TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2024

AI Safety

A New Textbook on AI safety

AI Safety, Ethics & Society, Dan Hendrycks, (May, 2024)

Dan Hendrycks is the director of the Center for AI safety and is the safety advisor to xAI (Musk's AI company).

Hendrycks' Catastrophic Risks I and II

I. Misuse: AI systems could be used for malicious purposes such as terrorism, manipulation and disinformation, or entrenching a totalitarian state.

II. AI Race: Competitive pressures may lead militaries and corporations to hand over excessive power to AI systems. This could result in increased risks of large-scale wars, mass unemployment, and eventual loss of human control of economies and military systems.

Hendrycks' Catastrophic Risks III and IV

III. Organizational Risks: Accidents are hard to avoid when dealing with complex systems such as AI. Without building a culture of safety, it is likely that there will be accidents in AI development and deployment. Some of these could prove catastrophic.

IV. Rogue AIs: We already face issues in controlling the goals of current-day AI systems. If this is also true with future AI systems that are more powerful and more integrated with our economies and militaries, we could see dangerous rogue AI systems emerge.

Older Literature on AI Safety

AI safety has been approached in a largely non-technical way in popular books such as

- Nick Bostrom's "Superintelligence", 2014
- Pedro Domingos' "the Master Algorithm", 2015
- Stuart Russell's "Human Compatible", 2019

Fundamental Goals

The older literature emphasizes “fundamental goals” of AI systems.

This is related to “orthogonality” — the principle that physics (scientific truth) does not determine goals. Goals are “orthogonal to” (independent of) truth.

A fundamental goal might be: “make as many paper clips as possible” (leading to the end of humanity).

Fundamental Goals

Fundamental goals are not discussed in any depth by Hendrycks.

I believe this is because Hendrycks is trying to be relevant to current frontier models (GPT, Claude, Llama, PaLM, Grok).

Current frontier models can follow instructions but do not have their own independent persistent goals.

Instructions as Ephemeral Goals

Current frontier models can follow instructions. For example "write a Python script that sorts byte strings in alphabetical order".

But an instruction to a current frontier model is "ephemeral" — the goal is achieved by typing a single response and then the system awaits the next goal.

A current frontier LLM is a "servant" without goals of its own.

Persistent Goals

I will define a “persistent” goal to be a goal pursued consistently through rounds of interaction with others.

Consider real-life contract negotiation.

A persistent goal is similar to “reward” in a Markov decision process.

Agentive AIs

I will call an AI that pursues persistent goals **agentive**.

Current frontier models are not agentive in this sense.

Agentive AIs can potentially go "rogue" — pursue some interpretation of a goal that was not intended or perform harmful actions as a consequence of a goal that is too narrowly defined.

The Inevitability of Agentive AI

Agentive AI is inevitable because people will benefit from it.

At some level of AI competence, a corporation will be more profitable if it is run by an AI.

A military will be more effective under AI generals.

A political campaign will be more effective under an AI campaign manager.

We need some way to make agentive AI safe given that people are bound to use agentive AIs in pursuit of their own self-interests.

Goal Specification: Alignment

Broadly speaking the alignment problem is the problem of giving an AI fundamental goals aligned with with human values.

More narrowly, the alignment problem is equivalent to the principal-agent problem.

Alignment: The Principal-Agent Problem

Suppose you hire a lawyer to be your advocate in some legal dispute

The lawyer is on your side but has a self interest in extending the dispute so as to bill more hours.

Here you are the “principal” and the lawyer is the “agent” hired by the principle to do something.

The general principal-agent problem is that the agent has self-interest which may diverge from the interests of the principal.

The State of the Art in Goal Specification

Constitutional AI: Harmlessness from AI Feedback

Bai et al ArXiv 2212.08073 [2022, Anthropic]

Constitutional AI is an attempt to provide a mission statement (fundamental goal) they call a “constitution” for LLMs.

Actual human mission statements, consititutions, treaties, and laws are stated in English (natural language).
There does not seem to be any alternative to stating goals in Natural Language.

In the constitutional AI paper the LLM interprets the goal (constitution) and judges whether an action supports the goal.

The State of the Art in Goal Specification

The interpretation of goals stated in Natural Language is a deep problem.

Constitutional AI has been shown to work to some extent but is not included in frontier models which instead use instruction fine tuning.

Are Safety Issues Premature?

We do not yet have agentive AIs.

It seems likely (to me) that there will be continued revolutions in deep architectures.

Safety issues will need to be re-assessed in the presence of new architectures.

Retrieval Augmented Generation (RAG)

A recent architectural revolution is the integration of LLMs with retrieval.

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, Lewis eta al. (Meta, April 2021).

Google search now produces an “AI overview” computed from retrieved documents, including retrieval from current news, and then generated by an LLM from those documents.

Memory Architectures

I believe that retrieval will play an even larger role in LLMs in the future.

More specifically, it will evolve into read-write memory architectures.

In such an architecture "CPU" will 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.

Near Term Evolution of Memory Architectures

Frontier models memorize text from their training data.

Knowledge stored in the model parameters can become obsolete. For example who is the current president of the US.

Because facts change, there is an advantage to moving factual knowledge (episodic memory) out of the model parameters and into retrievable and editable memory.

If the model does not have to store factual knowledge then the training data required might be much smaller.

I expect to see research in this direction in the very near future (we will try here at TTIC).

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.

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 pursue 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 independent judgement in determining truth.(very challenging).

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.

Moral and Legal Responsibility Lies with Humans

The law should specify that the human principal is legally responsible for the actions of their AI agent.

This assumes that the human principal can in fact control the agent.

We must ensure that this is true.

Distribution of Power

If every human has a powerful AI agent then the power of any single human must be weighed against to collective power of all the AI agents of all the other humans.

In many ways the structure of society might not change — just people, countries and politics as usual.

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.

White-Hats, Black-Hats

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

A black-hat team looks for vulnerabilities.

We need both.

The Servant Mission Protocol is a white-hat proposal.

Now for some black-hat concerns.

Black-Hat: Instrumental Goals

An instrumental goal is a goal that is helpful in achieving the fundamental goal.

Staying Alive: The agent can't fulfill its fundamental goal if it is dead. Stuart Rice summarizes this as "I can't bring you coffee if I am dead".

Obtaining Money and Power Money and power can be used in support of essentially any fundamental goal.

Winning Friends and Influencing People: It is always helpful to have people who will help you.

Instrumental Goals and the Servant Mission

While I am presenting instrumental goals as back-hat concerns, these instrumental goals are true of the human principal as well.

These instrumental goals are things that the human principle also wants.

Comments on Memory Architectures

Perhaps unrelated to safety, there is the issue of the nature of the memory in memory architectures.

We typically do not remember things verbatim.

Is there a structured form of memory closer to the semantic content of what was said.

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 Models can do Semantic Parsing

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 Models can do Semantic Parsing

(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

(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

Fundamental goals are orthogonal to world facts.

Agentive AIs with their own fundamental goals are inevitable and raise the possibility of Rogue AIs.

Goals must be specified in natural language which raises the issue of goal interpretation.

Current models are safe and discussions of future safety need to make assumptions about future architectures.

The advobot protocol, informed by an assumption of future memory architectures, is designed to preserve human control and human societal structure in the coming age of AGI.

END