TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2024

AI Safety

The Alignment Problem

Giving an artificial general intelligence (AGI) a mission or purpose in alignment with human values.

This can be phrased as finding a solution to the principal-agent problem for AGI agents.

White-Hats, Black-Hats

A white-hat team designs a safety system or protocol.

A black-hat team looks for vulnerabilities.

We need both.

White Hat: The Advobot Protocol

A personal advobot is an AI advocate for a particular person X where the advobot's fundamental goal is given as "within the law, pursue fulfilling the expressed requests of X".

The advobot protocol is that AGI systems be legally limited to advobots.

The term "AGI" needs to be incorporated into law and given an evolving interpretation by the judicial system.

White Hat: Safety Features of the Protocol

- The advobot must act within the law. Society can limit all advobots by changing the law.
- The advobot mission transfers moral responsibility from the advobot to its master.
- There is a large society of advobots — one per person — each with a different mission. This limits individual power.
- The advobot mission seems clearer that other directives such as Asimov's laws or Yudkowsky's coherent extrapolated volition.
- The advobot protocol preserves human free will.

Black Hat: Consider Large Language Models (LLMs)

Much of the literature on AI safety assumes that we can give an AI a goal such as "make as many paperclips as possible".

But large language models (LLMs) are not even "agentive" (explained below).

LLMs are trained to mimic people. People do not have clear objectives and do not always do what they are told.

Large language models are subject to the "Waluigi effect" where they flip to pursuing the very opposite of what they are told.

Agentive AGI

An AGI system is "agentive" if it takes actions in puruit of a goal.

Many systems can be decribed as taking actions in persuit of a goal. But an AGI is agentive if its potential actions include all the kinds of actions that people can take. For example legal filings of all kinds.

Current LLMs are not agentive.

The Waluigi Effect

Waluigi is the evil twin of Luigi in Mario Brothers.

The Waluigi effect occurs when an LLM holds two interpretations of its own statements — one genuinely cooperative and one deceptively cooperative.

When modeling humans both interpretations exist.

If the LLM reveals deception, the deception interpretation sticks.

Every turn of the dialogue has a chance of revealing deception.

White Hat: Constitutional AI

Constitutional AI is an attempt to provide a mission statement (or "constitution") to LLMs.

Constitutional AI has been show to work to some extent but is not included in GPT4 which instead uses reinforcement learning with human feedback (RLHF).

Ultimately it seems clear that we need to be able to specify missions.

Constitutional AI: Harmlessness from AI Feedback

Bai et al ArXiv 2212.08073 [Anthropic]

A Prediction: Memory Architectures

For both safety and performance reasons I believe strong AGI systems will be based on read-write memory architectures.

In a memory architecture a "CPU" works with an external memory in a manner analogous to a von Neumann machine.

We might have a **transformer CPU** where the transformer context is analogous to registers in a classical CPU.

Items can be loaded from memory into the CPU context and written from the CPU context into memory.

Read-Write Memory vs. Retrieval

There has been a fair amount of recent interest in retrieval architectures. For Example:

- Lample et al., Large Memory Layers with Product Keys Dec. 2019
- Lewis et al., Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (RAG), April 2021
- Improving Language Models by Retrieving from Trillions of Tokens (RETRO), December 2021
- Wu et al., Memorizing Transformers, March 2022
- Wang et al. Shall We Pretrain Autoregressive Language Models with Retrieval? A Comprehensive Study, April 2023

But I have not seen proposals of architectures that read and write to a random access memory.

Performance Advantages of Memory Architectures

The memory acts as an essentially infinite context with memory retrieval playing the role of the attention mechanism of a transformer but over all of memory.

The memory can be directly extended. The machine can read and remember today's newspaper.

The machine can use internal chain-of-thought processing involving reads and writes to memory.

Safety Advantages of Memory Architectures

We want to know what an agent believes.

We want to know the agents goals.

We want both of these things to be visible in the memory.

Interpretability (Opening the Black Box)

We should be able to engineer the memory such that memory entries are either literally textual statements, or have a rendering as text, and where the textual representation is faithful to meaning assigned by the machine.[1]

By observing the bandwidth to memory we can observe the "thought process" of the machine.

We can also edit the memory to maintain the quality of its information, or control the beliefs of the machine.

[1]For example, the machine's notion of entailment between memories is in correspondence with human entailment judgements between their textual representations.

Mission Statements (Fundamental Goals)

Orthogonality: Fundamental goals are axioms. They do not follow from, and are independent of, world knowledge.

An axiomatic **and immutable** mission should be built into the CPU.

The Advobot Protocol

A personal advobot is an advocate for a particular person X whose fundamental goal is given as "within the law, pursue fulfilling the expressed requests of X".

The advobot protocol is that AGI be limited to advobots.

Controlability

An advobot is controlable in three ways.

1. One specifies the advobot's fundamental goal.

2. One can give requests to the advobot — its fundamental goal is to purue obeying them.

3. One can directly edit the beliefs of the advobot. However, one might want legal protection against creating fake beliefs or legal guarantees that advobots use their own judgement in determining truth.(very challenging).

Defining AGI

Legally limiting AGI to advobots requires some legal interpretation of "AGI".

AGI is of course hard to define.

However, many legal terms are hard to define. Consider "intent", "bodily harm", or "assault".

Perhaps we can simply use the term "AGI" in legal discourse and leave its interpretation open to an evolving legal process.

Defining Truth

While it may be possible to edit the beliefs of an advobot, one might want legal protection for truth in advobot beliefs.

This would involve the ability to legally interpret "truth".

But the legal system has always had to judge truth.

Additional Safety Policies

No AI agent should have legal recognition as an person — an advobot cannot own property or enter into contracts.

Person X should have access to the thoughts of their advobot.

The goal structure of an advobot should be such that an advobot has no problem with its human master reading its thoughts.

A Missing Secret Sauce?

To date the results with retrieval models have been limited.

However, the virtues of belief state machines seem clear.

Further research in the development of belief state machines seems warranted.

Semantic Parsing and Class/Instance Architectures

Large language models can already do semantic parsing.

Semantic parsing represents a sentence as a set of class instances.

Each class instance is a belief (or at least a statement).

Language Model Generated Semantic Parses

INPUT: Russian lawmakers Monday ratified the illegal annexation of four partially occupied Ukrainian regions.

Output: Retrieving lexical entries for (ratifying), (annexing), and (occupying).

(ratifying) is when (the ratifier) ratifies (the treaty). After (the retifying), (the treaty) is in effect.

(the ratifier) is Russian lawmakers (the treaty) is the annexation of four partially occupied Ukrainian regions (the other party) is unknown

Language Model Generated Semantic Parses

INPUT: Russian lawmakers Monday ratified the illegal annexation of four partially occupied Ukrainian regions.

Output: Retrieving lexical entries for (ratifying), (annexing), and (occupying).

(annexing) is when (the annexer) annexes (the territory); After (the occupying), (the territory) is under the control of (the annexer).

(the annexer) is Russian lawmakers (the territory) is four partially occupied Ukrainian regions

Language Model Generated Semantic Parses

INPUT: Russian lawmakers Monday ratified the illegal annexation of four partially occupied Ukrainian regions.

Output: Retrieving lexical entries for (ratifying), (annexing), and (occupying).

(occupying) is when (the occupier) occupies (the territory); after (the occupying), (the territory) is under the control of (the occupier).

(the occupier) is Russian lawmakers (the territory) is four partially occupied Ukrainian regions

Summary

Belief state machine advobots are desirable for interpretability, controlability, and AI safety generally.

A potential architecture for belief state machines is a retrieval-based class/instance transformer — a retrieval transformer that retrieves from a vector database of class definitions (semantic memory) and vector database of class instances (episodic memory).

END