http://dx.doi.org/10.14236/ewic/EVAC18.3

On the Cruelty of Computational Reasoning

Alexander Hogan
ETIC Lab
Newtown, UK
alex@eticlab.co.uk

Kevin Hogan
ETIC Lab
Newtown, UK
kevin@eticlab.co.uk

Christian Tilt
ETIC Lab
Milton Keynes, UK
chris@eticlab.co.uk

We seek, firstly, to demonstrate the cruelty of current computational reasoning artefacts when applied to decision making in human social systems. We see this cruelty as unintended and not a direct expression of the motives and values of those creating or using algorithms. But in our view, a certain consequence nevertheless. Secondly, we seek to identify the key aspects of some exemplars of AI products and services that demonstrate these properties and consequences and relate these to the form of reasoning that they embody. Third we show how the reasoning strategies developed and now increasingly deployed by computer and data science have necessary, special and damaging qualities in the social world. Briefly noting how the narrative underpinning the creation and use of AI and other tools provides them with power in neoliberal economies. Creating a disempowered data 'subject' in an inferior, economically and politically supine position from which they must defend themselves if they can.

Human Rights, Artificial Intelligence, Big Data, Society

1. INTRODUCTION

In this paper we discuss the consequences of applying algorithmic reasoning to and within human social systems. Our core argument is that algorithms in the form of logistic regression, Neural nets, decision trees and other forms of algorithmic reasoning are necessarily hierarchical in their impact as they are born of and applied to a hierarchical social system; they create a data (subject) that is located at the bottom of a hierarchy of power, removed from historical and personal subjectivity and essentially decontextualised for the purposes of machine reasoning. It is our contention that the application of pseudo cognition or (algorithmic) reasoning for whatever purpose will, in this context have results that are cruel.

Algorithms are widely used in purely commercial applications and both the range and scale of these applications has grown enormously in recent years. A similar pattern is being followed in their adoption by state and public agencies such as health systems, police forces and social security agencies. They have come to play a role in monitoring and even 'policing' humans and human behaviour. Such systems are frequently described as providing support for decisions made about critical and often intimate moments of people's lives but operational protocols have evolved in some settings to the point where they actually making the decisions.

When algorithms are recruited to positions of power within institutional interventions in complex social settings, we observe that they produce consequences, both intended and unintended, which are cruel or worse. This outcome, we argue, is not an accident, the product of poor design, or malice. Our position on the consequences of this contention are two-fold; that the unfortunate and unintended consequences will not be prevented by more rigorous methodology, a consequence we have termed the 'Law of Unintended Barbarity' and secondly that the use of algorithmic based decision making needs to be better understood in the wider context at the outset in order to mitigate this effect. Addressing the motivations and goals that frame the problems that are chosen for applications, by whom and for what purpose, who it is exactly that will be subject to the results and how these effects will be monitored.

1. RECOGNISING BIAS

Bias can manifest itself in both ''positive'' and ''negative'' ways, we focus here on unjust discrimination as a negative outcome of systems incorporating algorithmic decision making. An important distinction needs to be made between bias and treating people differently for ''good''. For example, it is arguably justifiable to treat people already in debt higher interest rates even though this does mean they are being discriminated against as a group. We follow Howard and Borenstein (2017) who suggest that applications should consciously acknowledge that an individual or group is going to be treated differently. We suggest that cases where such conscious recognition does not occur and no explicit attempt is made to mitigate the consequences in concert with both decision makers

and impacted individuals are examples of unjust discrimination arising from negative bias.

First we discuss biases arising internally from project design and execution; the bias arising from (necessary) selection in the sources of data included in systems for their training and development and the bias arising from the institutional structures that inform both the formulation of technical problems and acceptable solutions. Second we discuss those sources of bias implicit in any sociotechnical system; the types of bias inherent in human beings interacting with complex decision support tools such as complex algorithms whether based on AI tools or not.

2.2 Bias in Health Insurance

Virginia Eubanks in her book “Automating Inequality” describes the difficulties encountered by the author and her partner after being struck with intense healthcare needs only days after changing jobs and hence health insurance providers in the US. She ran afoul of an algorithm designed to look for fraudulent behaviour in customers.

In making a significant claim so soon after starting on a plan and for her partner, someone in need of significant pain medication known to be abused, she was identified by the system as a highly probable case of fraud. This in turn set in motion a set of circumstances which created a terrible and frightening experience.

Not only was she flagged as a suspect and faced with expensive and dangerous consequences - removal of coverage at a critical moment - but the systems she encountered for administering her health insurance and health-care subsequently escalated and persisted with the ramifications of her misclassification. This continued even after she had “corrected” the initial mistake. No ‘entry point’ existed for her to put right what should never have gone wrong and the further she went from the original error, encountering different platforms for acquiring and paying for health-care, talking to different representatives of various services etc. the greater the difficulty in catching up with the ramifications of the initial mis-classification.

As Eubanks observed, she was greatly helped by her experience, education and support networks in overcoming this nightmare at what was already a moment of great stress. Had she not been in possession of these things (as many people are not) her experience could have been so much worse.

Although ostensibly providing a service for those in need of health insurance the individual patient is not the focus of this system accept insofar as they may endanger profitability. The focus here is the detection and prevention of fraud. If nothing else the system should have been able to better accommodate the outcome of the case being flagged and resolved. Not only was the victim not able to undo the damage caused by the fraud warning, but it seems likely if not certain that the system was also not able to ‘learn’ from this mistake. A narrow focus for the project, without due concern for the patients and their loved ones as necessary stakeholders and ‘users’ resulted in an experience which undermines the purpose of the service itself.

2.1 Bias in Public Service: improving teachers

A serious example of the Law of Unintended Barbarity was explored in detail by Cathy O’Neil in her book Weapons of Math Destruction(O’Neil, 2016). An algorithm designed for quantifying the effectiveness of teachers in a Virginia school district was deployed as the basis for firing a fixed percentage of low performing teachers annually. This process was designed to force teachers to improve their performance and claw their way into the appropriate performance percentile and hence save their jobs.

The algorithm designed was based mostly on exam results - an easily obtainable and useable metric - ignoring any duties and responsibilities a teacher may have undertaken that could not be cheaply and reliably measured. As a result, teachers with excellent peer reviews and many years experience found themselves unemployed. On investigating their situation, evidence was found that they had taken on classes who’s grades likely been inflated in previous years (perhaps on behalf of less scrupulous teachers gaming the system as they knew how the algorithm was likely to work). It is also easy to imagine a situation where a student in this district (yet again poor) in having to deal with all the other stressors of life they encounter, did not place particular importance on end of year exams, especially relative to their more privileged peers. One unintended consequence was that many of the sacked teachers were able to find new jobs, over the border in Virginia in a more prosperous school district with no such algorithmic regime. In short good teachers were lost and the community as a whole demoralised.

Once again such an intervention is only possible if the ‘subjects’ of the project - working teachers, are decontextualised, denied their history of prior performance and the project designed without including all of the important stakeholders. The project goal of improving educational outcomes is not itself a reason why teachers were not included in the project design or indeed a richer definition of

'good performance' defined and operationalised within the project. It is possible for example, to conceive of a project utilising the same technologies that provided support to teachers in order that they might have been better able to teach. The first intention however, was that the tool developed in this project would allow management to manage teachers in a way that 'objectively' sidestepped their history of achievements, skills and commitment. In this case the choice of a metric for teacher performance embedded in the new system both reflected and intensified bias already inherent in the notion of measuring educational outcomes using simple quantitative measures alone.

2.3 Bias in public service: improving the courts system

In recent years a number of predictive policing systems have been implemented in the USA and more recently in the UK. They have quickly been subject to a degree of criticism from both academic and journalistic sources, not least with respect to the contention that they have been designed in such a way as to reflect the bias of extant policing methods with respect to crime reporting and management (Joh, 2017). We discuss here another application of AI techniques in the criminal justice field, in this case to the question of sentencing and the decisions underpinning the use of probation orders.

Building on a tradition of applying psychometric methods in predicting outcomes for offenders a private company, Northpointe created a system for correctional offender management profiling for alternative sanctions (COMPAS). This tool has been used in various jurisdictions for managing defendants pre-trial and post trial for sentencing. The use of COMPAS has been challenged and criticised in the courts, press and academic papers (Angwin et al.,2016; Chouldechova, 2017). The system has been ably defended in terms of Its technical quality for predicting recidivism in the USA and against claims that it is biased against defendants from non-white ethnic backgrounds (Flores, et al. 2016). It is however reasonable to observe that the COMPAS system is twice as likely to predict blacks as being at a high risk of reoffending when they do not in fact do so in comparison to white defendants; 44.9% of blacks were rated as a high risk of reoffending but did not and 23.5% of whites were rated as having high risk for reoffending when they did not go on to commit another crime. Conversely, white defendants are far more likely than blacks to be predicted not to reoffend (44.7%) when they do in fact go on to reoffend and this is compared to 28% of black defendants who having been labelled low-risk, go on to reoffend Angwin et al., 2016). Detailed research examining this phenomena suggest that in cases where two groups (as is the case with black and white crime) systematically differ in the rate at which they offend and reoffend then it is a mathematical impossibility not to create differences in the false positive and false negative rates for those two populations with a tool such as COMPAS (Chouldechova, 2017). Bias in this case is not an unhappy coincidence or the consequence of technical failings in the design of the system, it is an inherent property of the system.

The advent of COMPAS and similar systems reflects the underlying logic driving the development of algorithmic tools in many public services in the USA and UK. In discussing the development and application of COMPAS the company behind the system expressed the concerns of their customers that are addressed by the product; "In overloaded and crowded criminal justice systems, brevity, efficiency, ease of administration and clear organization of key risk/ needs data are critical." (Northpointe, 2013).
Similarly, in a review of the implementation of a similar system in the UK the logic underpinning Durham Constabulary's harm assessment risk tool (HART) was described; "As is common across the public sector, the UK police service is under pressure to do more with less, to target resources more efficiently and take steps to identify threats proactively" (Oswald, Grace, Urwin and Barnes, 2017 p1). They go on to characterise the what a better outcome would be as the system is deployed; "a more efficient use of police resources in a landscape of more consistent, evidence-based decision-making".
Missing from these accounts of developing and designing AI based interventions in the police and criminal justice system is the effort expended to seek out the views of those who might use or be the subjects of such systems. Nor indeed can we find a discussion of the opportunities that such technologies might bring for aiding the workers; police, probation officers and officers of the courts with the problems they experience. Least of all are we presented with evidence of efforts to ascertain what those subject to such systems felt about its utility and the ways which they impact upon their lives. No evidence is presented as to how the lessons learned by the staff deploying the system or those subject to it can be utilised to improve the system and its outcomes. In short a cheaper, more consistent version of what has gone before is created but with even less opportunity for those who work with or are subjects of the system to understand it, complain about its consequences or to improve matters. Even if it is possible to improve the reliability of systems with improved data preparation and collection, the creation of tools to more consistently and reliably discriminate against 'subjects' of the system is by no means certain to enhance society.

2.4 Bias in Human Operators of complex systems

As the study of human operators in machine systems accelerated after the second world war and towards the end of the twentieth century, two related cognitive tendencies were noted as key features in many system failures. Accidents and systemic failure can occur where people and automation, especially decision support systems work together and people behave in accordance with a bias which leads people to place too much reliance upon systems which they are also biased to believe to be more powerful than they are. The first bias takes the form of an overreliance on machine systems manifested as low levels of vigilance and suspicion with regard to to the state of machines. In short human operators quickly come to rely on machines, fail to monitor their performance and 'trust' them so much that they fail to react even to inappropriate and unusual outputs from the system. This will lead to missed cues concerning dangerous and expensive failures and missed opportunities when things go badly wrong.

The second form of bias is a natural tendency on the part of system users to ascribe greater power (efficiency and effectiveness) and autonomy (a limited need for supervision) to automated aids than to other forms of advice (Markus, 2017).
Perhaps most worrying is the considerable body of research evidence that shows how powerful this effect is, it has been found in many different settings and shown to influence both naive and expert users, from trainees to expert pilots. The problems arising from this effect are exacerbated both by high levels automation and higher levels of reliability of the tool or aid. In effect as the system becomes more reliable or consistent then so too does human performance in the system degrade. The research output of the Cognitive Ergonomics community has made it clear over time that training and explicit instructions to verify the output of automatic decision aids cannot be deployed to prevent bias and overreliance in users. Ultimately the ability to successfully monitor the behaviour of cognitive decision aids or rather system outputs, depends very largely upon the extent to which individual users feels themselves to be responsible for the overall performance of the system. If the system is large, amorphous and without clear lines of responsibility and individual accountability the decision aid(s) will ultimately be run without supervision - despite attempts to prevent this.

In order to appreciate the systemic nature of these biases and their impact upon any systems containing both humans and decision support tools it is important to recognise as suggested by Parasuraman and Manzey (2010), that "complacency and bias are observed in highly trained and expert machine operators, as well as novices, and cannot be attributed to a lack of knowledge or skill" (Parasuraman and Manzey, 2010). Markus (2017) describes how automated decision aids contributed to the destruction of the lending market in the US amidst the crash of 2006/7. What happened was that the lending decision support algorithm had became a tool for use by novices rather than a decision support tool for experts. Novices with different capabilities and motivations behaved in different ways from the original planned users of the system. To make matters worse, the participation of any human staff in the control loop diminished as incentives, production pressures, and cognitive biases effectively came to discouraged supervision and intervention in what became entirely automated decisions (Markus, 2017).

In light of the certainty of human bias in complex systems incorporating algorithmic decision making we can expect serious operational problems. Unless project management and evaluation is considerably improved and widened beyond the technical scope of the software tools themselves, then failures are highly likely.

2. ALGORITHMS AS ECOLOGY: DEBT MANAGEMENT PLANS

The final type of bias we wish to describe arises from the increasing numbers of people who find themselves, often unwittingly, subject to multiple actors deploying a variety of algorithmic tools to 'manage' their lives. The results of such interactions are a prime example of the phenomenon of unintended cruelty. Our example came from a commercial research project to produce decision support applications for a large firm in the financial sector. In this project we learned from having to explicitly understand the relationship between individual behavior, personal indebtedness and algorithmic reasoning systems. It was simple enough to create an AI product to predict a particular behavior at a certain point but we needed to undertake far more research to understand what was going on. The project required that we process a large data set comprising the actions taken by people and the companies they interacted with as they struggled to manage personal debts; loans, overdrafts, store cards, unpaid bills and other similar forms of debt.

On investigating we found, counter to our initial expectations, that this data did not reflect the UK's national picture for the distribution of debt. There weren't more women than men as clients (despite there being far more women in significant debt in the general population), young people were under-represented in the sample and older people over-represented. Most important, it was the profile of the creditors to whom the debt was owed - and not the amount of debt - that predicted what actions the individual's ultimately took in order to 'deal with' their distress. People who owed money to debt collection agencies and in fact individuals who owed money to three or more debt collection

agencies made up the majority of cases. Unwittingly because they were unaware of one anothers actions, debt collection agencies when acting at the same time with the same person where causing them to simply give up and admit that they could no longer support the situation.

We realised that what we were looking at was a data set of debtors comprised almost entirely of people who had sought out help with their situation because of the strategies utilised by their various creditors (and agents who had subsequently taken ownership of their debt) to retrieve money. Once the initial owner of the debt, a bank or car loan company had applied an algorithm to decide to sell that debt on to a collection agency and the debt agency had subsequently applied algorithmic reasoning to decide on how to follow up their 'customer'. Behaviours such as seeking a county court judgement, appointing bailiffs and other tactics including selling on the debt themselves. At the point where individuals found themselves dealing with multiple such companies they would most likely come forward to ask for help in constructing a debt management plan.

We had access to a customer base of more than 60,000 people in debt but what predicted their financial struggles and how they sought to resolve them was not how much they owed but who they owed it to.
So homogenous was the data, with respect to creditor profile, that it quickly became obvious that we were not the first to have algorithmically processed these people (or rather their data) and likely they were a hyper-selected population. They were the product of repeated financial decision-making machines ruling them unfit for further profit and thus handing them off to the next rung on the ladder as customers (debts) from which debt purchasing companies could with increasingly unpleasant measures, continue to extract value. At this point people were very often incapable of helping themselves and this not just because of the level of debt and repayments but the simple impossibility for many of even operating the system requiring them to understand and negotiate the many and complex demands being made of them.

In this case we see how when multiple actors in contact with individual people (customers) deploy algorithms to manage their affairs the impact upon those people of the cumulative impact can be terrible in their consequences. As the banks and other lenders automate the decision to minimise their exposure by selling on personal debt and debt purchase companies manage their relationship with their customers using similar systems to manage the debt to best advantage the cumulative impact on customers can impact health and wellbeing, family life and loved ones. Finally the creation of a selected sub-group who have been and are being managed by algorithms creates a powerless, vulnerable group who are defined by their status as

3. REFERENCES

Angwin J, Larson J, Mattu S, Kirchner L. (2016) Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. A vailable online at *www.propublica.org/article/machine-bias-risk*-SSRN: *https://ssrn.com/abstract=3020259*

Markus, L. (2017) Datification, Organizational Strategy, and IS Research: What's the Score?, *The Journal of Strategic Information Systems*, Volume 26, Issue 3, Pages 233-241,

O'Neil,C.(2016) *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* Crown Publishing Group New York, NY

Oswald, M., Grace, J., Urwin, S. and Barnes, G., (2017) Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality. Information & Communications Technology Law, Forthcoming. Av a i l a b l e a t S S R N : *https://ssrn.com/ abstract=3029345 or http://dx.doi.org/10.2139/ssrn. 3029345*

Parasuraman, R., Manzey, D.H., 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration,". *Human Factors* 52 (3), 381–410.

Chouldechova A. (2017) Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments *Big Data*.Volume: 5 Issue 2

Dieterich,W Oliver,W Brennan,T. (2013) Predictive Validity of the COMPAS Reentry Risk Scales Northpointe *https://epic.org/algorithmic-transparency/crim-justice/EPIC-16-06-23-WI- FOIA-201600805-MDOC_ReentryStudy082213.pdf* retrieved May 2018

Eubanks,V. (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* New York, NY: St. Martin's Press

Flores AW, Bechtel K, Lowenkamp CT. (2016) False positives, false negatives, and false analyses: A rejoinder to ''machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. Available at: *http://www.crj.org/page/-/publications/ rejoinder7.11.pdf*

Howard, A and Borenstein,J. (2017) The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. *Sci Eng Ethics.* 2 0 1 7 S e p 2 1 . d o i : 1 0 . 1 0 0 7 / s11948-017-9975-2. [Epub ahead of print]

Joh, E. (2017) Feeding the Machine: Policing, Crime Data, & Algorithms. William & Mary Bill ofdebtors.

They have none of the resources

Rights J. (2017 Forthcoming).. Available at personal, intellectual or financial that may be required to understand or militate the circumstances they find themselves in. Not least because no other single actor in this ecosystem has the information to fully comprehend what it is that is happening to these people.

This phenomenon in which the combination of algorithmic reasoning used as tools in a cumulative and combinatory fashion creates both ethical and regulatory problems that no single agency or actor is in a position to address. The moment is rapidly approaching where no member of society, because they have to have a credit rating, a bank account, health and police records and are subject to many, many observations of their Internet behaviour can avoid being the subject of multiple, cumulative and interacting implications of systems that they cannot see, or indeed in many cases even be aware of.