Additional notes for module 3

Recapping the ethical frameworks

In the videos, we talked about four ethical theories or frameworks. These were:

1. Utilitarianism: which tells us to *maximize the good consequences* of our actions, in terms of maximizing wellbeing (reducing harm and promoting good). Utilitarianism (which is a form of 'consequentialism') says that this maximizing calculation is the ONLY moral rule we should follow, even if it means that some people must be harmed to produce the Greatest Happiness. In making a calculation about what produces best overall net wellbeing, we need to take into account both the *magnitude* of all the harms and goods/benefits and the *probability* of all the harms and goods/benefits. Whatever produces the best result in these terms is the right action – all other actions are wrong! Utilitarianism is popular, but some people object that it means we have to be unfair or unjust if that maximizes net wellbeing. Utilitarians would respond that fairness or justice just *is* maximizing the best consequences!

For more information: https://youtu.be/-a739VjqdSI

2. Deontology: which tells us to follow a *range* of rules like Act Fairly, Tell Truth, Don't Kill, Don't Deceive, Keep Promises, Be Generous, Show Gratitude, Be Kind, Make Amends when you do Wrong. Deontology tells us to intelligently apply these principles to our actions and, if they come into conflict, to try to balance them as best we can. For example, if a telling the truth would lead to GREAT unfairness, then a deontologist might say we ought to favour the rule "Act Fairly" and override the rule "Tell Truth". It's a matter of working out which rule or rules is/are the strongest in a given case. There is no algorithm to do this! Rather, you need to apply the rules as best you can, and be prepared to defend this against challenge. For example, you could argue in a particular case that the dishonesty is rather minor, whereas the unfairness if very serious – or vice versa.

Many deontologists would accept that sometimes it is permissible to harm individuals when the stakes are very high e.g. to kill someone if that is the only way to save thousands. However, unlike Utilitarianism, Deontology says that we *cannot* just try to maximize good consequences if this means overriding important rules like Act Fairly etc.

Immanuel Kant had a special and an even more strict kind of deontology. He said that some moral rules are absolute and must never be broken. Kant said that we should *always* treat people as ends in themselves, and never merely as means (even for good ends). This means we can never deliberately harm or use someone even if that is for good reasons. In particular, we should, said Kant, always respect other people's *autonomy* (ability to make decisions according to their own values). We can be said to *use* someone as means rather than as ends when we harm them and override their autonomy. For example, we would be using someone as means and not as ends in themselves if we lied to them, broke a clear promise, or exploited their weakness, even if we were aiming to produce some other good result.

For more information: https://youtu.be/wWZi-8Wji7M and https://youtu.be/8bIys6JoEDw (on Kant)

3. Virtue Ethics: which tells us to judge whether our actions are right or wrong by asking what a person of good virtue would do—that is, a person who is generous, just, courageous, benevolent (inclined to be kind and not harm people). Instead of telling us to maximize wellbeing or simply consult and follow moral rules, virtue ethics tells us to look to the *example of role models*, people of morally impressive character, who have these virtuous character traits and always avoid vice (greed, unfairness, selfishness, bad temper, recklessness, cowardice, etc.) and act wisely. These people show us which rules are important and *how they ought to be implemented* in given contexts. For example, some people would think that impressive and wise people like Gandhi, Buddha, Nelson Mandela, etc. are characters we should look to for moral guidance e.g. for the way they stood up to injustice. Such people (some say) had virtues like courage, honesty, kindness, and fairness. An example from AI might be Timnit Gebru in standing up to powerful companies to produce fairer algorithms. In contrast, perhaps you might say that Mark Zuckerberg has not always shown good character in running Facebook! We could also consider the good characters of people in books/movies, etc. Note that in Virtue Ethics, we don't necessarily need to point to such a person – we could simply ask: "what would a really just, kind, honest etc. person do in this circumstance?"

For more information: https://youtu.be/PrvtOWEXDIQ

4. Ethics of Care: which tells us to look at the various *relationships* we or others are in (parents, families, colleagues, fellow citizens, caring relations, employer-employee, etc). Ethics of Care, which derives from feminist ethics, says that we must be especially (though not exclusively) aware of people who are vulnerable, at risk, marginalized, or need help. We should pay close attention to *contextual features*, such as historical oppression, hidden bias, power imbalances, and dependent relationships. For example, women, children, trans people, the disabled, people of colour, etc. are often more vulnerable and/or have had histories of oppression or neglect. Also, some relationships involve power imbalances—here, the party with power needs to be particularly attentive to the way they are able to exploit or use those with less power. Ethics of Care says we should not just follow abstract rules, but we also need to use our feelings and emotions. For example, we can feel *compassion* for vulnerable people, a *sense of care* for them, and feelings of indignation and anger when people are treated unfairly. Ethics of Care says that traditional moral theories (like utilitarianism and deontology) has neglected the important role of feeling and emotion in ethics.

For more information: https://youtu.be/wvrNdr5L_5Y

There are also other YouTube resources.

Theory	Some key elements
Utilitarianism	Maximize net wellbeing (best overall balance of harm (e.g. pain, distress, loss) and benefit (e.g. pleasure, enjoyment, fulfilment). Everyone's interests are to count equally if they are similar interests.
Deontology	Follow key moral rule like Act Fairly, Avoid Doing Harm, Be Kind, etc. When rules conflict, work out which principle(s) should override. Kant's strict version of deontology: always treat people as ends in themselves, never merely as means: *never* deceive or exploit them or undermine their autonomy.
Virtue Ethics	Ask what a person of wisdom and excellent character (e.g. fair, kind, generous, honest) would do. Look to virtuous exemplars to interpret how to be fair, courageous etc. in particular situations.

Ethics of Care	Look at relationships and their context. Consider the vulnerable, the dependent, the marginalized and oppressed, and relationships of responsibility. Think about responses that involve feeling such as compassion, care, and a strong sense of responsibility.

The ethical theories give us different ways of working out what is right and wrong. Different philosophers will prefer different theories or will think that some theories are stronger than others. You don't need to decide which is the best theory; but just be aware that they do offer different ways of determining right and wrong that can conflict with each other. For example, a utilitarian may think that it is a moral duty to kill one person if that is the only way to save 3 people; whereas a deontologist may say that it would be wrong to break a moral rule against killing even if it would save more people. A virtue ethicist may agree, because (they might say) a just and benevolent person would not do that. You want to be sure that if you use *more than one* moral theory to justify a point of view, that those theories are not actually saying opposite things!!

There is also significant overlap between theories. For example, they can all talk about fairness, doing good, avoiding harm, etc. But: they may talk about these things in different ways. For example, virtue ethics tells us to look for guidance from just, kind, honest etc. people. Other theories also talk about justice etc. However, virtue ethics tells us that the way to *interpret* these things is to ask what a person with virtue and good character would do. For example: Is it fair to use Northpointe's COMPAS algorithms to predict recidivism? A virtue ethicist would ask: Well, what would a person with an excellent and wise moral character say? Would a really just person approve of COMPAS? Would they want it banned? Or altered? (You may give different answers to this question, but the important thing is that you follow the recommendations of virtue ethics when you are using that theory—and similarly for the other theories).

Writing essays in Ethics

When doing ethics (as in this Assignment) you will be asked to make an *argument* for a point of view or for more than one point of view (e.g. that a particular kind of AI should or should not be developed or used in a certain way). This means you have to provide *good reasons* to back up your chosen point of view. You can use one or more of these <u>theories</u> to *justify* the point of view you have adopted. Since the theories use different standards to tell us how to determine what is right and wrong—e.g. utilitarianism has a different standard to deontology etc. (see above)—you will need to both *understand* what the theory says and also *decide* whether it might be useful to back up your chosen point of view. (Note that in this Assignment, you are being asked to adopt TWO points of view: one for and one against the use of the AI in the Options).

Furthermore, you can use one or more of the theories to *explain, justify and guide* the use of ideas or <u>principles</u> in AI ethics, such as trust, privacy, fairness, accountability, and safety. When you use an idea or principle like, say, 'fairness' or 'accountability' to back up your argument, you need to have an understanding of what fairness and accountability are and how they might apply to your case study. You can also use one or more of the ethical theories to show why fairness, accountability etc. are important and how they applies to the case.

Example. *AI soap dispenser that doesn't recognise darker coloured skin.*

You can argue, for example, that this technology is morally unacceptable with any of the four theories above.

Suppose your argument is: The AI soap dispenser should either be improved or replaced with a non-AI version. You might choose to claim (and note: this is just an example) that this is because it is biased in a way that makes it unfair. Then, you might justify this claim by arguing that (for example): (1) On utilitarian grounds it is unfair because it the soap dispenser causes harms x, y, z and leads to worse overall consequences than alternatives; (2) On deontological grounds it breaks the rules of Act Fairly and Avoid Harm; (3) On Ethics of Care grounds the soap dispenser is unfair because it further penalises a group (people of colour) who are already discriminated against in various ways; (4) On virtue ethics grounds: a person of virtue and fine character might say that *even if* the dispenser produces some benefits (e.g. it might reduce COVID-19 transmission since you don't have to touch the dispenser) it is still unfair to people of color.

This is only a simple example; and to write a good essay you would need to add a lot more *detail* to show how the theory you select backs up your position. Make sure the way you select and use the ethical theories of frameworks *really does* back up your position, and also explain how it does so. Avoid just throwing in any old reason to try to justify your position. Ask yourself: Could I defend this position if I was to discuss it with my tutor and my colleagues in a tutorial? Carefully select your reasons, including when you use the ethical theories. Writing ethics essays is about justifying your position with **good** reasons and explaining really *clearly* what those reasons are—and how they make your chosen point of view *stronger*. This takes practice.

For this soap dispenser example, we would (presumably) all agree that it is unfair to use it; after all, there are certainly other options. However, many other cases are less clear cut. Consider a medical application: laser technology that works better on lighter skin than darker skin. Perhaps there is no remedy or alternative to this: that is the nature of the lasers used. The only alternative is to remove this technology because it is unfair. However, it improves the outcomes of everyone regardless of their skin colour. Now, some people may argue that it would be unfair to use this technology. But others would argue that it is, despite its flaws, still fair to use it. We can then argue the point and provide reasons for and against. What would our four ethical theories/frameworks say? That is something you would need to explain and justify.

Note that in ethics essays it is common to use the first person ("In this essay, I will argue that…). When you select a point of view to argue, this point of view may be your own opinion, but it does not have to be

More detailed resources for the ethical theories can be found in the *Stanford Encyclopedia of Philosophy* and the *Internet Encyclopedia of Philosophy*, both freely available online. The latter has a section called "Ethics" for a general overview.

For a detailed explanation of how to write a good ethics or philosophy paper, see http://www.jimpryor.net/teaching/guidelines/writing.html