Toronto Womxn in Data-science, 2020. FairML workshop.

	Deontological ethics (duty ethics)	Utilitarianism (consequentialism)	Virtue ethics (teleological ethics)
Definition	• duty • ethics is about following the moral law • deontologists believe that ethics is governed by rules and principles and that we have a moral duty to adhere to these rules • universalise our thinking	• consequences • maximising expected utility • flexible to the circumstances of each decision • the most applied: the tech industry tend to be fond of it	• overall moral character • focused on ends or goals • organised around developing habits and dispositions that help persons achieve their goals and flourish as an individual • considers goodness in local rather than universal terms and emphasises not universal laws, but local norms • moral prudence • practical wisdom
Pioneers	Immanuel Kant	Jeremy Bentham and John Stuart Mill	Ancient Greek, Aristotle
Strengths	• established and clear ground rules (draw a line in the sand)	• equate well-being with wealth production and individual choices (close to American values) • seems somewhat quantifiable • maximising happiness is more accessible than lofty ideas of moral duty • particularly useful way to think about the implications of the new • balancing competing interests	• valuable guide to everyday moral choices • can be projected in technology itself
Weaknesses	• lofty • seen as occasionally obstinate • abstract	• insubstantive definition of "goodness" and the fact that it permits (and even invites) the consideration of particular problems in isolation from larger systems • may look too narrowly at who is affected by a given decision • sometimes struggles to protect individuals and minorities from oppression • let dubious behaviour slide if it doesn't cause any harm	• choosing and balancing appropriate virtue might be tricky

Typical questions	• How are rules applied to decisions? • What are the right rules? • What rules do we put in place in order to achieve our desired social goals? • What if everyone did what I'm about to do? • Am I treating people as ends or means?	• What is the greatest possible good for the greatest number? • What is the greatest possible balance of good over evil? • Am I maximising happiness for the greatest number of people? • Am I minimising pain?	• Who should I be? • What is the best form/ version of this particular thing, in these particular circumstances? • Would I be happy for my decision to appear on the front page of tomorrow's news? • What would someone infer about our character, hearing we made this decision? • What virtues am I demonstrating if I do this? or don't do this? • What values should a companion species live by? • How can it demonstrate them by its actions?

Fig. 2.9. Table of comparison of the three major schools of modern ethics (Bowles, C. (2018) Future Ethics. Hove, East Sussex: Now Next Press, pp.52–126).

General questions that could help your discussion further

- Can you identify the school of ethics behind the decision/idea?
- Can you think of unintended consequences that might arise in 1, 5, 10 years?
- What impact does your idea/solution have on different groups of people? Are they the same? Who fares worst in this system? Who benefits the most?
- Can you see any self-perpetuating biases that could arise as a result of this?
- What are the weaknesses and strengths of this decision/idea?
- Who are the stakeholders in this situation? Does it seem 'fair' when you assume the position of each stakeholder?
- Do a pre-mortem exercise on the idea.
  Imagine you rolled this out successfully and the decision has now impacted a lot of people positively. What went well? Can you think of the best-possible outcome?
  Now imagine the opposite, that it failed to have the intended impact? What led to it failing so terribly? Can you think of the worst-possible outcome?
- Suppose that the system failed and it has resulted in catastrophic results (the worst-case you just came up with). Who should be held accountable? Can anyone be held accountable? Why/why not?
- Who will want to abuse, steal, misinterpret, hack, destroy or weaponize what you've built? What rewards/incentives are inadvertently built-in for those people? How can we remove those rewards?

Now revisit your original idea and consider what you might change. Would you do anything differently? If so, what?

Toronto Womxn in Data-science, 2020. FairML workshop.

Scenarios

1. You are a product manager mandated with the task of implementing a fairness feature into your company's product offering. The designer in your team has come up with 3 different designs to detect, report, &/ measure fairness - a health report, a fairness slider and an auto update to push fairer models. What do you choose, if any?
    a. What kind of decision making does each approach drive?
    b. What decisions are taken-away/abstracted from the end-user in each?
    c. Who is accountable for outcomes of such a model/service?
    d. What are some downstream consequences that you can think of? What if there were malicious actors?
    e. How often would you monitor/update/revisit- the solution?
2. You are a developer tasked with implementing a fairness metric to detect biases in your models. The team-lead faced with the goal of ultimately delivering a well performing model, insists that you implement as many definitions as you can and measure your model(s) against all of them. You have since written an API to compute the scores for models over ~30 definitions of fairness. When it is time to make a choice on which model your team will deploy, the team makes a decision to choose the definition & metric that puts your model in best-light, i.e., the definition that reports a high score for fairness. What do you do?
    a. What are the potential consequences of this action on - the company and the customer?
    b. Imagine now that your company now made a public claim that your models are 'fair'. Would you do anything differently?
3. Your team has created a product that scores comments for toxicity, constructiveness, and healthfulness. Reddit, Twitter, and Facebook are all interested. Some journalists are worried that you'll destroy free speech and that's something a couple of your senior engineers have also voiced concerns about. What do you do?
    a. Do a pre-mortem exercise on the idea. Imagine you rolled this out successfully and the decision has now impacted a lot of people positively. What went well? Can you think of the best-possible outcome? Now imagine the opposite, that it failed to have the intended impact? What are the bottlenecks that would lead it to fail terribly? Can you think of the worst-possible outcome?
    b. Who will want to abuse, steal, misinterpret, hack, destroy or weaponize what you've built? What rewards/incentives are inadvertently built-in for those people? What can we do differently to remove those rewards?
4. You are a professor who specializes in computer vision. Several of your grad students have recently acquired millions of mugshots, annotated for the level of crime--so going from minor disorderly conduct stuff to murder. They'd like to use image processing to build a criminality detector. The police departments that provided them the mugshots are really really interested and one of their cousins works for a VC that would happily fund it. Given that there is a market for it, there are many labs you can think of who will be willing and able to build such a system even if you don't. What do you do?
    a. Should the government be allowed to regulate your work as a researcher?

   b. Who will want to abuse, steal, misinterpret, hack, destroy or weaponize what you've built? Can you think of safe-guards that you could build into the system?
5. Imagine you work for an insurance provider as a data-scientist. You've found that the intersection (i.e, postal-code) is one of the most important features that help in identifying risk. Your models drive a lot of value and are now on track to be launched in Atlanta. You notice pretty soon that certain communities are disproportionately affected by the introduction of your model. Upon further inspection you discover that Atlanta is highly segregated city and postal code serves as a very effective proxy for race. By law, it is illegal to discriminate against someone by race, but your model isn't using race as a feature. What would you do?
6. What is the role of transparency & explainability in AI? Consider Google's Duplex project, where the AI trained is so adept at holding a conversation over the phone that it could place reservations at a restaurant or book your flight tickets on your behalf. Should we tell people they're talking to a bot on the phone?
   a. In many countries, companies are obliged to declare it when a call is being recorded. Is it the role of the government to enforce this with companies to tell their users that they are talking to an AI?
7. What is the role of the government in regulating AI? Law moves a lot slower than technological advancements. How should governments balance between technological innovation/progress and ensuring the safety of its citizens?
   a. How often should governments revisit laws?
   b. Should governments trade privacy for more security or vice-versa?
   c. Should the government have an opinion on biases, fairness in public services?
   d. What could go wrong with all of these choices? Can you identify the communities that will be most severely affected/impacted by each of your choices?
8. There are several settings where algorithmic programs have been tested and shown to be biased in harmful ways. Let's consider 2 examples
   a. Targeted Advertising: Retailers that collect purchasing information from its customers can and do use this information to predict individual shoppers' choices and habits, and advertise accordingly. This targeting of advertisement can take other forms as well: it has been shown that Google shows ads for high-paying jobs to women much less frequently than to men.
   b. Sentencing Software: Judges in parole cases are using algorithms to determine the likelihood that a given individual, who is being considered for parole, will re-offend if released.In the case of one widely-used program, it was shown that black defenders are twice as likely to be flagged incorrectly as high-risk, whereas white defenders are twice as likely to be incorrectly flagged as low-risk

   In each of these scenarios - Who is accountable?
   - Is it the responsibility of the organization that supplies (biased) historic data, and should they be penalized, if the analytics produce biased recommendations?
   - Is it the responsibility of programmers who do not screen for biases in the data?
   - Is it the responsibility of the clients who buy and use analytics to check for bias?
   - Is it the responsibility of lawmakers and regulatory bodies to rule against, and checkfor, the use of biased analytics and algorithmic decision making?