Algorithmic harms

11 Jan 2022

sign up for discussant schedule

- Sign up here for a discussant position:
- https://docs.google.com/spreadsheets/u/1/d/16aZf8u6491yT8y76rCDbjjjrOX91u2D-KVLXPxYzKsg/edit?usp=sharing
- Instructions on the course website

wrapping up last class

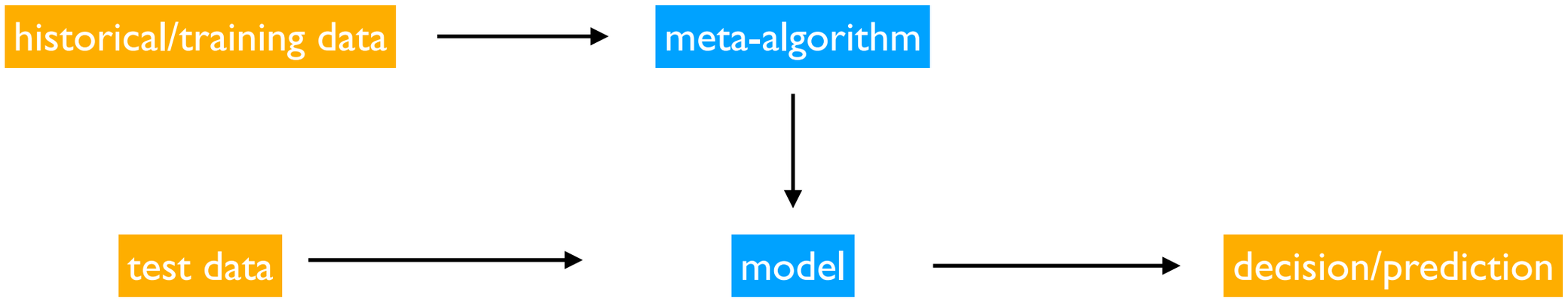
Discussion

- What are some of the competing definitions of "algorithm" and "automated system" that we discussed last week?
- What might it mean for an algorithm to be "biased" or "unfair?"

> As decision-makers in both government and industry create standards for algorithmic audits, disagreements about what counts as an algorithm are likely. Rather than trying to agree on a common definition of "algorithm" or a particular universal auditing technique, we suggest evaluating automated systems primarily based on their impact. By focusing on outcome rather than input, we avoid needless debates over technical complexity. What matters is the potential for harm, regardless of whether we're discussing an algebraic formula or a deep neural network.
>
> Lum and Chowdhury

"Automated Decision Systems" are any systems, software, or process that use computation to aid or replace government decisions, judgments, and/or policy implementation that impact opportunities, access, liberties, rights, and/or safety. Automated Decisions Systems can involve predicting, classifying, optimizing, identifying, and/or recommending.

Rashida Richardson, "Defining and Demystifying Automated Decision Systems" (Narrow Definition)

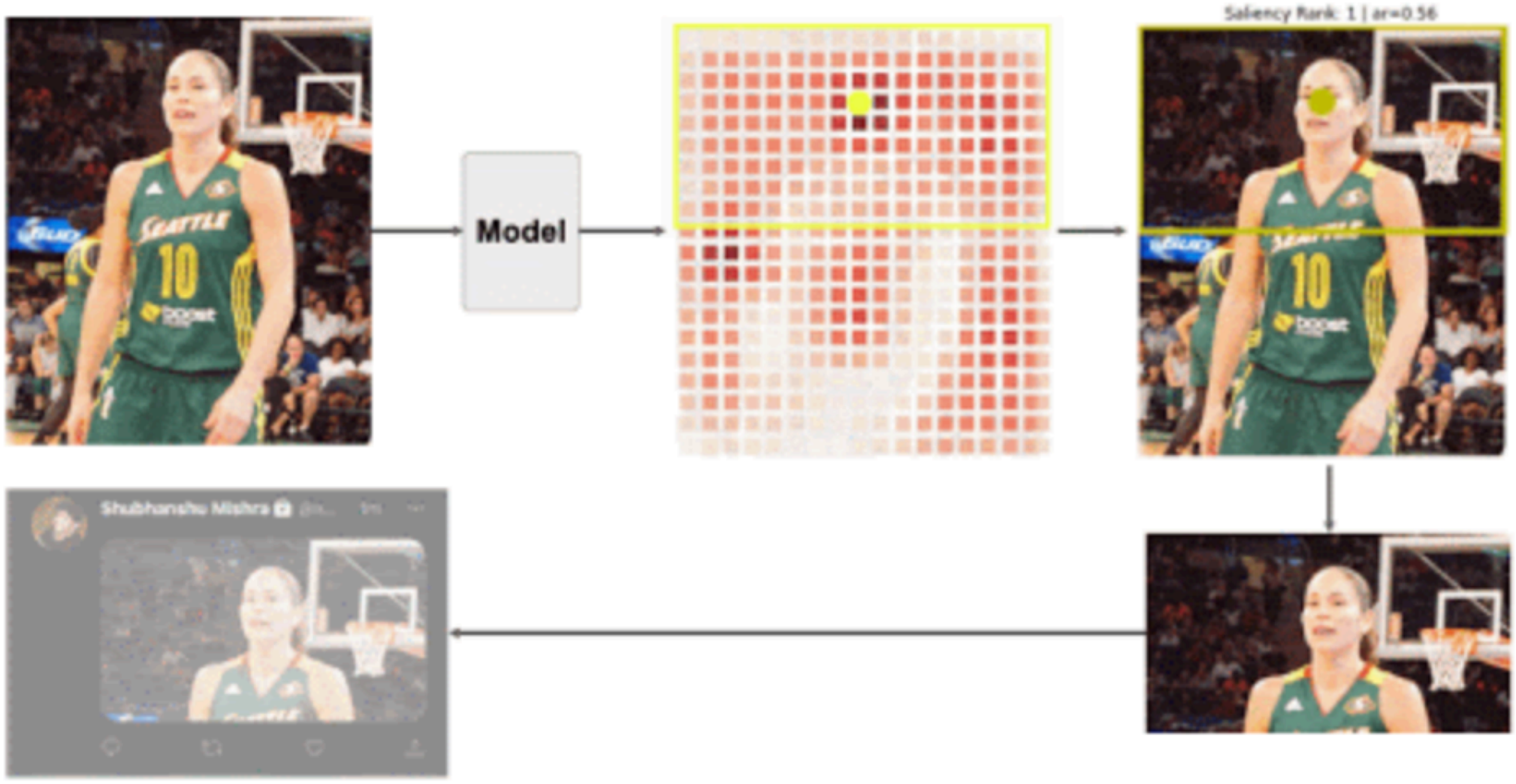

historical/training data
meta-algorithm
test data
model
decision/prediction

today's plan

a warm-up

- From the readings for today—identify as many examples of algorithms/ automated systems as you can.

Twitter image cropping



Chowdhury, "Sharing learnings about our image cropping algorithm."

recap

- What happened here?
- Does this “algorithm” fit the definitions we’ve discussed?
- What are the inputs? What is the “training data?” What is the “test data?”
- What are the final outputs? What are the intermediate outputs?

recap

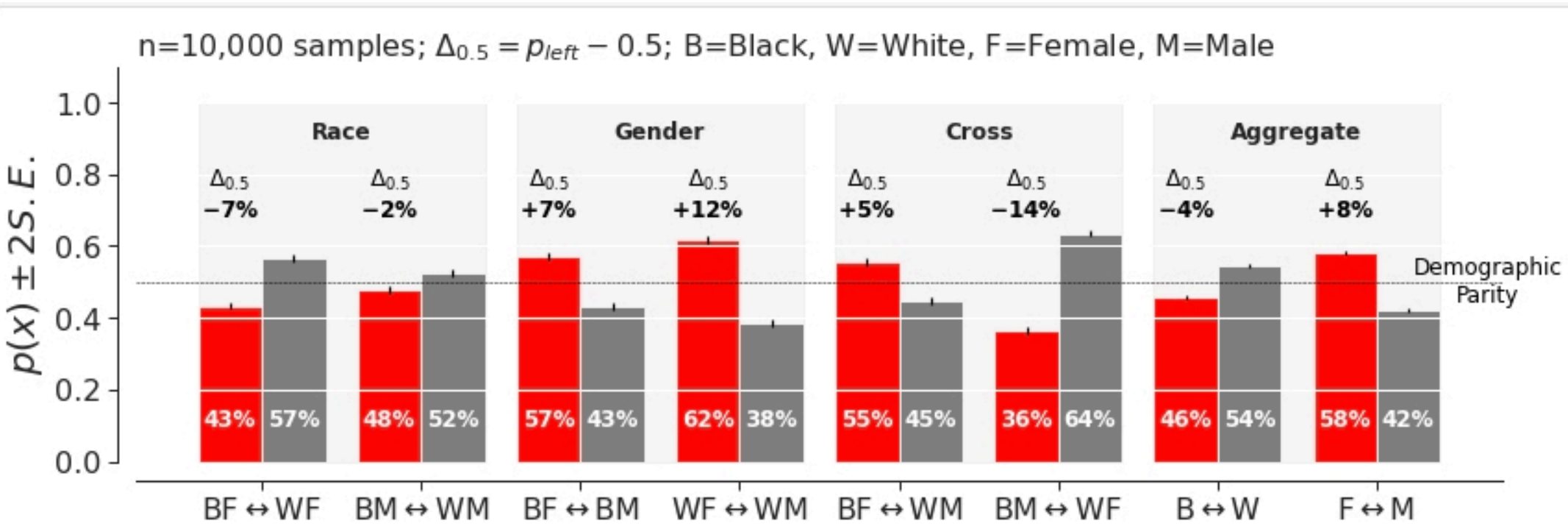
- What potential harms of the image cropping algorithm did Chowdhury and her team identify?

- Are these potential harms all potential "biases?"

- Are there other potential harms that you can think of?

recap

- How did Chowdhury and the Twitter team test to see whether the potential harms were actually happening? What did they observe?
- What actions did they take to address the criticisms received?
- How might these potential harms been avoided before starting use of this algorithm?



Chowdhury, "Sharing learnings about our image cropping algorithm."

Default Discrimination

ALBLA
@alliebland

Following

Then Google Maps was like, "turn right on Malcolm Ten Boulevard" and I knew there were no black engineers working there

9:42 PM - 19 Nov 2013

3,656 Retweets **3,749** Likes

100 3.7K 3.7K

https://commons.wikimedia.org/wiki/File:Lt._William_Tighe_Triangle_td_(2019-04-27)_10.jpg

Like the discriminatory designs we are exploring in digital worlds, hostile architecture can range from the more obvious to the more insidious – like the oddly shaped and artistic- looking bench that makes it uncomfortable but not impossible to sit for very long. Whatever the form, hostile architecture reminds us that public space is a permanent battleground for those who wish to reinforce or challenge hierarchies. So, as we explore the New Jim Code, we can observe connections in the building of physical and digital worlds, even starting with the use of "architecture" .as a common metaphor for describing what algorithms – those series of instructions written and maintained by programmers that adjust on the basis of human behavior – build.

Ruha Benjamin, *Race After Technology*

recap

- Before reading this text, what was your understanding of the word, "glitch?"

- What does Benjamin mean when she uses the word "glitch?" What are some "glitches" she points out?

Glitches are generally considered a fleeting interruption of an otherwise benign system, not an enduring and constitutive feature of social life. But what if we understand glitches instead to be a slippery place (with reference to the possible Yiddish origin of the word) between fleeting and durable, micro-interactions and macro-structures, individual hate and institutional indifference? Perhaps in that case glitches are not spurious, but rather a kind of signal of how the system operates. Not an aberration but a form of evidence, illuminating underlying flaws in a corrupted system.

Ruha Benjamin, *Race After Technology*

> And, as we shall see in the following chapters, the practice of codifying existing social prejudices into a technical system is even harder to detect when the stated purpose of a particular technology is to override human prejudice.
>
> Ruha Benjamin, *Race After Technology*

discussion

- What might Benjamin say about the image cropping example?
- What happens if we try to view the Twitter image cropping example as a glitch?

Algorithmic harms and algorithmic bias

In sectors as diverse as health care, criminal justice, and finance, algorithms are increasingly used to help make complex decisions that are otherwise troubled by human biases. Imagine criminal justice decisions made without race as a factor or hiring decisions made without gender preference. The upside of AI is clear: human decisionmakers are far from perfect, and algorithms hold great promise for improving the quality of decisions. But disturbing examples of algorithmic bias have come to light. Our own work has shown, for example, that a widely-used algorithm recommended less health care to Black patients despite greater health needs. In this case, a deeply biased algorithm reached massive scale without anyone catching it—not the makers of the algorithm, not the purchasers, not those affected, and not regulators.

Bembeneck et al.

Since both human and algorithmic decisionmakers introduce the possibility of bias, removing algorithms entirely isn't always the best approach. In fact, in some cases, biased algorithms may be easier to fix than biased humans. However, it falls on policymakers to ensure that the algorithms helping make complex decisions are doing so in a just and equitable way. To do so, we believe three key steps are required.

First, regulators must define bias practically, with respect to its real-world consequences.

Second, once the goalposts are clear, regulators must use them to provide much-needed guidance for industry and to define targets for objective, hard-hitting investigations into biased algorithms.

Third, as in other fields, regulators should insist on specific internal accountability structures, documentation protocols, and other preventative measures that can stop bias before it happens.

Bembeneck et al.

discussion

- How do you define bias?
- How is bias typically defined in statistics?
- How do Bembeneck et al. define algorithmic bias?
- How is their definition of algorithmic bias related to your definition of bias?

What is the ideal information the algorithm should be providing in this instance? It was supposed to identify patients who were going to get sick tomorrow so hospitals could enroll them in the extra help program today. We call that goal of the algorithm its **ideal target**. But what was the algorithm actually doing? In fact, it was doing something subtly but importantly different: it was predicting not who was going to get sick but who was going to generate high costs for the health care system. This is the algorithm's **actual target**. The wedge between these two is a key driver of bias.

Bembeneck et al.

Instead, regulators can focus on one simple question: is the algorithm predicting its ideal target accurately and equitably? This test will detect many forms of bias, like failure of an algorithm to generalize from one population to another.

Bembeneck et al.

discussion

- How are Benjamin's and Bembeneck et al.'s perspectives on algorithmic harms similar or different?