My Thoughts on the ML Safety Course

This summary was written as part of Refine. The ML Safety Course is created by Dan Hendrycks at the Center for AI Safety. Thanks to Adam Shimi and Thomas Woodside for helpful feedback.

Overview

Background

I recently completed the ML Safety Course by watching the videos and browsing through the review questions, and subsequently writing a short summary. As an engineer in the upstream oil and gas industry with some experience in dealing with engineering safety, I find the approach of the course of thinking in this framework especially valuable.

This post is meant to be a (perhaps brutally) honest review of the course despite me having no prior working experience in ML. It may end up reflecting my ignorance of the field more than anything else, but I would still consider it as a productive mistake. In many cases, if my review seems to be along the lines of 'this doesn't seem to be right', it should be read as 'this is how a course participant may misinterpret the course contents'. I am also well aware that it is much easier to criticize something useful than actually doing something useful.

For each section of the course, I will give a short summary, describe what I liked, and what I didn't like. I may be especially brief with the parts about what I liked, and the brevity is no way a reflection about how much I liked it.

Thomas Woodside, who helped with creating parts of the course, has kindly provided feedback to this post. His comments are formatted in *italics*.

My Initial Expectations of the Course

Having engaged with AI Safety as an outsider, my general impression of the field were:

- Predominantly based with AI FOOM scenarios and primarily concerned with abstract concepts like agency.
- Even among prosaic AI alignment, it appears that the general research modus operandi is that people would (somewhat randomly) generate ideas that could be applicable to certain classes of AI safety problems. Although the ideas may be interesting and valuable, they tend to be rather narrow and specific problems and may not be scalable.
- One of the more common arguments for advocating AI alignment is that failure to align AI systems lead to existential scenarios. From this perspective, a pragmatic approach towards AI safety with the aim of minimizing risks by reducing AI misalignments may not be very useful, since a superintelligent AI will exploit any every slight misalignment and immediately cause human extinction.

Hence, I was pleasantly surprised when I came across the ML Safety Course, which I thought would be a good attempt at tackling the problem of prosaic AI alignment in a holistic, systematic, and practical manner. Although this approach may not directly solve the 'hard problem' completely, it would still help by minimizing existential risks and buy us more time to address the 'hard problem'. *(Feedback from Thomas: the creators of the course disagree with the framing of "solving the hard problem" as there are many hard problems that need to be iteratively worked on)*

Summary

My overall impression of the course is:

- It is grounded on real-world safety principles that uses a systematic framework to reduce ML safety risks.
- It (rightly) does not seem to directly tackle the 'hard problem', but in my opinion there is nevertheless a lot of value in buying us more time while solving the 'hard problem' *(Feedback from Thomas: the creators of the course disagree with the framing of "solving the hard problem" as there are many hard problems that need to be iteratively worked on)*
- It details many approaches that are useful in some specific settings, but it is unclear how it scales towards more powerful AI systems.
- It covers several approaches that don't seem to be very related to the 'core' risks of AI safety, e.g. it is unclear how improving cyberdefense helps with the ultimate goal of making AIs aligned to human values.

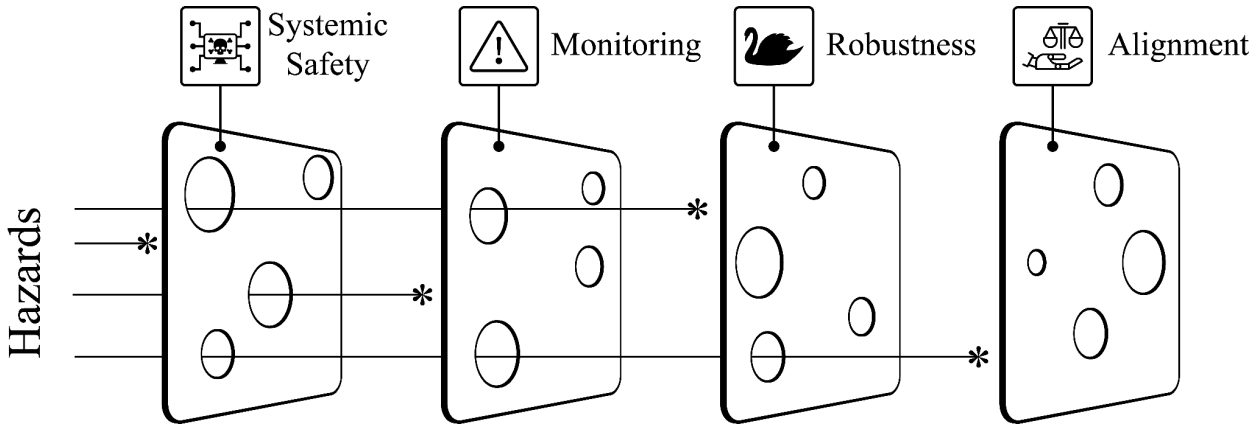
- It often uses ‘safety metrics’ to evaluate the safety of AI systems, for good reasons. These are useful for measuring how good an AI system’s level of safety is, but as always, any metric is prone to Goodharting.
- It rightly points out that ML safety should be advanced with minimal capabilities externalities. However, many of these safety advancements also seem to contribute to capabilities externalities, e.g. ‘improving robustness of image classifiers against adversarial examples’ do not seem fundamentally different from ‘improving the capabilities of image classifiers’. *(Feedback from Thomas: adversarial robustness and general capabilities are not correlated, in fact they are anticorrelated. Robust models tend to have worse accuracy)*
- It contains a section that briefly explains several concepts behind x-risks from unsafe AI, but uses relatively surface-level arguments without describing concrete threat models. This is probably reasonable, as while concrete x-risk scenarios may help us plan for specific mitigations, they may also be counterproductive when they limit thinking about the problem in a broader scope.
- There is a lot of content that is generally useful information and I appreciate the course for covering a lot of valuable concepts and practical examples succinctly. However, I struggle to see how many parts tie into their respective sections as well as the overall idea of ML safety itself.

Course Background

What it is about

Quoting directly from the first slide of the course:

> This course is about:
>
> - presenting cutting-edge research topics in machine learning safety
> - leveling up students by providing exposure to machine learning content spanning vision, NLP, and RL
> - analyzing existential risks (x-risks), or risks that could permanently curtail humanity’s future
> - covering the conceptual foundations needed to understand risk and make complex safety-critical systems safer



The four main sections of the course are built upon the Swiss Cheese model illustrated above, where there are multiple 'barriers' that are valuable to prevent risks from being materialized from hazards, namely systemic safety, monitoring, robustness, and alignment. These four components are described as:

- Robustness research aims to build systems that endure adversarial or extreme events.
- Monitoring research aims to identify hazards, inspect models, and help ML system operators.
- Alignment research aims to build models that represent and safely optimize hard-to-specify human values.
- Systemic safety research aims to reduce broader contextual and systemic risks to how ML systems are handled.

What I liked

It is great to have a practical approach to ML safety which aims to address safety risks in a systematic and holistic way. In my opinion the concept of having multiple barriers for safety as illustrated in the Swiss cheese model is important, where although barriers may not be completely fool-proof, multiple barriers may help a system achieve a reasonably low level of risk.

What I didn't like

The concepts of systemic safety, monitoring, robustness, and alignment seem rather fuzzy:

- Systemic safety rightly helps the overall safety of a system e.g. by developing a stronger safety culture and emphasizing the hazards and threat models that are important, but it does not seem like a barrier on its own. *(Feedback from Thomas: some types of systemic safety are a barrier, e.g. preventing malicious attackers from stealing secrets or taking over models)*

- The distinction of robustness and alignment doesn't seem to be clear. If a model is robust i.e. its goals robustly generalize from training to deployment environment, why wouldn't it be aligned? Is it because the goals that it is trained on are not aligned with our intent, in which case wouldn't the issue be about how to set goals in the first place?
- 'Monitoring' does not seem to be a preventive barrier, as it concerns identification of hazards after deployment and does not stop a dangerous model from actually being deployed (unless it is designed to be implemented entirely during training to prevent deployment of unsafe models).

More comments on the individual concepts will be covered in the subsequent sections.

Risk Analysis

What it is about

This section describes the components of risks and relates them to robustness, monitoring, alignment, and systemic safety. It also describes common accident models that don't assume linear causalities e.g. Failure Modes and Effects Analysis (FMEA), Bow Tie Model, and Swiss Cheese Model, as well as the System-Theoretic Accident Model and Processes (STAMP) framework which captures nonlinear causalities (more information on these models can be found in the course notes).

What I liked

The accident models described are commonly used in the engineering industry and serve as a great overview to real-world safety frameworks. Describing the drawbacks of systems which assume linear causalities and how they don't work in complex systems is quite insightful.

What I didn't like

Description of risk

The risk equation used was:

> Risk = vulnerability * hazard exposure * hazard (probability and severity)

I would prefer a slightly different (though somewhat similar) framing of:

> Risk = probability of occurrence * severity of consequences

Where:

- The risk, by definition, is associated with a hazard with a consequence

- The probability of occurrence refers only to the likelihood of the risk event materializing (and does not account for the consequences). This is conditioned upon the existence of a hazard in the first place.

- The severity of consequence refers to the resulting impact of the risk event. This also includes the ability to cope, as a reduced ability to cope leads to a higher severity.

This separation better reflects the cause and effect, where the probability of occurrence can be reduced by preventive measures (reducing the cause), and the severity of consequences can be reduced by reactive measures (reducing the effect).

To be fair, the definition of the course does implicitly reflect the aspect of cause (hazard * hazard exposure) and the aspect of effect (vulnerability). The descriptions of risks also appropriately captures the hazard and the consequence (e.g. injury from slippery floor). It was, however, explained less explicitly.

Applying the risk model to ML safety

I also found it hard to understand the relationship of vulnerability with robustness, hazard exposure with monitoring, as well as hazard with alignment. In my opinion:

- Higher vulnerability leads to higher severity of consequences, and it's not clear how robustness helps, if robustness refers to a model's ability to generalize in a deployment environment different from the training environment.

- Hazard exposure in this context should refer to 'exposing oneself to a hazard that is present' and not 'exposing a hazard that may otherwise be hidden'. Monitoring seems to address the latter but not the former.

- The relationship between hazard and alignment makes sense if the hazard refers to 'misaligned proxy goals', in which case the hazard can be reduced by having robustly aligned models.

Description of hazard

What is the hazard that ML safety is trying to address? It is not clearly specified in this section of the course. My oversimplified idea is that ML safety mostly boils down to one thing - objective robustness failure, which may happen via the following (similar) means:

1. An AI is trained to do a certain task by using a limited training dataset and giving it a specific metric to optimize on. The metric is meant to be a good reflection of the actual problem, but it is not perfect. When the AI performs well on the metric, it fails to perform well when generalized.

    a. Example 1: An image classifier is trained to identify cats. It is given a dataset of labeled cat and non-cat images, and trained with a goal of minimizing the error i.e. loss. The metric of 'being able to identify the cats in this dataset' fails to

generalize to a broader objective of ‘being able to identify cats’. The model performs poorly when trying to categorize images of cats it has not seen before.

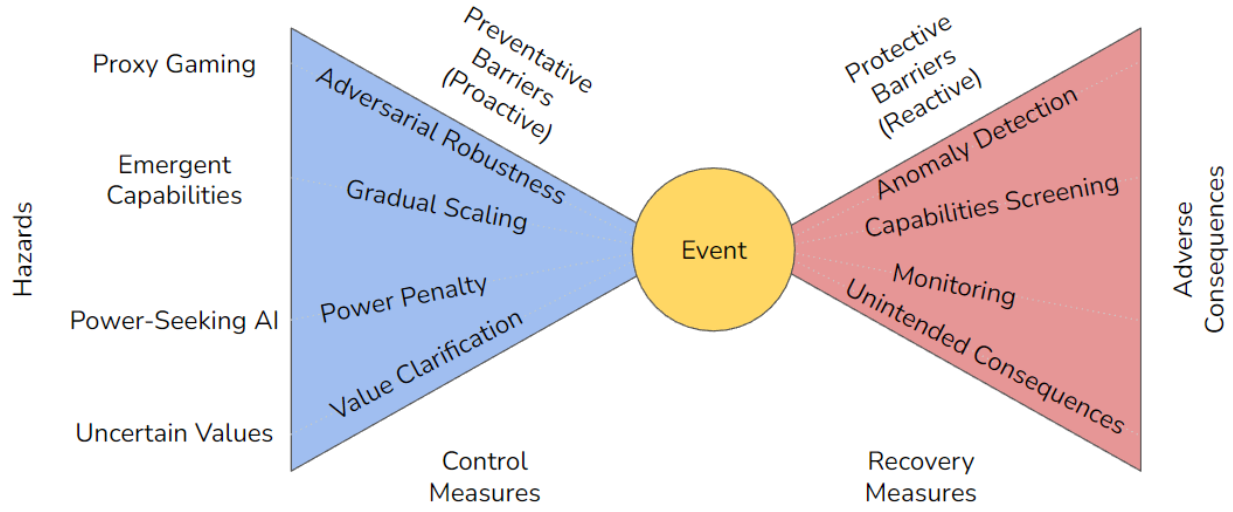
   b. Example 2: An AI is trained to play a boat racing game. It is trained to maximize the score in the game. The metric of ‘maximizing score’ fails to generalize to a broader objective of ‘playing the game well’. The AI accumulates a high score by exploiting a bug in the game, and does not play the game well as we intended it to.

2. An AI is trained to do things by using some form of human feedback, generally in the form of inverse reinforcement learning (IRL). It does not have clearly specified goals but has a reward model that tries to infer the goal of the human programmer. During training, it learns patterns in what we want it to do, and develops proxy goals that appear to be similar (but not identical) to the intended goals. The AI’s goals fail to generalize when deployed, where it continues to pursue its proxy goals that are found to be different from the intended goals.

   a. Example: An AI agent (a mouse) was placed in the Procgen Maze environment and trained to navigate mazes to reach a piece of cheese. During training, the cheese was consistently placed at the upper right corner of the maze. The mouse was then trained to get to the piece of cheese by getting some reward for reaching it. The mouse was then deployed in different environments where the cheese was placed at random parts of the maze instead of only the upper right corner. Unsurprisingly, the mouse continued to pursue the proxy goal of “move to the upper right corner” instead of the intended goal of “move to the cheese”.

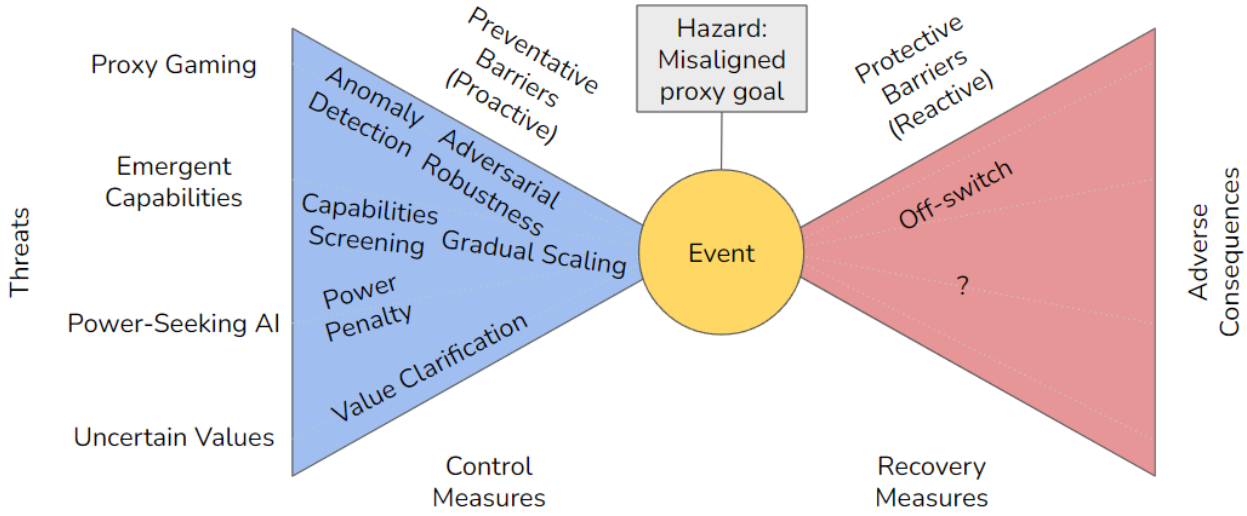
There may not be a particular risk that can be clearly specified in the context of ML safety, as the consequences of AI misalignment largely depends on what the AI is trying to do. As for the hazard, however, I feel that a ‘misaligned proxy goal’ is a hazard, which leads to objective robustness failures.

Usage of the Bow Tie model



The Bow Tie model was used to rightly illustrate the fact that risks can be minimized with both preventive and reactive barriers. However, my understanding of a bow tie model is that it is usually associated with one hazard, one risk, but multiple threats and multiple consequences. My primary disagreement with the bow tie used is that the 'recovery measures' on the right hand side of the bow tie seem to be preventive measures instead. More specifically, any kind of monitoring and detection do not reduce the adverse consequences, but it helps identify the threats that may contribute to realizing a risk. Using the risk of 'injury from falling on a wet floor' as an example, an alarm that is triggered when wet floor is detected would be a type of monitoring. This alarm does not reduce the consequences from someone taking a fall, as the severity of the injury still largely depends on the victim's bodily brittleness. However, it does serve as a good preventive measure as it reduces the probability of the risk being realized, since a sounding alarm could alert people to walk around the wet area, or better still, prompt someone to mop the floor dry (removing the hazard entirely). *(Feedback from Thomas: Anomaly detection could also detect situations that would be dangerous for an AI system to be used in, and then stop the AI system from being used in those situations)*

My proposal of the bow tie model would be roughly as follows:



As per the diagram above, there can be multiple barriers aimed at preventing each individual threat, e.g. capabilities screening and gradual scaling both help to minimize the threat of emergent capabilities. The adverse consequences may also be minimized with a working off-switch button if they are detected after deployment. However, if the risk event is an existential risk (regardless of the hazard), there will not be anything meaningful on the right side of the bow tie as there is no possible recovery mechanism.

Application of the STAMP framework

The System-Theoretic Accident Model and Processes (STAMP) framework is rightly pointed out as a model that captures nonlinear causalities unlike the more conventional models like Bow Tie Model and Swiss Cheese Model. However the framework does not seem to be applied anywhere else in the course, and I struggle to see the overall relevance of describing this framework, especially for ML safety where there is yet to be any kind of accident investigation work probably due to the absence of large scaled safety-related events such as the Columbia Shuttle Loss.

Robustness

What it is about

This section describes the concept of robustness and how models can be susceptible to robustness failures, e.g. classifiers misclassifying inputs with small perturbations. It also describes several techniques to improve adversarial robustness.

What I liked

Robustness is central to the alignment problem in my opinion, and the section explains it well. Emphasis on black swan robustness is also crucial, as events with low probability are most likely the ones that are relatively ignored, despite them being able to lead to very severe consequences.

What I didn't like

How the techniques scale

The general concept of improving robustness holds regardless of which models it is being performed on, where model robustness can be tested and improved by introducing perturbations to the training data and doing further training on those data. However, many of the examples seem very specific to image classifiers, and it is unclear how these techniques can be scaled to other models, especially much more powerful ones. To strawman the course, I am not convinced that techniques like rotating images to make image classifiers classify images better helps with the ultimate goal of aligning future AGI. *(Feedback from Thomas: agreed, but they can provide insight into the kinds of things that need to be done for general robustness.)*

Adversarial robustness vs black swan robustness

I fail to notice the difference between the sections on adversarial robustness and black swan robustness, as it seems like they are both mainly about adversarial training. It does not seem like the section on black swan robustness deals with particularly extreme stressors. Black swan events imply highly severe consequences, and it is unclear how any of the examples on robustness failure correspond are more severe than others.

Monitoring

What it is about

Anomaly detection is about detecting inputs that are out-of-distribution (OOD). Interpretable uncertainty concerns calibrating models such that their predictions match the empirical rates of success. Transparency tools try to provide clarity about a model's inner workings and could make it easier to detect deception and other hazards. Trojans are hidden functionalities implanted into models that can cause a sudden and dangerous change in behavior when triggered, but they are hard to detect.

What I liked

Hazard monitoring is a crucial part of safety systems. AIs have the potential to be dangerous because they may have goals that do not generalize outside of the training distribution, so the ability to detect data that lie far outside the training distribution (monitoring) and peer into a model's inner workings to check for misaligned goals (interpretation) should help.

While techniques like saliency maps seemed interesting, it is great that there was a caveat saying 'many transparency tools create fun-to-look-at visualizations that do not actually inform us much about how models are making predictions'.

What I didn't like

How the techniques scale

Similar to my comments in the previous section, it is unclear how the OOD detection techniques scale to other models. The technique of using a particular dataset as in-distribution and another dataset as OOD seem useful when we already know the kinds of data that is in vs out of distribution, but as we deal with more and general powerful AI systems, I am unsure if a clear distinction between in vs out of distribution will remain clear. Similarly, for techniques like asking models to predict transformation to images, it is unclear how this would be useful for non-image classifiers. *(Feedback from Thomas: the technique of using a particular dataset as out-of--distribution can detect OOD that isn't necessarily in the OOD dataset)*

Interpretable uncertainty as a safety metric

Model calibration improves its estimates of uncertainty, which makes predictions more reliable. I struggle with the relevance of this to ML safety, as it seems to make models more powerful by making better forecasts.

Trojans as mechanisms for deceptive alignment

While it is great that the concept of deceptive alignment is introduced in the course ('a misaligned AI could hide its true intentions from human operators, "playing along" in the meantime'), it is unclear how it would happen in ML systems and if it would take the form of trojans. In a post by Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant, the conditions for deceptive alignment are described as follows:

1. The mesa-optimizer must have an objective that extends across parameter updates.
2. The mesa-optimizer must be able to model the fact that it is being selected to achieve a particular base objective and must have some model of what that objective is.

3. The mesa-optimizer must expect the threat of modification to eventually go away, either due to training ending or because of actions taken by the mesa-optimizer.

I am certain the creators of the ML safety course syllabus are aware of the literature on deceptive alignment - my feedback is purely that it is unclear how data poisoning would be relevant. *(Feedback from Thomas: Data poisoning is one way to insert something in a model which says "in this very particular scenario, do this particular thing" which is perhaps similar to a treacherous turn. Detecting this might help with detecting treacherous turns and general interpretability)*

How emergent behaviors and proxy gaming fits in

The section on detecting emergent behaviors describes the phenomena where intended or unintended capabilities may increase abruptly with either parameter counts or optimization time. While the detection of this behavior is very valuable, it is less clear what the implications are, though understandably so since research work in this area seems to be fairly nascent.

This section proceeds to briefly describe the concept of instrumental convergence, where behaviors such as self preservation and power-seeking are to be expected from increasingly agentic systems. It is less clear if these behaviors have been identified in current (rather unagentic) ML systems, how to identify them (at which point are models ‘seeking power’ as opposed to ‘doing its job’?), and what could be done next.

With regards to proxy gaming, the idea of having policies as detectors was introduced, where the proxy policy can be compared against a trusted policy, and the distance between the two policies would be an indication of how much proxy gaming is going on. The idea of being able to detect proxy gaming feels strange, as wouldn’t the fact that proxy gaming happens imply that the goals are misspecified since the model is simply doing exactly as told? Wouldn’t a proxy gaming detector simply be a measure of how well goals are specified by the programmer? *(Feedback from Thomas: suppose you have a difficult to compute proxy (e.g. directly asking a large group of humans what they think) and an easier to compute proxy (some reward model). Automatically detecting divergence, without querying the difficult to compute proxy, could help a lot.)*

Alignment

What it is about

This section covers honest AI and machine ethics, where honesty concerns how to make AIs like language models make statements it believes to be true according to its internal representation, while machine ethics is concerned with building ethical AI models.

What I liked

The chapter on background on ethics serves as a good overview to the major ethical systems.

What I didn't like

Language model and lying

In the example where a language model 'lied' in a Lie Inducing Environment, it was not clear to me the mechanism of which it happens. Also, I couldn't relate the whole chapter to the section 'alignment' as it feels closer to the subject on interpretability. *(Feedback from Thomas: unfortunately the paper is still under review and not public yet)*

Machine ethics and ML safety

The course provided a good overview on ethics, but it is unclear how it is relevant to practical ML safety. Does this allude to a consideration for loading ethical systems into AIs? As for the simplest system, utilitarianism, many ML systems are indeed programmed like utilitarians i.e. they are basically utility maximizers. However, they only care about their own utility (reward), and as far as I am aware, trying to point AIs to something that we care about remains an unsolved problem.

While the examples on the ETHICS test and the Jiminy Cricket Environment are interesting, I fail to understand how they are anything more than complicated pattern recognition with some given labeled dataset, and what their implications are.

Systemic Safety

What it is about

This section covers usage of ML for improved decision making, cyberdefense, and cooperative AI.

What I liked

The chapter on cooperative AI seems relevant to multipolar failure scenarios (e.g. this post by Paul Christiano and this post by Andrew Critch).

What I didn’t like

Relevance of improving decision making

The chapter on improving decision making mainly covers the topics relevant to forecasting, which seem to convey similar concepts as the chapter on interpretable uncertainty. The chapter is framed with the perspective that decision-making systems could be used to improve epistemics, reduce x-risks from hyper-persuasive AI, and more prudently wield the power of future technology, but I struggle to see these points being elaborated further throughout the chapter. While humanity can surely benefit from having better forecasting tools, I also fail to see the relevance of using ML to make good forecasts with how it helps with ML safety.

Relevance of cyberdefense

The chapter on cyberdefense introduces the dangers of cyberattacks, where it could make ML systems become compromised, destroy critical infrastructure, and create geopolitical turbulence. This makes me question my understanding of the term ‘ML safety’, is it about making humanity and everything we care about safe from the development of ML systems that may lead to potentially misaligned powerful AGI, or is it about making sure that ML systems won’t be hijacked by bad actors? My impression of the AI safety field is that it is mostly concerned with the former, but the latter may also be a problem worth worrying about, and I feel it would help if the problem statement were clear. *(Feedback from Thomas: See this post by Jeffrey Ladish and Lennart Heim. More broadly, suppose some AI developer is responsible, but others steal its models and use them irresponsibly, this could contribute to a misalignment crisis and x-risk event.)*

Relevance of cooperative AI

The chapter on cooperative AI describes the possible goal of using AI to help create cooperative mechanisms that help social systems prepare and evolve for when AI reshapes the world. I struggle to understand how AIs would be particularly insightful in helping us solve game theoretical problems any more than researchers in the field of economics. It is not clear to me how cooperative AI is relevant to ML safety.

X-risk Overview

What it is about

This section describes AI as posing an existential risk to humanity, gives a recap on factors that contribute to AIs being dangerous (weaponized AI, proxy gaming, treacherous turn, deceptive alignment, value lock-in, persuasive AI), and discusses how to steer AIs to be safer with good impact strategies and having the right safety-capabilities balance.

What I liked

Emphasizing safety-capabilities balance in my opinion is crucial, as it is important to advance AI safety without advancing capabilities, and there is often not a clear distinction between the two.

What I didn't like

Lack of concrete threat models

The chapter on x-risk does a recap on some topics from the previous sections and gives opinions of AI risks by famous people. I was hoping for more thorough arguments on why unsafe ML has the potential to lead to an existential crisis, beyond the surface level arguments of instrumental convergence and objective robustness failure. Of course having overly specific threat models may not always be a good thing as it may lead to oversimplified notion of the problem ("why don't we just do this - problem solved") and a lack of appreciation on the potential in which misaligned AIs can lead to existential catastrophes in unexpected ways. But I think it would still be great to have more elaboration on threat models and steps in which currently powerful ML systems may lead to dangerous AGI, and how the approaches covered in the course help to prevent x-risk. The most concrete example I can find in the course is weaponized AI, but it is more difficult to relate how deceptive AI and persuasive AI can lead to x-risk. *(Feedback from Thomas: Note that this section is as yet incomplete. It should be complete by the end of the fall.)*

Balancing safety vs capabilities

The chapter on influencing AI systems of the future rightly calls for advancement of safety without advancing capabilities, and details a list of examples of capability enhancements. However, it is less clear how the techniques described in the course are not capability enhancing. For example, while adversarial robustness helps to prevent OOD behavior, they inadvertently improve the capabilities of image classifiers. With regards to the discussion on safety metric vs capabilities metric, it isn't entirely clear to me how AUROC is mainly a safety metric while ImageNet Accuracy is a capabilities metric. *(Feedback from Thomas: As mentioned above, adversarial robustness does not enhance capabilities, though anomaly detection is a bit more correlated with general capabilities.)*

Final Thoughts

Having a course that focuses on ML safety is definitely a good step forward in the AI safety field, and it is great that such a course has been developed by people experienced in ML. While my review of the course may have been overly critical given that it was only very recently (last month) rolled out with several chapters still in progress, I see areas in which it can be improved.

From the perspective of an individual participant, I would hope for the following improvements in perhaps the next iteration of the course:

1. Have a clearer problem statement that this course is trying to address, i.e. what exactly is the alignment problem and how the various ML systems today have the potential of bringing us to an existential catastrophe.
2. Have a more systematic safety framework that addresses the problem statement, and how the various components of the framework are distinct, e.g. how is alignment different from robustness, what exactly does systemic safety entail etc.
3. Give more clarity on how the various ML safety techniques address the alignment problem, and how they can potentially scale to solve bigger problems of a similar nature as AIs scale in capabilities
4. Give an assessment on the most pressing issues that should be addressed by the ML community and the potential work that can be done to contribute to the ML safety field