The Ethics of Algorithms

I opened a discussion on the ethics of algorithms recently, with a small thing about what algorithms are, what can be unethical about them and how we might start mitigating that. I kinda sorta promised a blogpost off that, so here it is.

Algorithm? Wassat?

Al-Khwarizmi (wikipedia image)

Let's start by demystifying this 'algorithm' thing. An algorithm (from Al-Khw?rizm?, a 9th-century Persian mathematician, above) is a sequence of steps to solve a problem. Like the algorithm to drink coffee is to get a mug, add coffee to the mug, put the mug to your mouth, and repeat. An algorithm doesn't have to be run a computer: it might be the processes that you use to run a business, or the set of steps used to catch a train.

But the algorithms that the discussion organizers were concerned about aren't the ones used to not spill coffee all over my face. They were worried about the algorithms used in computer-assisted decision making in things like criminal sentencing, humanitarian aid, search results (e.g. which information is shown to which people) and the decisions made by autonomous vehicles; the algorithms used by or instead of human decision-makers to affect the lives of other human beings. And many of these algorithms get grouped into the bucket labelled "machine learning". There are many variants of machine learning algorithm, but generally what they do is find and generalize patterns in data: either in a supervised way ("if you see these inputs, expect these outputs; now tell me what you expect for an input you've never seen before, but is similar to something you have"), reinforcement-learning way ("for these inputs, your response is/isn't good) or unsupervised way

(“here’s some data; tell me about the structures you see in it”). Which is great if you’re classifying cats vs dogs or flower types from petal and leaf measurements, but potentially disastrous if you’re deciding who to sentence and for how long.

Simple neural network (wikipedia image)

Let’s anthropomorphise that. A child absorbs what’s around them: algorithms do the same. One use is to train the child/machines to reason or react, by connecting what they see in the world (inputs) with what they believe the state of the world to be (outputs) and the actions they take using those beliefs. And just like children, the types of input/output pairs (or reinforcements) we feed to a machine-learning based system affects the connections and decisions that it makes. Also like children, different algorithms have different abilities to explain why they made specific connections or responded in specific ways, ranging from clear explanations of reasoning (e.g. decision trees, which make a set of decisions based on each input) to something that can be mathematically but not cogently expressed (e.g. neural networks and other ‘deep’ learning algorithms, which adjust ‘weights’ between inputs, outputs and ‘hidden’ representations, mimicking the ways that neurons connect to each other in human brains).

Algorithms can be Assholes

Algorithms are behind a lot of our world now. e.g. Google (which results should you be shown), Facebook (which feeds you should see), medical systems detecting if you might have cancer or not. And sometimes those algorithms can be assholes.

Here are two examples: a Chinese program that takes facial images of ‘criminals’ and maps those images to a set of ‘criminal’ facial features that the designers claim have nearly 90% accuracy in determining if someone is criminal, from just their photo. Their discussion of “the normality of faces of non-criminals” aside, this has echoes of phrenology, and should raise all sorts of alarms about imitating human bias. The second example is a chatbot that was trained on Twitter data; the headline here should not be too surprising to anyone who’s recently read any unfiltered social media.

We make lots of design decisions when we create an algorithm. One decision is which dataset to use. We train algorithms on data. That data is often generated by humans, and by human decisions (e.g. “do we jail this person”), many of which are imperfect and biased (e.g. thinking that people whose eyes are close together are untrustworthy). This can be a problem if we use those results blindly, and we should always be asking about the biases that we might consciously or unconsciously be including in our data. But that's not the only thing we can do: instead of just dismissing algorithm results as biased, we can also use them constructively, to hold a mirror up to ourselves and our societies, to show us things that we otherwise conveniently ignore, and perhaps should be thinking about addressing in ourselves.

In short, it's easy to build biased algorithms with biased data, so we should strive to teach algorithms using ‘fair’ data, but when we can’t, we need to use other strategies for our models of the world, and can either talk about the terror of biased algorithms being used to judge us, or we can think about what they’re showing us about ourselves and our society’s decision-making, and where we might improve both.

What goes wrong?

If we want to fix our 'asshole' algorithms and algorithm-generated models, we need to think about the things that go wrong. There are many of these:

- On the input side, we have things like biased inputs or biased connections between cause and effect creating biased classifications (see the note on input data bias above), bad design decisions about unclean data (e.g. keeping in those 200-year-old people), and missing whole demographics because we didn’t think hard about who the input data covered (e.g. women are often missing in developing world datasets, mobile phone totals are often interpreted as 1 phone per person etc).
- On the algorithm design side, we can have bad models: lazy assumptions about what input variables actually mean (just think for a moment of the last survey you filled out, the interpretations you made of the questions, and how as a researcher you might think differently about those values), lazy interpretations of connections between variables and proxies (e.g. clicks == interest), algorithms that don’t explain or model the data they’re given well, algorithms fed junk inputs (there's always junk in data), and models that are trained once on one dataset but used in an ever-changing world.
- On the output side, there’s also overtrust and overinterpretation of outputs. And overlaid on

> that are the willful abuses, like gaming an algorithm with ‘wrong’ or biased data (e.g. propaganda, but also why I always use “shark” as my first search of the day), and inappropriate reuse of data without the ethics, caveats and metadata that came with the original (e.g. using school registration data to target 'foreigners').

But that's not quite all. As with the outputs of humans, the outputs of algorithms can be very context-dependent, and we often make different design choices, depending on that context, for instance last week, when I found myself dealing with a spammer trying to use our site at the same time as helping our business team stop their emails going into customers' spam filters. The same algorithms, different viewpoints, different needs, different experiences: algorithm designers have a lot to weigh up every time.

Things to fight

Accidentally creating a deviant algorithm is one thing; deliberately using algorithms (including well-meant algorithms) for harm is another, and of interest in the current US context. There are good detailed texts about this, including Cathy O'Neill's work, and Latzer, who categorised abuses as:

- Manipulation
- Bias
- Censorship
- Privacy violations
- Social discrimination
- Property right violations
- Market power abuses
- Cognitive effects (e.g. loss of human skills)
- Heteronomy (individuals no longer have agency over the algorithms influencing them)

I'll just note that these things need to be resisted, especially by those of us in a position to influence their propagation and use.

How did we get here?

Part of the issue above is in how we humans interface with and trust algorithm results (and there are many of these, e.g. search, news feed generators, recommendations, recidivism predictions etc), so let's step back and look at how we got to this point.

And we've got here over a very long time: at least a century or two, to back when humans started using machines that they couldn't easily explain because they could do tasks that had become too big for the humans. We automate because humans can’t handle the data loads coming in (e.g. in legal discovery, where a team often has a few days to sift through millions of emails and other organizational data); we also automate because we hope that machines will be smarter than us at

spotting subtle patterns. We can’t not automate discovery, but we also have to be aware of the ethical risks in doing it. But humans working with algorithms (or any other automation) tend to go through cycles: we’re cynical and undertrust a system tip it’s “proved”, then tend to overtrust its results (these are both part of automation trust). In human terms, we're balancing these things: