Algorithmic Fairness – Feedback Loops

Marcello Di Bello - ASU - Fall 2021

Our goal is to understand how feedback loops work, focusing on predictive policing, but the concept can be generalized to other domains.

Other expressions with a similar meaning are: self-reinforcing process; vicious circle; self-fulfilling prophecy; self-referential process. Slightly different, but still closely related, are the ideas of echo chamber and ideological polarization.

General ideal

A feedback loop is a process that repeats itself through multiple iterations over time. At each iteration, the output of the process serves as input for the next iteration.

For example, consider the application of the function $f(...) = (...)^2$ onto itself. The feedback loop takes the following form: $f(...) = (...)^2$; $f(f(...)) = ((...)^2)^2$; $f(f(f(...))) = (((...)^2)^2)^2$; etc. The formula for compounding interest, $p(1+i)^n$, is another example, where n is the number of years, i the interest rate, and p the principal.

If you start with 10k, you will have almost 200k—a 20 times increase—after 30 years even if you added no money in the process assuming a 10 percent return each year. Money begets money at an ever increasing pace.

Compounding interest is a good feedback loop for those with money. We often think of feedback loops in a negative way, though.

Example 1: US News's college ranking

> U.S. News's first data-driven ranking came out in 1988 ... as the ranking grew into a national standard, a vicious *feedback loop* materialized ... the rankings were self-reinforcing. If a college fared badly in U.S. News, its reputation would suffer, and conditions would deteriorate. Top students would avoid it, as would top professors ... The ranking, in short, was destiny (p. 53).[1]

[1] O'Neil (2016), Weapons of Math Destruction, Chapter 3.

Besides the feedback loop above, another seemed to take place:

> ... college administrators ... worked hard to improve each of the metrics that went into the score ... when you create a model from proxies, it is far simpler for people to game it ... proxies are easier to manipulate than the complicated reality they represent. (p. 53)

Are there two feedback loops going on and feeding on one another here?

Example 2: PredPol

> ... a pernicious feedback loop. The policing itself spawns new data, which justifies more policing. And our prisons fill up with hundreds of thousands of people found guilty of victimless crimes. Most of them come from impoverished neighborhoods, and most are black or Hispanic. So even if a model is color blind, the result of it is anything but. In our largely segregated cities, geography is a highly effective proxy for race (p. 87).[2]

Here the feedback loop leads to an exacerbation of racial inequalities. How does this work exactly? Can we unpack the mechanism?

[2] O'Neil (2016), Weapons of Math Destruction, Chapter 5.

Attempting to define "feedback loop"

There are different types of feedback loops, each operating in different ways and with their own structure. Bad feedback loops seem to have in common the following:

Is anythig missing from the three items?

(i) iteration of the same process multiple times

(ii) using outputs as inputs at the next iteration

(iii) departure from reality or distortion of the intended goal that worsen as the process continues to be iterated

A feedback loop generates no learning and no progress, only the illusion of learning and progress

Predictive Policing

Predictive policing can be place-based or person-based. We will focus on place-based predictive policing such as PredPol. In both instances, the goal is to predict the occurrence of crime based on historical data and deploy police resources *proactively*. The historical data used for predicting policing typically include arrests as well as citizens calls about crime incidents. One assumption of predicting policing, and of machine learning in general, is that trends in the historical data will repeat themselves in the (near) future.

PrepPol uses three predictors: past type of crime, location and time of crime.

A common criticism of predictive policing is that it does not predict crime, but rather, policing itself:

> even the best machine learning algorithms trained on police data will reproduce the patterns and unknown biases in police data. Because this data is collected as a by-product of police activity, predictions made on the basis of patterns learned from this data do not pertain to future instances of crime on the whole. They pertain to future instances of crime that becomes known to police. In this sense, predictive policing ... is predicting future policing, not future crime.[3]

[3] Lum and Isaac (2016), To Predict and Serve? Significance

Lum and Isaac's claims

1. Illicit drug use was evenly distributed in Oakland according to estimates based on the 2011 National Survey on Drug Use and Health (NSDUH). However, drug arrests in Oakland in 2010 were heavily concentrated in poor and minority neighborhoods. So drug arrests do not reflect illicit drug use.

See Figure 1(a) and 1(b), p. 17.

2. Suppose PredPol had been used instead of traditional policing. Drug arrests, once again, would have been concentrated in roughly the same areas. Racial disparities in drug arrest would still remain, despite drug use being evenly distributed across racial groups.

See Figure 2(a) and 2(b), p. 18.

3. Whenever police officers are given an incentive to arrest when they patrol a neighborhood, predictions about volume of crime at future locations will even more clearly point to the neighborhoods previously singled out as crime hot spots.

See Figure 3, p. 19.

Only the last claim is squarely about feedback loops.

Brantingham's response

Jeffrey Brantingham is a UCLA anthropology professor inventor of PredPol.

> In practise, the majority of hotspot and place-based predictive policing algorithms focus not on arrests, but on crimes predominantly reported to the police by the public ... Thus, the goal is to send police resources to areas where crimes have been reported by victims, thus preventing future crimes in those areas. While a feedback loop for reported crime may be possible, in this case the self-reinforcement is towards places where citizens are placing calls for service (p.2).[4]
>
> ... the analyses do not provide any guidance on whether arrests themselves are systemtically biased ... The current study is only able to ascertain that arrest rates for black and Latino individuals were not impacted, positively or negatively, by using predicting policing (p. 4).

[4] Brantingham et al (2018), Does Predictive Policing Lead to Biased Arrests? Results From a Randomized Control Trial, Statistics and Public Policy, 5(1): 1-6. Another important study is Mohler et al (2015), Randomized Controlled Field Trials of Predictive Policing, Journal of the American Statistical Association, 110(512): 1399-1411.

Simulating feedback loops in predicting policing

More details about the simulation are available in my note on Simulating Feedback Loops in Predictive Policing

Suppose police want to learn about shares of crime from two areas, A and B. As a matter of fact, each area contributes 50 units of crime, but the police initially believe that one area contributes more crime than the other, say 20:80. Police collect information about where crime is happening every day by patrolling. Suppose police patrolling correctly detects 10% of the crimes happening in an area. There are no false positives. Police update their beliefs about shares of crime by area with information from patrolling every day. Will the police eventually learn the true shares of crime by area?

Scenario 1: Police decide the area to patrol by flipping a weighted coin. Side A and side B of the coin have a probability corresponding to the share of crime the police believe is happening in A and B respectively. Depending on the outcome of the coin flip, police patrol one area per day. In this scenario, the police will *not learn* that crime shares are 50:50 even though patrolling is unbiased and correctly detects crime at equal rates in both areas.

Scenario 2: Instead of flipping a coin, police send resources in proportion to how much crime they believe is happening in an area. They will split their presence by area accordingly. In this scenario, police will approximately learn the true shares of crime by area.

The claims are of course limited to a particular simulation and its underlying assumptions.

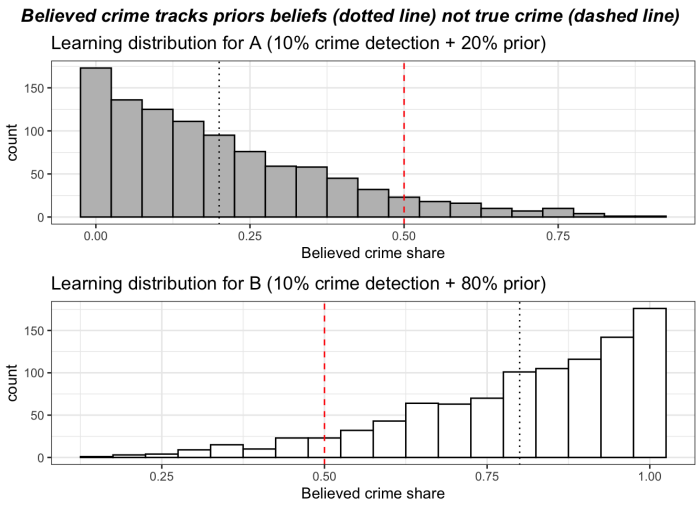

Figure 1: Scenario 1 with coin flip. Police will tend to learn that the shares of crime are similar to what they originally believed. In fact, they will often enough learn that the shares of crime are more extreme than they originally believed.

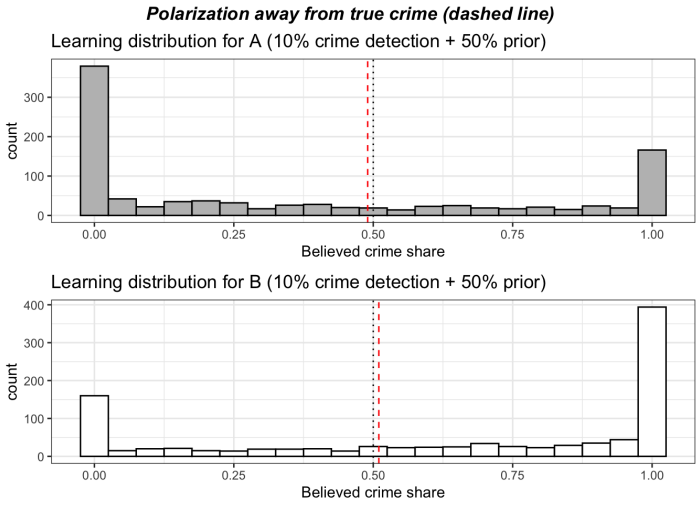

Figure 2: Scenario 1 with coin flip. The police will believe that crime is mostly concentrated in the area that contributes slightly more crime than the other. The area that contributes 51 percent of the crime will be believed to contribute almost the totality of crime. This is a winner-takes-all phenomenon.

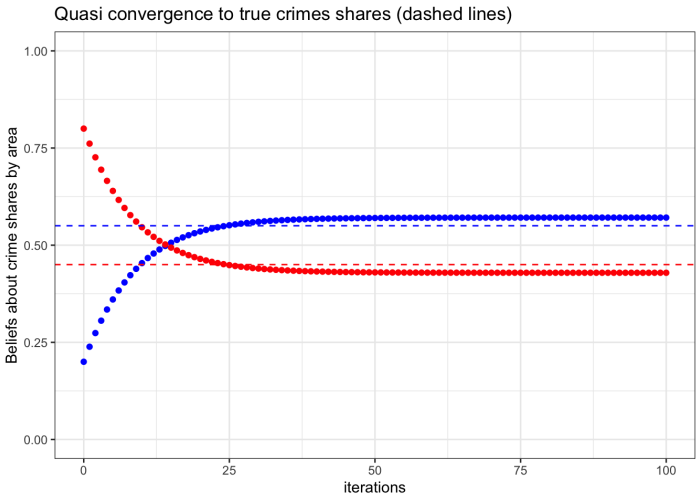

Figure 3: Deterministic proportional allocation of police resources. Police beliefs about shares of crime converge toward values relatively close to the true values of 45:55 even if prior beliefs about crime were fer way from true value.