1 Review

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.

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.

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.

4. "Cooperative AI is about getting multiple agents to produce positive outcomes. Therefore, 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.

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.

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

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.

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.

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.

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.

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

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.”

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 we improve monitoring, malicious actors could inspect models with transparency tools, find their weaknesses, and exploit the models of benevolent actors. If we improve alignment, malicious actors could flip the sign of the human value function and direct models to be maximally harmful; malicious actors could use deception detectors to get better at deception. If we improve systemic safety, malicious actors could use forecasting tools to better anticipate the future and achieve their nefarious plans.” Argue against this comment.

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

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.

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

17. What is an example of a negative sum game?

18. Analyze the dilemma in the matrix below. What is the Nash equilibrium? 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?
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? 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?

		Other’s Choices	
		Pacifist	Aggressor
Own Choices	Pacifist	Peace, Peace $(5, 5)$	Defeat, Victory $(-100, 10)$
	Aggressor	Victory, Defeat $(10, -100)$	War, War $(-50, -50)$

The following is the payoff matrix of the game Chicken:

		Player 2	
		Keep Going	Swerve
Player 1	Keep Going	$-100, -100$	$2, -2$
	Swerve	$-2, 2$	$0, 0$

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

20. Is there a dominant strategy in the chicken game?

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 payoff matrix achieve Nash Equilibrium?

22. Consider the following simple card game. There are two colors for the cards, red and black. Each card has a number on it. Player 1 is given a red 5 and a black 5, while player 2 is given a black 5, a red 3, and a red 2. The game they are to play is the following: at a given signal the players simultaneously expose one of their cards. If the cards match in color, player 1 wins the (positive) difference between the numbers on the cards; if the cards do not match in color, player 2 wins the (positive) difference between the numbers on the cards played. Construct a payoff matrix for this game.

23. “Pretend that you’re driving on the highway and you see a chair in the middle of the road. You had to turn out of the way to avoid hitting the chair with your car. You consider stopping your car, getting out, and moving the chair out of the way. On the other hand, you’re late to work, and it’s not your fault that the chair is in the road. You have to figure whether to move the chair or not.” (A question from the course *Model Thinking*.) This is an example of a prisoner’s dilemma game, a collective action problem, a common pool resource problem, or none of these?

24. Using the categories described in the class, categorize the following research goals as one of these five: Robustness, Monitoring, Alignment, Systemic Safety, General Capabilities

    (a) Object Detection

    (b) Detecting Anomalies

    (c) Code generation given imprecise human instructions

    (d) Making AI honest

    (e) Detecting model dishonesty

    (f) Making models generally more accurate

    (g) synthesizing Trojan triggers

    (h) cleansing models of Trojans

    (i) AI for brainstorming decision-making considerations

    (j) Improving optimization of objective functions

    (k) Making objective functions less vulnerable to optimizers

    (l) Improving a model's ability to compress data distributions