Fragility > accuracy

Incentive problems for machine learning systems

Tom Slee

January 14, 2019

Algorithms as actors in equilibrium.

Contents

The troubles with sorting

Algorithms and sorting

Modern software classifies, filters, scores, and ranks people for many purposes: advertising, social services, hiring, imprisoning, and much more. This essay uses the word "sorting" as an umbrella term for all these activities.

The recent explosion of activity in machine learning shows how many problems can be thought of usefully as sorting problems. Image recognition is obviously sorting: this is a dog, that is a cat. But hiring is sorting: this person has the attributes of a successful employee, that person does not. Advertising software sorts people into target audiences.

YouTube recommendations sort videos by their expected viewing time. Playing Go involves sorting game positions into winning positions and losing positions.

After a period during which machine-learning based sorting was promoted as free from human bias and more objectively evidence-based than human decisions, there has been an upswing in concern about fairness, bias and transparency in machine learning [1], [2], [3], [4]. These important concerns have since been acknowledged by many practitioners; they have led to research into statistical criteria for fairness, mechanisms for explaining machine-learning results, and more. There is much still to do, and concerns about fairness and bias will continue to be important.

Beyond statistics

The concern about fairness and bias imagines a machine learning system as a camera, recording and portraying some aspect of the external world. It asks: does the system portray the external world fairly and faithfully?

Like other sorting methods, machine learning does not just portray the world, they change it. In the words of somebody, machine learning is "an engine, not a camera". Introducing a new sorting system changes the incentives for those on the receiving end, who are being sorted. People care about how they are categorized, so they respond.

The income taxation system, for example, sorts people into tax brackets. People respond by altering their behaviour in order to avoid being put into a high bracket. YouTube's recommender system sorts videos by predicted viewing time: producers respond by watching their views intensely and tuning what they do to what works. Google's search results sort web sites by "relevance" for a query, and the industry of search-engine optimization has grown up to help web site owners tune their sites for best exposure.

The dynamics do not stop there, of course. The sorting system itself must respond in turn to those responses, and so on until some kind of equilibrium is reached. Governments close tax loopholes, YouTube adds new systems to detect undesirable videos (whatever that means) that are nevertheless promoted by its recommender, and Google revises its search algorithms to combat the efforts of the SEO industry.

The dangers of sorting

As machine-learning becomes ubiquituous it is important to go beyond the static analysis; we must think about machine learning systems as active agents and ask what effects they have on the world they describe. There have certainly been many analyses that to take this view, but so much of the "AI ethics" debate has focused on bias and transparency that it may be worth collecting together some of what is left out in the static analysis. That is what I try to do here.

The incentive-driven dangers of machine learning are different to the problems of bias and unfairness. Some covered here include:

1. "Lesser evil" decisions. When our actions may lead to bad scores on sorting systems, we may choose to avoid beneficial actions. If contact with the social services is going to help create a profile that may lead child services to my door, I may avoid accessing programs that would help me. (Eubanks)

2. Bad targets. If police profile racial minorities when (for example) searching for drugs based on the claim that those minorities have higher offending rates, the result may be to increase overall crime rates, as those communities who are not searched take advantage of their freedom to increase their rate of offence. (Harcourt)

3. Out of control media. YouTube recommender systems prompt video producers to produce massive numbers of "spam" videos as an efficient way to gain traffic, and hence advertising income. (Bridle)

4. Lost expertise. Machine learning on health records may become the norm for some

diagnostic procedures. These diagnoses are based on the recorded observations of doctors seeking a diagnosis, but may change the role of doctors into data entry professionals. The quality of recored observations may change. (Froomkin)

5. New and arbitrary rules. Sorting systems are inevitably accompanied, sooner or later, by additional rules to prevent "gaming" the system. These rules may be important in our lives, but may have little legitimate authority or justification behind them.

6. Loss of autonomy. As sorting systems become ubiquitous, experimental and adventurous behaviour that may be perfectly legitimate becomes increasingly damaging.

These dangers tend not to appear when a sorting system is first introduced, but only later when it has displaced other forms of decision making. At that time, we risk being trapped without the ability to go back to where we were.

Responding to machine learning

Investing in a score

Consider a machine learning system S that maps features $\vec{x}$ of target individuals onto some output y, where the features may be specified (as in classical machine learning or predictive analytics) or may be discovered by the system itself (as in deep learning).

Suppose a change in output δy for a target brings a benefit to the target of δu. The target individual can make an investment δc in changing one of the elements of $\vec{x}$, x_i. The result of this investment will be:

$$\delta u = \frac{du}{dy} \cdot \frac{\partial y}{\partial x_i} \cdot \frac{\partial x_i}{\partial c} \cdot \delta c \qquad (1)$$

This investment is worthwhile only if $\delta u > \delta c$, or:

$$\frac{du}{dy} \cdot \frac{\partial y}{\partial x_i} \cdot \frac{\partial x_i}{\partial c} > 1 \tag{2}$$

What this means is that worthwhile investments are those for which:

- The outcome is important to the person being sorted.
- The outcome depends significantly on the feature.
- The cost of changing the value for a feature is not too large.

Responses to sorting systems have been given several names. Harcourt uses the word "elasticity" from economics [5]; Espeland and Sauder use "reactivity" [6]; [7]. Here I will use the word "elasticity".

Examples

Some examples may help. [Need to say why some of these are problematic.]

If a feature leads to a better outcome, there is an incentive for those being sorted to invest in that feature. The economics of signalling comes from the idea that some features (like university degrees) can be obtained more easily by some people than by others, and that employers want to hire those who can obtain one. The response to a sorting on degrees is a **separating equilibrium** in that it separates the sorted into identifiable groups. But more on this below. Insurance is built around "actuarial" practices that try to identify those who are good risks from those who are bad risks.

Harcourt: screening may lead to more crime [5].

Allegheny County Department of Health and Security in Indiana (?), USA adopted a machine learning algorithm called FTSA, which would determine if a child is potentially at risk of neglect or abuse. With a high enough risk score, child services would be called in to possibly take action. The FTSA risk score can change the entire life course of a child and its parents: there is a great incentive for parents to change behaviour that may flag their child as high-risk.

The vendors of many analytics applications keep the weights for the various factors secret to prevent their systems being gamed. Knowledge of which features matter the most makes it easier for targets to game the system by investing in changing just those features that make the most difference.

Appearance in Google search results is an important way for many businesses to reach potential customers. The outcome matters to the businesses, but finding ways to improve their ranking is costly. However, so many businesses are affected that an entire search-engine optimization industry grew up to help individuals and businesses improve their Google search ranking. The industry lowers the cost of changing the features that affect sorting outcomes.

There have been a number of attempts to algorithmically score individuals by their on-line influence. The best known of these to date was the Klout score, which measured the size and membership of a user's social media network (Twitter followers, LinkedIn contacts and so on) and ranked individuals between 1 (low influence) and 100 (a high "Klout"). But ultimately the Klout score was never taken seriously and Klout closed in 2018. The outcome was not important to the targets, and few people invested time or effort to improve their score. The rumoured Chinese social credit score, however, may be a reputation system that matters more.

It is easy to slip into thinking of this investment as "gaming the system" but that discussion comes later. For now, let's postpone any moralistic assessment and concentrate simply on incentives, noting that some forms of optimization are encouraged by system owners while others are discouraged. Uber's reputation system can be thought of as a simple sorting system using the five-start rating from each customer as input. Uber encourages its drivers to offer bottled water to ensure a good rating, but discourages offering money.

Accuracy and incentives

The problems of incentives are different to those of accuracy.

One might think that accuracy and elasticity are inversely related: the closer a machine learning system is to being completely accurate, the less any one target can do to improve their score. If this were true then the problems of elasticity would be solved in the same way as the problems of fairness; eliminating bias and ensuring fairness together would also solve the problems of elasticity. But this is not the case.

One reason is that "accuracy" for (supervised) machine learning is defined over some domain of data. Generalization beyond the training, validation and test set is always a challenge, and extrapolation beyond those sets provides opportunities for "optimization".

Notably, deep learning (DL) seems to generalize well in some dimensions – in part because of the use of stochastic gradient descent as an optimization procedure – but the results are often fragile in other dimensions. Well-known examples include photos that are indistinguishable to human eyes but which are categorized differently by DL systems, "single pixel" modifications, photos of animals in unusual environments (goats in trees, or being held by humans), three-dimensional models that are badly identified from all angles (tortoise gun). The same fragility is displayed by speech recognition.

Loosely speaking, DL systems enable optimization over a massive number of "features". Rule-based systems that rely on human input may use a few hundred features, and classical machine learning may use thousands of features, but the DL-based YouTube recommender systems "learn approximately one billion parameters and are trained on hundreds of billions of examples" [8]. The more parameters, the bigger the "attack surface": each parameter provides an opportunity for investment.

Ranking systems present the biggest incentives for over-investment. Ranking systems are zero-sum games for the participants being ranked: one person's move up the ladder is another person's move down. If being moving up the ladder brings real rewards, there is the possibility of an arms race between participants. Like the Red Queen in Alice in Wonderland, one must run as hard as one can to stay still.

Exit, voice, and sorting systems

Machine learning is a sibling of policy-based and "actuarial" systems. From evidence-based sentencing to insurance rates to the policy-specified scripts of call center operators, they are all driven by statistics and data.

Independent human decisions are different: clinical diagnosis, renting an apartment to an applicant, venture capital investments — these rely on individual judgement and reflection.

For those on the receiving end of these decisions, how can we react to decisions we think are unfair? In the case of human decisions we can exit: go somewhere else. We can get a second opinion, we can go to a different apartment, we can apply to another investor. The whole point of data-driven decisions is that they scale consistently so we may have fewer options. If low Google rankings are making our site invisible, if a machine-learning-based radiology program makes a diagnosis we doubt, we cannot easily go somewhere else. Instead, we appeal: we choose voice over exit.

There is, of course, a spectrum. Human decisions are not independent of culture, as many people of colour have found when trying to rent a place to live, and so exit is not always an easy option. And we can ignore some machine-learning decisions (for example: the Klout attempt to define an online "reputation score" was widely ignored until it folded). Still: the shift from human decisions to automated decisions is also, broadly speaking, a move from a regime of exit to a regime of voice as deciding agents have become scalable.

Any one individual human can provide only a few recommendations per day in response to other people's queries, but Siri or Alexa can make millions, and so the dynamics of "word of mouth" has changed.

It is because of the difference in scale and because of the change from a regime of exit to a regime of voice that the flaws of an algorithm matter more than those of any one human. Appeals can be heard only if there is transparency.

We have seen already that the incentive to "optimize" our outcomes depends on how much the outcome matters to us. This depends, in turn, on whether exit is a reasonable possibility or not. A consequence is that problems arising from bad incentives will not show up in pilot projects or in the early days of a new system. They show up when the system reaches scale, when it matters to more people, and matters to them more. Nobody would care about the Facebook newsfeed if it wasn't that it reaches billions of people. So pilot projects and comparisons of "accuracy"—however necessary they are—will never be sufficient to detect incentive problems with machine learning systems.

Incentive problems are often presented as unintended consequences, or unanticipated consequences. But perhaps we can anticipate some, and make decisions with them in mind. We can learn from previous generations of data-driven decisions.

Added note to be included somewhere

It remains a continuing debate how—or whether—to make data-driven systems accountable for bad individual decisions. Sometimes these are described as bugs to be remedied when possible. At other times the system owner may respond if driven to by the prospect of damaging publicity. And sometimes a system owner will insist on being judged statistically. When a Tesla running on "autopilot" crashed into a truck and killed its driver, CEO Elon Musk was quick to defend his company. Human-driven cars crash too, and it is unreasonable to expect that automated vehicles will not also crash. "Musk has said on multiple occasions that Teslas are almost '4x better' than average cars, with only one fatality per 320 million miles in cars equipped with Autopilot." (Bhuiyan)

It's super messed up that a Tesla crash resulting in a broken ankle is front page news and the ~40,000 people who died in US auto accidents alone in past year get almost no coverage. #+END$_{\text{QUOTE}}$

Categorizing, scoring, and ranking

Large-scale data-driven sorting systems did not start with computing, and we can learn about responses from studies of previous systems.

Categorizing

Responses are, it turns out, ubiquitous. People care about all kinds of sorting. They care about being categorized even when the effort seems purely descriptive and non-judgemental.

For a long time, the cause of death recorded on the public death certificate in the Netherlands was the same as that entered into the statistical records. The system was changed in 1927 when it was decided that the cause of death for records purposes was to be recorded *separately* from the entry on the death certificate. With this seemingly innocuous change "There was a considerable increase in Amsterdam of cases of death from syphilis, tabes, dementia paralytics, ... and suicide" [9] p 141.

The change makes sense once we know a bit more about the two records. The public death certificate is visible to the families of the deceased and potentially to others who knew them. Prior to the change, those who filled out the cause of death avoided naming causes that may have added to relatives' pain or caused embarrassment at a difficult time. Once the two entries were separated, they could enter these freighted causes of death into the statistical record without any awkwardness.

The example comes from Geoffrey Bowker and Susan Leigh Star's 1999 book *Sorting*

Things Out, which contains a wealth of others. Classifications of diseases may seem to be purely scientific in nature, but they govern access to resources and treatement for those who are being classified. Classification of race was, obviously, massively important in apartheid South Africa. The identification and classification of job responsibilities was central to the professionalization of nursing. Classification should, they argue, be reclassified (p 319): no longer purely descriptive efforts, they are "powerful technologies" that form part of the infrastructure of the world about us (p 319). Their study is made difficult by their everyday and seemingly banal nature: "Delving into someone else's infrastructure has about the entertainment value of reading the yellow pages of the phone book" (p 321). Still, there are "dances between the classifier and classified" to be observed. It's not just that "the map is not the terrain", but that the map and terrain "co-construct" each other.

Most classification schemes find a place for things that are left over: a "miscellaneous" drawer or a folder labelled "Other". These deserve particular attention, holding as they do things (or people) that "remain effectively invisible to the scheme" (p 325).

Scoring

If categorization inevitably introduces responses, scoring schemes do more. The categories now have an order.

Bernard Harcourt's 2006 book *Against Prediction* [5] argues against the use of "actuarial" or statistical methods brought into the law enforcement and justice systems under the banner of "evidence-based" practices and inspired by the profiling practices of the insurance industry. He argues that not only do people respond actively to these methods, but that their responses may undermine their professed goal.

Harcourt critiques econometric models of racial profiling. Does a disproportionately high rate of searches of minority motorists reflect "efficient discrimination" resulting from the desire to maximize the number of successful searches, or raw racial prejudice (p

112)?

You can think of efficient searching as a two-step process: assign a risk score to each motorist and search those who score above a certain threshold. If more minority motorists happen to score highly, then maybe that's just a reflection of the fact that they offend more often. If the "hit rate" is the same across racial groups, the different search rates are justified and not evidence of prejudice.

People respond to search rates. That minority drivers may recognize their increased risk and offend less as the chances of being searched increases, while majority group drivers may relax and take the chance of carrying drugs, knowing they are unlikely to be interfered with. There is an "elasticity" in the offence rate for each group.

Harcourt argues that the appropriate goal for searches is to lower the overall crime rate as much as possible and shows that, if the two populations respond to changes in search rates differently (which is plausible, given the differences in socioeconomic conditions), then the "equal hit rate" criterion is a bust. If the elasticity of minorities is less than that for majorities then the drop in crime from increased surveillance of minorities is more than made up for by the increase in crime in the relatively unbothered majority population. The best way to minimize crime, says Harcourt, is to search randomly.

Ranking

Of all the classification systems in use, ranking systems–where people (or their works) are ranked in an order from highest to lowers—are where the incentives become biggest. And nowhere are the consequences of ranking methods more peculiar than the ranking of US law schools by the US News and World Report (USN).

The USN ranking of law schools is described by Espeland and Sauder in their book "Engines of Anxiety" [7]. The authors describe how these rankings, produced by an unimportant magazine with no particular expertise in what makes a good educational institution, have come to dominate the life of US law schools. Deans obsess over moves up or down a few ranks. Administrations invest in those areas that they think will bring the biggest ranking improvements, while neglecting other areas.

The law schools are a perfect storm of incentives and the ranking system has become a trap.

Opting out of the rankings is not easy. There is (for whatever reason) no real competitor to the USN rankings. Students pay close attention to them and use them to make their application choices, so if law schools want to get the best students, then they have no choice but to play the game. Employers use the rankings to separate good students from others, some claiming they will take students only from "tier one" schools, so the students also find it difficult to opt out.

The rankings are also, to some extent, self-fulfilling. One of the most important parts of a good educational experience is the quality of the other students you study with.

Kieran Healy describes the dynamic in a review of Lauder and Espeland's book [?]:

> The academic legal establishment did not so much fall into this trap as become entangled in it. Like a fly touched by the thread of a spider's web, they were at first only lightly caught up, but then found that each move they made in response only drew them in more tightly.

The tyranny of educational rankings is not limited to law schools. There are a number of different rankings of "the world's top universities" and so no one ranking carries the weight of US N&WR, but these rankings have also become influential. Espeland and Sauder describe (p 183) how a position in the global rankings has become important in universities' competition to attract international students. Governments have also used the rankings as justification for new educational initiatives.

French universities have historically been organized differently to those in other countries (particularly the US and the UK), with many small institutes rather than a relatively few large ones. This model does not translate well into the ranking schemes, and so French universities have been poorly represented. Now the French university system is reorganizing itself: consolidating institutes into larger to boost their ranking, a reorganization driven not by any educational considerations, but simply by the dictates of influential rankings. The Universite Paris-Saclay, for example, has been constructed from twenty separate institutes. (Again, one could ask how this behaviour translates into the language of "gaming": see below for more on this).

The consequences of investment in rankings were emphasized in a 2015 commencement address by Yale Professor of Law Daniel Markowitz [10]. He starts by congratulating his audience: "you are sitting here today because you ranked among the top 3/10ths of one percent of a massive, meritocratic competition; and one in which all the competitors conspicuously agree about which is the biggest prize… For your entire lives, you have studied, worked, practiced, trained and drilled; and then you've been inspected, and finally—you made it here, after all— selected." This in contrast to the "old boy" network-driven admissions procedure of a few decades prior, when "A mid-century graduate recently reported that he came here after Jack Tate (then Dean of Admissions) told him at a college fair (straightway, and on the basis of a single conversation) 'you'll get in if you apply.'"

The meritocratic ranking of applicants is not, however, a simple advance in equality:

> [A]lthough it was once the engine of American social mobility, meritocracy today blocks equality of opportunity. The student bodies at elite colleges once again skew massively towards wealth: students from households in the top quarter of the income distribution outweigh those from the bottom quarter by 14 to 1… The skew towards wealth at the most elite universities is almost inconceivably greater still.

The monetary investment required to achieve what is needed for admission to Yale Law School is now far beyond the reaches of most Americans. Marlkowitz continues: "American meritocracy has thus become precisely what it was invented to combat: a mechanism for the dynastic transmission of wealth and privilege across generations. Meritocracy now constitutes a modern-day aristocracy, one might even say, purpose-built for a world in which the greatest source of wealth is not land or factories but human capital, the free labor of skilled worker."

The USN rankings are indeed "engines of anxiety". As we will see with internet systems, thinking of them in terms of "accuracy" in any statistical sense is to miss their biggest effect, which has been to reorganize and shape the structure of the industry.

Simplification

James C Scott.

The ethics of response

Tax avoidance. Payola. Teachers (O'Neill).

Moral hazard and adverse selection: principal-agent problems.

A new age for sorting

Digital technology has brought about a new age of data-driven sorting.

In public sector decision-making concerning individuals, judicial decisions, social program eligibility, and immigration decisions are among those that have been increasingly automated, moving from a rule-based, policy-driven form of systematic sorting to increasingly automated systems. The moves have been driven by promises of cost-effectiveness and efficiency, standardization, and objectivity (as we shall see below).

The move to digital record keeping provided a source of data for sorting in the form of business analytics. The two are natural complements: the more data is available, the more the reason to use it; the more we move to analytics and data-driven decisions, the more reason to maintain complete, consistent, and standardized data. Corporate databases are one source of such analytics. In the healthcare field the adoption of electronic records has led, as we shall see, to experiments using data-driven predictions to guide healthcare decisions.

The move to analytics drives further the move to what was called "business process re-engineering".

The classical story of the free market supply chain is the story of the pencil. As Milton Friedman said in *Free to Choose*: "There's not a single person in the world who could make this pencil." He goes on to list all the different parts that make up the whole: the lead, the wood, the paint, the brass ferrule and so on, before concluding that "It was the magic of the price system: the impersonal operation of prices that brought them together and got them to cooperate, to make this pencil, so you could have it for a trifling sum."

But companies now believe they can do better than the price system. By tracking the supply chain in ever-incresing detail, they can track every component, identify shortages, anticipate price changes, and more. Data-driven decisions make the supply chain "visibile" to management, and permit consistent, predictable operation from source to sale.

New data sources are appearing all the time. Internet of Things (IoT) devices feed into "data lakes"; smart cities promise to provide minutely detailed portraits of city activities; "quantified self" enthusiasms have morphed into massive resources of personal wellness data, from FitBit records to Apple Watch heart rate monitoring; DNA records from 23AndMe, Ancestry, and more feed into drug discovery and diagnostics.

For the largest digital platform companies, any activity on their platforms produces new streams of data, and this wealth of data puts FAANG (and Microsoft) in a category of their own. Facebook's intimate measuring of our attention, Google's detailed tracking of our location and activities, Amazon's insights into sales of potential competitors through its marketplace and AWS businesses.

In addition, new techniques are continually driving algorithms, applications, and whole industries based on this data.

Deep Learning

Digitization has been with us for decades, but since the 2011 "big bang" of the ImageNet competition, neural-network based deep learning techniques have sparked a new generation of powerful "AI" or machine learning techniques, which are producing new applications at a startling rate as unstructured data becomes increasingly useful as a resource. Major steps forward in speech recognition and image recognition are leading to innovations in conversational agents, augmented reality, self-driving cars, and robotics. The headline-grabbing advances of Alpha Zero and DeepMind promise a new era as automated systems can now handle tasks thought to be uniquely intellectual (chess and Go) as well as those "common sense" tasks (is that a dog in the picture?) that have long eluded AI researchers.

Even supposedly unstructured problems (image recognition) can be solved by computerized means. Intuition can now be reproduced at scale. It doesn't always work—there are plenty of crap, badly-thought-out deep learning systems out there—but the best are amazing. Conversation is now at scale: Siri, Alexa, and Google Home now have millions of distinct personal conversations per day.

While it is worth reminding ourselves that deep learning is essentially pattern matching on a huge scale, one of the new recognitions coming from deep learning is just how many of our own intuitive decisions are also pattern matching at heart. Not only the daily tasks, but also the seemingly intellectual tasks. Good chess players have a feeling for strong chess positions and good moves that comes, at least in part, from a wealth of pattern-matching experience. We do not yet know how far this will go.

The adoption of deep learning has been phenomenal. Google "has undergone a fundamental paradigm shift towards using deep learning as a general-purpose solution for nearly all learning problems" [8].

Pervasive sorting

As these sorting machines extend their reach we see new topologies of decision-making systems.

One topology is the chain of sequential systems, one feeding into the next. From child services to healthcare to hiring to educational opportunities and credit opportunities to policing and sentencing and bail, the major contours of whole lives are becoming shaped by the recommendations or guidance of machine learning systems. As these systems become coupled together the incentives for any one action become reflected in constraints on our next actions.

Another topology is the raft of systems that feed off a core of data. The opportunities we are aware of, the ways we see ourselves represented in the world, the prices we pay for services, immigration decisions: the list goes on. The way we present our "personal brand" through our daily activities (with "online" and "offline" activities increasingly coupled) affects not just one decision, but many, and the incentives to curate our presentation carefully are growing steadily.

With all these new systems, it is finally time to look at how we respond to digital sorting, and at the problems these responses may produce.

Responding to digital sorting

As would be expected, the responses to categorizing, scoring, and ranking systems that we saw above play out in the face of digital sorting. This section makes an examples-based survey of the forms responses take.

Allegheny FSTA

In her book *Algorithmic Inequality* [11], Virginia Eubanks explores the ways that automated decision systems affect those who come into contact with Americal social services agencies. One of those systems is the Allegheny Family Screening Tool (AFST), an analytics application used to predict child abuse or child neglect at the time of birth, to alert child services to children who may be at risk.

AFST was developed in New Zealand by economists Rhema Vaithianathan and Emily Putnam-Hornstein. From a starting set of over 200 variables, they built a predictive model based on 132 demographic and personal history variables, including such things as the length of time a parent has been on benefits, the mother's age, single parent status, and so on. Among these variables is other contacts with Allegheny social services for parenting support programs.

Eubanks says that the "AFST oversamples households that rely on public assistance programs", in part because "the AFST only has access to data collected on families using public services, not on those that access private resources for parenting support."

Eubanks provides an in-depth critique of many aspects of the program. Here I think of the program as a scoring system. The score has a large effect on the lives of those who score high enough to warrant the attentions of child services, there are variables that have an effect on the outcome, and those being scored may be able to change those inputs. As Eubanks writes: "Targeting 'high risk' families might lead them to withdraw from networks that provide services, support, and community… AFST might create the very abuse it seeks to prvent. It is difficult to say a predictive model works if it produces the outcome it is trying to measure." [11], p 169.

Allegheny County DHS disputed Eubanks's "incorrect assumption that the more public benefits (SNAP / TANF) a family accesses, the higher the AFST score. For 45% of families, receipt lowers the score" [12]. The AFST tool, the DHS claims, went through an extensive process including "careful procurement, community meetings, a validation study, independent and rigorous process and impact evaluations and an ethical review, which concluded that not only was use of the AFST ethical but also that not using it might be unethical because of its accuracy."

Eubanks replied that without seeing the details of the model, which is not public, it is not possible to know for sure under what conditions contact does lead to a higher score [13]. [I thought there was a personal story in Automating Inequality: check this.] In the absence of a more concrete assertion from DHS, Eubanks' claim is not refuted. But even a perception that taking advantage of parenting assistance programs may endanger custody of one's child could be a real incentive to avoid those programs.

A lesson from the AFST program is that, as in the case of causes of death in the Netherlands, data is always coupled to its use. Changing the use changes the incentives around the collection of the data, potentially with harmful consequences.

Sorting and the professions

Social workers

We usually expect that sorting system owners do not want users to game the system, but there are exceptions. A paper from Canadian legal scholar Jennifer Raso describes one such case in depth [14].

Ontario Works is the welfare program for the province of Ontario. For many years, caseworkers classified applicants into deserving or undeserving of social assistance. There were claims of prejudice from welfare rights advocates; the auditor general claimed that caseworkers were unduly generous, an opinion that found favour with fiscal conservatives. Over a few decades, caseworkers' judgements were increasingly legalized and formalized and specified as front-line decisions became reviewable by courts.

In 2014 a new Social Assistance Management System (SAMS) was introduced. SAMS requires caseworkers to fit applicants into categories listed in drop-down menus (this is a rules-based decision system) and it generates decisions based on those categories. SAMS was intended to "de-centre caseworkers as decision makers" and to provide an objective, transparent, and consistent assessment. The jobs of front-line workers were deskilled as they became data entry workers rather than social workers, although less so than in some other jurisdictions, as the vendor (IBM) claimed that "SAMS would ostensibly make easy, routine benefits decisions, and free up workers to meet with more complex clients" [14] p 8.

The end result has been a bit more complex. Raso writes that "To say SAMS failed to function as promised is an understatement." Even after initial problems were solved, the software still makes decisions that do not match real-world circumstances. For example, SAMS makes its own inferences about who is a family member, and "may make sole-support mothers dependent on previous household members, such as former intimate partners or their parents, even where caseworkers have reviewed relevant evidence and determined that these individuals do not live together".

SAMS also encodes strict interpretations of some requirements, and in response "workers find ways to redirect SAMS so that their clients receive the benefits they are entitled to according to flexibly-worded Ontario Works legislation".

> [C]aseworkers learn how to expertly translate and input client data to produce decisions that more closely match their interpretation of clients' needs and welfare laws. The ways in which workers "manipulate the system" to

> produce a particular decision, though common knowledge among their colleagues, are black boxed to program managers, auditors, and benefits recipients.

There are gaps between SAMS and the Ontario Works legislation it implements, and caseworkers "often sophisticatedly balance the contradictory program purposes that legislators leave unresolved", or which SAMS and the legislation leave ambiguous, by entering fake data into fields that SAMS requires to be filled. SAMS fails to capture the complexity of applicants' lives, and caseworkers become experts at mediating with the system to provide what they believe their clients need.

Journalism and the justice system

Medicine and Deep EHR

If there is one area in the world of work where Deep Learning (DL) techniques are being investigated it is surely healthcare. The two major stories are the use of supervised learning CNN image recognition techniques to diagnose conditions, and the use of DL techniques on electronic health records (EHR) as a means of screening. It is the latter use case of most interest.

The application of DL to EHR promises to make another change in the nature of professions, driving them further towards data entry. There is a phrase in software engineering: yak shaving, which can mean something like procrastination: doing a less useful task in order to avoid a more important task. Keeping perfect patient records at the cost of actually treating patients used to be an example. Records should be kept reasonably, but the focus should remain on the patient. The map is not, after all, the territory. But now that is changing and maintaining the digital image of a patient may be the most important part of a doctor's job.

A review of "Deep EHR" research [15] starts by emphasizing the appearance of a new data set. The US HITECH act of 2009 provided $30B incentives for hospitals and doctors to adopt EHR systems, and in the decade since adoption in hospitals has grown by 9-fold to 84% and in doctors' practices it has doubled to 87%.

EHRs are large and complex documents. For any patient they include demographic information, medical history, images, clinical notes and more. The HITECH act standardised reporting using codes from the ICD-10-CM and ICD-10-PCS standards, which together comprise more than 140,000 codes of remarkable detail,[16]. There are, as just one tiny example, sixteen different codes under the topic of "shoulder stability" [Cabitza].

Even without considering the freeform clinical notes and medical images, this is a very sparse and high-dimensional data set: just the kind of thing for which DL techniques have proved move effective than previous methods. In one investigation [17], the investigators note that "Traditional modeling approaches have dealt with this complexity simply by choosing a very limited number of commonly collected variables to consider." The volume of data is remarkable: on average each patient's record had over 200,000 discrete pieces of data to be analyzed, and the study employed a data set with over 46 billion tokens of EHR data. It may not be surprising that, while the data came from two American hospitals, 31 of the 36 authors are from Google. The study used the data to predict mortality, readmission likelihood, length of stay, and the likely diagnosis at discharge for over 100,000 patients. The authors note that "using the entirety of a patient's chart for every prediction does more than promote scalability, it exposes more data with which to make an accurate prediction". The success of the study is in part because "personalized predictions... leverage many small data points specific to a particular EHR rather than a handful of common variables".

Paper-based patient "charts" or medical histories were traditionally used as monitoring tools, to track selected items from visit to visit, or from day to day. With the introduction of Electronic Health Records the focus of patient documentation changed: they became used for "institutional priorities" [16], including billing to insurance companies. They also provide a foundation for oversight of physician practice, part of "the increasing emphasis placed on enumerated documentation to facilitate external evaluation and verification of processes and procedures." As a consequence they have become more standardized and more consistent.

EHRs introduced new incentives regarding what is recorded. As with, for example, the Ontario Works program discussed above, clinicians may "strategically classify diagnoses—among the many specific options—to assure they meet coverage" [16].

The ICD codes are complex, but are neither precise nor unambiguous. There are inherent uncertainties: patients misstate symptoms and doctors may not trust patients' self-reporting.

EHR data remains often incomplete for at least two reasons. It is mainly used in conjunction with the patient him- or herself, as a supplement to ongoing conversations. Data entry itself is far from free, so that data entry is selective. The absence of a body-mass-index recording may not mean the number is unimportant, simply that the physician can see that it is not a problem for a particular patient and so may decide not to record it. As an ethnographic (?) study of physicians notes [16]:

> When clinicians enter a consultation room, they situate the computer in front of them as they face the patient. They attend to the computer continually throughout the consultation, working on the electronic health record (EHR)— opening and closing pages, clicking check boxes, and typing notes—stopping only for the briefest of physical examinations before returning to the computer. When consultations are over, the clinicians return to their desks with the laptops and type away to complete required entries in the EHR, sometimes for long hours after their patients have gone.

As with Ontario social workers, doctors have had to come to turns with a change in their professional identity as EHRs have been introduced, and some have made their expertise with the new systems a matter of professional identity [18].

And now comes another change. If EHRs are to be adopted as a predictive tool the standard for their completeness and consistency changes, and the incentives around their maintenance changes too. It is not clear if, from a purely statistical point of view, EHRs can support these "secondary uses". A recent survey of the topic [19] identified a set of problems:

- Incompleteness – missing information;
- Inconsistency – information mismatch between various or within the same EHR data source;
- Inaccuracy – non-specific, non-standards-based, inexact, incorrect, or imprecise information.

These problems are not, we should emphasize, problems at all unless we consider the EHR from a point of view of "data quality". But if EHRs become tools for predictive use, we will see a further change in the role of doctors: they will become even more subsidiary to external decision making (a loss of professional prestige and identity), will focus more on creating and maintaining a thorough digital representation of a patient's health.

Some concerns have been raised in [20] and elsewhere. Deskilling is an obvious one: if the nature of the physician as a profession changes, the quality of EHR data may decrease rather than increase.

Reputation systems and inflation

Automated essay scoring

Automated marking of essays has long been a goal of education technologists.

YouTube and social media incentives

YouTube now uses a deep-learning algorithm to generate recommendations for its videos (what to watch next) [8], having previously used a more traditional matrix factorization technique [21]. Both optimized for predicted watch time.

Video producers looking to increase their viewership must follow the behaviour of the recommender system, producing more of what works and less of what fails.

One video creator, Johnny Tanber, told Buzzfeed News in 2017 that "In terms of contact and relationship with YouTube, honestly, the algorithm is the thing we had a relationship with since the beginning. That's what got us out there and popular... We learned to fuel it and do whatever it took to please the algorithm." [22]

The same article quotes Davey Orgill, who left his job to make superhero parody videos, and whose channel reached 2 million viewers before being shut down. He argued that "the platform is responsible for encouraging... objectionable, sexual, and violent superhero content ostensibly oriented toward children... 'YouTube blames it on these people that were doing it, but for a year their algorithm pushed this content... People were doing it because it was creating millions and millions and millions of views. They created a monster." The algorithms would promote videos that the company later claimed "violated our Community Guidelines"

It's not only "ab initio" algorithms. Google also uses a network of contractors who rate search quality and so help train the AI. "We use search raters to sample and evaluate the quality of search results on YouTube and ensure the most relevant videos are served across different search queries" says the company [23]. It's "fauxtomation" at work.

In reply, "The company also promised to apply its 'cutting-edge machine learning' that it already uses on violent extremist content to trickier areas like child safety and, of course, said that it plans to have more than 10,000 human reviewers evaluating videos on the platform in 2018."

Views are a challenging metric though. Many are fake [24]. Close enough to risk an "inversion"

Two lessons: the idea of gaming the system misses much of what is happening. Better simply to think in terms of incentives. An algorithmic governance system (recommender system in this case, together with various filters and human raters) creates an environment that rewards certain kinds of behaviour. Some of this behaviour goes against public or company "guidelines". The platform then needs a second level of governance to handle those who follow the rules of the algorithm. It *inevitably* creates a problem of dubious material; and seeks help to manage it. Accuracy here plays little role. The recommender systems may be tuned for eyeballs while the guidelines are built around values. If the gap is wide, maybe the recommender system needs a rethink. Or maybe we should look again at what we have created: the claims that recommender systems would come without side effects.

Also, the company has two contradictory sets of rules: the algorithmic rules and a set of values-based rules or principles. If there is a gap, the company is free to navigate how it handles it. First take the money, then decide that the algorithm got it wrong. Blame the content producer and present itself as a noble actor trying to maintain standards.

James Bridle's essay "Something is wrong on the internet" [25, **?**] shows how the algorithm has worked when it interfaces with automated or semi-automated video production.

Responding to responses: ethics and the intention gapn

Facebook and Airbnb and Yelp: ill-defined values and algorithms. Facebook: blaming people for following incentives that they create.

How do we think about the gap?

Who gets to learn from mistakes? Autonomy

Systems learn from mistakes: the neural paradigm.

Speech as exploration, not statement of opinion.

Learning from mistakes vs predictive accuracy. The static individual and the ability to change.

Only some people can exit. Innovation and experimentation at the platform level replaces innovation and experimentation at the level of the managed.

Outlawing user actions

Keeping the system clean by punishing individual action.

Yelp again

Andrew Ng and autonomous vehicles.

Acceptance

Essay writing as good writing.

Reputation systems.

Gaps as bugs: whose responsibility?

Seeing the world through the lens of the software development process.

Taking ownership of the gap. Work with us to improve. I don't know about insoluble, but certainly underestimating the scale of the problem because they did not recognize that they are introducing new incentives. deep learning systems: will they always be vulnerable? What distinguishes the unexpected from the malicious? g

End Matter

References

[1] Frank Pasquale. *The black box society: the secret algorithms that control money and information*. Harvard University Press, Cambridge, 2015.

[2] Cathy O'Neill. *Weapons of Math Destruction: How big data increases inequality and threatens democracy*. Crown Random House, 2016.

[3] Solon Barocas and Andrew Selbst. Big Data's Disparate Impact. *California Law Review*, 104:671, 2016.

[4] Safiya Umoja Noble. *Algorithms of Oppression: how search engines reinforce racism*. New York University Press, 2018.

[5] Bernard E. Harcourt. *Against Prediction*. University of Chicago Press, 2006.

[6] Wendy Nelson Espeland and Michael Sauder. Rankings and Reactivity: How Public Measures Recreate Social Worlds. *American Journal of Sociology*, 113(1):1–40, July 2007.

[7] Michael Sauder and Wendy Nelson Espeland. *Engines of Anxiety: Academic Rankings, Reputation, and Accountability*. Russell Sage Foundation, 2016.

[8] Paul Covington, Jay Adams, and Emre Sargin. Deep Neural Networks for YouTube Recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems*, RecSys '16, pages 191–198, New York, NY, USA, 2016. ACM.

[9] Geoffrey C. Bowker and Susan Leigh Star. *Sorting Things Out: Classification and its Consequences*. The MIT Press, 1999.

[10] Daniel Markowitz. A New Aristocracy (Yale Law School Commencement Address), May 2015.

[11] Virginia Eubanks. *Automating Inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2017.

[12] Allegheny County DHS. Statement in response to "Automated Inequality" by Virginia Eubanks, 2018.

[13] Virginia Eubanks. A response to Allegheny County DHS, February 2018.

[14] Jennifer Raso. Displacement as Regulation: New Regulatory Technologies and Front-Line Decision-Making in Ontario Works. *Canadian Journal of Law & Society / La Revue Canadienne Droit et Société*, 32(1):75–95, April 2017.

[15] Benjamin Shickel, Patrick Tighe, Azra Bihorac, and Parisa Rashidi. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604, September 2018. arXiv: 1706.03446.

[16] Linda M. Hunt, Hannah S. Bell, Allison M. Baker, and Heather A. Howard. Electronic Health Records and the Disappearing Patient. *Medical Anthropology Quarterly*, 31(3):403–421, 2017.

[17] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine*, 1(1):18, May 2018.

[18] Adam Reich. Disciplined doctors: The electronic medical record and physicians' changing relationship to medical knowledge. *Social Science & Medicine*, 74(7):1021–1028, April 2012.

[19] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary Use

of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translational Bioinformatics*, 2010:1–5, March 2010.

[20] A. Michael Froomkin, Ian R. Kerr, and Joelle Pineau. When AIs Outperform Doctors: Confronting the challenges of a tort-induced over-reliance on machine learning. SSRN Scholarly Paper ID 3114347, Social Science Research Network, Rochester, NY, November 2018.

[21] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, and Dasarathi Sampath. The YouTube Video Recommendation System. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 293–296, New York, NY, USA, 2010. ACM.

[22] Charlie Warzel and Remy Smidt. YouTubers Made Hundreds Of Thousands Off Of Bizarre And Disturbing Child Content. *BuzzFeed News*, December 2017.

[23] Davey Alba. YouTube Has A Massive Child Exploitation Problem. How Humans Train Its Search AI Is Partly Why. *BuzzFeed News*, December 2017.

[24] Michael H. Keller. The Flourishing Business of Fake YouTube Views. *The New York Times*, August 2018.

[25] James Bridle. Something is wrong on the internet, November 2017.