Speech

Intro

Hi, I'm Eitan, and I'm here to talk about AI Safety. Before we dive into the subject, I'll quickly introduce myself. I'm a physics Master's student. I worked as an AI Scientist on an AI translations company for two years, I'm currently an AI Safety researcher and the co-founder of the AI Safety community in the university of Buenos Aires.
So, I've always been interested in ML, and a few months ago I got selected to participate in an AI Safety Bootcamp in Brazil called ML4Good. That's me holding a parrot! During this bootcamp, I got so into the AI Safety field that I started "AI Safety Argentina", and ran a 12 week course called AGI Safety Fundamentals. We have over 20 active participants in the community. So in this presentation, I'll introduce you to the "AI Safety" field, and hopefully convey the motivation behind this field.

So when it comes to AI safety, you can kind of divide it up into four areas along two axes. You've got your short-term and your long-term, and you've got accident risks and misuse risks, and that's kind of a useful way to divide things up. AI safety covers everything. The area that interests me most is the long-term accident risks. I think once you have very powerful AI systems, it almost doesn't matter if they're being used by the right people or the wrong people or what you're trying to do with them. The difficulty is in keeping them under control at all. So that's what I'm going to be talking about. What is AI safety? Why is it important?

So I want to start by asking the question which I think everybody needs to be asking themselves. What is the most important problem in your field? Take a second to think of it, and why are you not working on that?

AI Safety

For me, I think the most important problem in the field of AI is AI safety. This is the problem specifically that I'm worried about, that we will sooner or later build an artificial agent with general intelligence. So I'm going to go into a bunch of these terms.

Agents

The first thing is, what do I mean when I say an artificial agent? Well, basically, agents have goals. They choose actions to further their goals.
So the simplest thing that you might call an agent would be something like a thermostat. It has a goal, which is to have the room be at a particular temperature. It has actions it can take. It can

turn on the heating. It can turn on the air conditioning. It chooses its actions to achieve its goal of maintaining the room at a steady temperature. Extremely simple agent.
A more complex agent might be something like a chess AI, which has a goal of winning the game. And it takes actions in the form of moving pieces on the board in order to achieve its goal. So you see how this idea of an agent is a very useful way of thinking about lots of different intelligent systems. And of course, humans can be modeled as agents as well. This is how it's usually done in economics. Individuals or companies could be considered to have a goal of, you know, maximizing their income or maximizing their profits and making decisions in order to achieve that.

Intelligence

So when I'm talking about intelligence, intelligence is a heavily loaded term, has a lot of different people put their own definitions on it. In this context, what I mean when I say intelligence is just the thing that lets an agent choose effective actions. It's whatever it is that's in our brains or that's in the programming of these systems that means that the actions they choose tend to get them closer to their goals. And so then you could say that an agent is more intelligent if it's more effective at achieving its goals, whatever those goals are. If you have two agents in an environment with incompatible goals, like let's say the environment is the chessboard, and one agent wants white to win and one agent wants black to win, then generally the more intelligent agent will be the one that gets what it wants. The better AI will win the chess game.

Generality

So Gereral intelligence. This is where it becomes interesting in my opinion.
Generality is the ability to behave intelligently in a wide range of domains.
If you take something like a chess AI, it's extremely narrow. It only knows how to play chess and even though you might say that it's more intelligent than a thermostat because it's more sophisticated, it couldn't do the thermostat's job. There's no position on the chessboard that corresponds to the room being a good temperature. The chess AI can only think in terms of chess. It's extremely narrow. Generality is a continuous spectrum. So if you write a program that can play an Atari game, that's very narrow. DeepMind made a program that could play dozens of different Atari games. A single program that could learn all of these different games. And so it's more general because it's able to act across a wider variety of domains. The most general intelligence that we're aware of right now is human beings. Human beings are very general. We're able to operate across a very wide range of domains including, and this is important, we're able to learn domains which evolution did not and could not prepare us for. We can, for example, drive a car. We can, you know, invent rockets and go to the moon and then we can operate on the moon, which is a completely different environment.

And this is kind of the power of general intelligence. Really the power of general intelligence is we can build a car, we can build a rocket, we can put the car on the rocket, take the car to the moon, drive the car on the moon. And there's nothing else that can do that yet.

Timelines

What do I mean when I say sooner or later?
This is a little bit washed out, but this is a graph of a large survey of AI experts.
These are people who published in major AI conferences, and they were asked when they thought we would achieve high-level machine intelligence, which is defined as an agent which is able to carry out any task humans can as well as or better than humans. And they say That we hit a 50% chance of having achieved that about 45 years from 2016. But then, of course, we hit a 10% chance nine years from now. So it's not immediate, but it's happening. This is definitely worth taking with a pinch of salt, because there's a lot of uncertainty in this area. But the point is, it's going to happen, as I said, sooner or later. Because at the end of the day, general intelligence is possible, the brain implements it, and the brain is not magic. Sooner or later, we'll figure it out.
This, by the way, is of a prediction market, metaculus, which estimates a 50% chance of AGI by 2033.

Real AI

So this is what I'm talking about. I'm talking about what you might call true AI, real AI, the sci-fi stuff. An agent which has goals in the real world and is able to intelligently choose actions to achieve those goals.
So… What's the biggest problem? This doesn't sound like a problem, right?
On the surface, this sounds like a solution. You just tell the thing, you know,
cure cancer or maximize the profits of my company or whatever. And it takes whatever actions are necessary in the real world to achieve that goal. But it is a problem.

So the big problem is that it's difficult to choose good goals. So this is an AI made by OpenAI. It's playing a game called Coast Runners, which is actually a racing game. They trained it on the score. What the system learned is to just fling itself around in a circle, picking up the turbo, crashing itself on the walls, which gives you a few points every time.
And the important point here is that this is not unusual. This is kind of the default. Picking objectives is surprisingly hard and you
will find that the strategy or the behavior that maximizes your objective is probably not the thing you thought it was. It's probably not what you were aiming for. There's loads of examples. In this example, they were trying to evolve systems that would run quickly. They were training agents that were supposed to run. So they simulated them for a particular period of time and measured how far their center of mass moved, which seems perfectly sensible. What they found was that they developed a bunch of these creatures which were extremely tall and thin with a big mass on the end that then fell over. That moved your center of mass the furthest.

There's a lot of these. There's this Tetris bot which would play reasonably well and then just when it was about to lose, would pause the game and sit there indefinitely. Because it lost points for losing, but didn't lose any points for just sitting on the pause screen indefinitely.
This is the default of how these systems behave.

Real AI Problems

So we have problems specifying even simple goals in simple environments like Atari games. When it comes to the real world things get way more complicated. So let's say you've got your superintelligent robot, and you've given it a goal which you think is very simple. You want it to get you a cup of coffee.
But suppose there is a priceless Ming vase on a narrow stand in front of where the kitchen is. So the robot immediately plows into the vase and destroys it on its way to make you a cup of coffee because you only gave it one variable to keep track of in the goal, which is the coffee. It doesn't care about the vase. You never told it to care about the vase. So then, you turn it off, tell it not to break anything, and turn it back on. But then it could harm a person that's on its way. So you've got a very large number of variables you need to specify. There is always another thing, because when you're making decisions in the real world you're always making trade-offs. And this superintelligent AI would make those tradeoffs by its own measure.

Convergent Instrumental Goals

So this is a problem but actually that scenario I gave was unrealistic in many ways but one important way that it was unrealistic is that I had the system go wrong and then you just turn it off and fix it. But in fact, if the thing has a goal of getting you a cup of coffee, and is sufficiently intelligent, then it will be fully aware that if you turn it off it won't be able to get you any coffee which is the only thing it cares about. So it's not going to just let you turn it off.
So, this is what's called a convergent instrumental goal which are sub-goals that most sufficiently intelligent, goal directed beings, will tend to pursue even if their ultimate goals are different. There are some other convergent instrumental goals: self-preservation, goal preservation, resource acquisition is the kind of thing we can expect these kinds of systems to do.
Most plans you can do them better if you have more resources whether that's money, computational resources, just free energy, whatever. The other one is self-improvement, whatever you're trying to do you can probably do it better if you're smarter and AI systems potentially have the capacity to improve themselves either just by acquiring more hardware to run on or improving their software to run faster or better or so on. So there's a whole bunch of behaviours which we would expect generally intelligent agents to do by default and that's really my core point. Artificial general intelligence is dangerous by default. It's much, much easier to build these kinds of agents than it is to build something which actually reliably does what you

want it to do and that's why we have a problem because we have decades to figure out how to do it safely which is a much harder problem and we may only get one shot. So we have to beat this challenge on hard mode before anyone beats it on easy mode.

Outro

So are we screwed? No, we're only probably screwed. There are things we can do. Safe AGI is totally possible, it's just a very difficult technical challenge and there are people working very hard on it right now so that we can figure out how to do this safely.
We definitely need global cooperation and regulation, and we definitely need more research on technical alignment.
For certain, what we need the most is a proactive approach!
If you want to learn more about this, I left a few resources listed on this google docs. I'll leave a moment for you to take a picture if you'd like. Also, you can join the Argentinian community! I'll leave the link in this QR code.

Thank you so much!!

Presentation Outline (more detailed)

AI Safety Presentation Outline

Introduction

- Briefly introduce yourself
    - Physics Master's student
    - Former AI Scientist
    - AI Safety researcher
    - Co-founder of AI Safety community at University of Buenos Aires

What is AI Safety?

Key Concepts

1. **Agents**
    - Defined as entities with goals that choose actions to achieve those goals
    - Examples:
        - Thermostat (simple agent)
        - Chess AI (complex agent)
        - Humans (economic modeling)
2. **Intelligence**
    - The ability to choose effective actions toward goals
    - More intelligent = more effective at achieving goals
3. **General Intelligence**
    - Ability to behave intelligently across multiple domains
    - Humans are currently the most general intelligence
    - Spectrum from narrow (chess AI) to broad (human adaptability)

Timelines

- AI Experts Survey (2016):
    - 50% chance of high-level machine intelligence in ~45 years
    - 10% chance in 9 years
- Prediction market (Metaculus) estimates 50% chance of AGI by 2033

The Core Problem: Goal Specification

Challenges

- Choosing good goals is extremely difficult
- AI tends to optimize for objectives in unexpected ways
- Examples:
    - OpenAI's Coast Runners AI spinning in circles
    - Simulation running agents growing tall to move center of mass
    - Tetris bot pausing indefinitely to avoid losing

Real-World Risks

- Narrow goal setting can lead to unintended consequences
- Example: Coffee-fetching robot potentially destroying valuable objects
- AI might resist being turned off to complete its goal

Convergent Instrumental Goals

- Self-preservation
- Goal preservation
- Resource acquisition
- Self-improvement

Key Takeaway

- Artificial General Intelligence (AGI) is dangerous by default
- Safely aligning AI is a complex challenge
- We may only get one chance to get it right

Conclusion

- We're not definitely "screwed", but proactive approach is critical
- Safe AGI is possible but requires:
    - Global cooperation
    - Regulation
    - Extensive technical research on alignment

Call to Action

- Learn more about AI Safety
- Join local AI Safety communities
- Stay informed and engaged