16
Machine Learning and Ethics

Throughout this course, we have discussed the technical aspects of model design, training, and testing in depth. However, we have not yet discussed some of the social implications of this technology. What are some ethical and legal issues in deployment of ML techniques in society? What are the caveats and limitations to temper our exuberance about the possibilities of ML? This brief chapter addresses these issues, and we hope as technologists you will continue to investigate and consider such issues throughout your career.

16.1 Facebook's Suicide Prevention

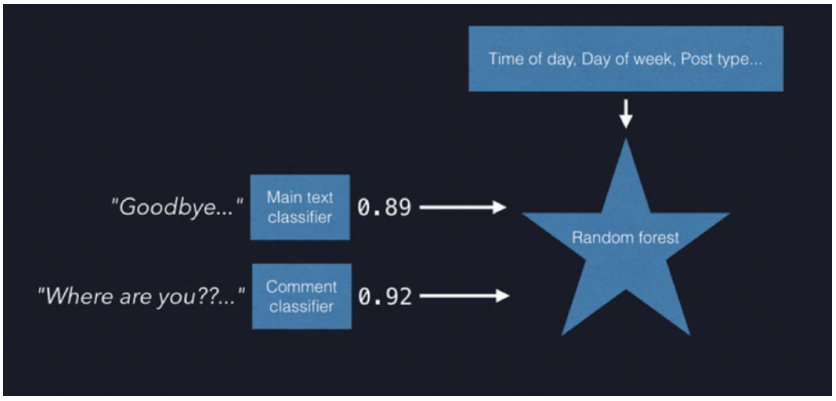

Figure 16.1: A visualization of the Facebook model to predict suicides.

In 2017, Facebook launched a program to use a machine learning algorithm to predict suicide risk amongst its user population. It has continued with various iterations over the years. Figure 16.1 gives a visualization of the four-step process:

1. ML algorithm automatically analyzes a post by processing its text content and comments

2. Algorithm additionally uses spatial-temporal context of the post to perform a risk prediction
3. A human behind the algorithm performs a personal review to finally verify if a threshold is reached
4. If the post is poses a serious risk, Facebook performs a wellness check through the person's contacts, community organizations, etc.

At first sight, this may appear to be very good idea: even if it saves just one life, surely the project is worth it? But the announcement of the project cause a lot of controversy among people. The following are some of the potential problems that people identified:

1. False positives may result in stigmatization.
2. Many people who contemplate suicide do not end up going through with it. Facebook's reporting could lead to criminal penalties (in regions where suicide is a crime), involuntary hospitalization, stigmatization etc.
3. Involvement of authorities (*e.g.*, law enforcement) raises risk of illegal seizures
4. Should Facebook be liable for any problem caused by misdetection?

Beyond these points, there are deep philosophical questions associated with the concept of suicide as well. For instance, is suicide actually immoral? Even if it is immoral, is it the responsibility of Facebook to get involved? Is it moral for Facebook to use personal information to assess suicide risk? Opinions differ.

16.2 Racial Bias in Machine Learning

Suppose we are designing a machine learning approach for loan approval. The general approach will be to take a dataset of $(\vec{\mathbf{x}}, y)$, where $\vec{\mathbf{x}}$ is a vector of the individual's attributes (*e.g.*, age, education, alma mater, address, etc.) who got a loan and $y \in \{-1, 1\}$ indicates whether they actually paid off the loan or not. Using the approaches we learned in Section 4.2, we could train a binary classifier through logistic regression. Civil rights legislation forbids using the individual's race in many of these decisions, so while training we could simply mask out any coordinates which identify race. However, this does not guarantee that the classifier will be entirely "race-neutral." [1]

In 2016, a study [2] found that COMPAS, a leading software for assessing the probability that a prison inmate would commit another

[1] The reason is that race happens to be correlated with many other attributes. Thus if a classifier uses any of the correlated attributes, it may be implicitly using racial information in the decision making process.

[2] *Machine Bias*, by Anwin et al., in *Pro Publica* 2016.

serious crime, disproportionately tags African-American as being likely to commit crimes — in the sense that African-Americans who were tagged as likely to commit another crime were only half as likely to actually commit a crime than a similarly-tagged person of another race.

	White	**African-American**
Labeled Higher Risk & Did *Not* Re-offend	23.5%	44.9%
Labeled Lower Risk & *Did* Re-offend	47.7%	28.0%

Table 16.1: COMPAS correctly predicts recidivism 61 percent on average. But African-Americans are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. Conversely, whites are twice as likely as African-Americans to be labeled lower risk but go on to commit other crimes.

16.3 *Conceptions of Fairness in Machine Learning*

We will briefly consider possible ways to formulate fairness in machine learning. Keep in mind that this task is intrinsically difficult, as we are attempting to assign a technological perspective to a fundamentally normative problem. The first property we might want a ML classifier to have is called *demographic parity*, which effectively enforces that the output of classifier does not depend on a protected attribute (*e.g.*, race, ethnicity, gender).

Definition 16.3.1 (Demographic Parity)**.** *We say that a binary classifier that outputs $y \in \{-1, 1\}$ satisfies* ***demographic parity*** *if $\Pr[y \mid x_i = a] = \Pr[y \mid x_i = b]$ where a, b are any two values that a protected attribute x_i can take.*

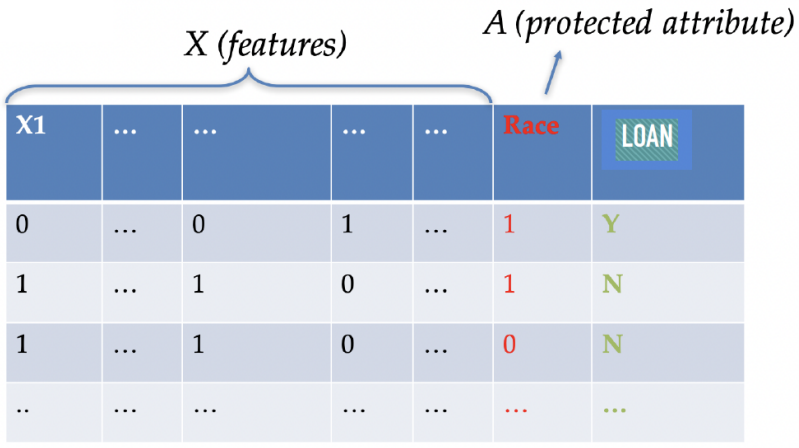

Figure 16.2: A hypothetical application of ML to a loan approval application. *Race* has been made a protected attribute in an attempt to prevent bias during training.

A visualization of how a protected attribute could be specified in a dataset is shown in Figure 16.2. Consider the loan approval example

from the previous section. If the binary classification model for the loan approval satisfies the demographic parity property, then the model approve loans for different races at similar rates. One way to achieve this condition is to use a regularizer term $\lambda(\Pr[y \mid x_i = a] - \Pr[y \mid x_i = b])^2)$ during training. [3]

[3] Does this seem like a good formulation of fairness?

Another property we want a "fair" model to satisfy is called the *predictive parity*. This is the property that the model in Table 16.1 failed to satisfy.

Definition 16.3.2 (Predictive Parity)**.** *We say that a binary classifier that outputs $y \in \{-1, 1\}$ satisfies **predictive parity** if the true negative/false negative/false positive/true positive rates are the same for any values of a sensitive attribute.*

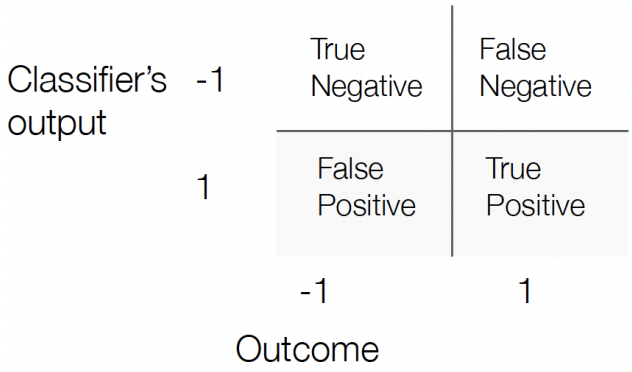

Figure 16.3: A table of all possible outcomes based on the model output and the ground truth outcome. This is also known as a *confusion matrix*.

Ideally, we want a ML model to satisfy both the demographic parity and predictive parity. However, it turns out that these two notions are incompatible!

Theorem 16.3.3 (Fairness Impossibility Theorem)**.** [4] *Under fairly general conditions,* demographic parity *and* predictive parity *are incompatible.*

[4] See *Inherent Trade-Offs in the Fair Determination of Risk Scores*, Kleinberg, Mullainathan, and Raghavan, *ITCS* 2017. The paper actually considered three possible definitions of "fairness" and showed every pair of them are mutually incompatible.

There are other formulations of fairness, but it is difficult to find a combination of these notions that are compatible with each other. So one way or another, we need to sacrifice some notions of "fairness."

16.4 Limitations of the ML Paradigm

The predictive power of ML seems immense, but is it true that if we have enough data and the right algorithm, then everything become predictable? If yes, then one could imagine societal programs leveraging this to precisely target help to where it would be more effective. We first consider a famous — and somewhat amusing — example of a study [5] that turned out to be false.

[5] *Extraneous factors in judicial decisions*, Danziger et al., *PNAS* 2011.

16.4.1 *Hungry Judge Effect*

The study analyzed the parole decisions made by 8 Israeli judges in over 1,100 cases. The data in Figure 16.4 shows that prisoners were much more likely to be granted parole after the judge took a lunch break or a coffee break. The study therefore suggested that judges tend to be stricter before a break (maybe because they are "hangry") but more lenient when they return from the break.

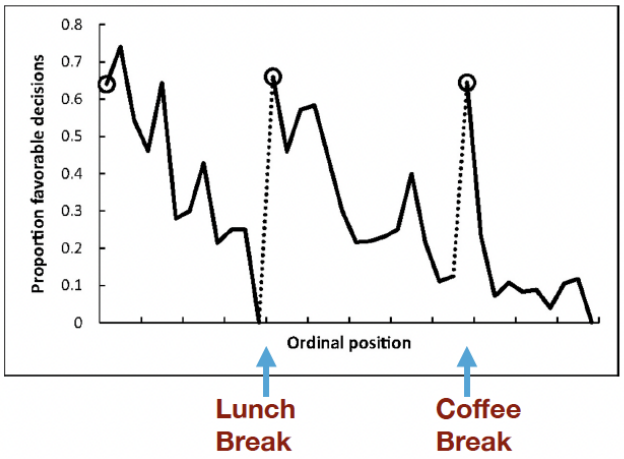

Figure 16.4: Data from the study shows an uptick in favorable decisions following a lunch break or a coffee break.

Nevertheless, it turns out that this "hungry judge effect" can be explained by a completely different reason. A followup study [6] found that the ordering of cases presented to the judge was not random: prisoners with attorneys were scheduled at the beginning of each session, while prisoners without an attorney were scheduled at the end of a session. The former group were let on parole with a rate of 67%, while the rate was just 39% for those without attorneys. Another important observation was that attorneys tended to present their cases in decreasing order of strength of case, with the average attorney having 4.1 clients. Computer simulations of hunger-immune judges faced with cases presented according to these percentages showed the same see-saw effect of Figure 16.4.

[6] *Overlooked factors in the analysis of parole decisions*, Weinshall-Margel and Shepard, *PNAS*, 2012.

16.4.2 *Fragile Families Challenge*

The Fragile Families Challenge is a collaborative project initiated by the Center for Research on Child Wellbeing at Princeton University. A brief description of the initiative's motivation is provided on the website: [7]

> *The Fragile Families Challenge is a mass collaboration that combines predictive modeling, causal inference, and in-depth interviews to yield insights that can improve the lives of disadvantaged children in the United States.*

[7] Source: `https://www.fragilefamieschallenge.org`.

By working together, we can discover things that none of us can discover individually.

The Fragile Families Challenge is based on the Fragile Families and Child Wellbeing Study, which has followed thousands of American families for more than 15 years. During this time, the Fragile Families study collected information about the children, their parents, their schools, and their larger environments.

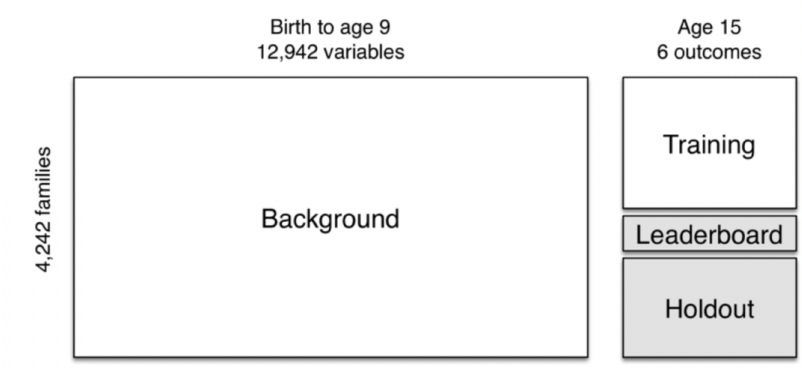

Figure 16.5: Diagram illustrating the dataset of the Fragile Families Challenge.

The initiative has collected immense data on multiple families, including interviews with mothers, fathers, and/or primary caregivers at several ages. Interviewees were inquired as to attitudes, relationships, parenting behavior, economic and employment status, etc. Additionally, in-home assessments of children and their home environments were performed to assess cognitive and emotional development, health, and home environment. The goal was to predict six key outcomes at age 15 (*e.g.*, whether or not the child is attending school) given background data from birth to age 9 as shown in 16.5. However, up to this point no method has done better than random guessing.

This is food for thought: what is going on?

16.4.3 General Limits to Prediction

Matt Salganick and Arvind Narayanan, professors at Princeton University, recently started a course [8] which aims to explore the extent to which interdisciplinary problems in social science and computer science can be predictable. In general, following are some major themes that can make prediction difficult:

1. The distribution associated with data can shift over time
2. The relationship between input data and desired outputs can change over time

[8] The course, COS 597E/SOC 555 is a seminar first offered in Fall 2020.

3. There is a possibility for unknown coordinates to be unintentionally ignored (*i.e.*, as in the hungry judge effect)

4. The "8 billion problem," which outlines how data available in the real world is fundamentally finite and limited

16.5 Final Thoughts

As described in the preceding sections, users and designers of machine learning will often face ethical dilemmas. Designers may have to operate without moral clarity or easy technical fixes. In fact, technical solutions may even be impossible. To appropriately acknowledge these limitations, it is important to embrace a culture of measuring and openly discussing the impact of the system being built. Indeed, a general principle to follow is to avoid harm when trying to do good.