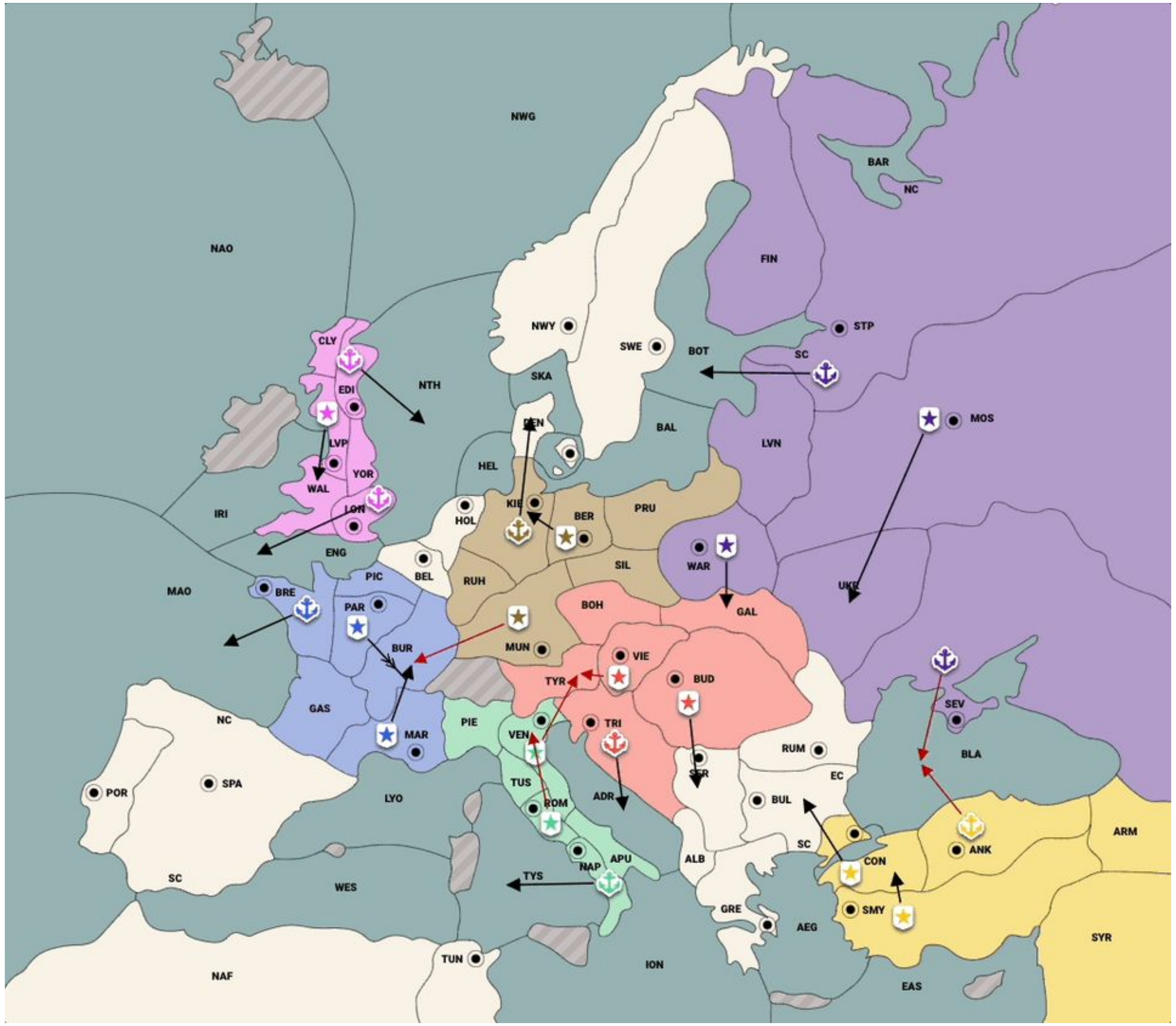
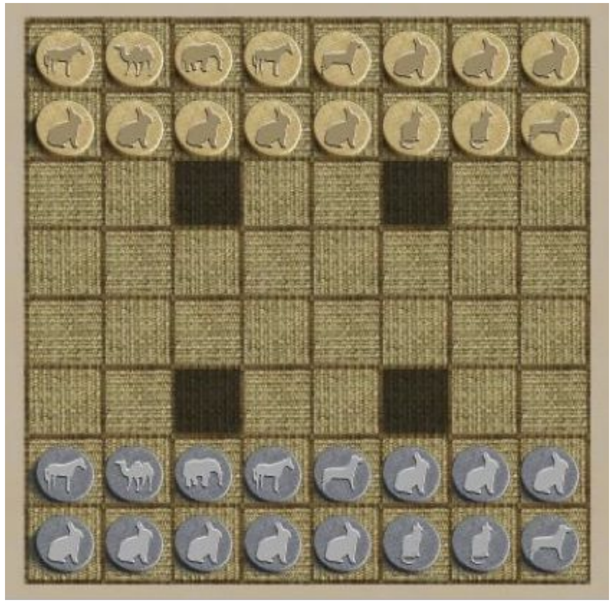
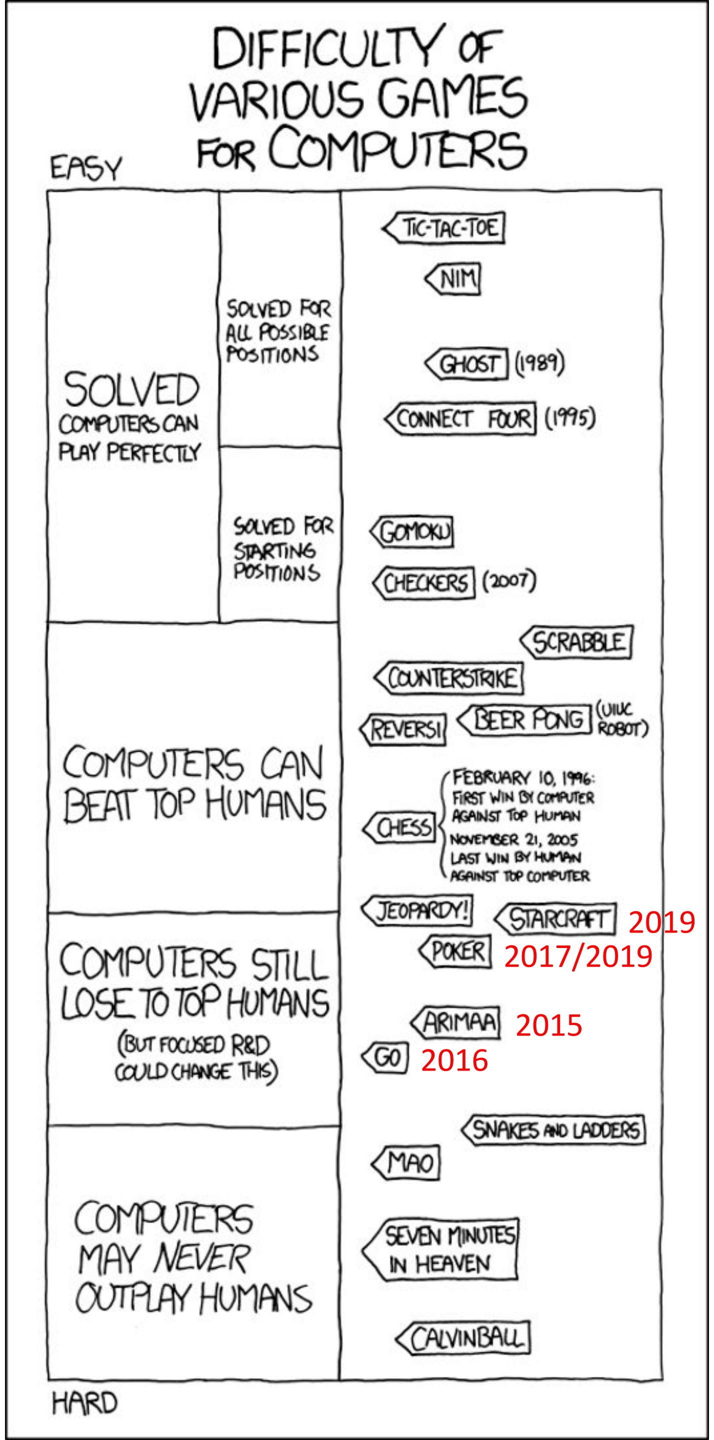
Learning to Cooperate and Compete via Self Play



Noam Brown



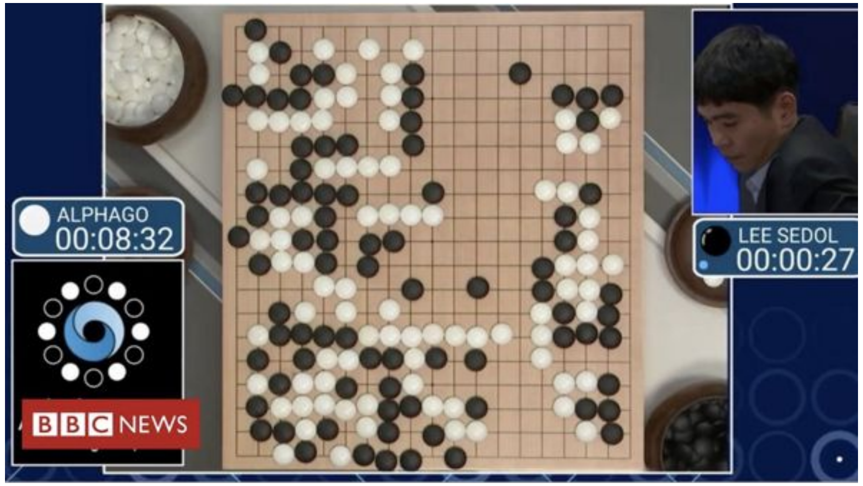

DIFFICULTY OF VARIOUS GAMES FOR COMPUTERS
EASY
SOLVED COMPUTERS CAN PLAY PERFECTLY
SOLVED FOR ALL POSSIBLE POSITIONS
TIC-TAC-TOE
NIM
GHOST (1989)
CONNECT FOUR (1995)
SOLVED FOR STARTING POSITIONS
GOMOKU
CHECKERS (2007)
COMPUTERS CAN BEAT TOP HUMANS
SCRABBLE
COUNTERSTRIKE
REVERSI
BEER PONG (UIUC ROBOT)
CHESS
FEBRUARY 10, 1996: FIRST WIN BY COMPUTER AGAINST TOP HUMAN
NOVEMBER 21, 2005 LAST WIN BY HUMAN AGAINST TOP COMPUTER
JEOPARDY!
COMPUTERS STILL LOSE TO TOP HUMANS (BUT FOCUSED R&D COULD CHANGE THIS)
STARCRAFT
POKER
ARIMAA
GO
COMPUTERS MAY NEVER OUTPLAY HUMANS
SNAKES AND LADDERS
MAO
SEVEN MINUTES IN HEAVEN
CALVINBALL
HARD

2019

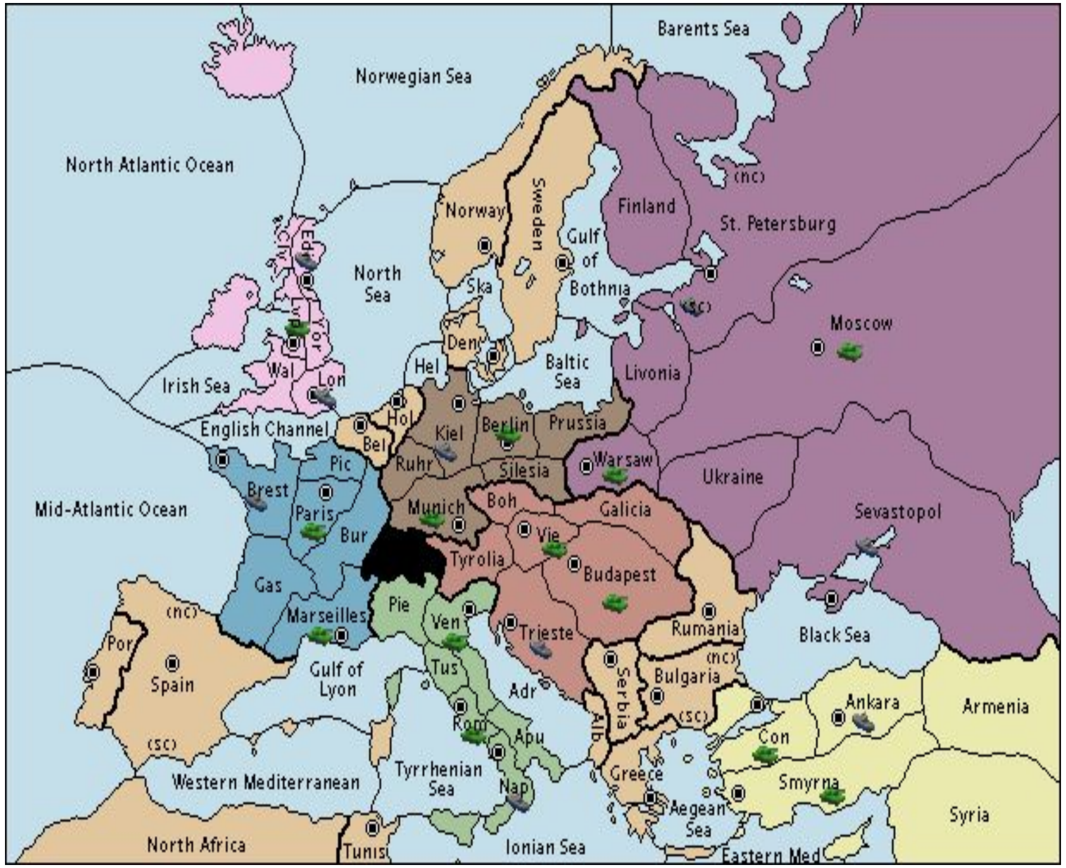
2017/2019

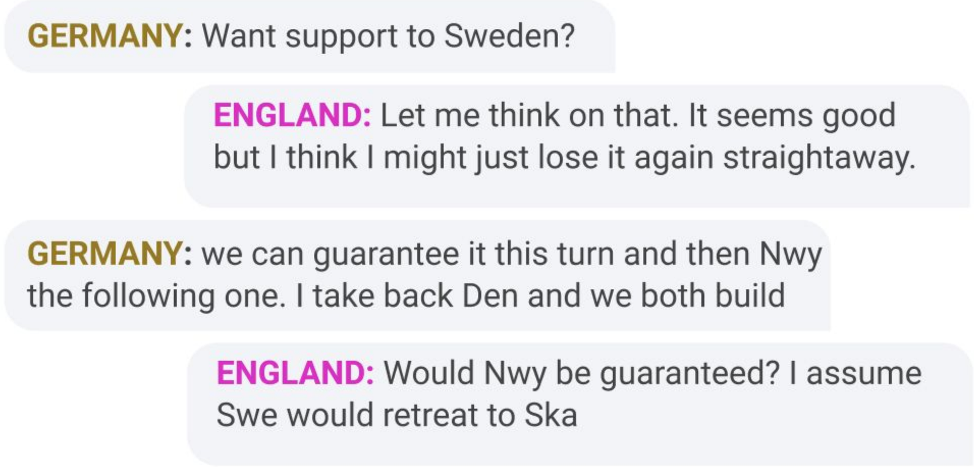
2015

2016



GPT-2
The scientist named the population, after their distinctive horn, Ovid’s Unicorn.





GERMANY: Want support to Sweden?

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

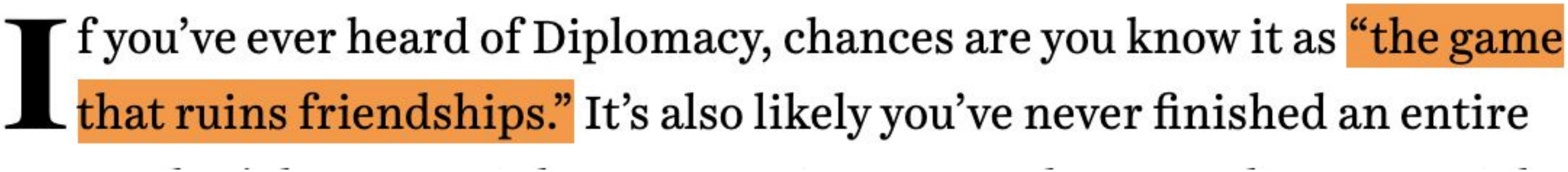
GERMANY: we can guarantee it this turn and then Nwy the following one. I take back Den and we both build

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

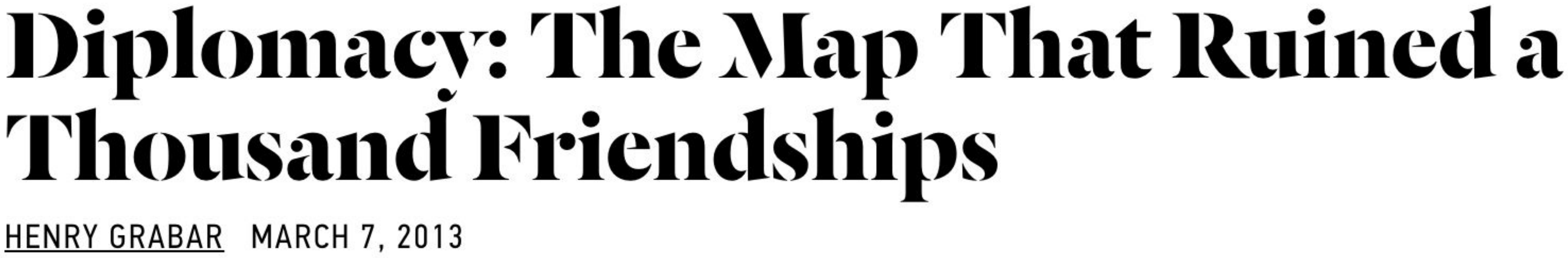
- A popular strategy game from the 50s
  - 7 players trying to conquer Europe in WW1
  - JFK and Kissinger’s favorite game
- Each turn involves **private natural language negotiation**
- Moves are done simultaneously

  e.g. F CLY - NWG, A DEN H, F SKA S A SWE – NWY, …
- 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 work from MILA, DeepMind, ourselves, others

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

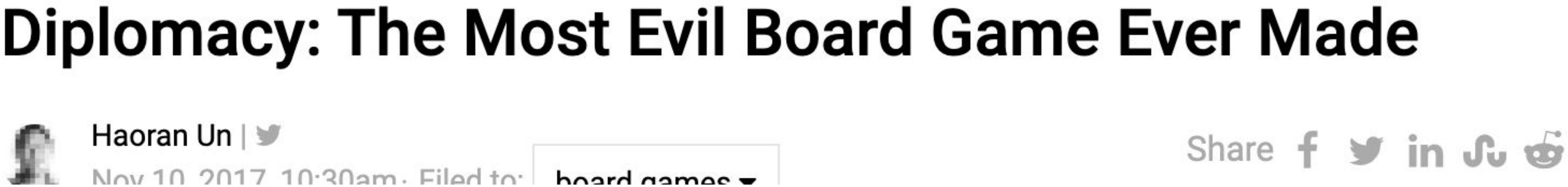


If you've ever heard of Diplomacy, chances are you know it as "the game that ruins friendships." It's also likely you've never finished an entire



Diplomacy: The Map That Ruined a Thousand Friendships

HENRY GRABAR MARCH 7, 2013



Diplomacy: The Most Evil Board Game Ever Made

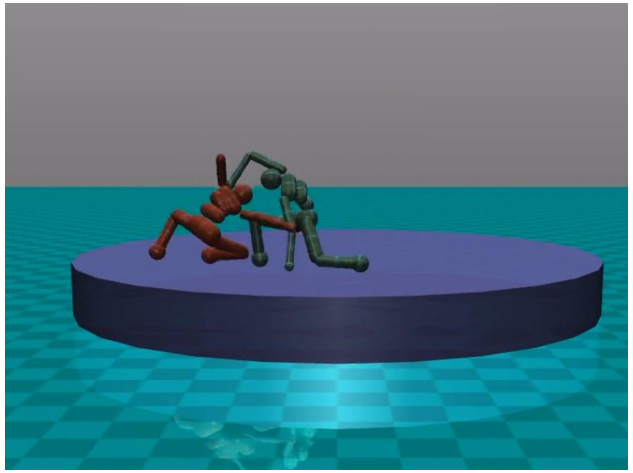
Haoran Un |

Nov 10, 2017, 10:30am · Filed to: board games

“Diplomacy is ultimately about **building trust** in an environment that encourages you to not trust anyone.”

-Andrew Goff
3-Time Diplomacy World Champion

Self-Play in 2p 0-Sum Games



OpenAI 5 (Bot) is GODLIKE with a double kill!

OG.Topson bought back

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 in 2p0sum: each player’s strategy is optimal given the other player’s policy

In balanced games, playing minimax ensures you will not lose on average

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

	Round 1	**Round 2**	**Round 3**
Us	🪨	🪨	🪨
Best Response	📄	📄	📄

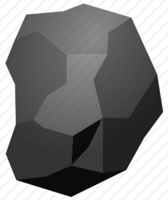
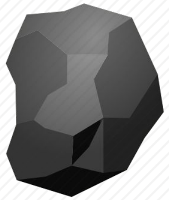
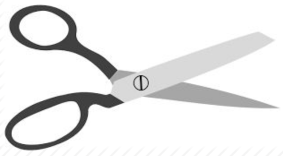
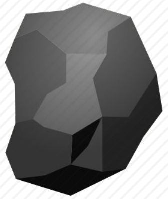
Our Exploitability = 1

Minimax Equilibrium

Minimax Equilibrium in 2p0sum: each player’s strategy is optimal given the other player’s policy

In balanced games, playing minimax ensures you will not lose on average

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

	Round 1	**Round 2**	**Round 3**
Us			
Best Response			

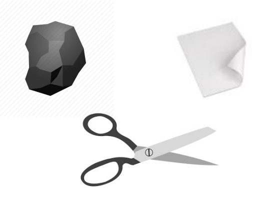
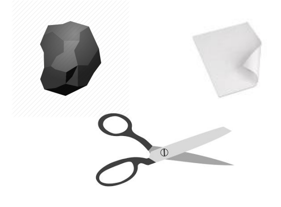
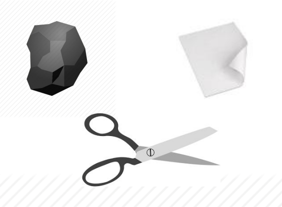
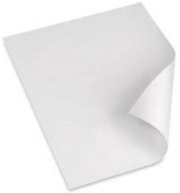
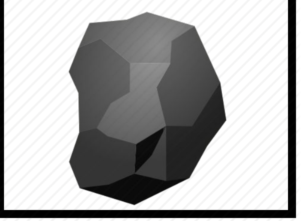
Our Exploitability = 1

Minimax Equilibrium

Minimax Equilibrium in 2p0sum: each player's strategy is optimal given the other player's policy

In balanced games, playing minimax ensures you will not lose on average

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

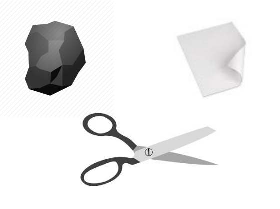
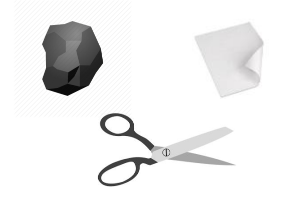
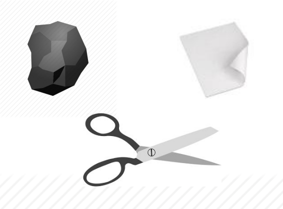
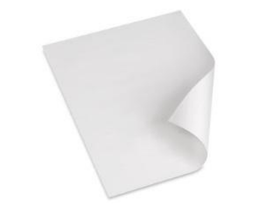
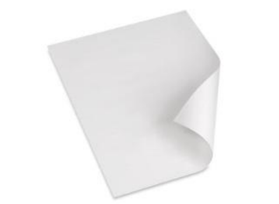
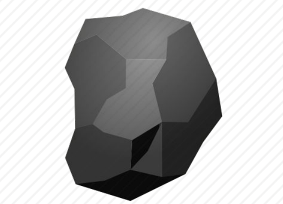
	Round 1	**Round 2**	**Round 3**
Us			
Best Response			

Our Exploitability = 0

Minimax Equilibrium

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

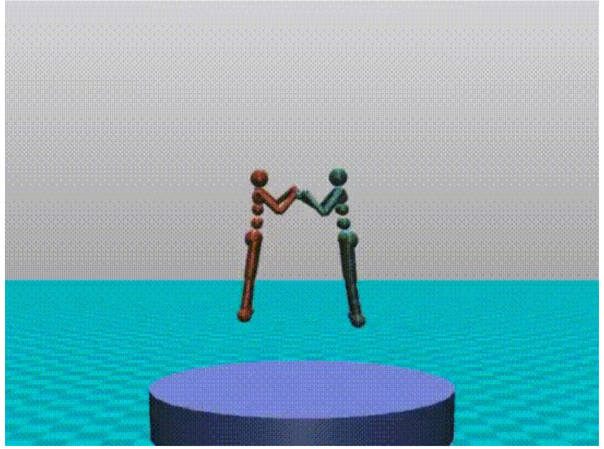
-Ryan Fee’s Poker Strategy Guide

	Round 1	Round 2	Round 3
Us			
Best Response			

Our Exploitability = 0

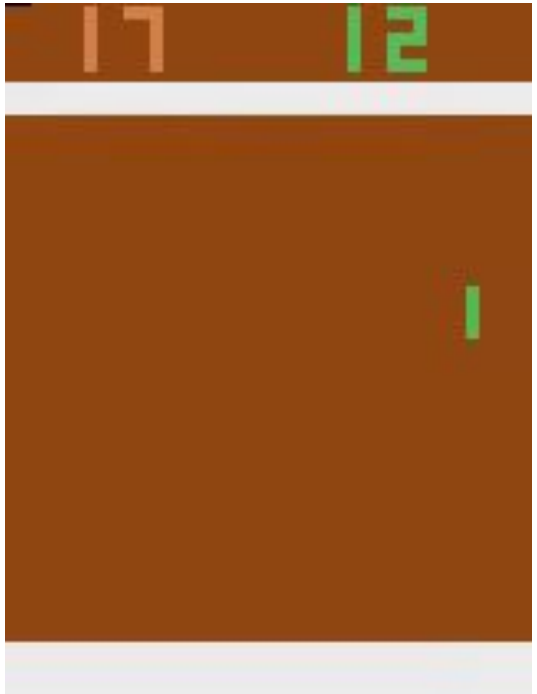
Self-play in two-player zero-sum games

- 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**

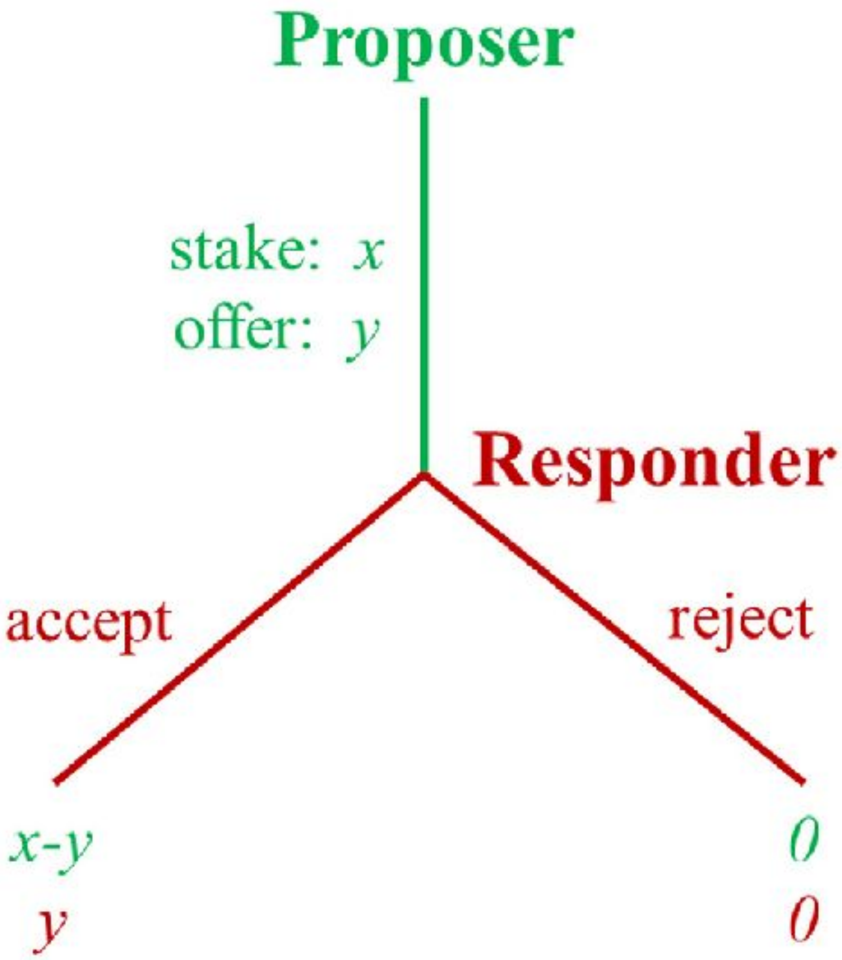
Question: Why is self play limited to two-player zero-sum games?

Answer: Because outside two-player zero-sum games, unlimited memory and compute isn’t enough. You may need human data as well!

Ultimatum Game

- Alice is given $100
- Alice must offer $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

THAT WAS AN INCREDIBLY COUNTERINTUITIVE RESULT TO NOBODY BUT ECONOMISTS.
THE HUMANS AREN'T DOING WHAT THE MATH SAYS. THE HUMANS MUST BE BROKEN.

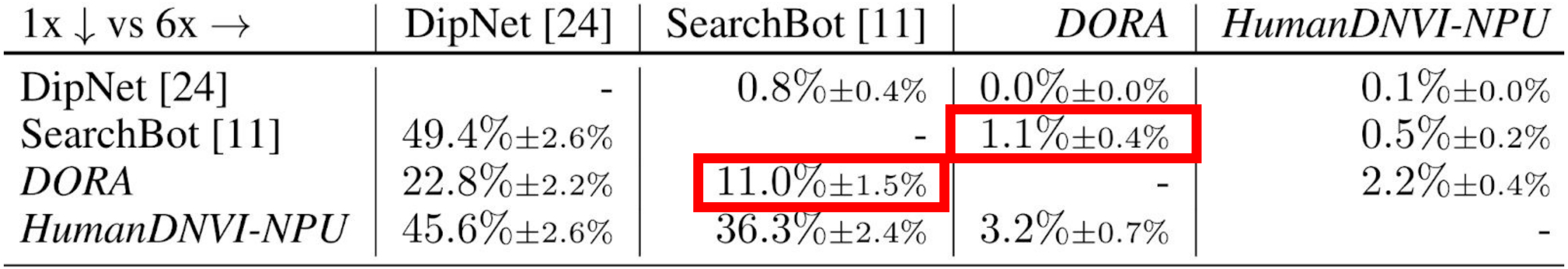


DORA: No-press Diplomacy from Scratch [1]

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

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

- Performance with bots in 7-player no-press Diplomacy:



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

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

piKL- Human-regularized RL and planning

(Jacob et al. 2022)

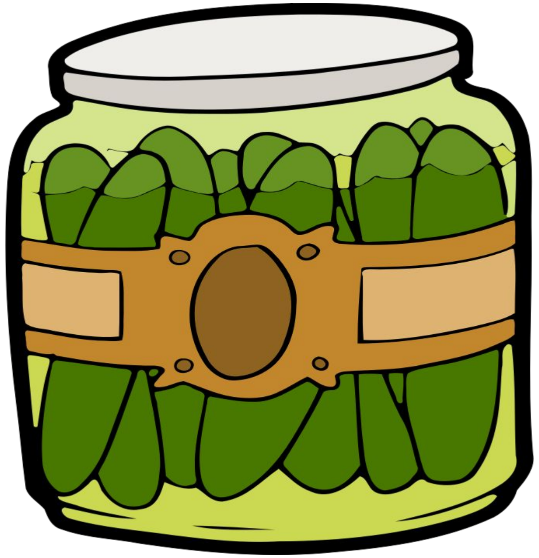
Idea: Given **anchor policy τ** from human imitation learning, when optimizing policy **π,** optimize the regularized utility:

$$u(\pi) = EV(\pi) - \lambda D_{KL}(\pi||\tau)$$

λ is the **anchor strength**:

- **λ** = 0: self-play from scratch
- **λ** = infinity: human behavioral cloning
- Choosing **λ** in-between gains benefits of both.

Results: Significant policy improvement while maintaining high human compatibility.



CICERO Plays with Humans

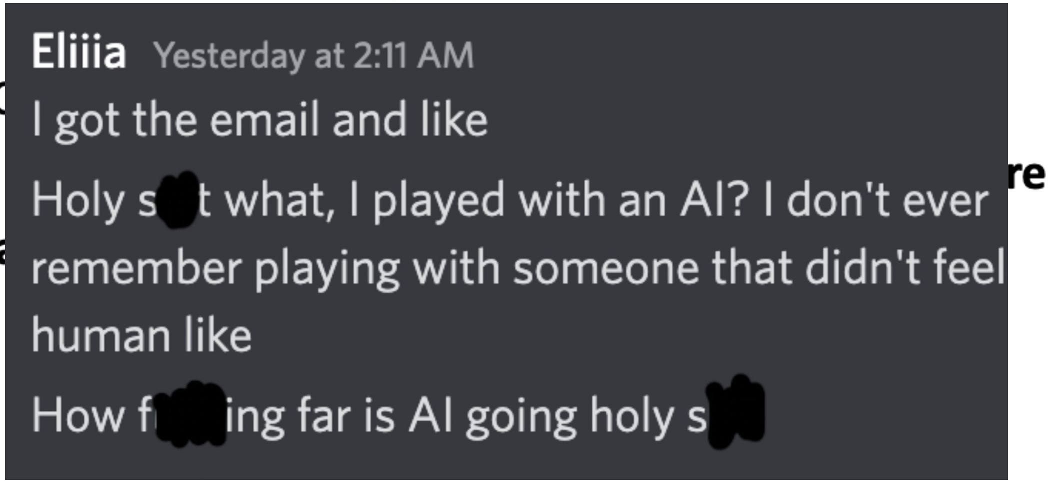
We entered CICERO anonymously in an **online Diplomacy league**

CICERO **was not detected as an AI agent** after **40 games** with 82 unique players *, sending and receiving an average of **292 messages per game**.

CI

of **re**

tha



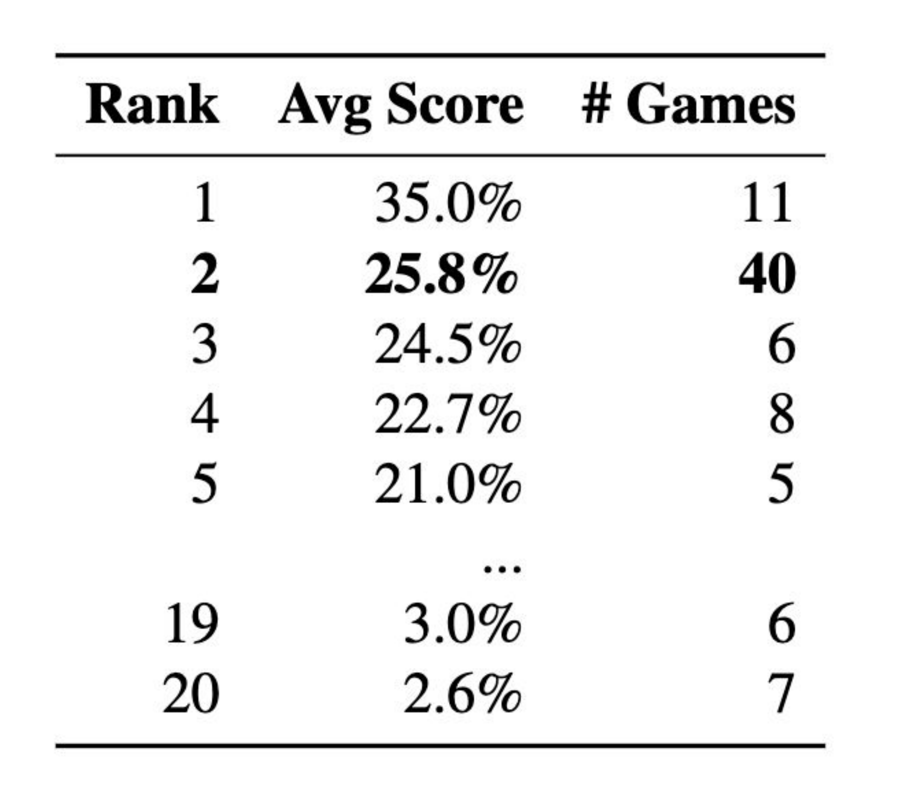
Eliiia Yesterday at 2:11 AM

I got the email and like

Holy s[redacted]t what, I played with an AI? I don't ever remember playing with someone that didn't feel human like

How f[redacted]ing far is AI going holy s[redacted]

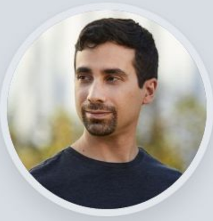
* One player mentioned in post-game Discord that they were suspicious that our account was a bot after a game, but didn't follow up about it



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

FAIR Diplomacy Team

Anton Bakhtin


Noam Brown

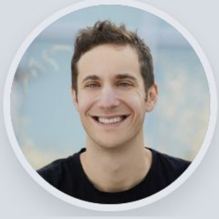
Emily Dinan

Colin Flaherty

Jonathan Gray

Hengyuan Hu

Athul Paul Jacob


Adam Lerer

Mike Lewis

Alexander Miller

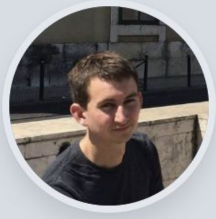
Adithya Renduchintala

Weiyan Shi

David Wu

Hugh Zhang

Gabriele Farina


Daniel Fried

Andrew Goff

Mojtaba Komeili

Minae Kwon

Karthik Konath

Sasha Mitts

Stephen Roller

Dirk Rowe

Joe Spisak

Alex Wei

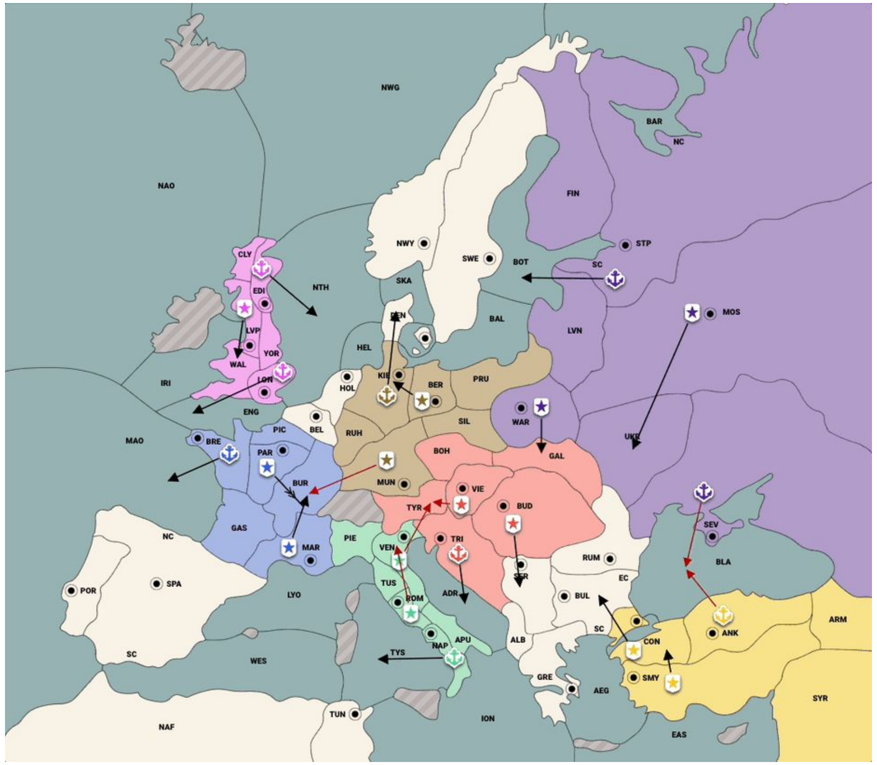
Markus Zijlstra

Recap

- Sound self play will compute a minimax equilibrium in any two-player zero-sum given sufficient memory and compute

- Outside two-player zero-sum games, self play isn't enough

- Self-play with KL regularization toward a human imitation policy (i.e., piKL) works well in general-sum games!



- See our papers for details:
  - Mastering the Game of No-Press Diplomacy via Human-Regularized Reinforcement Learning and Planning. Bakhtin et al. ICLR 2023.
  - Human-Level Performance in the Game of Diplomacy by Combining Language Models with Strategic Reasoning. FAIR et al. Science 2023.

- Code and models (along with those of our work in full-press): Diplomacy with dialogue) available at:

`https://github.com/facebookresearch/diplomacy_cicero`