COMING UP WITH A MORE FOCUSED THESIS

status quo

"Streetlevel algorithms" **make the wrong decisions in marginal cases** precisely because they **dont understand their task**, let alone underlying goals, and inevitably they make decisions that advance their tasks, **but not the goals**.

1 OPTION 1: WRONG DECISIONS IN MARGINAL CASES

"Street–level algorithms" make the wrong decisions in edge cases, marginal situations, and novel circumstances, … [al2: lol where do i go with this? this just seems like a self–evident statement, begging a "why though?"]

What changes?

Nothing we're good lol
*Honestly I just have no idea what to do with this if **this** is the core argument.*

2 OPTION 2: DON'T UNDERSTAND THE TASK

"Street–level algorithms" don't understand the *nature* of their task, and as a result they make choices that may be as good as random when they encounter cases that challenge them to make decisions that are consistent with the spirit of the task.

What changes?

YouTube. YouTube's classification system encounters novel things like LGBTQ content and doesn't understand what to do with it in the context of the task it's been assigned. **As a result, it makes a decision that we should probably call "arbitrary", and just goes with it, essentially creating a policy.**

We need to design systems that either recognize that they're seeing something new (which won't be perfect) or we need to design more opportunities for people to reach someone who more deeply understands the nature of the classification so that they can get a more substantive judgment

Algorithmic bail–setting. Bail recommendation systems don't really understand what they're doing; they're just calculating various probabilities of a defendant returning for their court date, given previously seen characteristics. As a result, it makes decisions that are totally ignorant of other concerns that go into deciding bail. When it encounters reasonably unfamiliar situations, the recommendation system basically makes the wrong call, not knowing exactly what it's doing but effectively "going through the motions" that look kind of right.

It's difficult to encode the nature of such a nebulous task into the system. What we need to do in this case, for now at least, is to scale back what the system offers; for example, rather than indicating a suitable bail level, return predictive probabilities that a defendant will appear for their court date, given various bail levels, and specifically make clear the embeddings that were generated in classifying the defendant, at least so that a judge can determine on their own whether the system has sufficiently grasped the nature of the task.

Crowdwork. Pass. I think I'm just not excited about talking about crowdwork.

3 OPTION 3: THE DIFFERENCE BETWEEN THE TASK AND THE GOAL

"Street–level algorithms" don't understand the *purpose* of their task, and as a result they fail to see opportunities to *fail* at their tasks in service of the underlying goals, such as when the circumstances of a case warrant exceptions, or "discretion".

What changes?

YouTube. YouTube has designed a system that classifies content as "*containing sexual content* or *not*". It's good at that specific task, and would almost certainly satisfy engineers' evaluative metrics. What it's not good at is connecting its task to any overarching **goal** that the engineers or anyone else has. It doesn't understand that there are types of content (and contexts surrounding content) that arguably make content on YouTube suitable for a broader community. As a result, YouTube's classification system has no sense of specific instances that it should take a *positive* classification of content (in other words, identify "sexual content") and nevertheless *look the other way.*

Understanding when to exercise discretion in this context is enormously challenging, because the nature of art is that these artists will always be trying new things, finding novel ways to entertain and be creative; and people will find novel, creative ways to be novel and creative. That being said, it's feasible for classification systems to recognize when *some* circumstance

surrounding the content it's classifying is uncharacteristic, and provide either the person being acted upon or some other agent to evaluate the decision, either confirming or overruling it, ultimately adding more nuance of the system.

Algorithmic bail–setting. If the goal of the criminal justice system is to impart justice, then we may need to look to the difference between what these algorithmic systems claim to do and what we expect agents in such a position to do. Put another way, we need to look at the difference between an agent that accurately predicts the likelihood of a defendant appearing for their court date (independent of underlying racial, economic, historical factors), and an agent that carefully aims to right some of the injustices that have systematically harmed people of color and other minority groups.

Corbett-Davies and Goel point out one seemingly important issue at the heart of algorithmic bail recommendation systems: that the notion of "fairness" itself is (1) somewhat problematic, (2) open to interpretation, and (3) ultimately a value judgment that humans must make. Engineering a system that computationally generates a "fair" (or "just") outcome for people at the margins may be impossible. We should refocus (and, going forward, much more seriously reflect on) our intent and specifically our **goal** when we design a system like this so that its actions align with the goals of our community.

Crowdwork. Ugh pass.

4 OPTION 4: MARGINAL DECISIONS BECOME POLICIES

"Street–level algorithms" have to make decisions about every situation that comes their way; those decisions are sometimes well–grounded by the data that trained the system, but sometimes not. Regardless of how well it handles making those decisions, the choices it makes effectively become policies applying to those cases.

What changes/is new?

Algorithmic bias in search. Engineers have designed systems to do tasks like create networks representing which pages point to each other, what each page on the internet contains on it, and more. In the mainstream (main effect? what's the word for stars that are all kinda normal?), sites get the normal, appropriate treatment. But sites at the margins (not in traffic, but in *nature*) might get misclassified or mishandled, and the result

Algorithmic bias in news feeds?

REFERENCES

[1] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).