Organisation for Economic Co-operation and Development

DAF/COMP/WD(2023)74

Unclassified

English - Or. English

7 July 2023

DIRECTORATE FOR FINANCIAL AND ENTERPRISE AFFAIRS
COMPETITION COMMITTEE

Auditing as policy – Note by Cathy O'Neill

14 June 2023

This note was submitted by Cathy O'Neill, Data Scientist and CEO of ORCAA (O'Neil Risk Consulting & Algorithmic Auditing) to serve as background material for Item 4 of the 140th OECD Competition Committee meeting on 14-16 June 2023. This paper was not commissioned nor vetted by the OECD Secretariat; the opinions expressed and arguments employed herein are exclusively those of the author and do not necessarily reflect the official views of the Organisation or of the governments of its member countries.

More documents related to this discussion can be found at
https://www.oecd.org/competition/algorithmic-competition.htm

Antonio CAPOBIANCO
Antonio.Capobianco@oecd.org, +(33-1) 45 24 98 08

JT03523161

This document, as well as any data and map included herein, are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

ORCAA

Auditing as Policy

May 2023

ORCAA has an algorithmic audit methodology that we have been deploying and refining for the past 5 years. We often engage with regulators and lawmakers who want to craft policies that mandate high-quality algorithm audits to mitigate harms from deployed algorithmic and AI systems. This memo outlines an approach.

Q1: "What methods can competition authorities use to investigate algorithms?" *(e.g., what technical methods may work best, particularly if you have access to the underlying algorithm/source code and all relevant data, which a competition authority is likely to have when engaging in an antitrust investigation)*

When auditing, we want to ask, in a specific algorithmic context, for whom does this fail? That means we make a list of stakeholders and potential algorithmic failures.[1] For example, if it's a hiring algorithm, we might worry about false negatives, i.e. perfectly qualified candidates that are being mysteriously rejected. We will also want to subdivide stakeholder groups according to concerns; so if we are working in the context of an antidiscrimination law, we might worry specifically about protected classes having higher rates of false negatives.

Two high-level observations:

1. This works pretty well for a narrow context like a hiring algorithm, but not so well in a broad context like ChatGPT or the Facebook newsfeed algorithm. It's true: those huge examples are not auditable, at least not comprehensively. We should instead write down examples of high impact and high risk stakeholders and concerns -- for instance, disordered eating and related mental health harms to teens from social media use -- and try to deal with those specific pairs one at a time. That's literally the best we can do.
2. The audit process is context-specific. The same algorithm in a different setting could have totally different stakeholders and concerns. Consider a risk score for Type II diabetes; in scenario A your doctor has it and helps you stay well, and in scenario B a potential employer has it and uses it as a pretext to not hire you. Clearly audits would look very different in these two scenarios.

An important consequence of these observations is that no regulation can be very explicit about what an accountable algorithm will look like. We liken it to designing the cockpit of an airplane. It's a process, and the result is different from the dashboard of a car or a train. To design an algorithmic auditing policy, think of three steps:

[1] At ORCAA we use a framework called the Ethical Matrix to do this

1. Identify potential harms to stakeholders

What risks do we need to be aware of, to fly this safely?

The first step of setting up an audit policy is to declare what the audits will be looking for. In our proposed approach, the answer is a list with items of the form: "It would be a problem if [X stakeholder group] had a worse result in terms of [Y outcome]." Often the X's are legally- protected classes, and the Y's are regulated interests of consumers or constituents (e.g. equity in the provision of credit, or in access to employment). The list will of course depend on the scope of the policy, and the authority of the public actor, in question.

We note that the EU AI Act has a good feature in this regard: it performs a triage by importance, I.e. risk level. That's critical, because there are lots of things that can go wrong; we want the most important dials in our cockpit first, that have to do with life and death.

2. Define metrics for harms

What dials do we need in the cockpit, to monitor these risks?

Taking the list from the first step, translate each stakeholder-harm item into a measurement(s) that uses data. This forces specificity. Back to the example of a hiring algorithm: suppose the algorithm gives each candidate a "fit score" from 1-100, which is used to sort candidates into Low/ Medium/High tiers. Hiring managers then choose which candidates to interview. Suppose we worry about gender bias. Should audits measure the average fit score given to women vs men?[2] The share of women vs men in each tier (Low/Medium/High)? The proportion of women vs men that are asked to interview? All of the above? If so, that would be three "dials" in the gender equity panel of the cockpit. It is important to think through what each metric captures and what it misses. This step also specifies the data needed from the audit target. For instance, these measurements might require candidate records that include the fit scores given by the algorithm.

3. Establish reporting standards

What's the procedure for checking the dials and course-correcting?

The last step is to outline rules and processes for reporting and interpreting the measurements. These can address:

[2] To compare outcomes by gender, we need to know (or at least have a good guess of) the gender of each individual. Likewise for race and other protected classes. Since such data is often not collected, especially in contexts where discrimination between protected classes is illegal, this can be a project unto itself, involving inference, self-report surveys, or merging with other datasets.

- Parties involved: What entities or algorithms are subject to the audit requirements? Who can conduct an audit? Who is the audience for audit reports?
- Timing: How often should audits be conducted? How often must reports be submitted? What is the timeline for review of audit results?
- Thresholds: Per the previous step, the audit must include certain measurements. Are there thresholds of acceptability for these?
- Remediation: If the audit report finds a problem, or if the audit is not satisfactory, what are the consequences?

The recent settlement between Meta and the Department of Justice related to discrimination in the display of housing advertisements is a good example that fits on this framework: it specifies a harm (race and gender bias in the display of certain ads); a way to measure it statistically with data; and a cadence and structure for reporting, with specific performance thresholds to meet.

Q2: "What specialist skills and knowledge do competition authorities require to investigate algorithms?" *(e.g., what is the composition of your team when engaging in an algorithmic audit, and how resource intensive is it, how long can it take etc.)*

You will need an ethical review board with lots of stakeholder representation to build the original list of stakeholders and potential failures. After you've made your list, you will need domain experts and possibly experimental design experts to identify the right measurements and design the experiments/analyses behind each cockpit "dial", help from the internal data science team to implement the cockpit, and statistics experts to decide whether the results are meaningful.