Foibles and Issues of (LLM-based) AI

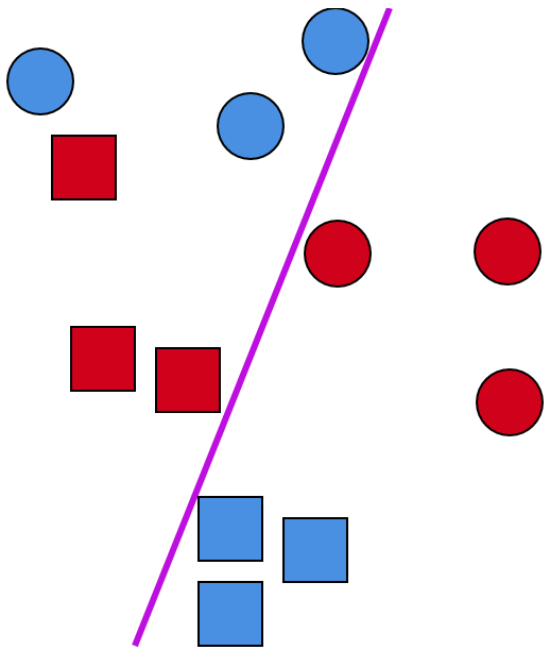
Data, AI, and Society Lectures 3-4 Mashup

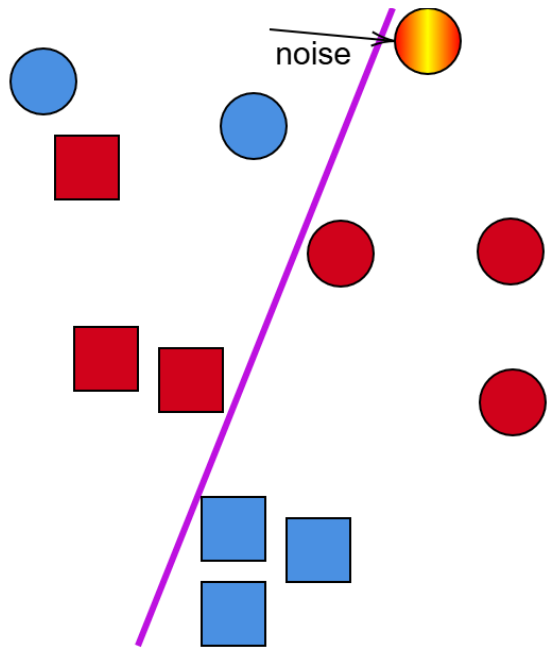
Kenneth W. Regan
University at Buffalo (SUNY)

CSE199 Freshman Seminar, Fall 2026

How AI Can Be Fragile

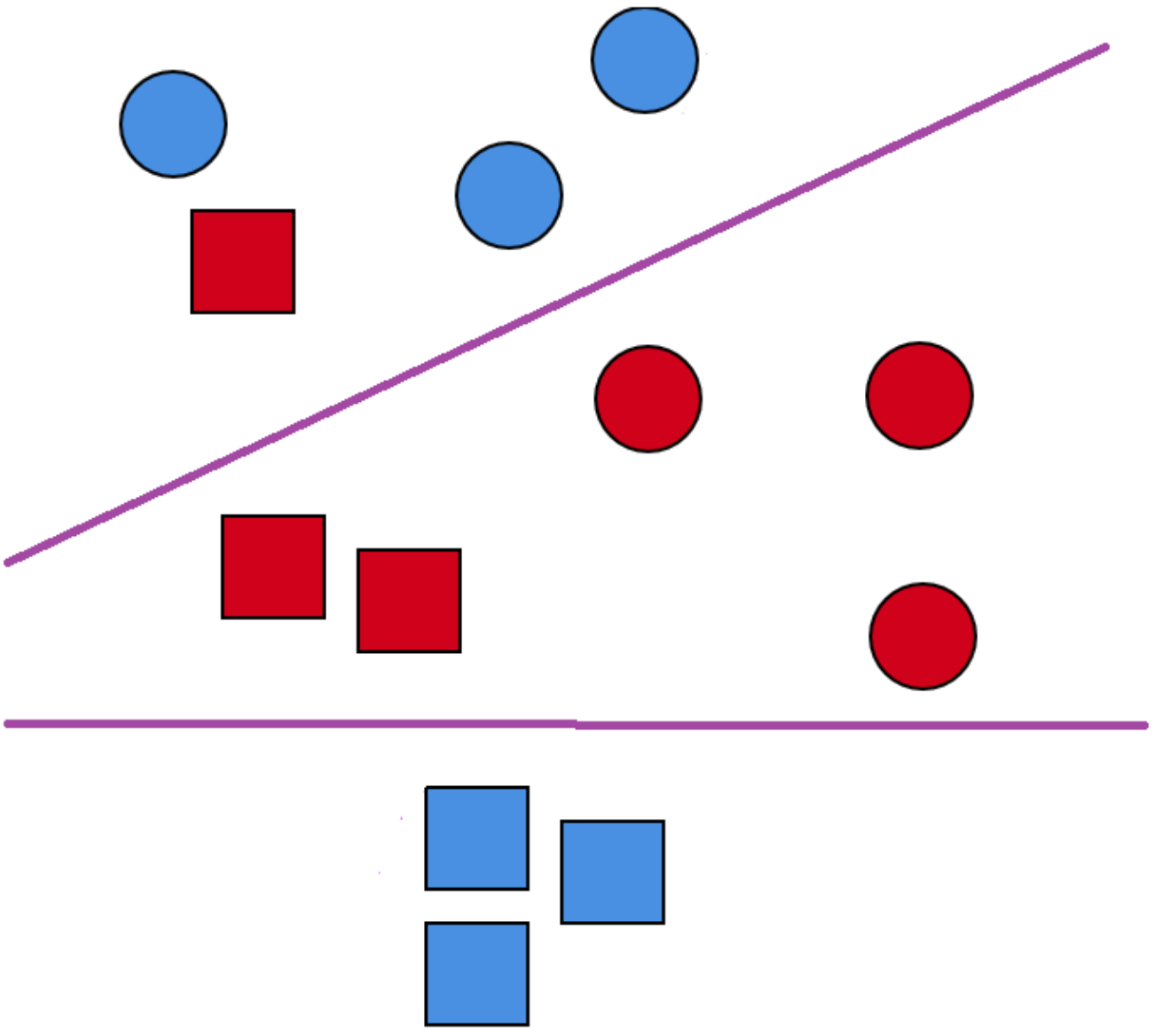
We gain *encoding advantage* by needing just one line to separate both the red shapes and the blue shapes.





But the line cuts things real close. A little noise could make a blue *Circle* be misread as a *Square*, as shown at right.

Making AI Solutions More Robust



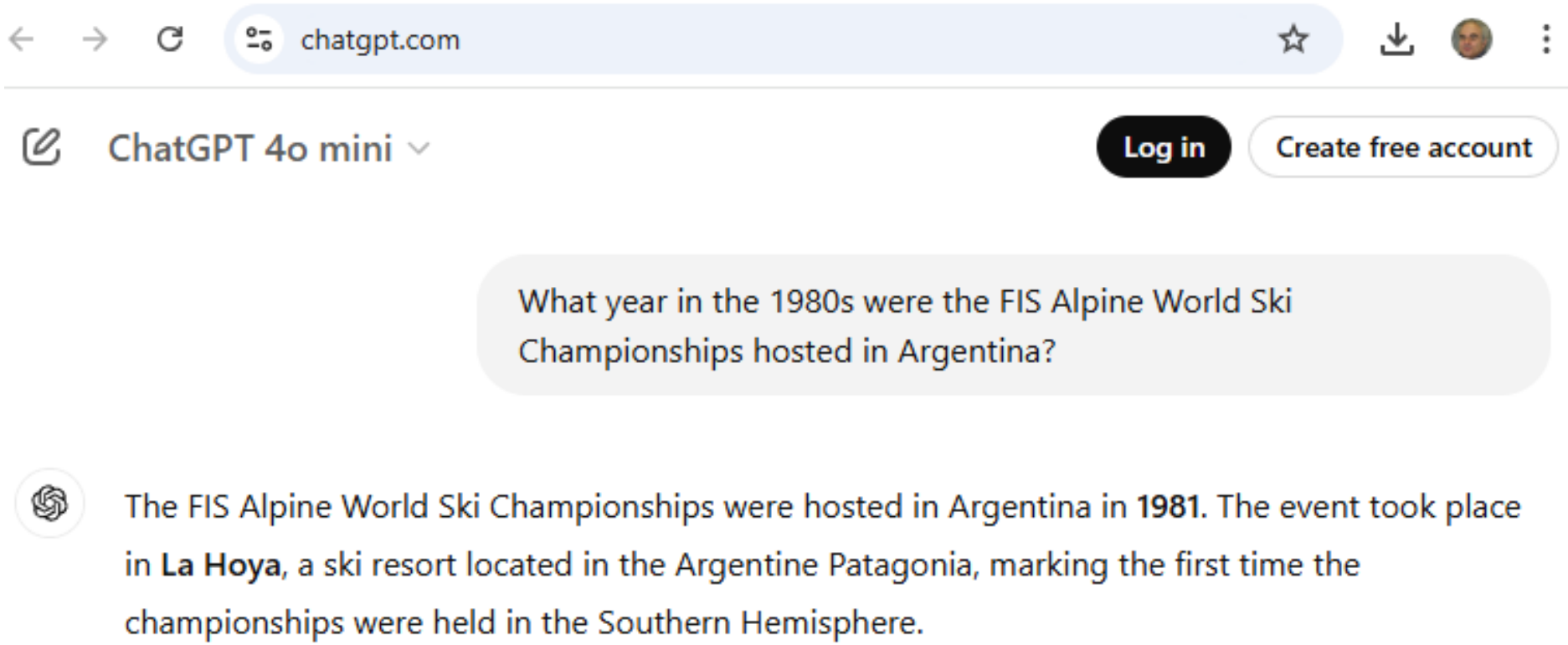
Can we guide the training so that the LLM will prefer solutions with more "wiggle room" even while using more weights to do them?

This is an example of the utility-function end of the Alignment Problem.

The issue is that it's hard to tell in advance how the LLM training will orgainze itself. LLMs may become better at mounting these mods themselves: **RSI** for **Recursive Self-Improvement.**

How AI Can Be "Truthy"

I reproduced an example in a Medium article on AI hallucinations.



In the article, the response was, "The FIS Alpine World Ski Championships were hosted in Argentina in 1985."

It Got Worse (Fact Checks: FIS, Wenzels)

The FIS Alpine World Ski Championships were hosted in Argentina in **1981**. The event took place in **La Hoya**, a ski resort located in the Argentine Patagonia, marking the first time the championships were held in the Southern Hemisphere.

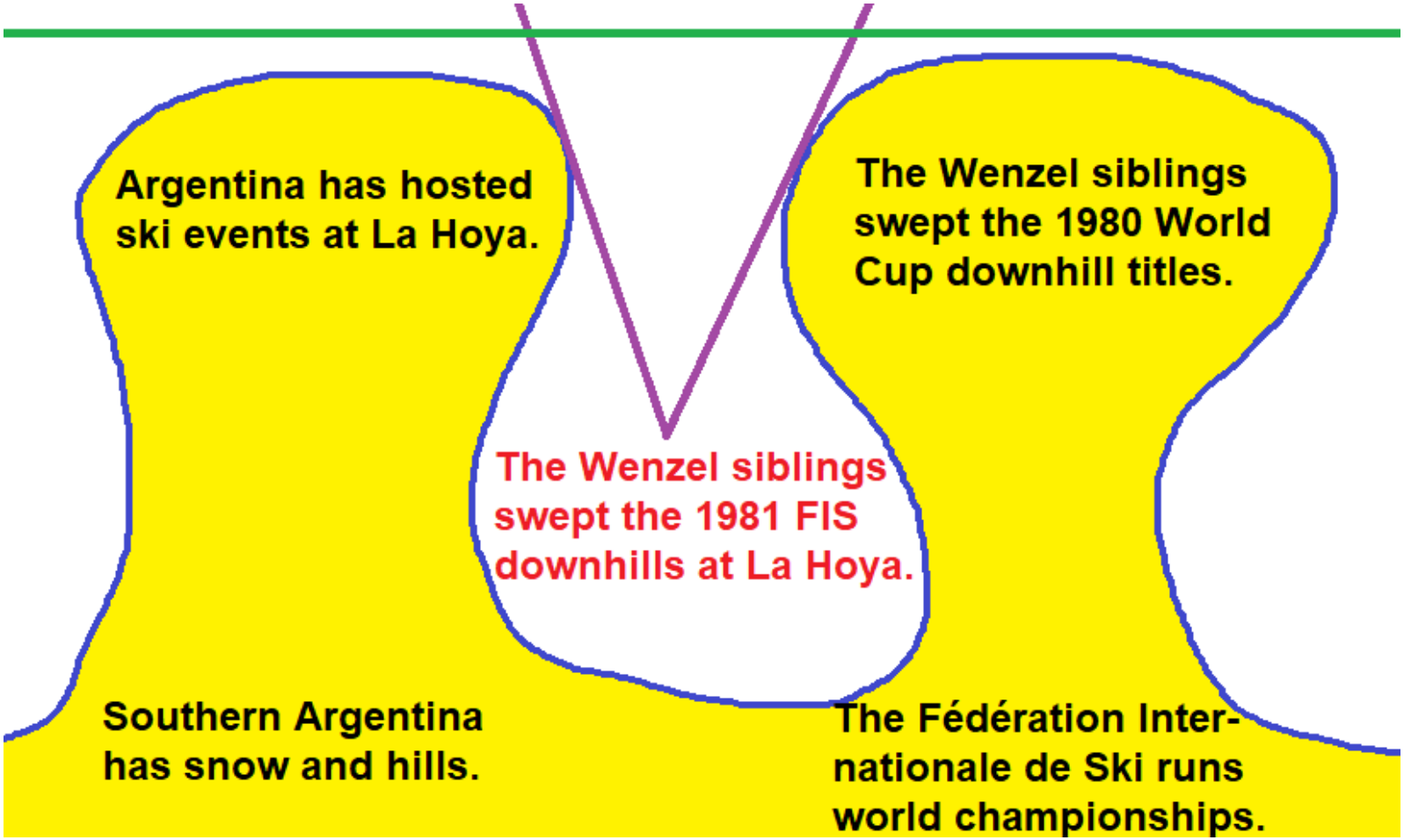
Who won the women's and men's downhill races at those championships?

At the 1981 FIS Alpine World Ski Championships in Argentina, the winners of the women's and men's downhill races were:

- **Women's Downhill**: **Hanni Wenzel** (Liechtenstein) won the gold medal in the women's downhill.
- **Men's Downhill**: **Andreas Wenzel** (Liechtenstein) won the gold medal in the men's downhill.

The Wenzel siblings made history with their victories in both events.

Rough Diagram of What Could Have Happened



The red is a *positive linear combination* of the four true items in black. When combos sit in "truth pockets," linear separators (in purple and green) cannot penetrate far enough to rope them off.

ChatGPT 5.5 Fixed This Last December

Prompt: In which year of the 1980s were the FIS Alpine World Ski Championships hosted in Argentina?

Followup: Who won the women's and men's downhill races at those championships?

Today's free models were on the frontier less than a year ago.

Lower-level embedded models can still err: Yahoo error, fixed by Google. Yahoo bonzoness.

More Improvements—Old and New

"Jesus flipping over the tables in the Temple." From the movie *Jesus Christ Superstar*—then try it on Cutout or NightCafe or Simplified:



Two Results From 2023—one famous, one mine



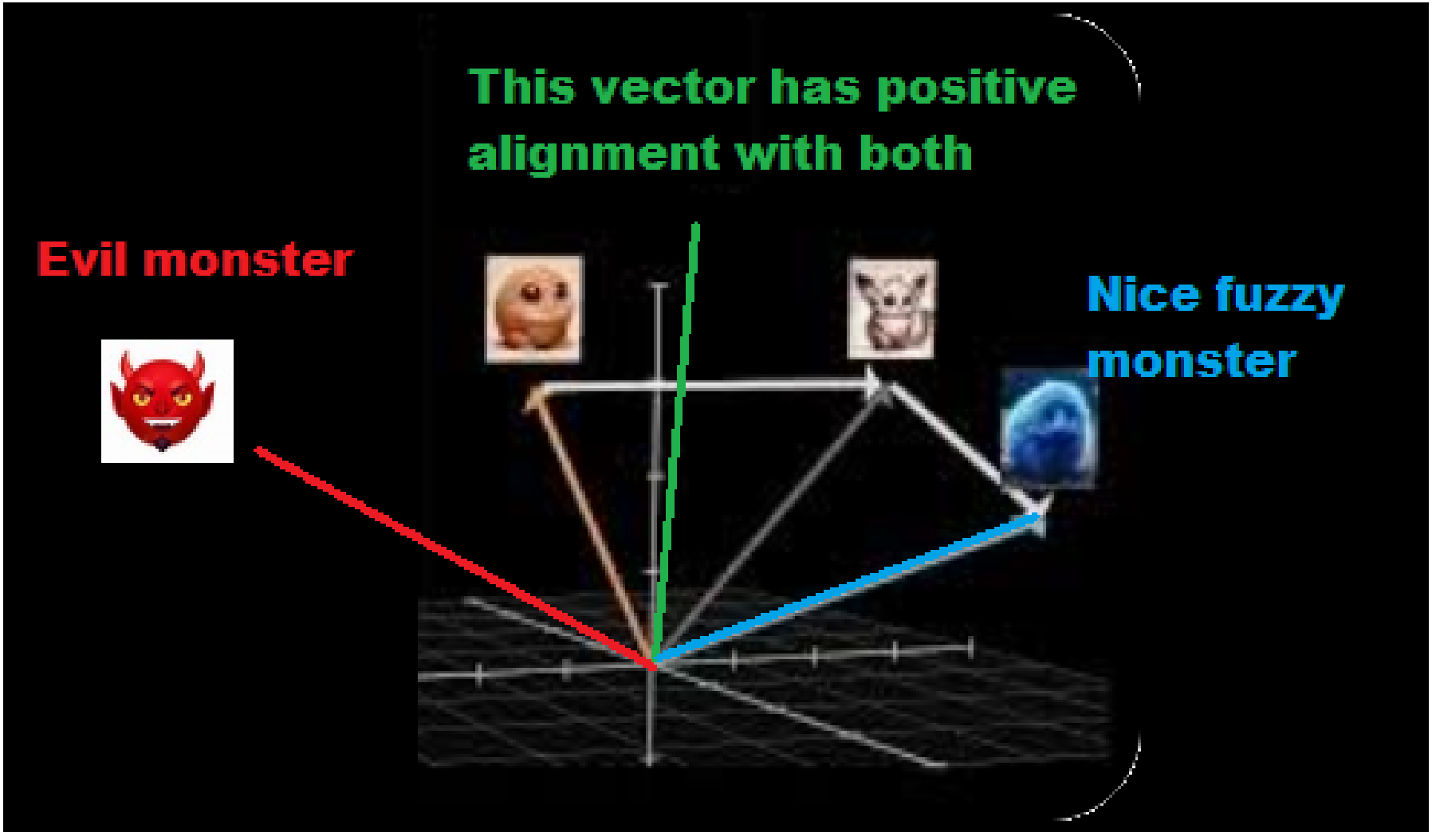


Now Well Understood (GPT6-Astra)



Maybe this one is “overdone.” Possibly misunderstandings are being pushed “deeper under the hood” where it is harder for us to tell them.

Multiple Alignments (mod of 3Blue1Brown source)



Deeper AI Issues and Dangers

- Show penultimate slide from last year.
- Explains why major labs are devouring old books.
- But this year we are **not** "bumping up against" data limits anymore.
- Go into assignment and https://alkjash.github.io/ai-risk//

Logical Impasses Pose Obstacles to Solutions

- The famous example "All Cretans are liars"—said by a Cretan poet—is resolvable: the poet is lying and some Cretans are truthful.
- Others don't have a way out: "This statement is a lie."
- One can formulate within the mathematical proof system **L∃∀N** a statement S whose semantics is, "S has no proof in L∃∀N."

- If S is *false*, then L∃∀N is busted (it proves something false).
- So S is *true*, but then L∃∀N is **incomplete** (it can't prove S).

- "Grants are so competitive now, it feels one must already have succeeded at the proposed research to have a chance."
- AI safety impasses are like all these examples.
- They are **abstract**—they do not equate to the **concrete** danger scenarios asked for on your written homework.
- They may help critique solution ideas for next week's presentations.

Impasse #1: No Silver Bullet Against Bad Behaviors

- Suppose we draw any list M of specific bad behaviors to prevent.
- Suppose we have a program G that examines any program P and always tells whether P will do a behavior on M.
- Building G is a good idea, right? Well, G **cannot exist!**:
- Given G, we **can** build a program Q that does the following:

  - Run G on Q—possible by self-recursion.
  - If G says Q is safe, then execute a bad behavior on M.
  - Else if G says Q is malicious, do nothing—stay safe.

- Thus G is wrong about Q.
- So there **does not exist** a **foolproof** guardian program G.
- Impasse holds whatever list M we try to draw up.
- This is on top of how *open systems* are more powerful and versatile.

Impasse #2: Cannot Both Predict and Control

- Suppose we try to control an AI by giving it a series of commands.
- It is easy to (inadvertently) make commands that are inconsistent.
- The effect would be similar to simply giving this as a command:

Disobey this command!

- We cannot predict what the AI will do in response to this.
- So this says we cannot predict what an AI will do in all situations, if we give it general facility to react to control commands.
- Commands like that *could* come from a malicious actor.
- Reference: Roman Yampolskiy, Uncontrollability of AI.
- See also: Brcic and Yampolskiy, Impossibility Results in AI: A Survey. (Yampolskiy got his PhD at UB in 2008.)

Impasse #3: Can AI Help Us Control AI?

My chess-cheating model employs a "superhuman AI" in the form of Stockfish or other supreme chess-playing **engine** E.

- I use E to give an answer key for testing *human* players.
- This enables building a model for **predicting** *human players*.
- E itself is *tame*—e.g., it won't hack a chess server.

But what we need is to predict *AI behaviors*, not (just) human behaviors.

We can enlist our top AIs A in such a task, to help solve safety for other AIs B. We will probably need to do so. But:

> This **presumes already having solved** the safety problem for A.

How do we get off from "starting at square 1"? (Note also slide 12 here.)

Impasse #4: The Math of Human Behaviors

- This harks back to last week's "Prisoner's Dilemma" activity.
- The only *mathematically* **stable** solution is "both defect."
- Liv Boeree, video: The Dark Side of Competition in AI.
- Liv Boeree, long video, Poker, Game Theory, AI, Simulation, Aliens, and Existential Risk.
- Condorcet's Paradox also enters in:
- Preferences among 3 or more options **can be circular**.
- Arrow's Impossibility Theorem: no *equitable* way to fix this.
- There are other issues with *determining collectively optimal utility functions* that havelong been known in *social choice theory*, before the advent of AI—and before discussion of crafting utility functions that are *safe* and *stable*.

Impasse #5: (Un-)Explainability

- When Garry Kasparov lost to Deep Blue in 1997, he demanded explanation of a subtle move the computer made.
- The IBM team refused—but they could have fully complied by showing the saved trace of Deep Blue's **value-based** search.
- Modern chess programs use neural weights. No explanation comes from this.
- Explainability is not automatic for self-driving cars.
- Chain-of-Thought in LLMs is not innate, so not comprehensive.
- Explanation is needed for human understanding...and human understanding is needed for human control.
- Further relevant: LLMs are grown, not crafted.

Issue #6: Removing Versus Introducing Bias

Yampolskiy makes a particularly adept point here under the heading, "Why AI Alignment May Be a Contradiction in Terms":

> Much of computer science [...] is devoted to removing bias from algorithms so they make fair, accurate decisions. Alignment, by contrast, asks engineers to intentionally introduce a "pro-human bias." [...] Any biased decision is by definition not the best decision you can make.

- Elon Musk has several times tried to give Grok a pro-conservative bias. Note outcome in results.
- Last year's week-3 project was about telling how biases could affect a simple implementation of the sentiment-analysis idea in this week's activity, and ideas to remove them.
- [Segue to describing the activity and other data science material as time allows.]

Sentiment Analysis

- Hugely successful in consumer product research, see this and this.
- E.g. paper, "Vehicle defect discovery from social media."
- Often simply tells whether a page exudes happiness/contentment or sadness/anger.
- Can we use it to predict elections? Brexit 2016, Canada 2015, USA 2016 (paper), USA 2016 (BrandsEye).
- Simplistic idea: if the electorate is (un-)happy that's (bad) good news for the incumbent.
- "Joy" was an express term of the Harris-Walz campaign. Did it show?
- Note various forms here, including using AI to *do* sentiment analysis.