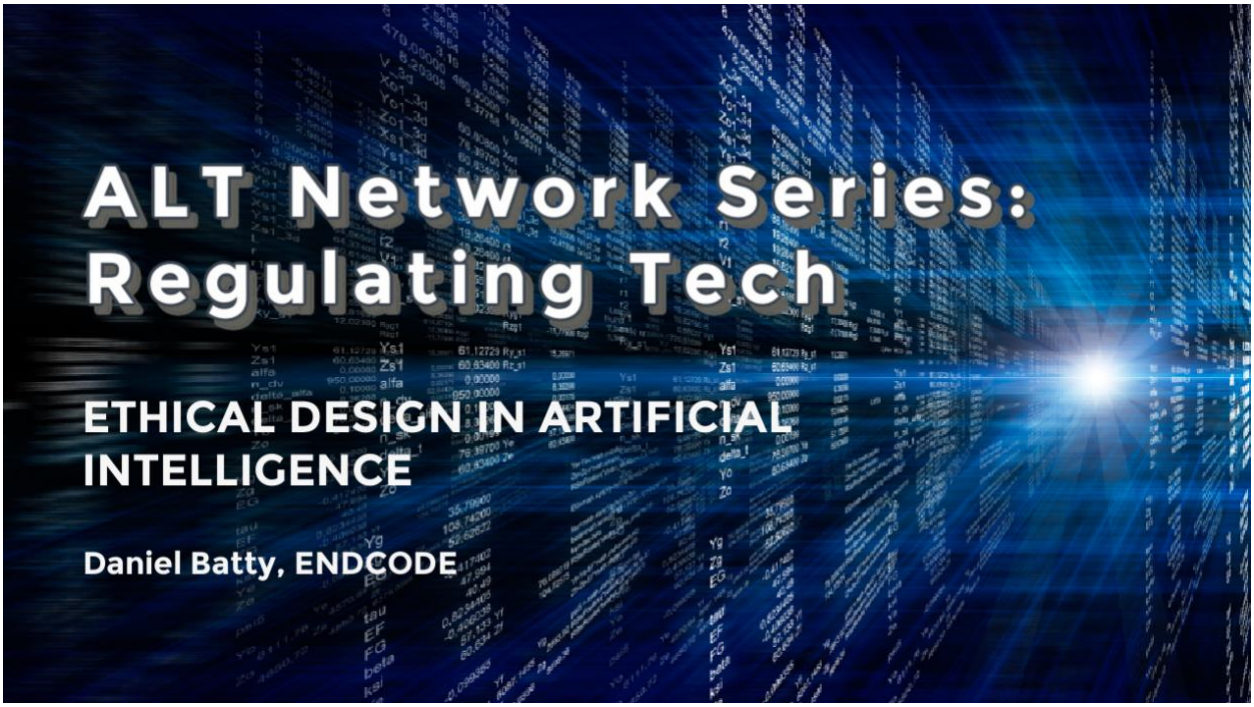

Ethical Design in Artificial Intelligence

Ethics could be described as the study of the human condition that has been ongoing for thousands of years in all societies, documented earliest by the ancient Greek philosophers. As human beings we have the unique characteristic of exceptional intelligence which has allowed us to not only appreciate our lives, but also the lives around us. Through our long and bloody histories, we, as humanity have arrived in the 21st century with an almost unanimous view of what it means to be an ethical person. We espouse the traits of integrity, honestly, empathy and fairness as ideal characteristics to which all human beings should strive. However, these metaphysical traits cannot easily be converted into computer code.

Artificial intelligence ("AI") is a hallmark of emerging technology in what is now known as the fourth industrial revolution. AI is not one single innovation but refers to many different fields of computer science including machine learning and deep learning. What all aspects of AI development have in common is the simulation of human intelligence within computers. Similar to human beings, computers learn by learning information, and like humans, the information inputted and processed by an AI will affect its decision-making.

Over the last decade, we have heard many examples of AI systems acting unethically through one form of discrimination or another. In 2019 Adair Morse of the Haas School of Business released a paper titled Consumer-Lending Discrimination in the FinTech Era which found that AI algorithms in the mortgage sector were unintentionally discriminating against black and Latino homeowners by making inferences from big data processing. In 2014 Amazon made use of an automated application review algorithm to sort through resumes submitted for engineers and

coders. The algorithm was trained by processing the resumes of the current engineers and coders who were largely male. The result was that the algorithm made an inference that Amazon preferred male candidates and would reject the resumes of female candidates.

When discussing AI and ethics, there are two primary schools of thought: machine ethics which assess the ethical behaviour of machines, and human ethics which assess the ethical behaviour of the developers, users and adopters of the machines.

While there are many examples of AI algorithms acting in a biased or unethical manner, it should be remembered that AI are not as intelligent as us. They do not think and reason beyond a given task. They do not rationalise their actions according to a moral code or a greater belief in fairness and equality that guides much of human decision-making. An AI algorithm cannot be racist or sexist in the same way a human can. Much of the bias, including the two examples above are unintentional on the part of the developers and simply reflect wider societal bias.

So how then, do we develop ethical AI algorithms? Unfortunately, as these systems become increasingly more complicated and find patterns in ever increasing big-data databases, they are likely to keep making unintended connections that result in unfair decision making.
In 2019 the European Union's High Level Expert Group on Artificial Intelligence developed guidelines which describe in part, four "ethical imperatives" that developers should strive to achieve -

1) Respect for human autonomy:
   Human beings have a fundamental right to liberty - and this includes freedom of thought and conscience. When respecting human autonomy, an AI developer should ensure that the algorithm does not unfairly coerce, deceive or manipulate the end user.

2) Prevention of harm:
   AI systems should not create or exacerbate harm. This entails an analysis of the specific use case for the AI system. Where vulnerable groups are expected to be users, beneficiaries or recipients of the system, special attention should be given to ensure that the outcome sought does not adversely affect these groups.

3) Fairness:
   AI should be substantively and procedurally fair in interactions with humans. Substantive fairness means taking steps to guard against unfair bias and discrimination. Ideally AI systems can be used to enhance fairness in equal access and opportunity regarding education, goods and services. Developers should be mindful of a balance between competing objectives in the development and use of the AI system. While procedural fairness means the ability of the end user to seek remedies and to contest any decision made by an AI system.

4) Explicability:
   AI developers should be transparent and open regarding the capabilities and use of the AI system. Any decisions made by an AI that affect a human being should be explained.

It should be remembered that ethics are principles that we strive to achieve in order to better our lives, the lives around us and reflect the values to which we hope our broader society will aspire. Giving credence to these principles in the development of AI is simply the next evolution in the quest of building a better world.

Daniel Batty
Legal Policy Advisor
EndCode