Accountability in AI

The paper "*The responsibility gap: Ascribing responsibility for the actions of learning automata*" raises the question of who should be held responsible for the actions of a machine. In the context of a responsibility gap, the paper mentions four primary types of learning automata - symbolic systems, connectionist architectures (including reinforcement learning systems), genetic algorithms (including genetic programming), and autonomous agents (mobile and immobile).

The author argues that as the AI technology advances, there is a gradual change in the scope of responsibility of the programmer. This is true in the sense that the outcomes for predictive algorithms such as regression, largely depend on the choice of target variables, dependent variables, transformations, and the dataset itself. All these factors are well within the scope of the developer/programmer. However, a programmer can never be sure of what learning algorithms such as those employed in autonomous vehicles, robots, recommender systems, etc. "learn" when they are deployed in dynamic environments. For example, in UK, a couple of researchers from University of Cambridge developed a deep learning model to diagnose coronavirus, which was trained on a dataset that included scans of patients, some of whom were scanned while lying down and some while standing up. Turns out, since the patients who were more seriously ill were scanned while lying down, the algorithm learnt to predict serious covid risk from a person's position.

There have been many incidents where the algorithms have led to some error. For instance, Amazon's facial recognition technology, Rekognition, mistakenly matched prominent athletes such as Super Bowl champion Duron Harmon of the New England Patriots, Boston Bruins forward Brad Marchand, and 25 others from New England to a database of mugshots in a test arranged by the Massachusetts part of the American Civil Liberties Union (ACLU), thereby classifying them as criminals. In this case, the personal cost of being misclassified is still marginal, given the celebrity status of the individuals, but consider a similar case from Feb 2019, where another false facial recognition match led to a Black man's arrest. Nijeer Parks had to spend 11 days in prison for a crime he never committed.

Leonelli, in the paper "*Locating ethics in data science…*" mentions that "…all individuals need to take some responsibility for potential implications, in relation to their specific roles". Well, it could apply to some contexts, for instance, in a reinforcement learning mechanism such as Roomba, a developer could set an incorrect reward, say, to cover the longest distance which could then make the algorithm incorrectly learn to maximize the distance anyhow, thus leading to unsatisfactory cleaning or jittery navigation. But, in the above situations, how can someone be held responsible if an algorithm ends up learning or identifying something incorrectly. Incentivizing individuals to consider ethical and moral implications of their data, codes, software etc. may improve the quality of the deliverables, but it will still not impact what an algorithm ends up learning in the real world. After all, in an ever-changing dynamic world, there are only so many scenarios that can be tested and accounted for during the development phase. Matthias correctly points out this ambiguity by stating that "In the same degree as the influence of the creator over the machine decreases, the influence of the operating environment increases."

Consequently, it is not right to blame an individual or a group of individuals for a machine's actions over which they have no control after a certain point. While explainable AI may alleviate this responsibility gap to some extent, achieving complete non-ambiguity could be particularly difficult.