Review, week 8

David McSharry

1. An honest power seeking model could manipulate us by selecting what truths it tells us, and with holding truths like what it's explicit plans are. This especially becomes a problem as AI gets more intelligent than us and we find it harder to determine it's motivations. An example of this could be a politician who is technically honest, but skirts questions when they are inconvenient in a way that is not technically not lying
2. This may be the case for super intelligent AI when we cannot affect what it will do, but in the case of a subhuman AI this would help us determine if it is truly aligned before allowing it to scale. Honest AI would also be useful in the case of super intelligent AI since other AI could keep the super intelligent honest in AI in check better (in some sort of collaborative AI situation).
3. Although there will be a lot *less* scarcity in the future, resources in the universe (and certainly the observable universe) are finite, and so they are still scarce. There may be a period of low scarcity as AI pick low hanging fruit for a while but this will invariably be replaced by a period of scarcity as this low hanging fruit runs out.
4. A multiple agent system that is beneficial and safe does not mean that a single dangerous AI (or subset of AIs) cannot eventually overpower the the system. They could be less constrained than the AIs trying to keep the system helpful and safe and so could be advantaged to this end. An example of this can be seen in democracies where there is a coup and then the systems goals change quickly.
5. Morality is not easy to learn, this can be seen in the fact that different people and societies diverge in their morals, and over time morals have changed a lot, for example it being acceptable to own slaves in the past until fairly recently. There is no reason to expect that this moral development is finished and so it seems wrong to say that it is "easy". Also, just because you can learn morals doesn't mean that you will follow them, so even if morals were easy to learn we should still worry about AI acting immorally.
6. I expect modelling human values to be hard, but humans for example can build a fairly robust model of human values by example when they are young, and these values are probably not built into us since a child can learn drastically different values depending of what culture they have grown up in. Even dogs seem to be able to build a model of human values to some extent.
7. There a lot of "black boxes" we interact with in out everyday lives that we don't understand but trust. For example a car. When I drive I have a very poor understanding of how the car works but I still trust that it will work. This is because I am confident it has been thoroughly tested on a distribution of situations representative of those I will come across on the road.

8. The only thing you proved is that in the distribution of your interactions with the AI it behaves reasonably, this cannot be extrapolated beyond this distribution without good reason. For example you may have by chance missed the fact that whenever the model is interacted with in a certain specific way it behaves incorrectly. Therefore this cannot be a guarantee that the model will behave safely it the future.

9. Since this is a strong claim about all intelligent agents, an example of an intelligent agent that wants to cede power will suffice to disprove it. A mother is an example of one such agent who cedes power to her baby, giving everything she has to ensure the survival of the baby. Power seeking is often instrumental to other goals, but not to every goal.

10. Power seeking is instrumental to many other goals, if you have more power then you can better achieve your goal. Therefore a wide array of goals will result in an agent seeking power even if this was not desired.

11. Anomaly detection can be worked on without the aid of a super intelligence and has so far been tractable. The accelerating effect of people building super intelligence to solve anomaly detection is not very convincing, either they care about anomaly detection and so are aware of what building a super intelligence that is not robust to anomalies could do, or they don't care about anomaly detection and so this will not be a further incentive to build super intelligence.

12. This is not the case because anomaly detection is not predictive to what a models actions will be which is an important factor for building safe AI. Anomaly detection adds to safety, but is by no means sufficient.

13. With each of these fields of safety, in general an advancement will lead to some increase in capabilities, but this must be weighed against how much safer they allow us to make AI. In some cases this trade off won't be worth it, and in some cases it will so we must tread carefully, but to make no advancement would mean the blind advancement of AI would continue anyway (maybe marginally slower) without the tools to control and understand the AI which seems worse than the counterfactual.

14. The main utilitarian normative factor is that of the "goodness of outcome" of an outcome according to some utility function. The main deontological normative factors are constraints like "don't kill" etc.

15. If this were true it would suggest that ethics are purely relative and there are no true morals we can aspire to. I don't necessarily think it's true however, and it's certainly not self evident so I don't think this is a deadly blow to ethics. These (bee-like) behaviours promote utilitarian normative factors since they suggest that deontological normative factors are arbitrary and constraints like "don't kill" are not truly part of our moral framework, instead we simply try to maximise some utility function. In this case doing what every body else is doing a pretty good yet brain-dead way of maximizing utility as going against ones society often brings heavy downsides.

16. In war time, a deontological ethical framework may say not to kill, but a utilitarian framework may say that by killing people you are improving the world by removing them. This is a contradiction between these two frameworks.

17. An example of a negative sum game is war, where no matter what happens both sides lose out (in particular when one side doesn't have anything to loot like in a long drawn out war or where no land is won by either side).

18. The Nash equilibrium is war,war. Both parties regardless of where they originally started will choose peace, and they will both end up on peace,peace. In the commerce case there will be two Nash equilibria and so both war,war and peace,peace will be stable. In the case that parties are utilitarians peace, peace has the highest utility and so this is the equilibrium.

19. Swerve since -2¿-100

20. No, the best strategy depends on what your opponent will do
21. There is no Nash equilibrium, someone will always change their strategy given the other players strategy.
22. see attached page
23. none of these