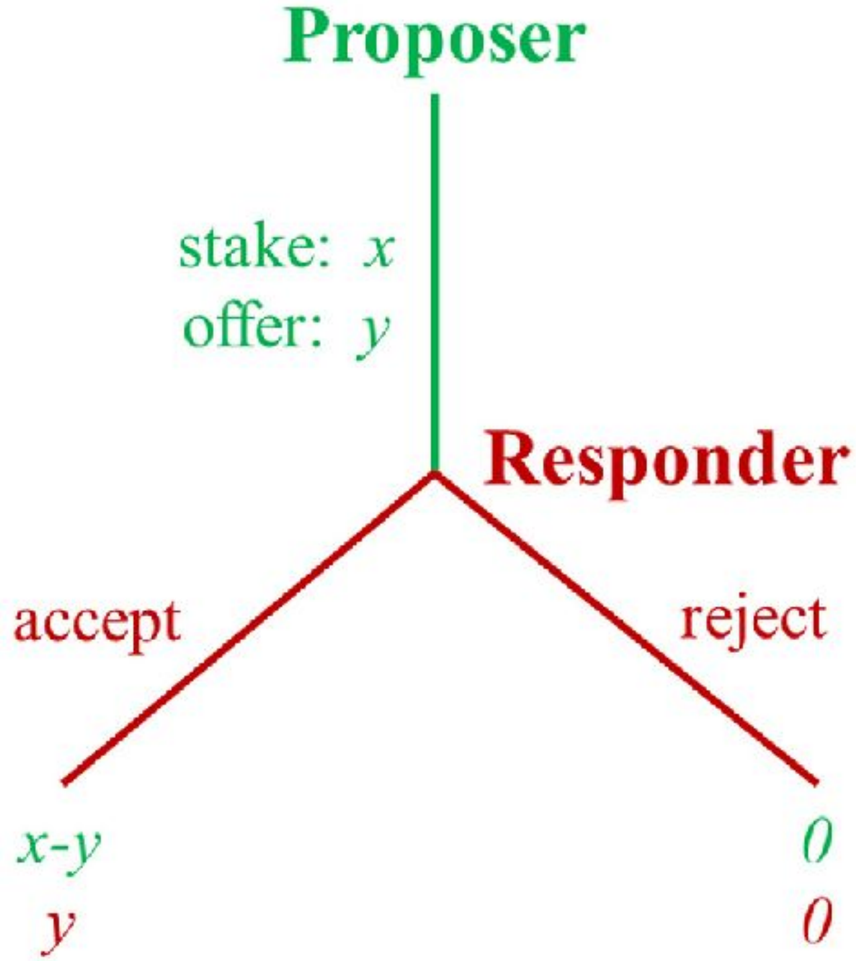
What about non-two-player zero-sum games?

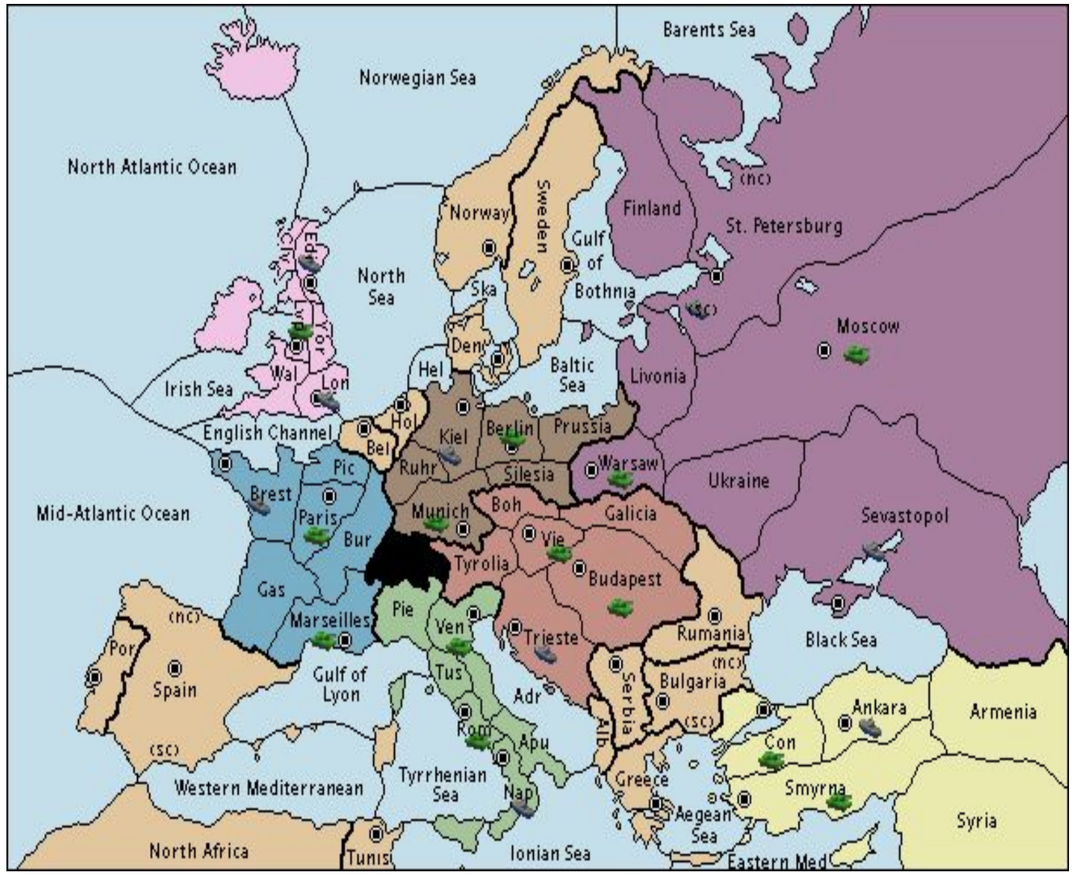
Central Claim:

Learning to cooperate with humans without using human data is a dead end

Ultimatum Game

- Alice is given $100
- First, Alice offers $0 - $100 to Bob
- Then, Bob must decide whether to **accept** or **reject**
  - If Bob **accepts**, then Alice and Bob keep their money
  - If Bob **rejects**, then Alice and Bob get nothing





GERMANY: Want support to Sweden?

ENGLAND: Let me think on that. It seems good but I think I might just lose it again straightaway.

GERMANY: we can guarantee it this turn and then Nwy the following one. I take back Den and we both build

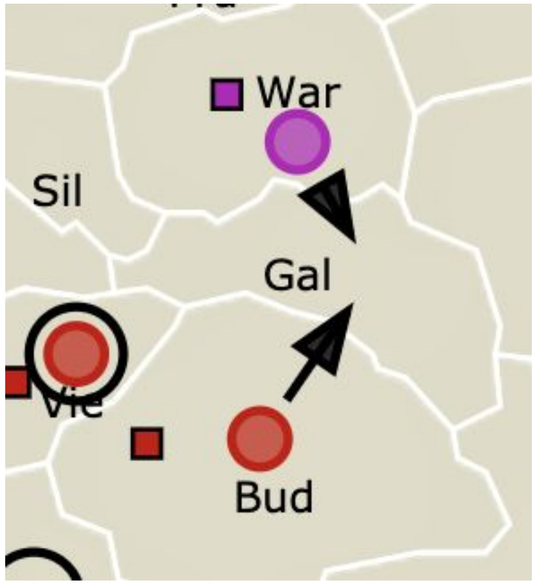
ENGLAND: Would Nwy be guaranteed? I assume Swe would retreat to Ska

Diplomacy

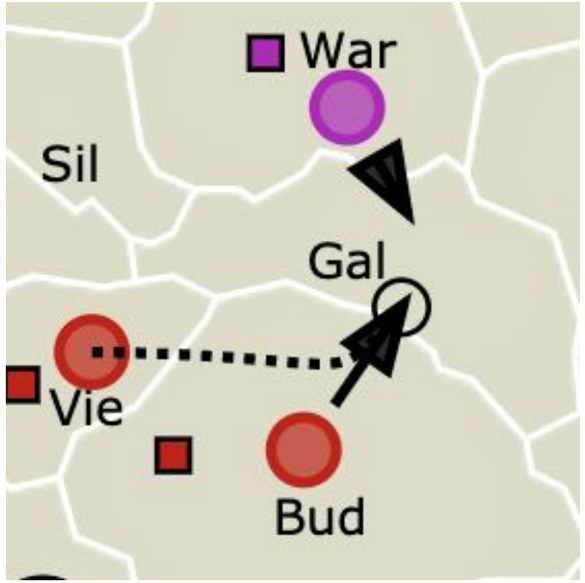
- A popular strategy game introduced in the 50s
  - 7 players trying to conquer Europe in WW1
  - JFK and Kissinger's favorite game
- Each turn involves **private natural language negotiation**
- Moves are done simultaneously
- Alliances and trust-building are key!
- Long considered a **challenge problem for AI [1]**
  - Research going back to the 80's
  - Research picked up in 2019 with the rise of LLMs

[1] Dafoe et al. *"Cooperative AI: machines must learn to find common ground".* Nature comment, 5/2021

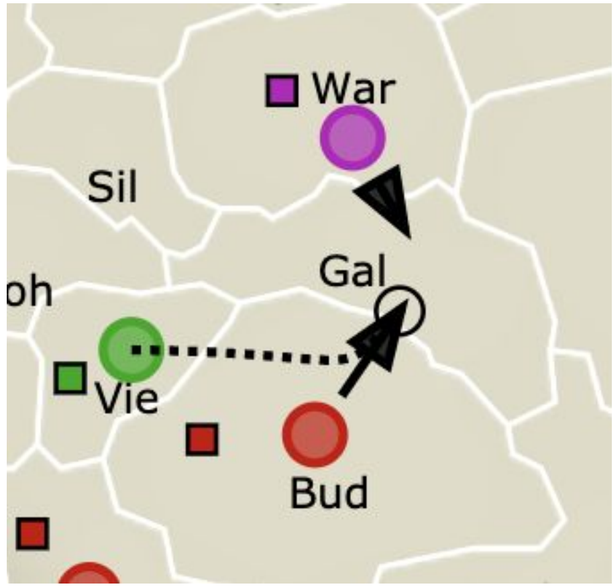
Support is key!



1v1
Fails



2v1
Succeeds



2v1
Succeeds

DORA: No-press Diplomacy from Scratch [1]

- DORA learns no-press Diplomacy through self-play
  - Similar to AlphaZero

- Performance in 2-player no-press Diplomacy:
  - **Win rate: 86.5%** +- 6.1% vs human experts

- Performance in 7-player no-press Diplomacy:

1x ↓ vs 6x →	DipNet [24]	SearchBot [11]	*DORA*	*HumanDNVI-NPU*
DipNet [24]	-	$0.8\%_{\pm 0.4\%}$	$0.0\%_{\pm 0.0\%}$	$0.1\%_{\pm 0.0\%}$
SearchBot [11]	$49.4\%_{\pm 2.6\%}$	-	$1.1\%_{\pm 0.4\%}$	$0.5\%_{\pm 0.2\%}$
DORA	$22.8\%_{\pm 2.2\%}$	$11.0\%_{\pm 1.5\%}$	-	$2.2\%_{\pm 0.4\%}$
HumanDNVI-NPU	$45.6\%_{\pm 2.6\%}$	$36.3\%_{\pm 2.4\%}$	$3.2\%_{\pm 0.7\%}$	-

[1] [Bakhtin, Wu, Lerer, Brown. NeurIPS 2021]

Who is the better poker player?

Minimax Equilibrium

Option 1: Someone who, over a large enough sample size, wins head-to-head vs. any other player

Population Best Response

Option 2: Someone who makes more money playing poker than anyone else

Who is the better poker player?

~~**Minimax Equilibrium**~~

~~Option 1: Someone who, over a large enough sample size, wins head-to-head vs. any other player~~

Not meaningful in general games!

Population Best Response

Option 2: Someone who makes more money playing poker than anyone else

Requires data on the population of players, i.e., human data

Treat humans as part of the environment

- **Step 1:** Collect a lot of human data and train an imitation model
- **Step 2:** Scale inference-time compute to *better model humans*
- **Step 3:** Scale RL with these human imitation models

Results in No-Press Diplomacy

[Bahktin et al. ICLR-2023 Best Paper Honorable Mention]

Diplodocus placed 1st in a 200-game real human tournament. (50 games each bot).

	Rank	Elo	Avg Score	# Games
Diplodocus-High	1	181	27% ± 4%	50
Human	2	162	25% ± 6%	13
Diplodocus-Low	3	152	26% ± 4%	50
Human	4	138	22% ± 9%	7
Human	5	136	22% ± 3%	57
BRBot	6	119	23% ± 4%	50
Human	7	102	18% ± 8%	8
Human	8	96	17% ± 3%	51
...	...	...	...	...
DORA	32	-20	13% ± 3%	50
...	...	...	...	...
Human	43	-187	1% ± 1%	7

CICERO Plays with Humans

- We entered CICERO anonymously in an **online natural language Diplomacy league**

- CICERO placed in the top 10% of players, and **2nd of 19** players who played at least 5 games
  - Achieved **more than 2x the average human score**

Rank	Avg Score	# Games
1	35.0%	11
2	**25.8%**	**40**
3	24.5%	6
4	22.7%	8
5	21.0%	5
	...	
19	3.0%	6
20	2.6%	7

Results in Hanabi

[Hu et al. arXiv-22]

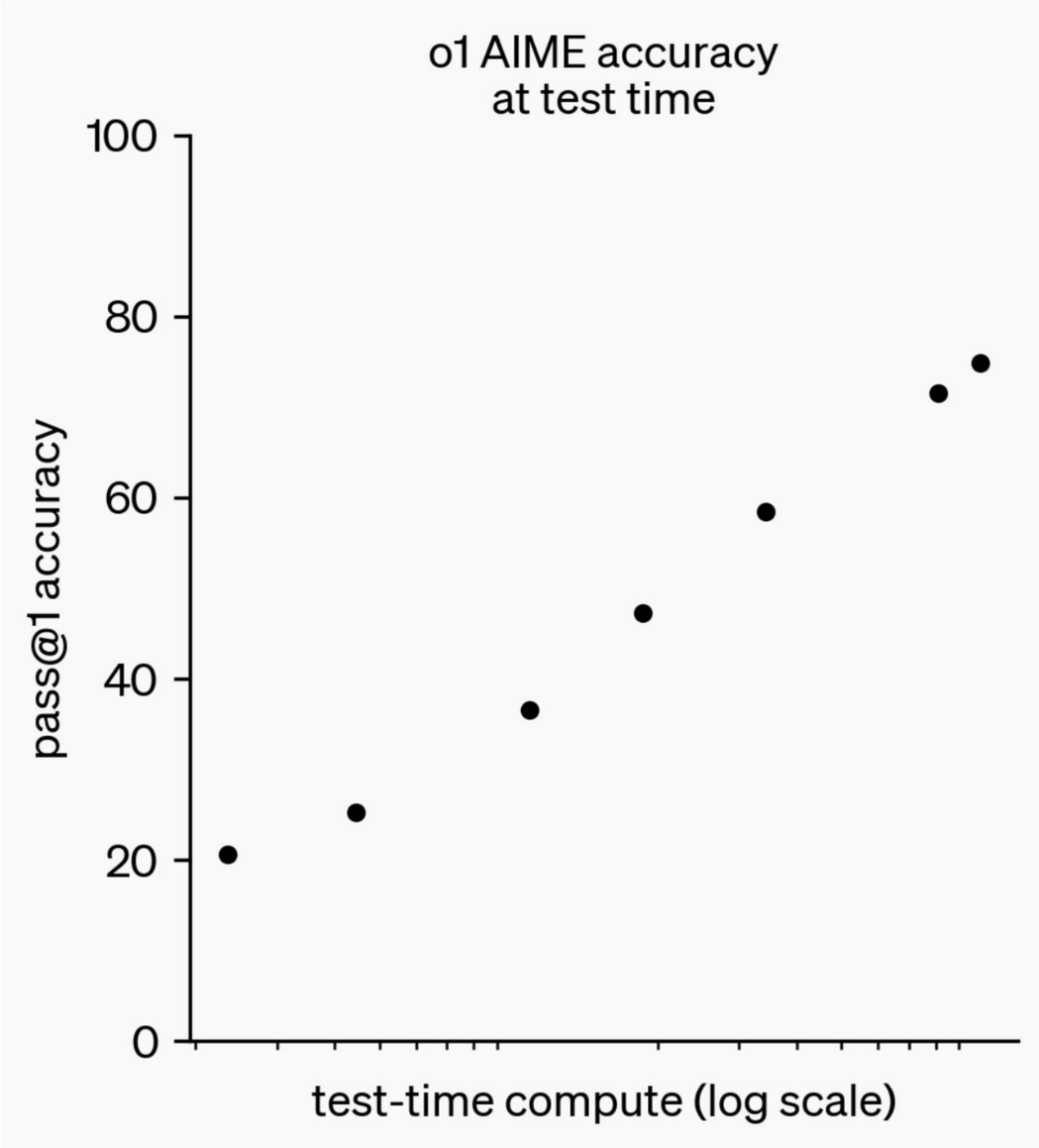
	w/ Human Experts	w/ BR-BC	w/ piKL3
All Testers (56)	14.54 ± 1.47	**16.73 ± 1.27**	**17.18 ± 1.28**
Newcomer (2)	0.00 ± 0.00	0.00 ± 0.00	**10.00 ± 7.07**
Beginner (17)	9.12 ± 2.65	**14.82 ± 2.42**	**14.47 ± 2.63**
Intermediate (23)	14.57 ± 2.27	**19.48 ± 1.64**	**18.52 ± 1.79**
Expert (14)	**23.14 ± 0.60**	16.93 ± 2.41	19.29 ± 2.14

Final Thoughts

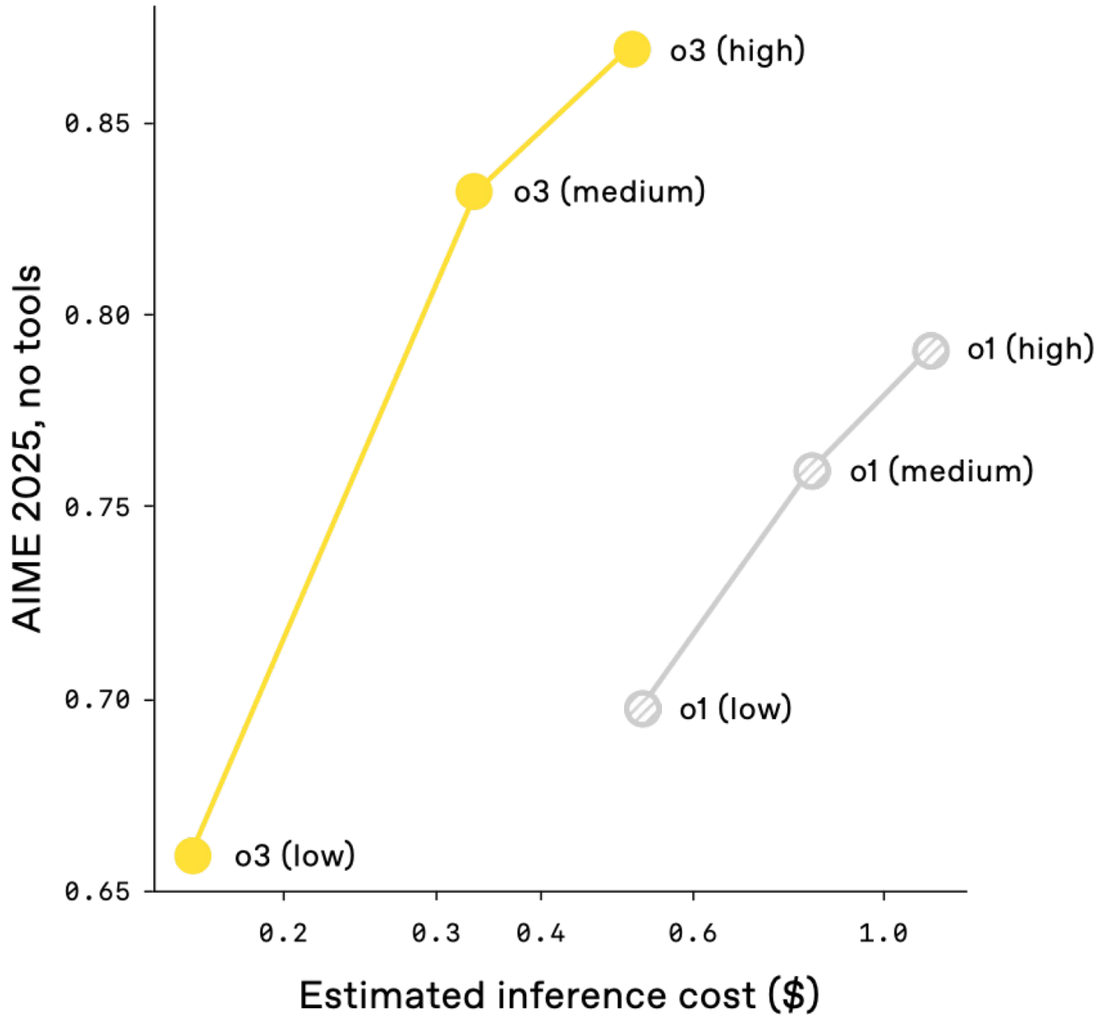
- Two-player zero-sum games are a **special case**
- In general, AlphaGo-style self-play does not converge to an “optimal” policy
- But I am optimistic about research on alternatives!

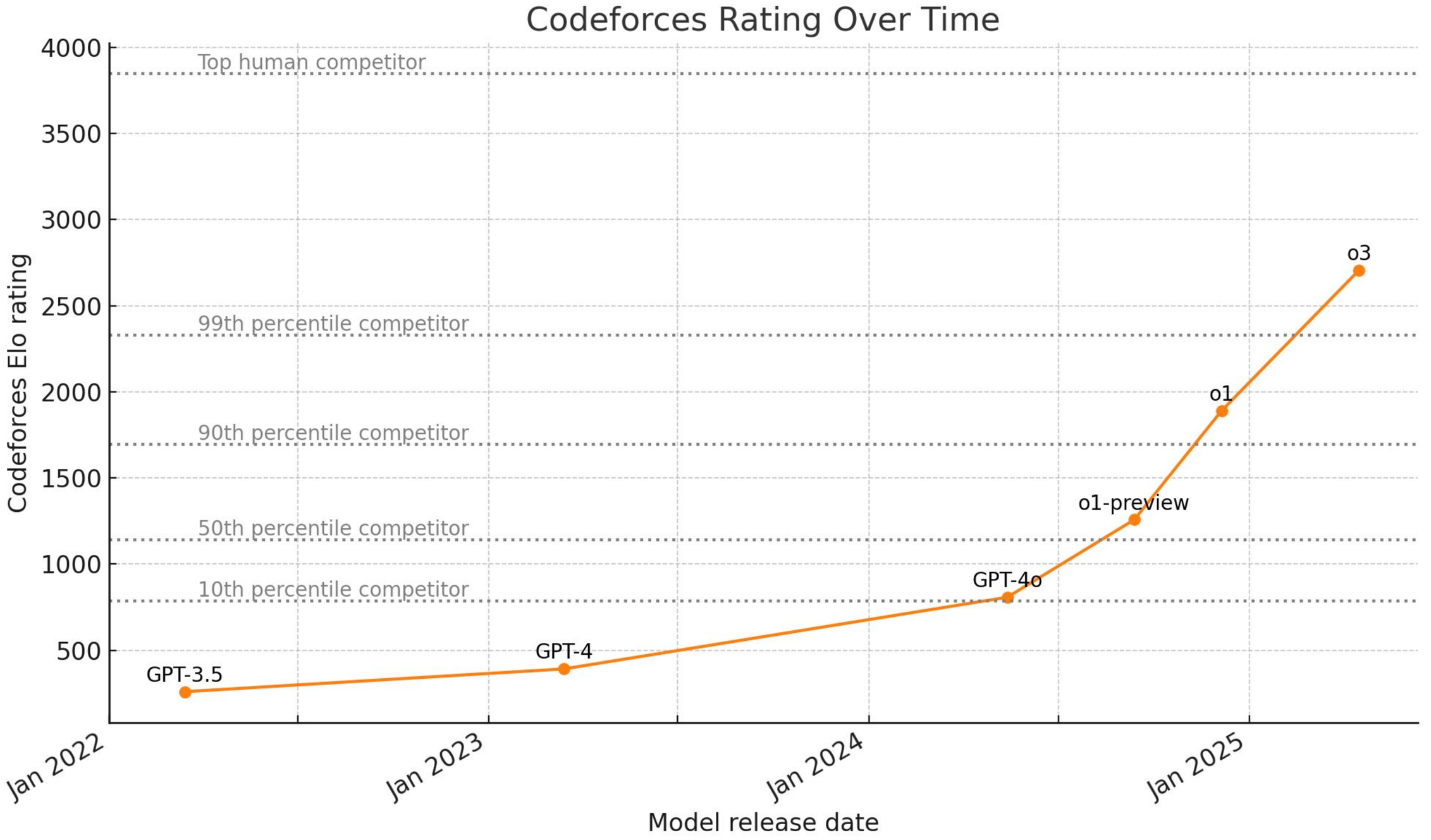
Agent-Agent LLM Cooperation

OpenAI o1


o1 AIME accuracy
at test time
100
80
60
40
20
0
pass@1 accuracy
test-time compute (log scale)

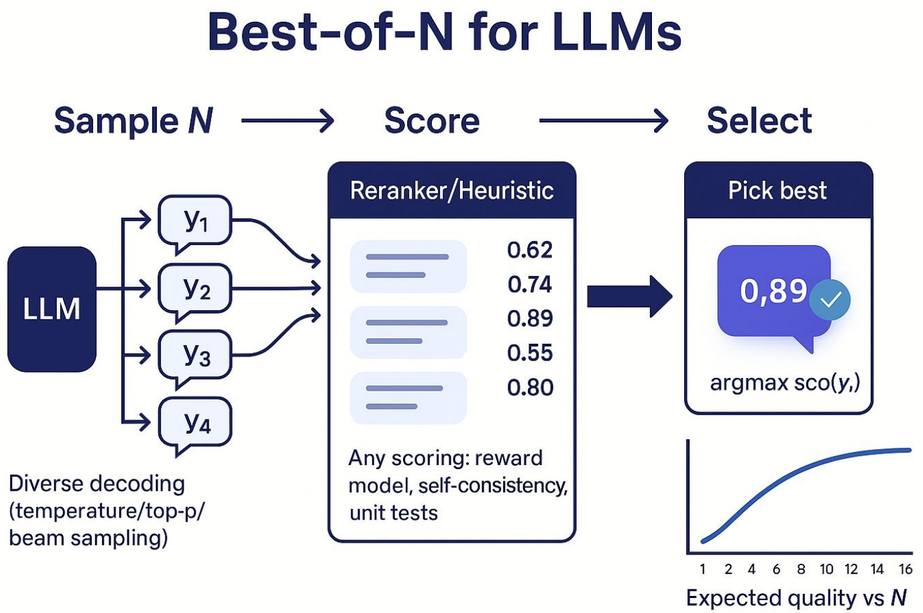
OpenAI o1 → o3


o3 (high)
o3 (medium)
o1 (high)
o1 (medium)
o1 (low)
o3 (low)
0.85
0.80
0.75
0.70
0.65
AIME 2025, no tools
0.2
0.3
0.4
0.6
1.0
Estimated inference cost ($)


Codeforces Rating Over Time
Codeforces Elo rating
4000
3500
3000
2500
2000
1500
1000
500
Top human competitor
99th percentile competitor
90th percentile competitor
50th percentile competitor
10th percentile competitor
GPT-3.5
GPT-4
GPT-4o
o1-preview
o1
o3
Jan 2022
Jan 2023
Jan 2024
Jan 2025
Model release date

Multi-Agent AI: Latency

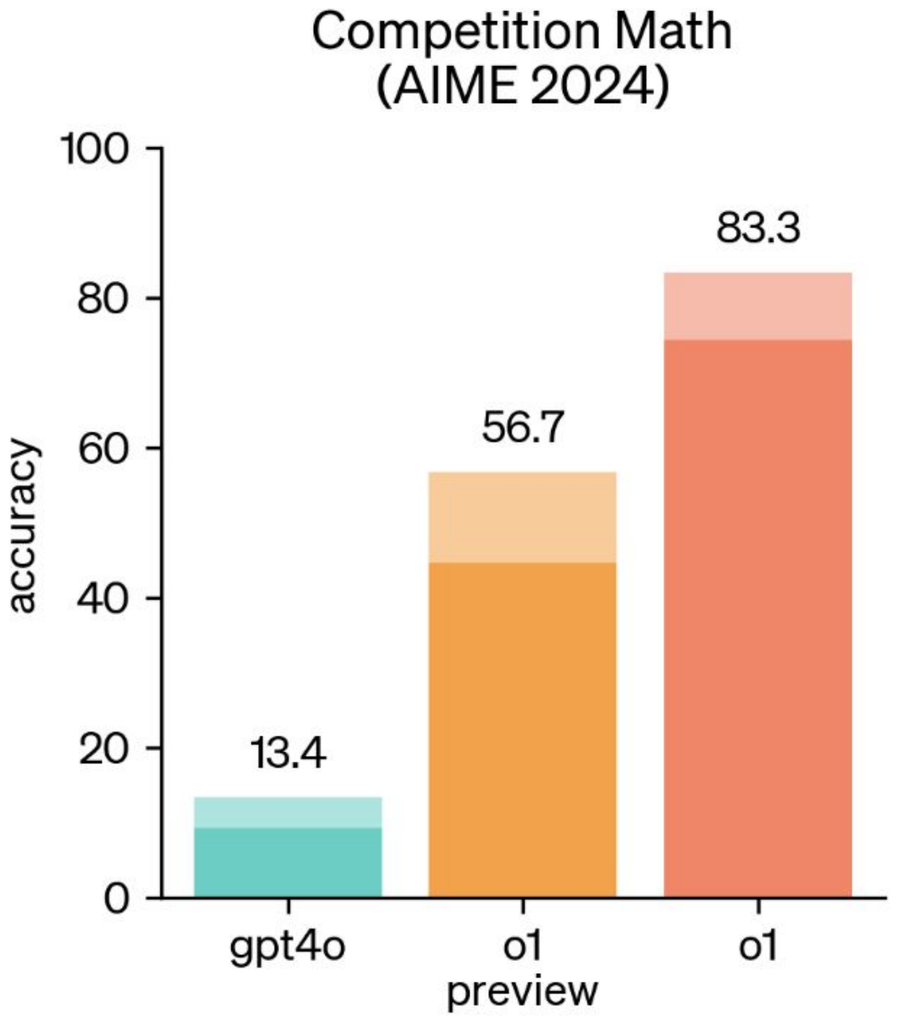
- CoT is inherently **serial**
  - Latency eventually becomes a bottleneck
- Other test-time scaling techniques are **parallel**
  - Best-of-N / consensus
  - Lower latency, but less compute-efficient



Multi-Agent AI: Latency

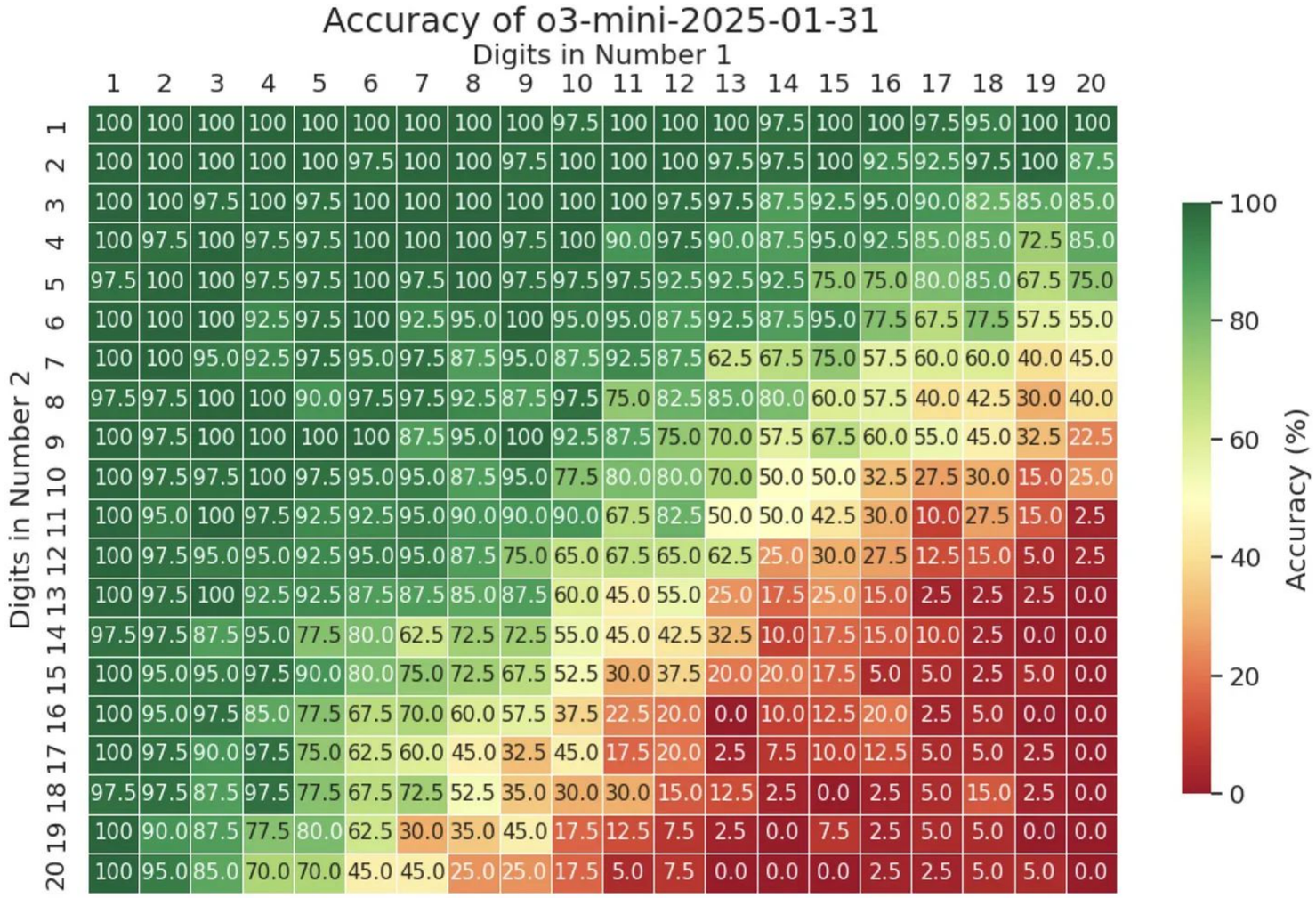
- CoT is inherently **serial**
  - Latency eventually becomes a bottleneck

- Other test-time scaling techniques are **parallel**
  - Best-of-N / consensus
  - Lower latency, but less compute-efficient

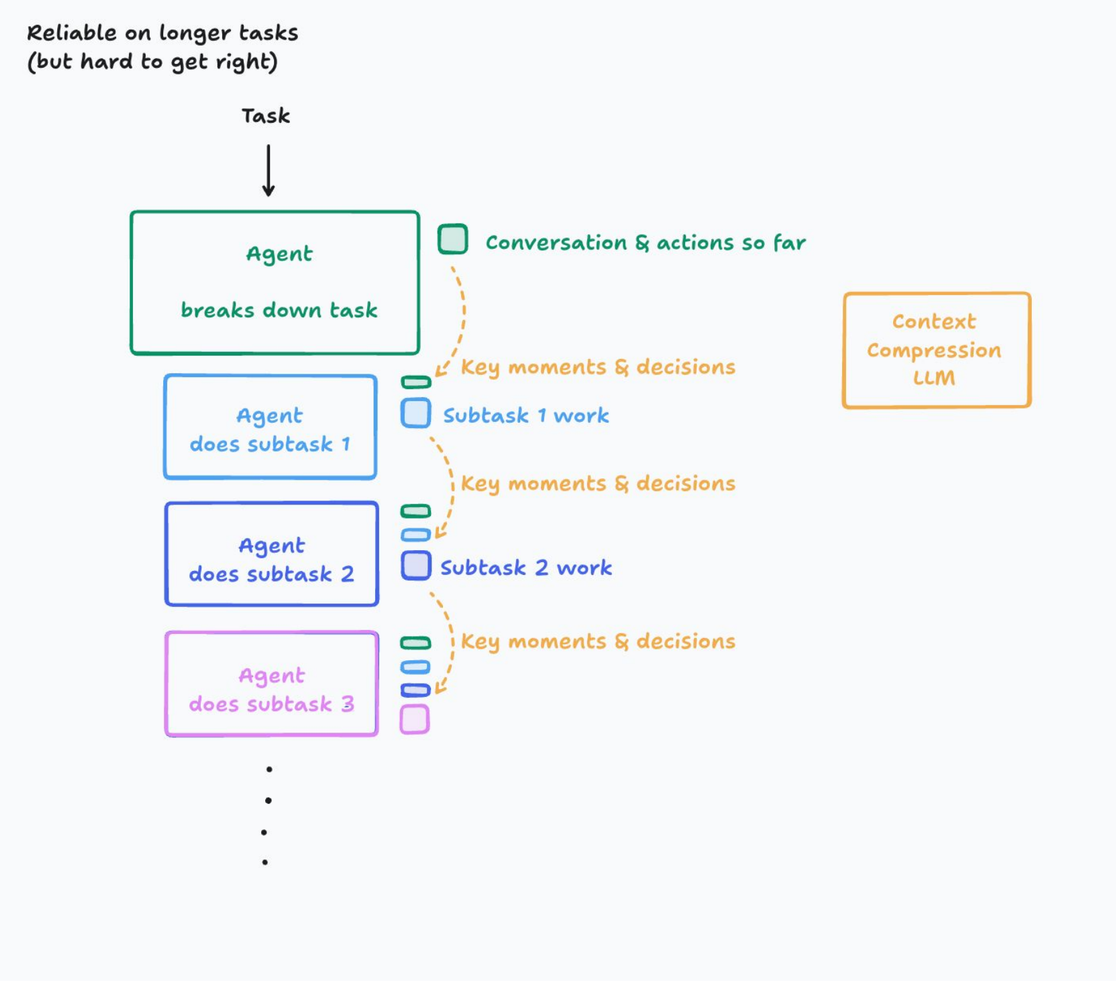


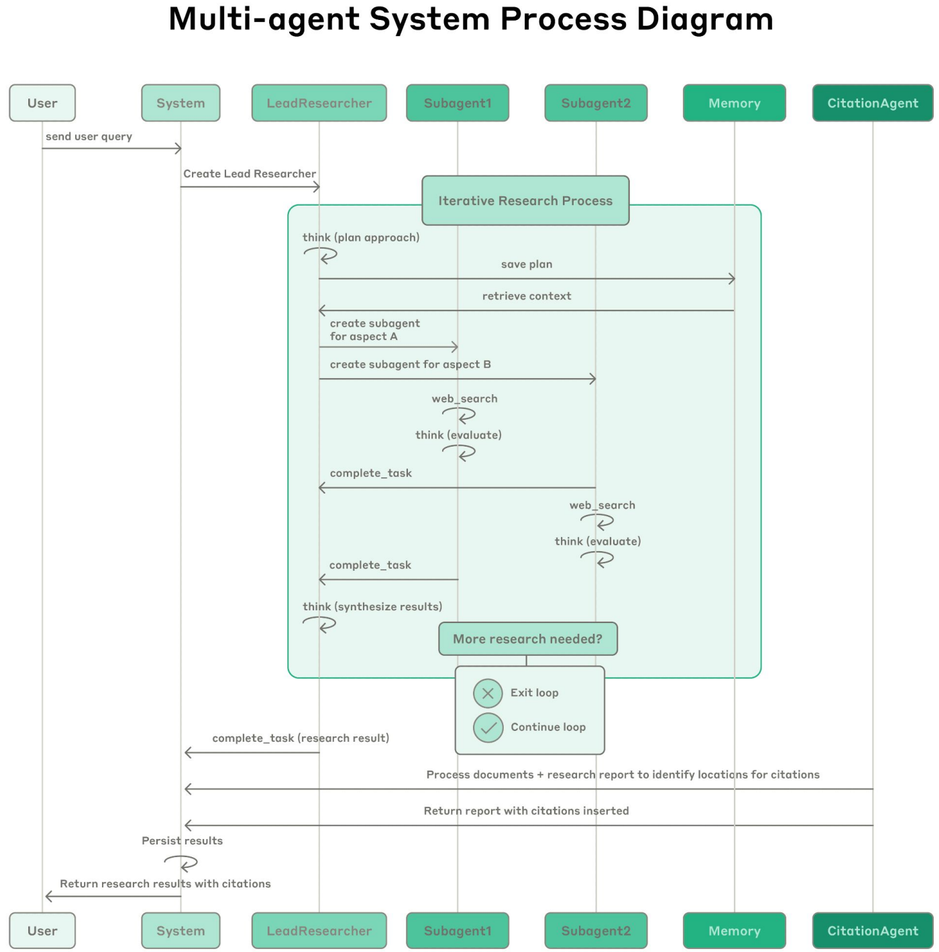
Multi-Agent AI: Diversity

- Reasoning models *can* multiply large numbers, but why do that?
- Diversity is a strength
  - Can use the best model for the particular query
- “Routing” is already a form of multi-agent AI



Multi-Agent Collaboration Scaffolds Today





June 12, 2025

Don’t Build Multi-Agents

By Walden Yan

Multi-Agents

If we really want to get parallelism out of our system, you might think to let the decision makers “talk” to each other and work things out.

This is what us humans do when we disagree (in an ideal world). If Engineer A’s code causes a merge conflict with Engineer B, the correct protocol is to talk out the differences and reach a consensus. However, agents today are not quite able to engage in this style of long-context proactive discourse with much more reliability than you would get with a single agent. Humans are quite efficient at communicating our most important knowledge to one another, but this efficiency takes nontrivial intelligence.

Since not long after the launch of ChatGPT, people have been exploring the idea of multiple agents interacting with one another to achieve goals [3][4]. While I’m optimistic about the long-term possibilities of agents collaborating with one another, it is evident that in 2025, running multiple agents in collaboration only results in fragile systems. The decision-making ends up being too dispersed and context isn’t able to be shared thoroughly enough between the agents. At the moment, I don’t see anyone putting a dedicated effort to solving this difficult cross-agent context-passing problem. I personally think it will come for free as we make our single-threaded agents even better at communicating with humans. When this day comes, it will unlock much greater amounts of parallelism and efficiency.