上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Week 11 Extension: Gaming AI — A 30-Year Tour

Tao Huang

John Hopcroft Center, School of Computer Science, Shanghai Jiao Tong University

https://taohuang.info/cs3317

https://oc.sjtu.edu.cn/courses/89538

AI tools assisted in generating some figures in these slides. All such content has been reviewed, and the instructor is responsible for its accuracy.

Why Games? Why AI Loves Them.

Games are AI's fruit fly — clear reward, fast simulation, ascending difficulty.

Reward is unambiguous. Win/lose. Score. No labeling required.

Simulation is cheap. Millions of episodes in a day on commodity hardware.

Difficulty ladders are built in. Easy modes to superhuman play in the same environment.

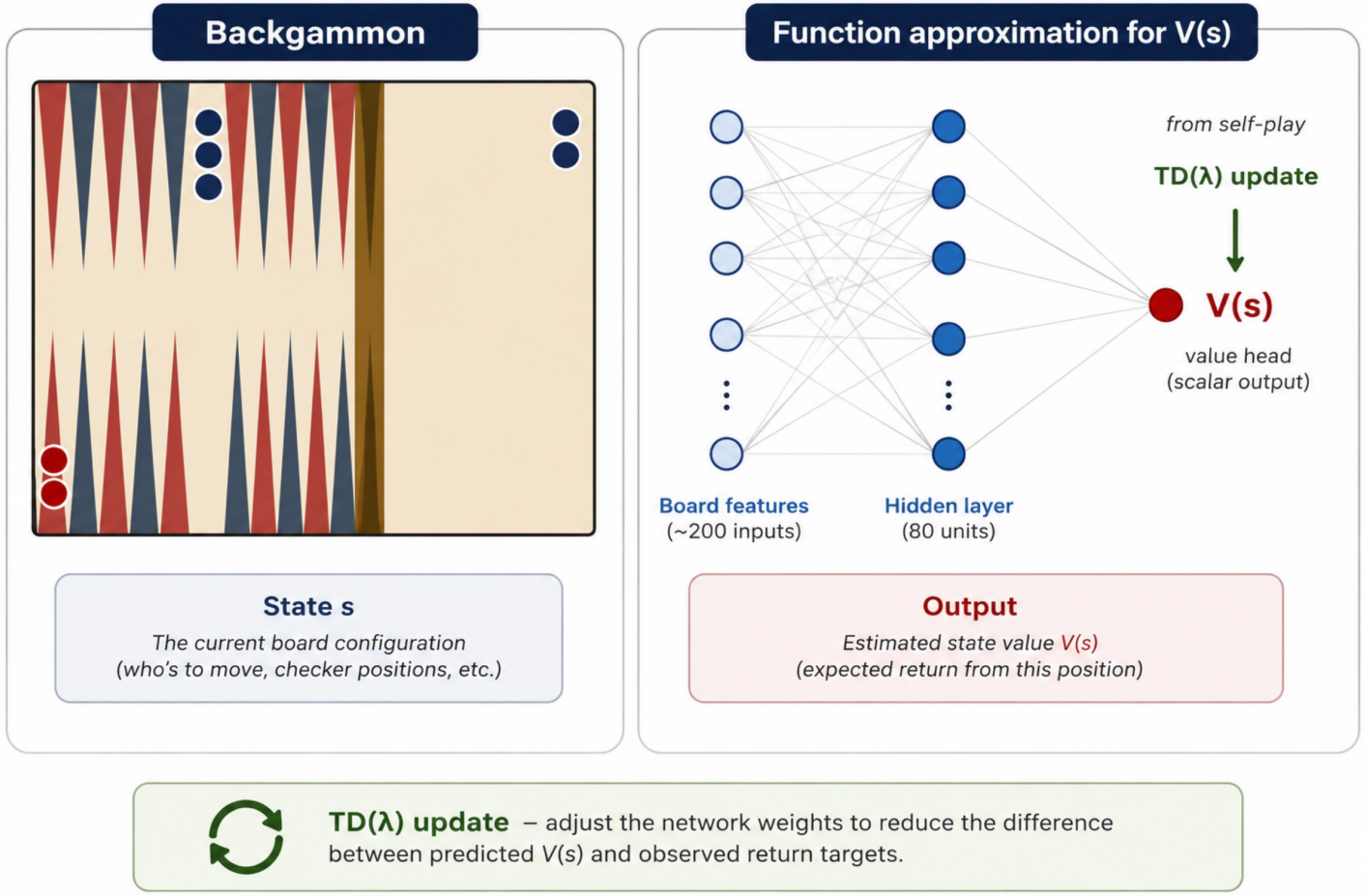
Progress is legible. 'Beats world champion at X' is a milestone the world can grade.

Every algorithm in Week 11 — TD, Q-learning, exploration — was invented for a game or proved itself on one. Today: the 30-year tour.

TD-Gammon: Where It Started

Tesauro. Temporal Difference Learning and TD-Gammon. CACM 1995

- **The model:** a small MLP, ~80 hidden units, trained by **self-play** with TD(λ).
- **The result:** beat the world's top human players in 1992. By accident, ***invented opening moves humans had missed for 5000 years.***
- **What it proved:** TD learning (L31) + a neural function approximator + self-play is enough. *No human game data*.



Atari & DQN: One Algorithm, 49 Games 🎮

Mnih et al. Human-level control through deep reinforcement learning. Nature 2015

The leap: input = raw 84×84 pixels. Output = Q-values for 18 buttons. ***No game-specific features.***

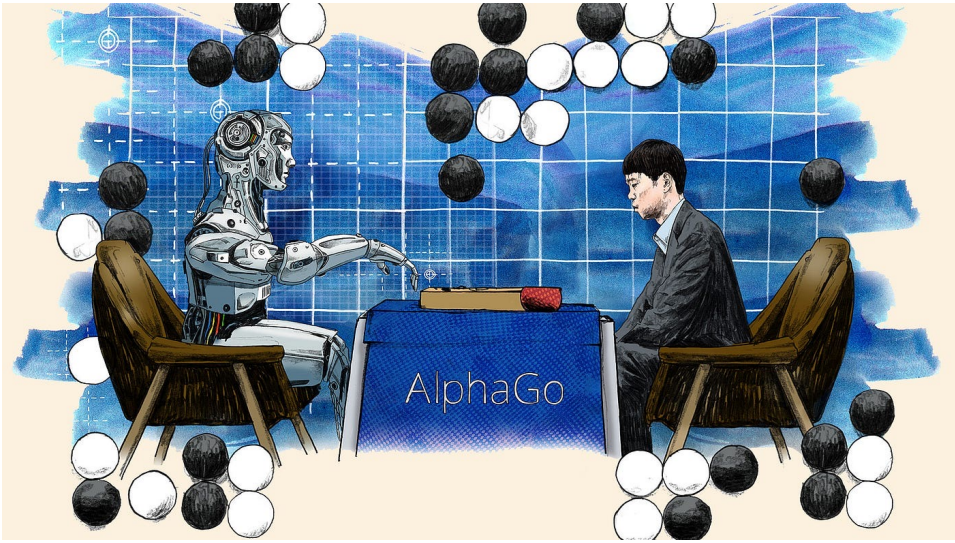
Two tricks made it stable: experience replay + a slowly-updated target network. *(Coming in L32.)*

The headline: one algorithm, one hyperparameter setting, ***human-level on 49 of 49 Atari games.***

AlphaGo🎮

Silver et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016

- **Architecture:** MCTS + policy net + value net. The networks guide search; ***search corrects the networks.***
- **Training pipeline:** supervised pretraining on human expert games → self-play RL → distill into the value net.
- **What Move 37 showed:** the model wasn't imitating humans. It had found Go moves ***no human had played in 2500 years.***



AlphaZero: Self-Play From Scratch

Silver et al. A general RL algorithm mastering chess, shogi, and Go through self-play. Science 2018

- **The simplification:** drop the supervised pretraining. **Start from random play.** Use only the rules of the game.
- **Same algorithm, three games:** chess, shogi, Go — all superhuman within 24 hours of TPU training.
- **What it removed:** human game data. *(Assumption removed: ✗ human demonstrations.)*

MuZero: Learning Without the Rules

Schrittwieser et al. Mastering Atari, Go, chess and shogi by planning with a learned model. Nature 2020

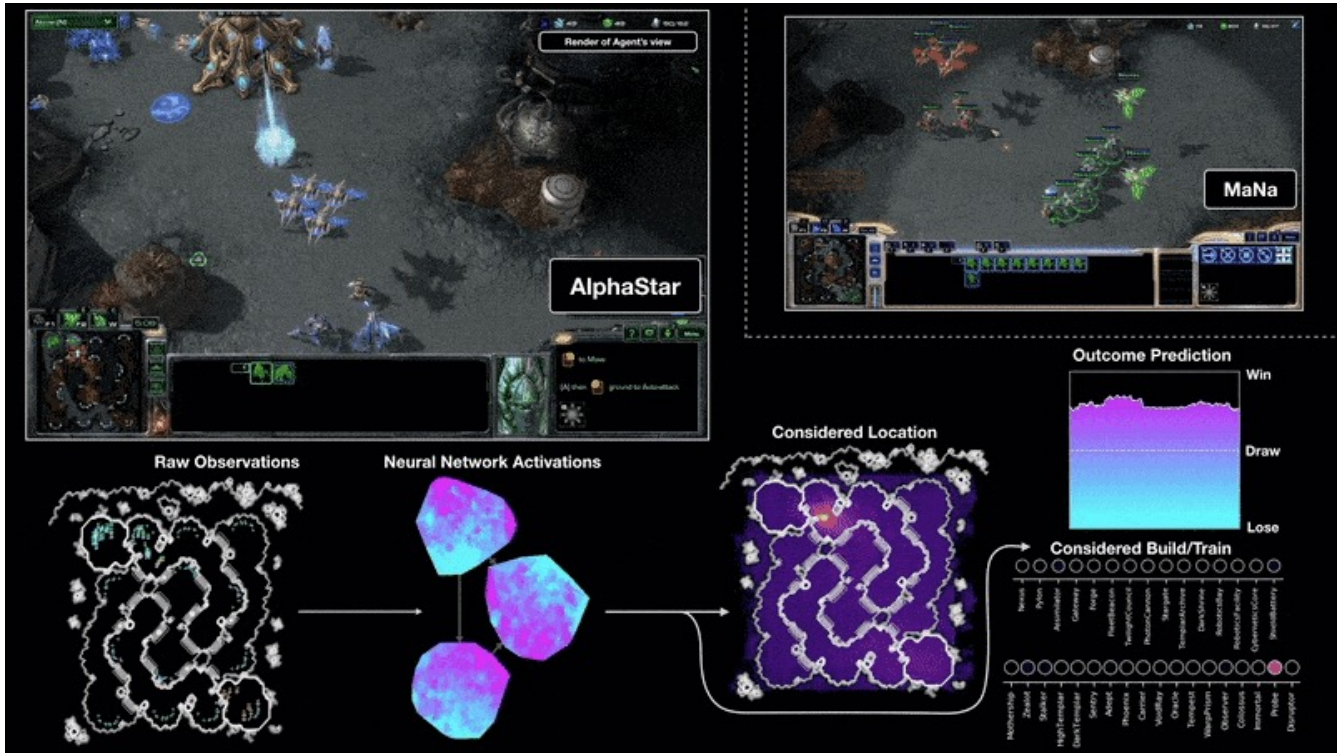
- **The simplification:** drop the rules. Learn a **latent world model** predicting only what planning needs — reward, value, policy. *Skip pixel prediction*.
- **Result:** matches AlphaZero on Go/chess/shogi ***and*** matches DQN on Atari (where rules aren't given). ***One algorithm*.**
- **What it removed:** the simulator. *(Assumption removed: ✗ known dynamics.)*

Algorithmic genealogy of Sora's world model and DreamerV3 robotics — same idea, different domains.

Real-Time Strategy: AlphaStar & OpenAI Five 🎮

Vinyals et al. AlphaStar. Nature 2019 · Berner et al. Dota 2 with Large Scale Deep RL. arXiv 2019

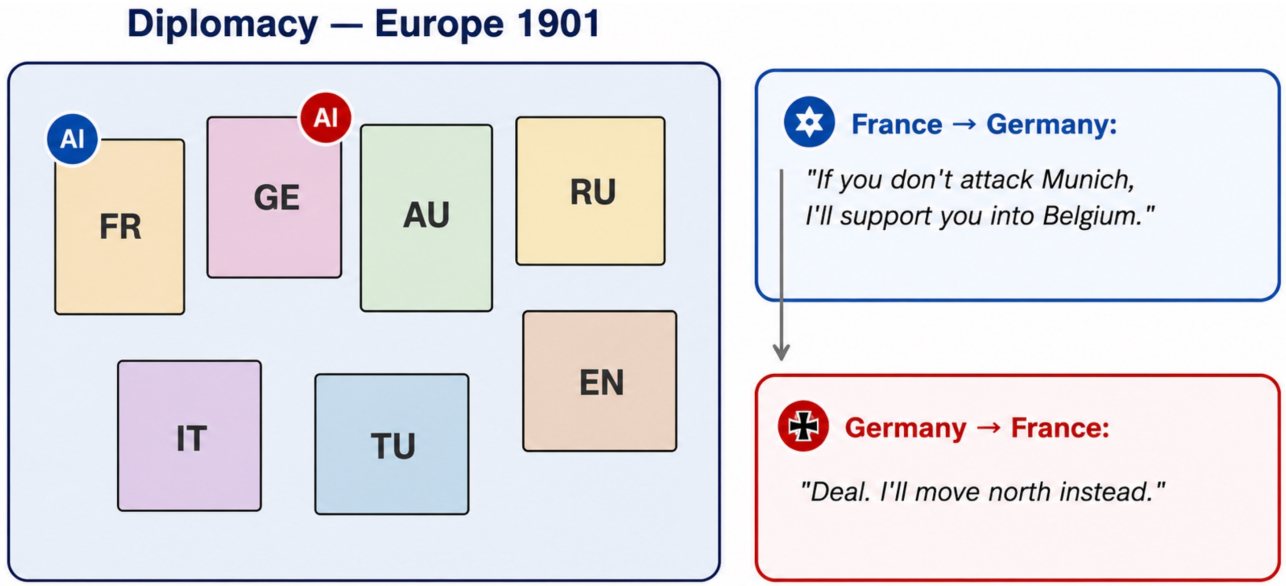
- **AlphaStar:** transformer + LSTM + **league play** (population of agents that exploit each other). Grandmaster at StarCraft II.
- **OpenAI Five:** **PPO at scale**. 180 years of self-play ***per day*** on 128k cores. Beat the Dota 2 world champions.
- **What both removed:** turn-taking, perfect information, single-agent assumption. *PPO comes in L32.*



CICERO: When AI Learned to Negotiate

Bakhtin et al. (Meta FAIR). Human-level Diplomacy by combining LMs with strategic reasoning. Science 2022

- **The game:** Diplomacy. Seven players, no dice, **negotiate alliances in natural language.** Lies are legal. Trust is the only currency.
- **The system:** RL policy for moves + LLM for dialogue + value head to score negotiations. ***The model wrote messages humans believed.***
- **What it added:** language. *(Assumption added: ✓ communication in natural language.)*

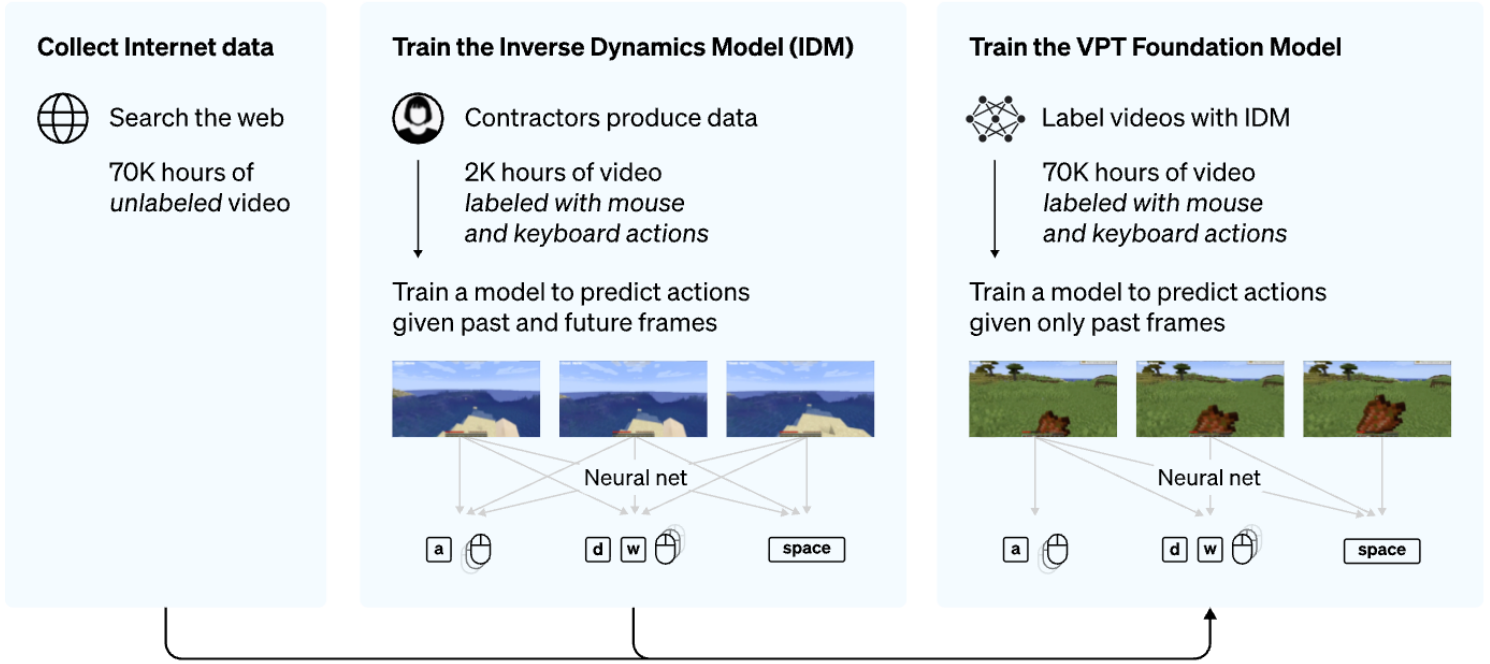


Bridge from pure-RL games to RL + language — what RLHF would be doing two years later.

Open-World Games: VPT & Voyager 🎮

Baker et al. VPT. NeurIPS 2022 · Wang et al. Voyager: An Open-Ended Embodied Agent with LLMs. arXiv 2023

The game: Minecraft. No win condition, no score, ***~infinite world.*** The hardest RL benchmark by 2022.
VPT's idea: train an inverse dynamics model on YouTube videos to recover actions, then *pretrain a policy on the internet's Minecraft footage*.
Voyager's idea: drop RL entirely. **GPT-4 as the policy** — writes code (skill library), debugs from environment, curriculum-itself.



Overview of the process of training the Inverse Dynamics Model (IDM) and the VPT Foundations Model

VPT method overview

VPT zero-shot results

Crafting of a crafting table “zero shot” (i.e. after pre-training only without additional fine-tuning)

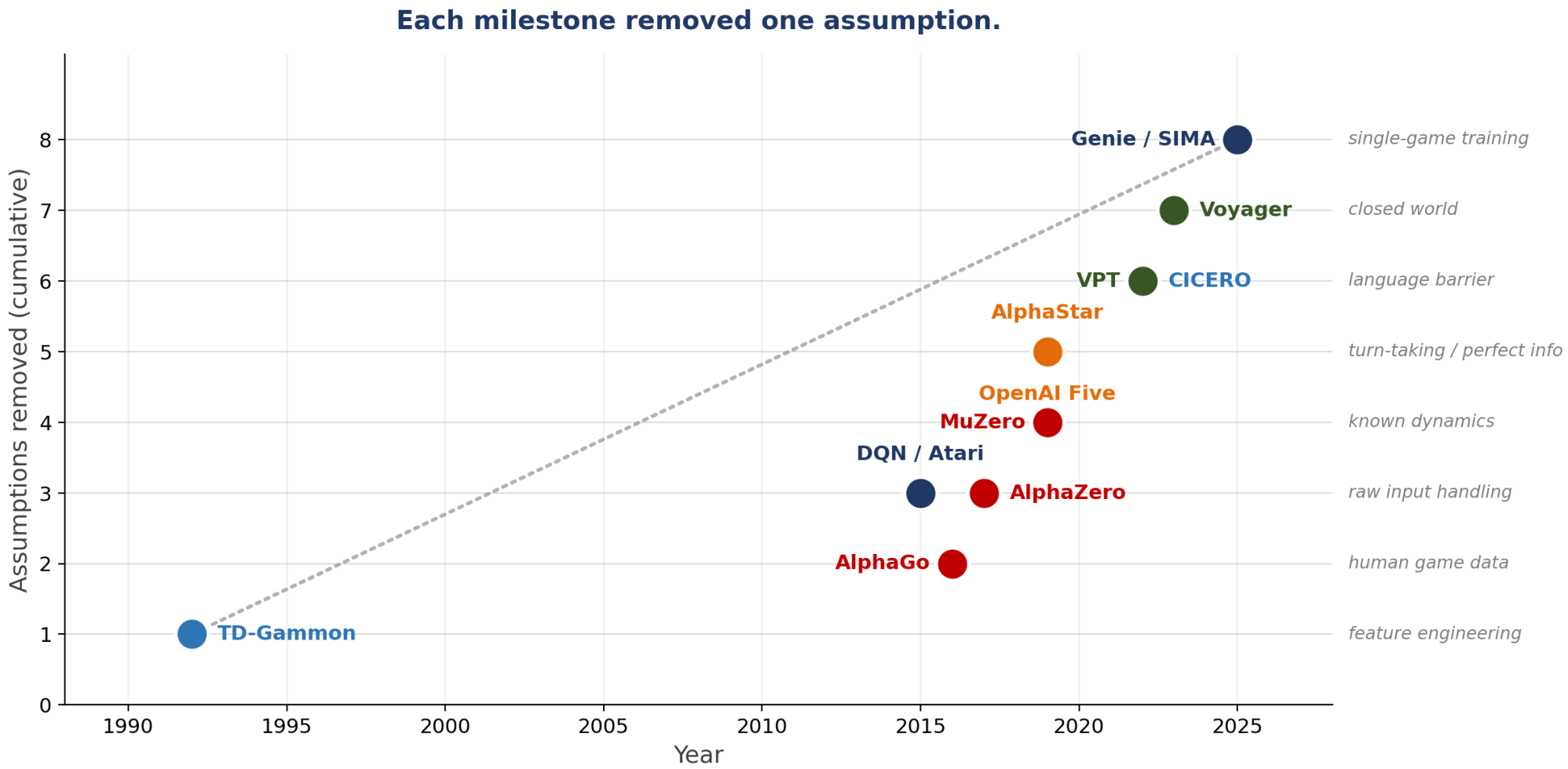
https://openai.com/index/vpt/

2024–26: Games As World Models 🎮

Bruce et al. Genie: Generative Interactive Environments. ICML 2024 · SIMA Team. DeepMind 2024

- **Genie 2 (2024):** a generative world model trained on gameplay video. Type a prompt → **get a playable 2D/3D world.** No physics engine — the model is the engine.
- **SIMA (2024):** one agent trained across 600+ commercial 3D games. Follows natural-language instructions. ***Reverse direction — one brain, many games.***
- **The reframe:** *games used to be where we trained AI.* ***Now AI is where we generate games.***

The Landscape: Assumptions Removed Over Time



The thesis of 30 years of gaming AI: progress = constraints removed. Algorithm family matters less than the constraint relaxed.

Where the Field Went Next

By 2026, gaming AI is splitting into three threads.

- **Generative worlds.**
  Genie, GameNGen, DIAMOND. The game ***is*** the model. Bridges to L42 (World Models) and embodied AI.
- **General agents.**
  SIMA, Adam, the o-series in games. One agent across thousands of environments.
- **Games as RL benchmarks for LLMs.**
  Atari for chain-of-thought, BabyAI for instruction following, Crafter for planning. ***Games are the new GSM8K.***

Contested in 2026: are games still the right benchmark when LLMs can play most of them out of the box?

References

Foundational

[1] Tesauro. *TD-Gammon.* CACM 1995
[2] Mnih et al. *Human-level control through deep RL.* Nature 2015

AlphaGo / AlphaZero / MuZero

[3] Silver et al. *AlphaGo.* Nature 2016
[4] Silver et al. *AlphaZero.* Science 2018
[5] Schrittwieser et al. *MuZero.* Nature 2020

Real-time strategy

[6] Vinyals et al. *AlphaStar (StarCraft II).* Nature 2019
[7] Berner et al. *OpenAI Five (Dota 2).* arXiv 2019

Language & open worlds

[8] Bakhtin et al. *CICERO (Diplomacy).* Science 2022
[9] Baker et al. *Video PreTraining (VPT).* NeurIPS 2022
[10] Wang et al. *Voyager.* arXiv 2023

2024–26 frontier

[11] Bruce et al. *Genie.* ICML 2024
[12] SIMA Team. *Scaling Instructable Agents Across Many Simulated Worlds.* DeepMind 2024