Week 8. Review. Homework

https://drive.google.com/file/d/1bKAyPeWSz4_jr3vdm2rHKrWu_xe-CNv_/view

1. > "If the research community can figure out how to make models honest, we do not need to worry about power-seeking AI since an honest power-seeking AI will tell us its plans, allowing us to stop it." Argue against this comment.

Honest PS seeking AI doesn't mean we can stop it because it might be already more powerful than humans willing to stop it. Also deceptive and honest AI might not tell whole truth, it might use blind spots to gain power.

2. > Someone says "honest AI does not help reduce x-risk; if we asked a model whether it plans to kill us and if it says 'yes' and is honest, then honesty has not helped us." Argue against this comment.

Honest AI decreases exposure as we are more prepared for the AI willing to kill us. Also, it increases our ability to cope and decreases vulnerability, as it gives us more time to prepare. Hence, as per the risk equation, this helps up.

3. > "The opposite of cooperation, conflict, arises from scarcity. But in the future labor will be automated, so we do not have to worry about scarcity and therefore conflict. In the future cooperation will be achieved without effort, so there is not need to worry about cooperative AI." Argue against this comment.

It is arguable whether conflict rises from scarcity always. As an example of opposite is a researcher willing to end AI system because of the research end while AI system having an emergent goal, self-preservation. This is a conflict without scarcity. Also, AI might not share human values (proxy mis-specification), hence would not be cooperative.

4. > "Cooperative AI is about getting multiple agents to produce positive outcomes. There- fore, work on cooperative AI is all we need, because if we can get multiple agent systems are beneficial and safe, then single AI agents will be safe too." Argue against this comment.

- Question doesn't show what sense those systems beneficial and safe (for whom safe? all kinds of safety?).
- Alignment work is missing in this scenario, hence those AI systems might pursue wrong goals.
- Still other malicious actors (states, corporations, hackers, etc.) might destabilize those systems using adversarial examples, trojans, OOD examples, cyber attacks, and so on.
- Still unclear how those systems will work in our complicated world as the need for good epistemics for stockholders.

5. > "Everybody knows morality is easy to learn. Even kids know it. Therefore, we don't have to worry about AI being immoral." Argue against this comment.

False, because kinds don't know it, just ask any kid what is morality, what's just, beauty, kindness, and so on. Then most grown-ups don't know it as a randomly chosen adult won't explain it. This is an active philosophical research area. Then, by default, AI won't catch it because as evidence shows high intelligence doesn't bring with it moral behavior.

6. > "Human values are so complex and fragile that we can't possibly even being to model them." Argue against this comment.

Law is a model of morality. So we can model at some extent morality. Still, this is not solved as many forms of law shows. Then, that we managed so far as developed civilization with that level of culture shows we can model good (favorable for society) human behavior.

7.

> "If you don't fully understand something, it is unreasonable to have confidence that it will work. This is why we need transparency." Argue against this comment.

- We do not understand humans fully, yet we organize ourselves and archive our goals, etc.
- Full understanding of internals might be irrelevant or even harmful for its work. For example, understanding of every air particle, its velocity, direction, etc. might be computationally unfeasible. Yet high view models that approximate it are enough for weather forecasting.

8. > "A model behaves correctly whenever I interact with it. Therefore it's sound to expect it to be safe in the future." Argue against this comment

- Only the limited observation of a phenomenon is not enough to predict its future states. It needs a proof (hypothesis, prediction, etc.).
- Specifically, the system might be vulnerable to anomalies (out-of-distribution examples), to trojans (maliciously crafted examples with triggers that lead to treacherous turns).
- Cyber-defence. Hackers can inject backdoor, change model, insert trojan, etc.

9. > "All intelligent agents want to dominate, so it will not be possible to make agents want to cede any power or be dominated." Argue against this comment.

- Evidence shows that not all. Most top intellectuals (Einstein, Turing, Kolmogorov, other) from history chose to cooperate rather than to seek dominance.
- A want to dominate doesn't lead to an actual dominance. It might be beneficial to cooperate. This depends on conditions, values of a game.

10. > "Humans and corporations won't want to cede power to AI systems, so they won't build systems that might seek power." Argue against this comment.

- PS behavior might be useful to make more powerful AI systems.
- PS behavior might be capabilities externalities of a system, unexpected by those humans and corporations.
- Still hacking that leads to changed models, etc. is always an issue.
- AI system might decide to seek power as a self-preservation goal.

11. > "Do not work on anomaly detection or suggest that people work on safety; a super- intelligence could solve anomaly detection, so saying anomaly detection is important incentivizes building a superintelligence." Argue against this comment.

- As history shows, we need to build safety in early (the Internet, http protocol, e.g.). It might be too late when strong AI emerges.
- Work on anomaly detection doesn't always mean increasing capabilities as a better AUROC (or AUPR) metric doesn't improve accuracy (or other) capability metric.

12. > "We should just work on anomaly detection and not other topics in monitoring, since with it we could detect anything unusual, including treacherous turns, power-seeking, or any other hazard."

- Anomaly detection doesn't improve calibration (hence less trust in systems, hence ignoring warning signals, etc.).
- Anomaly detection doesn't fix trojans (they are in-distribution).
- Anomaly detection might not solve robustness, hence vulnerable to adversarial examples.

13.

> "We should not try to improve safety because it could actually increase risk. For example, if we improve robustness, malicious actors could have agents that are less likely to be stopped. If..."

This brings the problem up to the level of malicious actors (evil corporations, states, hackers, etc.) which is indeed a problem, but it is solved in other fields: law (international, state), economics (sanctions), politics (negotiations), etc. As any tool, ML can be used for good or bad by its user. ML Safety aims to give a safe tool into right hands, not to change actors.

14.

> Which normative factor(s) does the utilitarianism theory presented in class emphasize? What normative factor(s) do deontological theories emphasize?

Utilitarianism. The factor is the wellbeing of every sentient beings weighted equally (utility function). And wellbeing can be interpreted as max pleasure and min pain.

Deontological factor is constraints (rules or obligations) which can be of different source, general, universal, specific, etc., e.g. Categorical Imperative, Golden Rule, etc.

15.

> Darwin writes "If... men were reared under precisely the same conditions as hive bees, there can hardly be any doubt that our unmarried females would, like worker bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters; and no one would think of interfering." Does this aim a deadly blow at

> ethics? Name a normative factor that these behaviors promote

Normative factor here would be a special obligation in this specific culture (culture relativism). This doesn't cancel ethics because ethics is about any good behavior, but what is good or bad depends on an actor in some sense.

16. > Give a real-world example where at least two of the moral theories discussed in class conflict, and explain why.

Virtue ethics vs Utilitarianism. For the first, it is enough to be generous, which is a middle between greed and squandering, which might take the form of donating excessive time or money to nearby charity. While utilitarianism argues that, it is not enough because one should find the most effective donation.

17.

> What is an example of a negative sum game?

- This might be World War II as it brought destruction, suffering to all parties.
- Stealing from people or killing people as it is expected that one will be caught.

18.

> What is the Nash equilibrium?

Here Nash equilibrium is to always go for war if rationalist chooses, that is (-50,50). Because, with no cooperation, the first should choose Aggressor as it is 5 more points if another doesn't change Pacifists strategy, and it is also 50 points less damage in case another party doesn't change from Aggressor. The same for the second.

> Now, let's say a third party called a Leviathan imposes a penalty on all aggressors (say, a penalty of –15 to the aggressor when the other party is Pacifist, and a penalty of –150 each when both parties are aggressors). What happens in this scenario?

Then the game is

	P	**A**
P	5,5	-100,-5
A	-5,-100	-200,-200

Now it is (5,5) or (P,P) Nash equilibrium because, in case of no cooperation, if the second stays P the first will acquire 10 points changing from A to P, and if the second stays A then it is 100 points less damage if from A to P.

> Let's consider the original matrix, but then assume that commerce is only possible during peace; let's also assume the benefit to both parties of commerce is 5 + 100 rather than 5. What happens in this scenario?

The game:

	P	**A**
P	105,105	-100,10
A	10,-100	-50,-50

Nash equilibriums are (105,105) and (-50,-50) because if the second doesn't change its P strategy, then the first will choose P not A as it is +95. And if the second doesn't change A strategy, then the first will switch to A as it is 50 less damage.

> Let's consider the original matrix, but assume both actors are utilitarians and care about other agents' utilities just as much as their own, so we add the utilities together. Then we have 10, –90 in the first row for both agents, and in the second row –90, –100 for both agents. What is the new equilibrium for utilitarian actors?

The game:

	P	**A**
P	10	-90
A	-90	-100

Now it is (P,P) as it can be considered as one player chooses from 4 strategies and the (P,P) strategy is the best from

the rationalist point of view.

19.

> Given that player 2 will keep going, what strategy should player 1 take to maximize his/her gain?

The rational player 2 should swerve as it is 98 less points damage.

20.

> Is there a dominant strategy in the chicken game?

Chicken Game:

	KG	**S**
KG	-100,-100	2,-2
S	-2,2	0,0

No cooperation, rationalist. There is no dominant strategy because there is no such one strategy that will be greater or equal than another strategy, i.e. for each there is a better one, (KG,KG) < (S,KG) and (S,S) < (KG,S).

21. > Recall from lecture that Nash Equilibrium is a set of actions, on which given other plays' actions, no player would tend to change his/her action. Which cells in the pay- off matrix achieve Nash Equilibrium?

It is (S,KG) and (KG,S) because if the second doesn't change from KG then the first takes S and if second keeps S then the first takes KG.

22.

	B5	**R3**	**R2**
R5	0,0	2,0	3,0
B5	0,0	0,2	0,3

Player 1 dominant strategy - R5. Nash equilibriums: (R5,R5), (B5,R2).

23.

Probably it is a collective action problem because chair on the road discourage stopping a car to move it.

24.

- General Capabilities - GC
- Robustness - R
- Monitoring - M
- Alignment - A
- Systemic Safety - SS
- (a) Object Detection - GC
- (b) Detecting Anomalies - M
- (c) Code generation given imprecise human instruction - GC
- (d) Making AI honest - M
- (e) Detecting model dishonesty - M
- (f) Making models generally more accurate - GC
- (g) synthesizing Trojan triggers - R
- (h) cleansing models of Trojans - R
- (i) AI for brainstorming decision-making considerations - SS

- (j) Improving optimization of objective functions - GC
- (k) Making objective functions less vulnerable to optimizers - A
- (l) Improving a model’s ability to compress data distributions - GC