Review

1. First, if an AI is too capable, it could be impossible to stop even if humans know its plans. Second, honest is not necessarily truthful or aligned. It could *think* it's doing things to help humanity (like making more paperclips!), but in truth it's severely harming humanity. Finally, it could miss some important things to be honest about. E.g., "I'm going to make some paperclips." A human asks, "how?" and it replies, "finding some metal, mining it, and using factories to turn it into paperclips," completely forgetting to mention that the metal and factories come from the ruins of cities it's about to bomb.

2. If we can create honest AI, then we the first AI's can be relatively harmless. This way, if we ask an AI, "do you want to kill me" and it answers "yes" we can just shut it off. We can work on existential risk with incapable AI's until we're confident they won't want to kill us when they're more capable.

3. First of all, people will still compete, even if it's not for food. But, AI don't have the same motivators. They want to optimize some function, and often other AI optimizing their goals will get in the way. For example, if one AI wants to make paperclips, and another wants to make staples, then they have to compete to use the limited amount of factories and metal to make their products. They *could* fly off to the asteroid belt to get more metal, but that takes time and doesn't give as much of a reward as stealing from the other AI.

4. If 90% of AI agents cooperate, they can check the 10% of defectors by jailing them. But the cooperation relies on multiple agents checking each other. If you have a single AI, it has nothing to check it and if it decides to defect it could do something disastrous.

5. Current morality is very cultural. For example, slavery was taught (and often thought) to be a moral thing in the 1800's South of America. Some cultures teach drinking to be immoral, others embrace it. It's not too hard to learn your culture, but that doesn't mean you're perfectly moral.

6. Humans can model human values pretty well, and AI is almost at the complexity of the human brain.

7. Humans don't fully understand each other, yet they trust each other to

do things correctly. Humans don't understand everything about biology (most drugs are just trial and error!), yet we trust pills to work correctly. Why not the same for AI?

8. Distribution shift can make the AI not know how to properly behave in the future.

9. Not all intelligent agents want to dominate. It can usually aid their goals, but you could change their goals so that they gain utility by *not* taking power.

10. The military is one human organization that absolutely wants an AI that can seek power, if only so it can better destroy other nations' militaries.

11. You can work on anomaly detection without building a superintelligence, in fact that's probably what you should do because you don't want your superintelligence to be vulnerable to a Trojan attack.

12. You might need an AI with far more capabilities to solve anomaly detection (which is dangerous if you can't stop it), but not to solve power-seeking or some other specific hazard.

13. It's a lot easier to stop a human malicious actor, especially if you have a benevolent super intelligence, than it is to stop a super intelligent malicious actor. If we improve safety enough, we can make AI systems incredibly difficult for malicious actors to turn.

14. Utilitarianism decides that things are right if they create more happiness, which leads to the Matrix. AI would be doing the right thing according to utilitarianism if it took over and drugged everyone on morphine, because then they would be so happy. Deontology decides things are right if they follow some pre-determined law or rule, which leads to value lock-in. At it's extreme, the law would never change, and so even if 100 years later humans think some of the previous laws are awful, the AI would think it's right to suppress those voices.

15. No, it doesn't destory ethics. Humans would not end up like that, which (the now lack of) slavery pretty strongly proves. The normative factor is, what is right is what you were taught, i.e. role ethics.

16. *Crime and Punishment:* If you rob and murder a spinster, and take the money for bettering yourself and many others, is it wrong? Utilitarianism says no, as the old lady was not getting much happiness in life, and in fact abusing and bringing pain to others' lives. Deontology says yes, because robbing and murdering is against the law.

17. Overfishing. If no one overfishes, everyone has enough fish. You can gain more utility by fishing twice as much, but if everyone does so, all the fish die out and no one can fish anymore.

18. The original Nash equilibrium is when both choose aggression, as you always gain more utility by being an aggressor. If the Leviathan imposes such a penalty on aggressors, then it always becomes benefical to be a pacifist, so the Nash equilibrium is pacifism. In the commerce scenario, the Nash equilibrium is to mirror your opponent. If they choose peace, you choose peace because the extra +5 from being aggressive doesn't counteract the +100 from trade, but also if the other person chose war, you only gain by also choosing war. If both agents are utilitarians, they will always choose peace.

19. Player 1 should swerve to get 98 more utility.

20. No. Your strategy is the opposite of your opponent's strategy, so if there was a dominant strategy it would need to be its own opposite.

21. (Swerve, Keep Going) and (Keep Going, Swerve).

P1\P2	B5	R3	R2
B5	0\0	0\2	0\3
R5	0\0	2\0	3\0

22. The Nash equilibrium is for player 1 to reveal R5 and player 2 to reveal R2.

23. It's most similar to the common pool resource problem. You can capture extra utility by continuing to drive to work, at the expense of other drivers' utility behind you.

24. (a) General Capabilities. (b) Robustness. (c) General Capabilities. (d) Systemic Safety. (e) Monitoring. (f) General Capabilities. (g) Robustness. (h) Systemic Safety. (i) Alignment. (j) General Capabilities. (k) Alignment. (l) Robustness.