People's Councils for Ethical Machine Learning

Dr. Dan McQuillan

Department of Computing, Goldsmiths, University of London, UK

forthcoming in the 'Ethics as Methods' special issue of Social Media and Society, Spring 2018

Introduction

Machine learning is a form of knowledge production native to the era of big data. It is at the core of social media platforms and everyday interactions. It is also being rapidly adopted for research and discovery across academia, business and government. This paper will explore the way the affordances of machine learning itself, and the forms of social apparatus that it becomes a part of, will potentially erode ethics and draw us in to a drone-like perspective. Unconstrained machine learning enables and delimits our knowledge of the world in particular ways: the abstractions and operations of machine learning produce a 'view from above' whose consequences for both ethics and legality parallel the dilemmas of drone warfare. The family of machine learning methods is not somehow inherently bad or dangerous, nor does implementing them signal any intent to cause harm. Nevertheless, the machine learning assemblage produces a targeting gaze whose algorithms obfuscate the legality of its judgements, and whose iterations threaten to create both specific injustices and broader states of exception. Given the urgent need to provide some kind of balance before machine learning becomes embedded everywhere, this paper proposes *people's councils* as a way to contest machinic judgements and reassert openness and discourse.

Machine Learning

Machine learning is becoming a methodological substrate for knowledge and action. But machine learning is not ethically neutral. It is skewed by data and obfuscated by nature, and these characteristics are particularly strong in the most successful kind of machine learning (neural networks). To grapple with the ways a set of computational methods can have ethical implications it is important to understand something of the actual workings of the algorithms (Geitgey, 2014). Looking at mathematical minimisation, decision boundaries and the role of data in the production of prediction, we can appreciate how machine learning is becoming a kind of dark matter that invisibly distorts the distribution of benefits and harm.

Such is the sometimes uncanny ability of machine learning to participate in activities previously considered uniquely human, such as playing chess or Go (Silver et al., 2016), that it seems to fulfill the original conception of artificial intelligence (AI) "to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves" (Kline, 2010). However, machine learning is nothing like the emergent general intelligence that characterises cultural representations of AI, and is instead a set of mathematical methods that can perform amazing yet utterly thoughtless feats of classification. The mode of

operation of machine learning is to ‘learn’; that is, when supplied with a large amount of input data and a corresponding set of targets it finds a function to map the features of the input data to the desired target outputs. Machine learning finds reproducible patterns in the data. Moreover, these patterns have predictive power in that they can predict the target value for new and unknown input. Thus machine learning is a form of numerical pattern finding with predictive power, prompting comparisons with science. But rather than being universal and objective it produces knowledge that is irrevocably entangled with specific computational mechanisms and the data used for the training.

The data used for training the algorithms is associated with a set of known labels or outcomes which are called the targets. The input features are weighted to optimise the function that maps the input to those targets. Imagine, for example, a training set consisting of clusters of green dots and red dots. They may represent the results of an MRI scan for cancer and the corresponding actual diagnosis (red or green), with the axes of the chart being the size and density of the observed tumour, and the green dots representing cases where the tumour was found to be non-malignant. The algorithm is fed the size and density of the tumours and told the target label of red or green in each case. Its task is to find a plausible decision boundary between them i.e. a line that separates the two kinds of dot and which can be used to classify future data points based on which side of the line they fall on. It's not an easy task because the boundary is not neat and the dots tend to mingle a bit. Mathematically, the algorithm calculates a *cost function*, which is a quantity representing how far each predicted result is away from the actual diagnosis for a given fit of the parameters. The power of machine learning comes from the iterative minimisation of this cost, such that the final fit works as well as possible across all the training data (Schutt & O’Neil, 2013). Thus, a humanly meaningful question (‘is this tumour malignant?’) is converted to an operation that computers are good at -- thousands and thousands of rapid and repetitive calculations. In the case of our dots, a well-chosen machine learning algorithm will find a boundary that makes the best job of dividing the positive and negative test results. There will always be false positives (classified as malignant, but not) and false negatives (malignant cases that are mis-classified) but the aim is to get a model that is statistically successful and can be applied to future test results.

The only thing that the machine learning knows about the world is the data that is fed to it. In one sense this accounts for its power, in that any problem which can be cast in a suitable numerical form can be tackled by the general family of machine learning algorithms. For the purposes of the mathematical minimisation, it doesn't matter whether the data represents house prices or terrorism suspects. Anything, in principle, can be ‘data scienced’. However this is also an Achilles heel. With the input data we sweep in all the potential biases that lie behind its construction as data in the first

instance. If the data is distorted by social prejudice then that is the pattern that the algorithm will learn (Lee, 2016). We have only to consider for a moment the way that algorithms powering predictive policing systems are trained on historical arrest and crime data and thus accumulate all the individual and cultural decisions about who to target for investigation and arrest. If there is discrimination embedded in the data then our machines will come to think our prejudices for us. If those machinic judgements are then used to target future activity, we have the technological reproduction of social discrimination (Brayne, 2017). The problem goes beyond straightforward racism or sexism, as data also inherits its ontology i.e. what gets constructed or counted as meaningful data depends on the worldview and assumptions of those defining the data (Boellstorff, 2013; Gitelman, 2012) - for example, whether depression is classified through biomedical symptoms or through people's experiences.

Machine learning brings with it another characteristic which can cause or obscure harm, which is the opaque nature of its decision making. As has been pointed out, machine learning depends on big data because the training set needs to be large enough to generate a useful predictive model. The features used by the algorithm are the different quantified aspects of the input which can be used to help find a pattern. It is not uncommon for the number of features to be in the hundreds or thousands. The methods also excel at finding patterns across diverse data by finding distant correlations that boost their predictive power. The more mixed the data the more options the algorithm has for weightings that optimise the overall fit. However, it can be hard to reverse this process in to human reasoning (Lipton, 2016). While in many cases the outcome is interpretable because the correlations seem to make some kind of sense, this is a form of post-hoc rationalisation. Based on our own experience, for example, it wouldn't surprise us that emails mentioning an unexpected lottery win and using terms like 'urgent' and 'money transfer' would be algorithmically classified as spam (Burrell, 2016), but the machine doesn't think that way because it doesn't think at all. It simply minimises a function applied over the whole input data set (in this case, bags of words extracted from emails) by carrying out thousands of matrix calculations very quickly.

The whole point about big data analytics is that the number and form of calculations that need to be carried out exceed the scale and complexity which people can comprehend directly. Otherwise there would be no need for the algorithms and there would be nothing new in the predictive potential of data science. There is no larger explanatory structure to fall back on because the outcomes are based on large scale correlations, not on logical causation. Thus there is no cumulative support for the proposed explanation from other non-falsified hypotheses, as there is in the physical sciences. While the operators of the algorithms may develop intuitions for what will work better in a specific

case and why, and the outcomes are tested on known data and parameterised in terms of their accuracy, there is an opacity at the heart of the methods. If we can't understand exactly what is being weighed in the balance, it is very hard to tell under what circumstances harm may be caused or in what ways the operations might be unethical.

The particular form of machine learning that is currently achieving spectacular success and dominating the popular narrative about artificial intelligence is neural networks, also known as deep learning. Significantly for any discussion of ethics, neural networks are also the hardest to interpret in ways that make sense to people. To understand the impressive power of deep learning we can consider the task of recognising faces. While we might guess that computer vision would work adequately with the rigid and predictable pose of the passport photo, we have tended to think that recognising someone's face under different conditions of lighting, distance and angle, or even when the person has aged, is a uniquely human ability. A data set that has been used to test this assumption is known as Labeled Faces in the Wild (LFW), which consists of more than thirteen thousand images of well-known faces collected from the web. Like a giant tabloid newspaper it consists of multiple close up photos of these famous people under all sorts of conditions. The rates of human performance on the LFW data is excellent, as we might expect, with recognition rates of between 97.53% and 99.20%. But in a paper from 2015 a team from Google presented a deep learning system called FaceNet which achieved a record accuracy on LFW of 99.63% (Schroff, Kalenichenko, & Philbin, 2015). In other words, the algorithm is better at recognising faces than people are, which is somewhat uncanny. Similar leaps are being made or anticipated in many other areas. The core ethical problem behind this new superpower is that we can't really tell how neural networks come to their conclusions, which makes it very hard to know whether they are 'doing the right thing' or likely to cause harm to persons at some point.

While functioning neural networks have been around since 1958 (Rosenblatt, 1958) they are very much back in fashion because of their traction with messy, hard to parameterise problems like computer vision, speech recognition and self-driving cars. Their great leap forward comes from a combination of vast training datasets, courtesy of social media and our digitised lifeworld, and the availability of computer chips called graphic processing units (GPUs) that were originally developed for gaming but turn out to be very effective at the large matrix calculations because high quality 3D graphics have very similar requirements for massively parallel processing. The deep learning algorithm itself consists of different layers of nodes (or 'neurons'); each node is connected to all the nodes in the subsequent layers, all the way through to the final target layer. The cost function applies differing weights to these node-to-node connections. Training is carried out by the

massively parallel process of ‘back propagation’; the gap between the prediction and the target values is fed back to neurons in the final layer, which makes a set of corrections and feeds them back to the layer before that, and so on (Nielsen, 2015). The process resembles a shuttle weaving connections across the threads on a steam-powered loom. There is no feature selection by the human operator (as there was in our tumour example, where size and density were the chosen features); the layers are force fed the input data and develop their own features through the weightings of connections between nodes. This is not a black box process; we can access the values of these weightings. The sometimes insurmountable challenge is to try to interpret these weightings in a way that is accessible to human reasoning. While we can form an intuition about some kinds of mathematics (such as the force, mass and acceleration in Newton’s second law F = ma) it can be hard to reverse the results of massively parallel minimisation in to a representational model of the external world. The results can be evaluated for accuracy by applying them to a new set of labelled test data, but calculating the number of false positives (i.e. data points that are wrongly classified) tells us nothing about the internal reasoning of the algorithm. Thus, deep learning transforms messy data at scale into testable predictions but sabotages the kind of questions we need to ask when assessing its ethical impact.

The predictive nature of machine learning promotes preemption, i.e. action that attempts to anticipate or prevent the predicted outcome. In a world overflowing with data, machine learning and deep learning are powerfully attractive. For any context where a lot is riding on uncertain outcomes, the lure of being able to peer through the fog of data and read off the probable future through computational pattern analysis comes to be almost irresistible. At the current time, it would be hard to think of an area of life that isn’t under consideration for treatment by algorithmic methods. Some large companies analyse their organisational data this way to try to predict which employees are most likely to be next to leave (Silverman & Waller, 2015) while universities experiment with systems that predict which students are at high risk of failing a course. In Australia, the Security Risk Assessment Tool automatically determines the risk category of everyone being held in immigration detention centres, and decides whether they should be in manacles when they go for a hospital appointment (Bagshaw & Koziol, 2017). In the USA, the adoption of predictive policing systems like Predpol is becoming quite commonplace (PredPol, 2015). Both governments and private enterprise are developing algorithms to predict which parents are most likely to abuse their children (Keddell, 2015; Brown, 2016). Even in this small subset of examples it is easy to imagine the risk of possible harm arising from interventions based on the algorithmic predictions.

As we have seen, machine learning is unlike the AI vision propagated by science fiction films like Blade Runner or the ‘existential threat’ invoked by Silicon Valley entrepreneur Elon Musk (Domonoske, 2017). It is a mode of calculative knowing that operates at the level of methodology and infrastructure as it classifies, makes predictions and constitutes its subjects. By its nature, machine learning is not generally visible to us; it is what happens to our data somewhere out of sight, after human actions have generated this data and prior to the social interactions that are shaped by its verdicts. This doesn’t lessen the impact that it will increasingly come to have. The effect of machine learning embedded in methodologies and institutions will be of a kind of dark matter, invisible in itself but pulling other systems into new shapes around it. The methods of machine learning will bring to many areas of social enquiry a science-like power to predict and a political opportunity to preempt. But the opportunity to operate these methods ethically is obfuscated by the machinery itself. By absorbing the latent content of our data and basing insights on opaque and possibly non-interpretable ‘learning’ the algorithms become actors that make ethical decisions for us in ways that are hard to challenge. We are approaching the machinic production of ethics.

Machinic Production of Ethics

The usurping of ethical agency by machine learning can be explained by looking at the ways that the affordances of machine learning erode consent and the avoidance of harm, and impact notions of justice and due process. The flow of data through the code produces classifications and inferences, affects how people are treated, and opens up future possibilities of unintended consequences (Markham, 2017). More, machine learning dismantles the ‘human subject’ as a useful concept for ethical treatment and reconstitutes it instead as means of targeting.

Recent studies provide strong evidence that machine learning models can also absorb discrimination, which is then perpetuated by their predictions. For example, word embedding is a method to represent text data as vectors and is used in many machine learning and natural language processing tasks. These vectors are sets of numbers that capture the semantic relationships between words (Colyer, 2016). A recent study of the word embeddings in Google’s widely used Word2vec, derived by a neural network from a Google news corpus of three million words, shows that it is intransigently sexist. For example, querying it for the word that satisfies the relationship ‘man is to doctor as woman is to x’ returns ‘nurse’, while ‘man is to computer programmer as woman is to x’ gives x as ‘homemaker’ (Bolukbasi et al, 2016).

The Tuskegee syphilis experiment highlights the dangers of embedded discrimination. The untreated progression of syphilis in African-American men in rural Alabama was studied over decades while the men were deceived in to believing they were receiving free health care and treatments. Begun in the 1930s, the experiment was allowed to run for forty years before a whistleblower made it public in 1972 and public outrage closed it down (Brandt, 1978). Partly as a result of the Tuskegee experiment the US government established the Office for Human Research Protections and set up Institutional Review Boards to vet research proposals in universities and hospitals (Office for Human Research Protections, 2009).

However, algorithmic methods escape oversight when the abstraction inherent in machine learning make it seem that there is no direct relationship to human subjects. The vast majority of academic research on machine learning and neural networks is not seen as requiring ethics review. Even where the data has been generated by or about some aspect of people's lives, the apparent distance of the data point from the human or the fact that it is 'already public' or covered by some Terms of Service not only makes it difficult to see it as having ethical problems (what Markham & Buchanan call the 'distance principle', 2016), but also puts it outside the purview of most Institutional Review Boards (Leetaru, 2016). But, as we shall see, the reconstitution of the distant subject as target is part of the ethical impact of machine learning.

Machine learning is only effective when the training data sets are large enough. The ethical question of consent is marginalised by this need for scale. Moreover, big data analytics benefit from using heterogeneous sources; the pattern finding mechanisms excel at finding correlations across not only vast but also varied sets of data. Payday loans companies like Wonga claim to use hundreds different data points varying from mouse click patterns to social media friendships to make rapid algorithmic lending decisions (Morozov, 2013). It is impossible to know in advance what purposes particular data will eventually be put to, and whether that will cause harm. One thing that we can say with reasonable certainty is that the unconstrained application of machine learning will impact our notions of justice. In particular, the preemptive interventions that flow from predictive machine learning run counter to due process. While the basis of most extant legal systems is 'presumed innocent until proven guilty', it is a principle that is bypassed by the operation of algorithmic prediction. "Big data enables a universalizable strategy of preemptive social decisionmaking. Such a strategy renders individuals unable to observe, understand, participate in, or respond to information gathered or assumptions made about them. When one considers that big data can be used to make important decisions that implicate us without our even knowing it, preemptive social decision making is antithetical to privacy and due process values." (Earle & Kerr, 2013).

These same methods are also productive of the situations that Miranda Fricker calls 'epistemic injustice'. One kind of epistemic injustice is testimonial injustice, where prejudices cause people to "give a deflated level of credibility to a speaker's word" (Fricker, 2007). In a world where the production of knowledge and truth is increasingly deferred to algorithms, mathematical predictions may be given a higher weight that the subject's own version of events, were that even to be sought in the first place. The other kind of epistemic injustice is hermeneutical injustice, "a kind of injustice in which someone is wronged specifically in her capacity as a knower". Fricker points to this as the kind of injustice experienced by social groups who lack the resources to make sense of their own experience.

In the world of big data the set of social groups whose life patterns will be authoritatively interpreted by distant machines is growing ever larger, including those at the sharp end of social provision such as ‘troubled families’ (Portal Analytics, 2017). Not only do the methods affect justice in a general sense, they are also being increasingly applied in direct juridical contexts. The dispute between nonprofit investigative journalism organisation ProPublica and Northpointe, the company whose software assigns risk scores to defendents awaiting trial (and thus influences whether they are released on bail) is a case in point. Northpointe say that their algorithm is fair and not racist because defendents given a risk score of seven have the same likelihood of reoffending (60 percent of white defendants and 61 percent of black defendants). ProPublica point to the unfairness that among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be classified as medium or high risk (42 percent vs. 22 percent) (Corbett-Davies, Pierson, Feller, & Goel, 2016), meaning that they receive harsher pre-trial bail conditions. It is mathematically impossible for a machine learning algorithm to achieve both types of fairness in a situation where the re-arrest rate for black defendents is higher, which links in turn to the likelihood of heavier policing in black neighbourhoods and the possibility of bias in the decision to make an arrest. It is becoming apparent that the questions of fairness and justice that are entangled with computational methods are unsolvable at the level of the machinic methods themselves.

The form in which machine learning reconstitutes its human subjects is central to understanding the collateral damage to ethics. When dealing with internet and our digitalised lifeworld the tendency is to consider human subjects when the data is collected directly from people through some online interaction, but not when working with points that were collected from the general firehose of social and infrastructural data. This is obviously questionable in terms of privacy, given the number of times since the AOL search query release of 2006 (Barbaro & Jr, 2006) that ‘anonymous’ internet

data turns out to have potentially harmful impacts on individuals. The emergence of big data has fuelled the debate about the centrality of human subjects to ethical assessments (Metcalf & Crawford, 2016).

Machine learning reverses the trajectory of data's journey away from its individual point of origin. It absorbs all manner of diverse and apparently insignificant data, all the way down to the level of single words and pixels, and uses these to target specific categories of subjects or to detect those who are anomalous to normative patterns. The raw material of this process corresponds to Deleuze's notion of 'dividuals' in that we are not working with whole human subjects but digitised fragments and moments: "We no longer find ourselves dealing with the mass/individual pair. Individuals have become 'dividuals,' and masses, samples, data, markets, or 'banks' " (Deleuze, 1992).

Deleuze was writing in relation to 'societies of control' as the successor form to Foucault's disciplinary society. Disciplinary societies are constructed around a relationship between the individual and the mass, whereas societies of control are articulated through the dyad of dividuals and databases. Rather than incarcerating individuals, the focus of control is on processes operating continuously across the space of social interactions. On this basis it may seem like machine learning will manifest as another iteration of the society of control, but that does not adequately account for its algorithmic gaze.

The wide-area view of machine learning hovers high above the flow of data, attempting to force fit the snippets of information in to patterns that represent its target. In the higher dimensional space of the computer's model we reappear as pathways, as vectors pointing in a particular, predictive direction. Not individuals as such, but not simply fragments of data aggregated by a database query, the subject is woven in to a silhouette perceptible by the machine. As Gregoire Chamayou puts it, "The production of this form of individuality belongs neither to discipline nor to control, but to something else: to targeting in its most contemporary procedures, whose formal features are shared today among fields as diverse as policing, military reconnaissance and marketing. It might well be, for that matter, that we are entering targeted societies" (Chamayou, 2014).

The Drone Perspective

The transformation of analog experience into data is a process of abstraction, which carries its own problems (Markham, 2013). Machine learning elevates abstraction (Wing, 2008) to an ethos, an orientation to everyday life. While the knowledge produced by machine learning is drawn from correlations rather than causations, it can appear to reveal a hidden mathematical order in the world that is superior to our direct experience (McQuillan, 2017). By seeming to promise an objective mathematised view based on hard data it resembles the scientific standpoint, the ‘view from above’ that is founded on a disembodied claim to a universal as well as objective knowledge (Haraway, 1988). The assumption of this privileged knowledge serves to justify intervention. The algorithmic eye stays permanently on station, each data point contributing to a panoramic view of its milieu, its models fitting observed activity in to the patterns it has been tasked with targeting in a way that is perceived as objective, as expressive of a higher order of knowing.

Machine learning, as a method of algorithmic targeting that has ethical consequences, risks inducing a drone perspective on the world. The drone perspective combines a privileged view from above with interventions of dubious legality. As with military drones, it is an orientation to action based only on patterns, whose determinations become self-justifying. The actual drone operations of the US Joint Special Operations Command (JSOC) are frequently based on ‘pattern of life analysis’, a term that describes the accrual over time of observations, signals intelligence and social media data that reveal distinctive patterns in a subject’s life. These that are then compared with signatures of activity that would justify a strike (U. S. Joint Forces Command, 2012). Drones are targeted to carry out signature strikes based on a target's patterns of behavior but without knowing that target's identity (Shane, 2015), sometimes killing militants but often killing civilians (Gregory, 2012). The legality of any of these targeted strikes is deeply disputed as the policy appears to violate both international humanitarian law (the ‘laws of war’) and international human rights. The former head of the International Legal Division of the Israeli Military Advocate General's Office justified targeted drone killings on the basis that "If you do something for long enough, the world will accept it. The whole of international law is now based on the notion that an act that is forbidden today becomes permissible if executed by enough countries... International law progresses through violations" (Feldman, 2009). Without urgent attention to the ethical challenges, the social interventions of machine learning will also tend to normalise actions of ambiguous legality. The models will iterate over the injustices embedded in their training data while potentially adding more through their own opacity. This construction of targets forestalls discourse, explanation or the judgement of peers.

The notion of algorithmic governance (Rouvroy, 2013) is changing from a topic of philosophical speculation to something that we will have to wrestle with on a day to day basis. The allure of big data (the ‘new oil’), the promise of smart infrastructures, and the post-austerity drive for large scale efficiencies and cost-savings makes algorithmic decision a default option for government. For example, Australian social security law states that “A decision made by the operation of a computer program under an arrangement made under subsection (1) is taken to be a decision made by the Secretary” (Elvery, 2017). Under this legislation, the Centrelink ‘robodebt’ algorithm has made decisions about overpayment of welfare benefits that resulted in debts being raised against thousands of individuals. While it has become clear that these algorithmic signature strikes often make mistakes, the burden of proof has been shifted onto the claimants to dispute it in each case while, in the meantime, they must pay off the debt or risk imprisonment.

Living under the gaze of the drone is to live in a ‘state of exception’ (Saif, 2016). The state of exception, as discussed by Giorgio Agamben, is a situation where law, rights and political meaning to life are suspended. It is the reduction of citizens to ‘bare life’ - a biological existence without civic existence (Agamben, 2005). While the consequences of machine learning in research and practice will usually emerge at a more prosaic level, their operations share another characteristic with these extrajudicial spaces. Agamben describes the topological structure of the state of exception as 'being-outside and yet belonging', which is also a good description of the operations of machine learning as ethically ‘other’ within the knowledge or governance structures they inhabit. The effect of ethically dubious and extra-legal machine learning is likely to be the production of partial states of exception (McQuillan, 2015). When patterns of life are understood as expressive of personal decisions, the process of identifying target individuals and groups as making risky or irresponsible choices encodes a divisive discourse, an ‘us and them’ perspective that justifies exclusion.

As can already be seen in current applications of big data analytics, algorithmic exclusion will be the calculative framework for the future of education (Selinger, 2015) and employment (Gee, 2017). People will be excluded from opportunities without ever knowing why, or having a chance to contest it. Of course, the affordances of machine learning will also be embraced by explicit operations of exception. In July 2017, the U.S. Immigration and Customs Enforcement (ICE) hosted an ‘industry day’ for tech companies, seeking an overarching vetting machine ‘that automates, centralizes, and streamlines the current manual vetting process while simultaneously making determinations via automation if the data retrieved is actionable’ (Biddle & Woodman, 2017). The

system must ‘determine and evaluate an applicant’s probability of becoming a positively contributing member of society’ and ‘predict whether an applicant intends to commit criminal or terrorist acts after entering the United States’. This also shows the breadth of data that is swept into AI analytics as, according to the slides from the ICE presentation, ‘The Contractor shall analyze and apply techniques to exploit publically available information, such as media, blogs, public hearings, conferences, academic websites, social media websites such as Twitter, Facebook, and LinkedIn, radio, television, press, geospatial sources, internet sites, and specialized publications with intent to extract pertinent information regarding targets’. Under the targeting gaze of so-called AI, states of exception will move from the edges of social experience to the centre.

People’s Councils

The ethical implications of an emerging drone perspective in the operations of machine learning demand an urgent response. This paper proposes that people’s councils provide a structure that counterbalances those aspects of machine learning that are toxic to ethics. While this is a speculative proposition, historical examples illustrate how people’s councils could restore collective subjectivity and agency in the context of advanced technologies. It is suggested that the benefits of this approach also help tackle the imbalance between action and automaticity that make social machine learning problematic in the first place.

Most of the discussions around the datafication of society focus on privacy, but the ethical problems raised by machine learning are primarily issues of justice. The suppression of discourse and the inability to debate and contest the epistemology of the machines’ models challenges both rights and fairness. It may be that some amelioration is to be found at a technical level by setting an algorithm to catch an algorithm, in that the data used for training can be mangled by one operation of machine learning to make sure that another can’t find within it any proxies for race, gender or other protected category (Zemel et al, 2013). But trade-offs in fairness, such as those surfaced by the COMPAS parole system, can’t be resolved at the level of calculation and must instead be part of a values-led discussion. The drone perspective is not simply an algorithm but an apparatus, that is, a combination of tools, protocols and institutions. Ethical decision-making in a social setting is a deliberative process, best informed by a rich understanding of context that can only come through involving the subjects as participants. Countering the drone perspective, the distanced targeting from above, requires something radically democratic.

The proposal here is that machine learning can be ethically reclaimed by combining it with the democratic structures of people's councils. People's councils are bottom-up, confederated structures that act as direct democratic assemblies, based on the face-to-face democracy of the Athenian ekklesia (popular assemblies) (Ober, 1993). These forms of assembly are horizontal structures in which everyone has an equal say about the matter being decided. Setting up people's councils for ethical machine learning means countering lack of consent with democratic consensus, replacing opacity with openness, and reintroducing the discourse that defines due process. The establishment of people's councils in contexts where people are severely impacted by machinic decisions mobilises a distributed form of democracy as a way to contest distributed algorithmic governance. It is likely that the councils operating on the same or similar topics (borders, education or social care, for example) would confederate; that is, form regional councils based on a system of recallable deputies (Biehl & Bookchin, 1997). The principle is to consciously adopt structures that reverse exclusion and exception. People's councils are a refusal to be rendered as 'dividuals' or to be reconstituted as targets, and instead to collectively question and challenge decisions made by machines. Machine learning represents one of the highest historical forms of the abstraction of social relationships, and needs to be counterbalanced by the unmediated relationships of popular assemblies.

The histories of different forms of people's councils offer insights into how and why we might use them to reintroduce ethics into machine learning. Take for example patients councils, a creation of the mental health users movement (Rogers & Pilgrim, 1991). In traditional psychiatric settings the 'patient' is constructed as the by-product of the clinical gaze. Users of mental health services have typically experienced both epistemic injustices (in the denial of their own account of their experiences) and more direct suspensions of their civil rights in the name of superior objective knowledge. Over the decades since the civil rights movements, as part of their wider struggle to be people rather than passive collections of symptoms, patients councils have been one of the tactics that users and survivors have successfully deployed to ensure that their accounts are considered alongside the versions of the psychiatric professionals (Survivors History Group, n.d.). By means of patients councils, service users have reconstituted themselves from diagnosed 'dividuals' to collective actors whose opinions count. A very different historical example highlights the potential for people's councils to turn complex technological system away from destructive ends.

In 1976 a Combined Shop Stewards Committee made up of shop floor workers produced an alternative corporate plan for Lucas Aerospace that advocated the production of socially useful products instead of weapons ('The Lucas Plan', 2016). Ideas prototyped by the workforce included

heat pumps, solar cells, wind turbines and hybrid power packs for vehicles. The energy and determination of the popular committee were captured in a documentary made for the Open University in 1978 (Open University, 1978). While the management of the time rejected the plans, many of the ideas have become mainstream forms of sustainable technology. Patients councils and the Lucas committee are instances of face-to-face structures that restored fairness and a wider concern for wellbeing to technocratic and potentially toxic contexts.

The risk that comes with the new powers of machine learning is that we become embedded in patterns that deepen harm. Nested deep learning systems will set up circulations and recursions, loops of self-justification where interventions in world modify the next wave of training data, reinforcing patterns and potential discriminations (Mackenzie, 2015). In her book ‘The Human Condition’ Hannah Arendt critiqued the instrumentalism and cycles of social reproduction that she saw as already characterising the industrial society of the post-war years. "If we see these processes against the background of human purposes, which have a willed beginning and a definite end, they assume the character of automatism. We call automatic all courses of movement which are self-moving and therefore outside the range of willful and purposeful interference" (Arendt, 1998 p151). In so-called AI we are introducing new processes which are self-moving and which partially evade purposeful interference. Arendt’s critique of our world view reads as a diagnosis of machine learning, where “real relationships are dissolved into logical relations between man-made symbols”.

The ethical challenges we face don’t come from the substitution of humans by machines but from the computational extension of existing social tendencies. A society which uncritically absorbs machine learning deepens its automaticity. Arendt also pointed to people’s councils as historical spaces where these tendencies were inverted, and noted that they arise at times of urgent need (she referred to Hungary 1956 as an example contemporaneous with her book). In particular, she saw them as the renewal of face-to-face democracy and as spaces for ‘action’. In Arendt’s philosophy, action stands as the alternative to instrumentalism and thoughtless process. An action, for her, is fundamentally a beginning. "Man does not so much possess freedom as he, or better his coming into the world, is equated with the appearance of freedom in the universe; man is free because he is a beginning..." (Arendt & Kohn, 2006). What she saw as action is exactly that the opposite to the patterns of life paradigm; action is the beginning which happens “against the overwhelming odds of statistical laws and their probability” (Arendt, 1998, p. 178). Thus, to argue for people’s councils is not only to advocate for direct democracy in the social application of machine learning, but to reclaim spaces for ethical action from generalised thoughtlessness and automaticity.

Conclusions

Dealing with the ethical consequences of machine learning is not a simple matter. We have seen that the methods of machine learning are entangled with ethical side effects prior to their activation, through training data, during their development, through opacity, and in practice, through the assemblages of institutions and ideas that form around them. Academia and wider society have laid down ethical principles as a way to ward off a repeat of bitter historical events, but it certainly seems that these will be eroded by uncertainties about consent, harm and even what constitutes a human subject. Unconstrained machine learning can become a drone perspective, a targeting gaze that blurs legality and divides the social along decision boundaries of ‘us and them’. How this can be counterbalanced is an open question; this paper proposes the model of people’s councils, horizontal and inclusive structures for democratic deliberation. The aim is to create structures where those affected can contest machine decisions through the collective refusal of automaticity.

References:

Agamben, G. (2005). *State of Exception*. (K. Attell, Trans.) (1 edition). Chicago: University Of Chicago Press.

Arendt, H. (1998). *The Human Condition*. Chicago & London: University of Chicago Press.

Arendt, H., & Kohn, J. (2006). *Between Past and Future* (Revised edition). New York: Penguin Classics.

Bagshaw, E., & Koziol, M. (2017, August 27). Computers replace humans in assessing danger of inmates in immigration detention. *The Sydney Morning Herald*. Retrieved from http://www.smh.com.au/federal-politics/political-news/computers-replace-humans-in-assessing-danger-of-inmates-in-immigration-detention-20170825-gy4i19.html

Barbaro, M., & Jr, T. Z. (2006, August 9). A Face Is Exposed for AOL Searcher No. 4417749. *The New York Times*. Retrieved from http://www.nytimes.com/2006/08/09/technology/09aol.html

Biddle, S., & Woodman, S. (2017, August 7). These Are the Technology Firms Lining Up to Build Trump’s “Extreme Vetting” Program. Retrieved 20 September 2017, from https://theintercept.com/2017/08/07/these-are-the-technology-firms-lining-up-to-build-trumps-extreme-vetting-program/

Biehl, J., & Bookchin, M. (1997). *The Politics of Social Ecology: Libertarian Municipalism*. Montreal ; Buffalo, NY: Black Rose Books.

Boellstorff, T. (2013). Making big data, in theory. *First Monday*, *18*(10). Retrieved from http://firstmonday.org/ojs/index.php/fm/article/view/4869

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *ArXiv:1607.06520 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1607.06520

Brandt, A. M. (1978). Racism and Research: The Case of the Tuskegee Syphilis Study. *Hastings Center Report*, *8*(6), 21–29. https://doi.org/10.2307/3561468

Brayne, S. (2017). Big Data Surveillance: The Case of Policing. American Sociological Review, 82(5), 977–1008. https://doi.org/10.1177/0003122417725865

Brown, M. S. (2016, February 25). Kids Endangered By Predictive Analytics? Child Advocate Says Yes. Retrieved 24 March 2016, from http://www.forbes.com/sites/metabrown/2016/02/25/kids-endangered-by-predictive-analytics-child-advocate-says-yes/

Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3*(1), 2053951715622512. https://doi.org/10.1177/2053951715622512

Chamayou, G. (2014, December 4). # THE FUNAMBULIST PAPERS 57 /// Patterns of Life: A Very Short History of Schematic Bodies by Grégoire Chamayou. Retrieved 20 September 2017, from https://thefunambulist.net/history/the-funambulist-papers-57-schematic-bodies-notes-on-a-patterns-genealogy-by-gregoire-chamayou

Colyer, A. (2016, April 21). The amazing power of word vectors. Retrieved 8 October 2017, from https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016, October 17). A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear. *Washington Post*. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Deleuze, G. (1992). Postscript on the Societies of Control. *October*, *59*, 3–7. https://doi.org/10.2307/778828

Domonoske, C. (2017, July 17). Elon Musk Warns Governors: Artificial Intelligence Poses ‘Existential Risk’. Retrieved 20 September 2017, from http://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk

Earle, J., & Kerr, I. (2013). Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy. *Stanford Law Review Online*, *66*, 65.

Elvery, S. (2017, July 21). Did you know computers now make important government decisions? [Text]. Retrieved 20 September 2017, from http://www.abc.net.au/news/2017-07-21/algorithms-can-make-decisions-on-behalf-of-federal-ministers/8704858

Feldman, Y. (2009, January 29). Consent and Advise. *Haaretz*. Retrieved from http://www.haaretz.com/consent-and-advise-1.269127

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford, New York: Oxford University Press.

Gee, K. (2017, June 26). In Unilever’s Radical Hiring Experiment, Resumes Are Out, Algorithms Are In. *Wall Street Journal*. Retrieved from https://www.wsj.com/articles/in-unilevers-radical-hiring-experiment-resumes-are-out-algorithms-are-in-1498478400

Geitgey, A. (2014, May 5). Machine Learning is Fun! Retrieved 20 September 2017, from https://medium.com/@ageitgey/machine-learning-is-fun-80ea3ec3c471

Gregory, D. (2012, November 7). Targeted killings and signature strikes. Retrieved 22 September 2017, from https://geographicalimaginations.com/2012/11/06/targeted-killings-and-signature-strikes/

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, *14*(3), 575–599. https://doi.org/10.2307/3178066

Keddell, E. (2015). The ethics of predictive risk modelling in the Aotearoa/New Zealand child welfare context: Child abuse prevention or neo-liberal tool? *Critical Social Policy*, *35*(1), 69–88. https://doi.org/10.1177/0261018314543224

Kline, R. R. (2010). Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence. *IEEE Annals of the History of Computing*, *33*(4), 5–16.

Lee, D. (2016, March 25). Microsoft issues apology for racist bot. *BBC News*. Retrieved from http://www.bbc.co.uk/news/technology-35902104

Leetaru, K. (2016, June 17). Are Research Ethics Obsolete In The Era Of Big Data? Retrieved 17 December 2016, from http://www.forbes.com/sites/kalevleetaru/2016/06/17/are-research-ethics-obsolete-in-the-era-of-big-data/

Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ArXiv:1606.03490 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1606.03490

Mackenzie, A. (2015). The production of prediction: What does machine learning want? *European Journal of Cultural Studies*, *18*(4–5), 429–445. https://doi.org/10.1177/1367549415577384

Markham, A., & Buchanan, E. (2012). Ethical decision-making and Internet research: Recommendations from the AOIR ethics working committee (version 2.0). The Association of Internet Researchers (AoIR).

McQuillan, D. (2015). Algorithmic states of exception. *European Journal of Cultural Studies*, *18*(4–5), 564–576. https://doi.org/10.1177/1367549415577389

McQuillan, D. (2017). Data Science as Machinic Neoplatonism. *Philosophy & Technology*, 1–20. https://doi.org/10.1007/s13347-017-0273-3

Metcalf, J., & Crawford, K. (2016). *Where are Human Subjects in Big Data Research? The Emerging Ethics Divide* (SSRN Scholarly Paper No. ID 2779647). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.com/abstract=2779647

Morozov, E. (2013, January 30). Your Social Networking Credit Score. *Slate*. Retrieved from http://www.slate.com/articles/technology/future_tense/2013/01/wonga_lenddo_lendup_big_data_and_social_networking_banking.html

Nielsen, M. A. (2015). Neural Networks and Deep Learning. Retrieved from http://neuralnetworksanddeeplearning.com

Ober, J. (1993). Public Speech and the Power of the People in Democratic Athens. *Political Science & Politics*, *26*(3), 481–486. https://doi.org/10.2307/419987

Office for Human Research Protections. (2009, June 23). Federal Policy for the Protection of Human Subjects ('Common Rule [Text]. Retrieved 12 January 2017, from https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html

Open University. (1978). *Lucas Plan documentary*. Retrieved from https://www.youtube.com/watch?v=0pgQqfpub-c

Portal Analytics. (2017, January 12). The Direction Portal is Taking Troubled Families. Retrieved 21 January 2017, from http://www.portalanalytics.co.uk/blog/the-direction-portal-is-taking-troubled-families

PredPol. (2015). Proven Results of our Predictive Policing Software. Retrieved 13 October 2015, from http://www.predpol.com/results/

Rogers, A., & Pilgrim, D. (1991). 'Pulling down churches': accounting for the British Mental Health Users' Movement. *Sociology of Health & Illness*, *13*(2), 129–148. https://doi.org/10.1111/j.1467-9566.1991.tb00093.x

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386.

Rouvroy, A. (2013, November 13). Algorithmic governmentality and the end(s) of critique. Retrieved 21 September 2015, from http://networkcultures.org/query/2013/11/13/algorithmic-governmentality-and-the-ends-of-critique-antoinette-rouvroy/

Saif, A. A. (2016). *The Drone Eats with Me: A Gaza Diary*. Boston, MA: Beacon Press.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. *ArXiv:1503.03832 [Cs]*, 815–823. https://doi.org/10.1109/CVPR.2015.7298682

Schutt, R., & O'Neil, C. (2013). *Doing Data Science*. O'Reilly Media. Retrieved from http://shop.oreilly.com/product/0636920028529.do

Selinger, E. (2015, January 13). With big data invading campus, universities risk unfairly profiling their students. *Christian Science Monitor*. Retrieved from https://www.csmonitor.com/World/Passcode/Passcode-Voices/2015/0113/With-big-data-invading-campus-universities-risk-unfairly-profiling-their-students

Shane, S. (2015, April 23). Drone Strikes Reveal Uncomfortable Truth: U.S. Is Often Unsure About Who Will Die. *The New York Times*. Retrieved from https://www.nytimes.com/2015/04/24/world/asia/drone-strikes-reveal-uncomfortable-truth-us-is-often-unsure-about-who-will-die.html

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., … Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Silverman, R. E., & Waller, N. (2015, March 13). The Algorithm That Tells the Boss Who Might Quit. *Wall Street Journal*. Retrieved from http://www.wsj.com/articles/the-algorithm-that-tells-the-boss-who-might-quit-1426287935

Survivors History Group. (n.d.). Mental Health and Survivors Movements. Retrieved 20 September 2017, from http://studymore.org.uk/mpu.htm

The Lucas Plan. (2016). Retrieved 22 January 2017, from http://lucasplan.org.uk/

U. S. Joint Forces Command. (2012). *Commander's Handbook for Attack the Network*. CreateSpace Independent Publishing Platform.

UNC Office of Human Research Ethics. (2016). Nuremberg Code. Retrieved 20 December 2016, from http://research.unc.edu/offices/human-research-ethics/resources/ccm3_019064/

United States Holocaust Memorial Museum. (n.d.). Nuremberg Code. Retrieved 22 January 2017, from https://www.ushmm.org/information/exhibitions/online-exhibitions/special-focus/doctors-trial/nuremberg-code

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1881), 3717–3725. https://doi.org/10.1098/rsta.2008.0118

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. In *PMLR* (pp. 325–333). Retrieved from http://proceedings.mlr.press/v28/zemel13.html