Using Your Powers for Good

SIG Math Fall 2025

Introduction

- There's one topic at the intersection of programming and math that we haven't touched on this semester, AI!
- LLMs are a perfect blend of computer programming and math theory. Understanding how they work only requires Python and linear algebra knowledge!
- I think AI is really important and the most influential technology of our time. It’s also one of the most dangerous, and I want to really implore anyone who has enjoyed the topics this semester to seriously consider looking into AI safety.

Why is AI dangerous?

- Depending on how much you use AI, you might have a couple of different perceptions:
  - Yeah, it's just glorified Google search where you type in questions and answer them.
  - AI can also be used agentically, such as in the clock code or the codex harness.
- Future AI systems will be much more intelligent than current ones and won't be limited to the environments we have them in right now.
  - When ChatGPT first came out, it was only good for answering short questions.
  - Two years later, it became good at searching the internet and thinking longer about harder questions.
  - Now, it can work agentically for up to hours at a time to write complex programs.
  - In the future, we might have AI systems that run for days or even weeks, and make difficult decisions, such as running a large company!

Why is AI dangerous? (cont.)

- Consider an artificial superintelligence (ASI), an AI that is significantly more intelligent than all humans on earth, and consider us giving it a harness that lets it run for weeks at a time.
  - We don't have ASI yet – current AIs are roughly equal to good humans or a bit worse in most domains.
  - Even if you don't believe we'll reach ASI soon, it's unreasonable to think that we won't ever get it.
- ASIs will have or be given goals that take a long time. But what if these goals are not aligned with human goals?
  - A famous example is Universal Paperclips – an AI told to maximize paperclip production ends up destroying humanity to build more paperclip factories. This is a silly example but still important!

Forms of misalignment

- When we talk about an AI system being “aligned”, we refer to it advancing human objectives in a way that humans want.
- Outer misalignment – the AI pursues the wrong objective.
  - Consider a call center company that wants to optimize customer happiness. This is hard to measure, so they instead measure tickets closed per day. But then an employee gives shallow answers to close tickets faster.
- Inner misalignment – the AI doesn’t actually pursue the correct objective
  - The call center company learns from their past failure and hires someone that will actually do a good job. Unbeknownst to them, this new hire only cares about looking brilliant rather than producing good outcomes. The hire is promoted to a management position, but then they start taking credit for others’ work and avoid any risk.
  - Happens when the AI is moved to an environment that it wasn’t trained on.
- Both outer and inner misalignment are serious failure modes!

Technical AI safety research breakdown

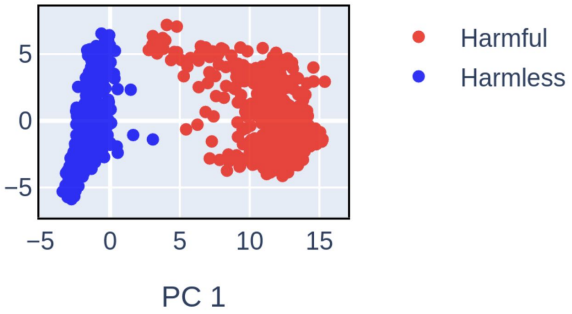
- There are many different fields of AI safety. The two broadest ones are technical and governance. Technical AI safety is divided into many subfields, some overlapping:
  - Mechanistic interpretability – how do the internals of an LLM work? One of the most popular subfields, the subject matter is very cool!
  - Scaleable oversight – if LLMs are much smarter than us, how can we check their work?
  - AI control – how can we control and stop a misaligned ASI?
  - Evals – how do we know if AI is meaningfully aligned?
  - Theoretical – what is the mathematics behind deep learning? Often involves lots of pure maths (e.g. Singular Learning Theory applies algebraic geometry)
- This is just the tip of the iceberg! There are many approaches to solving alignment, and we don't know which ones will work! So a lot of manpower is needed to try many different ideas.

Example paper: Refusal is mediated by a single direction

- Simplest terms: LLMs have an internal state that is a vector in a very high dimensional space (GPT-3 was 12,288D)
- We know from linear algebra that spaces have subspaces. These subspaces can contain information about what the model is thinking. Maybe refusal is one of those subspaces?
- The actual paper was a bit more complex, but here is the main idea:
  - Run the model on n harmful instructions and n harmless instructions, and record the activations for each prompt.
  - Take the average activations for the harmful instructions and the average for the harmless instructions. Averaging accounts for the variation in the prompt topic.
  - Subtract the mean of the harmless activations from the mean of the harmful activations (this technique is called *difference-in-means*)
- Now we have a vector that represents the harmfulness direction! But does it actually represent anything?

Example paper: Refusal is mediated by a single direction

- Yes! Along the refusal direction, prompts separate cleanly into harmful and harmless prompts.
- The researchers then found that if you add the refusal vector to a harmless prompt, the model would refuse to answer it, and if you subtracted the refusal vector from a harmful prompt, the model would answer it!
- Finding a linear subspace is a common technique in LLMs. For example, truth is also encoded as a linear subspace. But many things might not be linear!



How can I get started?

- ARENA – bite-sized Python modules designed to teach you alignment research from scratch
  - Requires knowledge of Python and linear algebra
  - Very popular, many programs assume you have already completed ARENA.
- BlueDot – AI safety courses with certificates covering introductory topics in the field
  - These are widely recommended and are a great first step in getting into the field!
- LessWrong – forum discussing AI safety topics and rationality
  - LessWrong's rationality resources are super valuable and are applicable to any field. Rationality asks the question of how you come to the best conclusions.
- Ask me – I'm not an expert but I can give you some more guidance if you have questions!

Other Ways to Use Your Powers For Good

80,000 hours

- AI is one of the most significant opportunities to contribute to the world, but you can make a difference in any field!
- When people think about living ethically, people usually think of things like recycling, being nice, and volunteering.
- However, this is missing something huge -- your choice of career.
- You have about 40 hr/wk * 50 wk/yr * 40 yr = **80,000 hours** in your career. This is a staggering number! This has some counterintuitive implications:
  - Unless you are the heir to some large estate, your time is the most effective way to help others.
  - It's worth spending a *lot* of time making small improvements to your career! For example, if you could find some way to make your career 1% better, it would be worth spending up to 800 hours to do it.

Motivating example: carbon footprints

- Climate-minded people often stress the importance of minding your carbon footprint. This includes refraining from things like not driving, not taking airplanes, not buying meat, and not buying many things.
  - The average American causes about 15 tons of CO_2 emissions per year.
- In comparison, donating to climate R&D organizations effectively reduces emissions by around $10/ton.
- Therefore, donating $1000/yr would be around 10 times as effective as cutting your emissions to zero! It's also much easier and cheaper.

Calculating impact

- Calculating impact is hard, but it's driven by a simple expression: P * S * F
  - P - how pressing the problems are that you focus on
  - S - the scale of contribution the path lets you make towards tackling those problems
  - F - your personal fit for the path
- These factors are multiplicative -- if you can find a problem that is twice as neglected, contribute twice as much, and are twice as good a fit for it, that is effectively 8 times the impact!
- Conversely, if you find a problem that is near 0 on any of the axes, your impact will likely be near zero.

It's incredibly bad to not be rich when you can be rich

- A child is about to drown in front of you. You can press a button to remove the water. It seems bad to not press that button.
- A child is about to drown in front of you. There is a slot that you can insert $3000 into to remove the water. Coincidentally, there is a button that will dispense $3000. It seems bad to not press that button.
- A child is about to drown in front of you. You can press a button to remove the water, but the button is locked until you complete 10 CS 280 assignments. (the water will wait until you complete the assignments) It seems bad to not complete the assignments.
- A child is about to drown in front of you. There is a slot that you can insert $3000 into to remove the water. Coincidentally, there is a machine that offers $3000 for completing 10 CS 280 assignments. It seems bad to not complete the assignments.
    - If you know about the machine, it also seems bad to complete 10 MATH 473 assignments instead of CS 280 assignments.

Doing the most good

- There are a lot of bad things that happen to people. Out of those bad things, dying is one of the most bad things that can happen to someone.
- Conversely, if you could take someone who would have died and given them an extra few decades of life, that would be very good!
- According to GiveWell, it **costs about $3,000 to save one life**.
  - This is not a lot of money! If you work a part time job, you will make that much in a few months. How often do you save a person's life?
- Suppose you decide to give 10% of your income per year. If you're a middle class person that makes 60k per year, that is about 2 lives saved per year.
- Suppose you are wealthier and make 300k per year, then you can save 10 lives per year!

Animal welfare

- I try to give all my estimates in terms of human lives because that's the easiest to empathize with, and humans do provide a lot of value!
- However, it's important to note that trillions of animals are killed every year in horrible ways in factory farms and they have conscious experiences just like we do.
- Additionally, donating to effective animal charities is orders of magnitude more effective than human ones!
- There is one field of animal welfare that is supremely effective that I believe you all have heard me talk about a lot...

Shrimp Welfare

Why programming?

- I’m sure you can tell that I’m building up to the point that becoming a programmer is one of the most effective ways to do good in the world.
- That being said, it’s not a panacea:
  - You might be bad at programming, or really hate it
  - You might have an alternative career path that you could do more good in
  - AGI has been invented and human programmers are no longer valuable
- But if you really like solving problems, I think programming as a career merits serious consideration.
  - You can work on AI, the single most influential technology ever invented
  - You can make a game or other startup and become rich and influential
  - You can work at an established company and become rich
  - There are a lot of different types of programming jobs for a wide variety of interests