Reading Response

The Ethics of Algorithms - Mapping the debate Mittelstadt (et .al)

This reading response will center on the paper *The Ethics of Algorithms: Mapping the Debate* (Mittelstadt1 et al. 2016). The author gives motivations for their paper through observations on how decisions and choices previously left to humans have now been increasingly delegated to algorithms.

The paper centers on algorithms with hard to explain decision-making logic. It does not discuss algorithms that automate tedious tasks such as printing a message on a screen multiple times. The main contribution it makes is a mapping of ethical concerns for these algorithms.

Six types of ethical concerns in particular are proposed by the author. These six concerns are: inconclusive, inscrutable, misguided evidence, unfair outcomes, transformative effects, and traceability. Inconclusive evidence is the result of probable but uncertain conclusions from machine learning algorithms. These conclusions are sometimes seen to have no connection with the data it utilized to reach them, insinuating the evidence is potentially inscrutable as well. Being so heavily reliant on the input data, conclusions can additionally be reached based on unreliable input data, misguiding the evidence. In a perfect scenario where all the evidence were perfect, there could still be unfair outcomes from the algorithm's actions that could be discriminatory. These activities can also lead to transformative effects for the rest of the society, for example motivating actions for humans based on the insights these activities generated.. Finally, traceability within machine learning algorithms should also be addressed on how to assign responsibility for negative outcomes. This is especially difficult due to a genuine mistake such as a lack of intentionality with a bug in the code or a true unethical bias from the developer.

A critical point

One critical point I would like to develop more about the reading is the traceability ethic concern. The reading defines traceability as: “a way to assign responsibility when there is a negative outcome”. It makes a strong argument that there should be a way to assign responsibility when harm is caused. Without a concern for responsibility, developers of algorithms are free to neglect all ethical concerns since they bear no consequence or responsibility when a negative outcome occurs.

The reading also makes a strong argument that it is rarely straightforward to identify who should be held responsible for the harm caused. The reading gives support for this through a programmer’s point of view. The task of dealing with complex voluminous amounts of code can easily lead to unintentional bugs in their code that may cause harm. This makes it difficult to assess if the programmer had an intentional means to cause harm or not.

The reading introduces a counter-example from existing work that states that learning algorithms should be considered moral agents with a degree of moral responsibility. This is an interesting design philosophy to take but ultimately a weak one. In order for artificial agents to bear moral responsibility, they need to first exhibit an understanding for ethics and morals which has not yet been achieved. Artificial agents do not exhibit the same level of maturity or intelligence as humans do, which further suggests that they should not be held to the same standard of moral obligation. Furthermore, shifting blame to artificial agents also allows humans to avoid responsibility for the initial design of these artificial agents and their learning algorithms.