**Module 2: Bias within algorithms (1 lesson)**

This module will reference and expand on the course referenced in the proposal (alan-turing-institute/bias-in-AI-course (github.com)). This module is expected to contain minimal overlap with the referenced course. However, we propose to keep this module so learners can choose this as a standalone course without necessarily having to complete the previous course. This will expand the audience for this course. Additionally, we can make this course optional for people who have completed the previous course.

1. Develop an understanding of the importance of fairness (e.g., as an intrinsic value (important in and of itself) or an instrumental value (important for what we can achieve with it));

Although fairness is often considered to be synonymous with bias mitigation and non-discrimination, there may be more to fairness than these other concepts described. There’s no question that discrimination is unfair, but that doesn’t mean the terms are equivalent. In what follows we will help you understand more about fairness and fairness in AI.

Intrinsic value has traditionally been thought to lie at the heart of ethics. Philosophers use a number of terms to refer to such value. The intrinsic value of something is said to be the value that that thing has “in itself,” or “for its own sake,” or “as such,” or “in its own right.” Instrumental value is the value of something to achieve some purpose. It is an “instrument” to achieve this purpose but has no value in and of itself.

Historical Background:

The question “What is intrinsic value?” is more fundamental than the question “What has intrinsic value?,” but historically these have been treated in reverse order. For a long time, philosophers appear to have thought that the notion of intrinsic value is itself sufficiently clear to allow them to go straight to the question of what should be said to have intrinsic value. Not even a potted history of what has been said on this matter can be attempted here, since the record is so rich. Rather, a few representative illustrations must suffice.

In his dialogue Protagoras, Plato [428–347 B.C.E.] maintains (through the character of Socrates, modeled after the real Socrates [470–399 B.C.E.], who was Plato’s teacher) that, when people condemn pleasure, they do so, not because they take pleasure to be bad as such, but because of the bad consequences they find pleasure often to have. For example, at one point Socrates says that the only reason why the pleasures of food and drink and sex seem to be evil is that they result in pain and deprive us of future pleasures (Plato, Protagoras, 353e). He concludes that pleasure is in fact good as such and pain bad, regardless of what their consequences may on occasion be. In the Timaeus, Plato seems quite pessimistic about these consequences, for he has Timaeus declare pleasure to be “the greatest incitement to evil” and pain to be something that “deters from good” (Plato, Timaeus, 69d). Plato does not think of pleasure as the “highest” good, however. In the Republic, Socrates states that there can be no “communion” between “extravagant” pleasure and virtue (Plato, Republic, 402e) and in the Philebus, where Philebus argues that pleasure is the highest good, Socrates argues against this, claiming that pleasure is better when accompanied by intelligence (Plato, Philebus, 60e).

Many philosophers have followed Plato’s lead in declaring pleasure intrinsically good and pain intrinsically bad. Aristotle [384–322 B.C.E.], for example, himself a student of Plato’s, says at one point that all are agreed that pain is bad and to be avoided, either because it is bad “without qualification” or because it is in some way an “impediment” to us; he adds that pleasure, being the “contrary” of that which is to be avoided, is therefore necessarily a good (Aristotle, Nicomachean Ethics, 1153b). Over the course of the more than two thousand years since this was written, this view has been frequently endorsed. Like Plato, Aristotle does not take pleasure and pain to be the only things that are intrinsically good and bad, although some have maintained that this is indeed the case. This more restrictive view, often called hedonism, has had proponents since the time of Epicurus [341–271 B.C.E.].[1] Perhaps the most thorough renditions of it are to be found in the works of Jeremy Bentham [1748–1832] and Henry Sidgwick [1838–1900] (see Bentham 1789, Sidgwick 1907); perhaps its most famous proponent is John Stuart Mill [1806–1873] (see Mill 1863).

Most philosophers who have written on the question of what has intrinsic value have not been hedonists; like Plato and Aristotle, they have thought that something besides pleasure and pain has intrinsic value. One of the most comprehensive lists of intrinsic goods that anyone has suggested is that given by William Frankena (Frankena 1973, pp. 87–88): life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one’s own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc. (Presumably a corresponding list of intrinsic evils could be provided.) Almost any philosopher who has ever addressed the question of what has intrinsic value will find his or her answer represented in some way by one or more items on Frankena’s list. (Frankena himself notes that he does not explicitly include in his list the communion with and love and knowledge of God that certain philosophers believe to be the highest good, since he takes them to fall under the headings of “knowledge” and “love.”) One conspicuous omission from the list, however, is the increasingly popular view that certain environmental entities or qualities have intrinsic value (although Frankena may again assert that these are implicitly represented by one or more items already on the list). Some find intrinsic value, for example, in certain “natural” environments (wildernesses untouched by human hand); some find it in certain animal species; and so on.

**Suppose that you were confronted with some proposed list of intrinsic goods. It would be natural to ask how you might assess the accuracy of the list. How can you tell whether something has intrinsic value or not?**

Now back to fairness. Is fairness an intrinsic value –that is, it is valuable for its own sake-or it is an instrumental value—that is, it is valuable because it leads to things that we find valuable. We clearly want fairness because it leads to things that we value. If people are treated fairly, then they are content and if treated unfairly they are not. In this way, fairness leads to acceptance or even stability of circumstances.

On the other hand, perhaps we value fairness for its own sake. Perhaps we seek to make things fair because a lack of fairness is wrong. It is not wrong because it leads to other things; perhaps it is wrong simply in and of itself.

1. The social implications of fairness within different domains;

We might be able to answer the question of whether fairness is an intrinsic or instrumental value by examining specific contexts. Why is fairness important in context such as policing and justice, education and healthcare? And what are the consequences of unfairness in these domains?

The UK College of Policing states “As a police service, we must show impartiality throughout all our dealings with colleagues, partners and members of the public. This is achieved by being unprejudiced, fair and objective. We consider different sides of a situation and ensure that each side is given equal consideration.”

Now imagine that a police force wants to use an AI tool for combatting human trafficking. How can developers ensure that the tool is fair and aligns with the College of Policing standard on fairness?

Reading: <https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/>

Reading: <https://www.latimes.com/california/story/2022-07-04/researchers-use-ai-to-predict-crime-biased-policing>

Unfortunately, unfairness, bias and discriminatory outcomes can occur in any context. Recently, the research community of fairness in ML has highlighted that ML systems can be biased against certain sub-populations, in the sense that they present disparate performance for different sub-groups defined by protected attributes such as age, race/ethnicity, sex or gender, socioeconomic status, among others. This occurs in the healthcare context as well.

Reading: <https://www.nature.com/articles/s41467-022-32186-3>

1. Identify conflict between fairness and other ethical values and ways to address and deal with them,

Even if we agree that fairness is a value that we ought to strive to implement in our AI tools, it doesn’t follow necessarily that it won’t come into conflict with other values. When building ethical AI, there are several values that are going to relevant to any given context. Sometimes these values conflict.

For example, it is possible for fairness to conflict with autonomy. During the Covid-19 pandemic, lockdowns were implemented in Europe and across the globe in an attempt to curb the spread of the virus. These lockdowns were applied to everyone equally regardless of risk. In this sense, they might be considered to be fair. But they also drastically inhibited people’s autonomy by prohibiting a whole host of activities involving movement or being with others. One could make the argument that the lockdowns should not have been applied equally to all. If a person leads a low-risk lifestyle by working at home and avoiding large crowds, why shouldn’t they also be allowed to travel or meet with more than 3 friends? Or is the fairness of the application of the laws more valuable than protecting autonomy?

Also during the pandemic, we witnessed a question with significant import: Who ought to ge the vaccine first? Is fairness the relevant value in this context? Was it fair for elderly people to get the vaccine first, or was a different value at play?

Reading: <https://www.nytimes.com/2020/12/05/health/covid-vaccine-first.html>

1. Explain why conceptualisation of fairness are important to algorithmic decision-making;

As we’ve seen throughout this module, we must understand fairness and we ought to strive to implement it into our development and use of AI tools. The consequences of not doing so can be devastating to individual rights and societal wellbeing.

Reading: <https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine>

Apply learnings of fairness to different use-cases from one of TRI’s own projects. For example, D4Fly which use AI for border security.

Case study: D4FLY was a research and innovation project funded by the EU Horizon 2020 programme “Secure societies – Protecting freedom and security of Europe and its citizens”.

The project focused on enhancing the quality and efficiency of identity verification at border crossings in all modalities: land, air, and sea by providing faster and more secure border control solutions.

The project consortium was formed by 19 partners from 11 European countries including universities, SMEs, research institutes and border control authorities.

Research topics included:

Biometric technologies for the verification of traveler’s identities on-the-move through 3D face, iris, and somatotype recognition

Smartphone applications for enhanced traveler verification

Thermal and multispectral imaging for counter spoofing

Advanced morphed face detection algorithms through Convolutional Neural Networks

Deep Neural Networks and advanced methods for impostor and document fraud detection

Computer vision algorithms for automated passport forgery detection

Ethical, legal, social impact assessment of the developed technologies

Trilateral Research led the ethics work for this project and sought to ensure the tools were developed and could be used fairly. It did this by appealing to regulations such as the European Charter on Human Rights, ethical theories and close collaboration with developers and end users.

In order to mitigate bias in the facial recognition tools, Trilateral asked developers to train the AI on as diverse a dataset as possible and to test and communicate the error rates of the system when used by individuals of diverse ethnic backgrounds and women.

Further, Trilateral established avenues to make end users aware of the error rates for these tools when used to verify the identity of individuals with diverse ethnic backgrounds and women. It was made absolutely clear that a negative verification result does not indicate any kind of criminal suspicion.

It is important to note that a focus on bias mitigation was included both in the development phase of the tool and in its deployment. The lesson is that ethical values must be included in both phases of AI creation and use.

Pre-reading:

<https://www.thoughtco.com/intrinsic-and-instrumental-value-2670651>

<https://www.vox.com/future-perfect/22916602/ai-bias-fairness-tradeoffs-artificial-intelligence>

<https://hbr.org/2020/10/ai-fairness-isnt-just-an-ethical-issue>

<https://levity.ai/blog/ai-bias-how-to-avoid>

<https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/>

Reflection Sheet

1. If you had to make a list of intrinsic values, which ones would you choose, and why?
2. Do you think fairness is an intrinsic or instrumental value? Why?
3. You read 2 articles about using AI in policing. Is it possible to create fair and non-discriminatory AI for predictive policing? Why, or why not?

Multiple Choice

1. Intrinsic value is...
   1. the value that a thing has “in itself,” or “for its own sake,” or “as such,” or “in its own right.”
   2. The value that a thing has for the purpose of achieving something else of value
   3. Hard to define
2. Can fairness conflict with other values?
   1. No
   2. Yes
   3. Yes, but fairness is always the most important ethical value
3. What did some philosophers think is intrinsically valuable?
   1. Pleasure/Happiness
   2. Sleeping a lot
   3. Nothing
4. Why is understanding fairness important to AI tool development?
   1. Because we are told so
   2. Because risks of unfairness occur in most societal contexts
   3. Because AI is inherently unfair

Task:

Write a three paragraph pitch to persuade a client to consider autonomous AI production because you have mitigated bias. Show that you understand what fairness is, how it might conflict with other values, but how it can still be achieved.

Task:

Re-read the Amazon case study and answer:

1. How could this have been avoided?
2. Using the principles learned in this module, make an argument that Amazon is culpable for neglecting bias.