On Amorality

Tufekci finds that the most terrifying part of machine learning algorithms is not that they are immoral, but that they are amoral. The algorithm is not domain aware, or at least it is not domain aware in an ethical sense. The "persuasion architectures" that make ML very attractive to virtually anyone who hopes to reach anyone in an individualized manner is that the algorithms work equally well regardless of the message being communicated. For example, Tufekci points out that the Trump administration has admitted to using Facebook posts crafted specifically to demobilize African Americans from voting. Of course, these are cases where there is a clear malicious intent, and we are aware of it, so this is not a dystopia we are unfamiliar with. However, she says, the scarier part is what happens beyond our, or even beyond the coder's knowledge. The most advanced ML algorithms are infamously "blackbox", meaning it is structurally very difficult to make sense of the partial progress in the system. Because of this, algorithms are evaluated not on the basis of whether or not they were able to discover the right characteristics, but on the percent of accurately categorized examples. Therefore, what the algorithm takes into account is a mystery, and this is often seen by computer scientists as a feature: The algorithm can detect patterns that humans cannot, and this is part of what makes them powerful. But the other side of the coin is that even when one wishes to, it is often not possible to figure out what the deciding mechanism of the algorithm is. Tufekci gives the example of targeting plane tickets to Las Vegas: It is uninteresting to us that a company might target single men in the ages of 30-40, because we are aware of this type of strategies, and in some senses, immune to it. However, an ML algorithm might just as well target people with bipolar disorder at the onset of a mania episode. Actually, the latter would be considered a more successful algorithm than the earlier, given its higher percentage of accurate sales. What makes this scary for Tufekci is that nobody in particular has to encode a morally dubious (at best) strategy such as this; the algorithm simply picks it up when it does its job well.

Humans, too, could reach the same conclusion, but we have an inner compass, or at least the fear of being held accountable for our actions, that automatically puts some possibilities out of reach for us. This is the banality of immorality; society has developed certain tools to notice it, combat it and punish it. However, amorality resists generalizations. We cannot mark a whole set of possibilities as "amoral" and treat them all in the same way. Amorality requires that we handle its cases one by one, individually. Of course, this goes against the building principles of algorithms using AI, where one of the goals is maximization of automatization. Then, what other tools do we have to bring ethical concerns into an amoral system?

The first instinct might be to fight fire with fire: Maybe we should encode content-restrictions into the algorithm and make a *moral ML algorithm* that makes accurate generalizations for us. Here, I see three challenges one could face: The ones concerning freedom of speech (*Should* we censor hate speech? Are we allowed to ban the Facebook group formed to gather for a racist rally?), ones concerning economic freedom (If you pay Facebook to get your ad to convince the most number of people, do I have a right to intentionally take something away from the algorithm that will make it worse?), and ones that are technical (How can I determine each item that factored into the algorithm, given the blackbox nature of most ML algorithms used in practice?) My current inclination is to look at some case studies, such as Facebook's and Twitter's "battles" with hate speech. Even some understanding cyberlaw might be necessary, although my impression is that the field of public policy is far behind technology.