Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making

Michael Veale
University College London
London, UK
m.veale@ucl.ac.uk

Max Van Kleek
University of Oxford
Oxford, UK
max.van.kleek@cs.ox.ac.uk

Reuben Binns
University of Oxford
Oxford, UK
reuben.binns@cs.ox.ac.uk

ABSTRACT
Calls for heightened consideration of fairness and accountability in algorithmically-informed public decisions—like taxation, justice, and child protection—are now commonplace. How might designers support such human values? We interviewed 27 public sector machine learning practitioners across 5 OECD countries regarding challenges understanding and imbuing public values into their work. The results suggest a disconnect between organisational and institutional realities, constraints and needs, and those addressed by current research into usable, transparent and 'discrimination-aware' machine learning—absences likely to undermine practical initiatives unless addressed. We see design opportunities in this disconnect, such as in supporting the tracking of concept drift in secondary data sources, and in building usable transparency tools to identify risks and incorporate domain knowledge, aimed both at managers and at the 'street-level bureaucrats' on the frontlines of public service. We conclude by outlining ethical challenges and future directions for collaboration in these high-stakes applications.

ACM Classification Keywords
K.4.1 Computers and Society: Public Policy Issues; H.1.2. Models and Principles: User/Machine Systems; J.1 Computer Applications: Administrative Data Processing

Author Keywords
algorithmic accountability; algorithmic bias; public administration; predictive policing; decision-support

INTRODUCTION
Machine learning technologies increasingly form the centrepiece of public sector IT projects, a continuation of existing trends of risk-based regulation [8] given an increasingly 'data-driven' flavour. Recently, deployed and envisaged systems have also found themselves under heavy fire from civil society [58, 3], researchers [72, 73, 80] and policymakers [31, 40, 39]. These models, often colloquially simply referred to as algorithms, are commonly accused of being inscrutable to the public and even their designers, slipping through processes of democracy and accountability by being misleadingly cloaked in the 'neutral' language of technology [82], and replicating problematic biases inherent in the historical datasets used to train them [5]. Journalists, academics and regulators have been recording different examples of algorithmic concerns. Those concerning race have taken centre stage, particularly in the US—ranging from discrimination in recidivism scores used in parole decisions [4, 13] to uneven demographic targeting in systems used for policing on the basis of spatiotemporal crime risk [70].

Bias or opacity are far from newly observed characteristics of computer systems generally [26, 11, 24], but these issues have been exacerbated by the rapidly expanding numbers of data-driven information systems entering decision-making processes. They are now increasingly seen within interdisciplinary research communities as highly salient failure modes of these systems, and relevant work seeking to address them is now commonly found in both major machine learning conferences and specialised side-events. The most well-known of the latter, the Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML), has fast become a convening venue and acronym for scholars, initially computer scientists and lawyers but increasingly ethnographers and political scientists, attempting to identify tangible ways by which algorithmic systems can become less undesirably discriminatory and opaque [66, 32, 78, 23, 17]. Techniques in discrimination-aware data mining attempt to define fairness (primarily in reference to anti-discrimination law [29]) and statistically assess it, assure it, or get closer to it [64, 32, 23]. Research stemming from the early days of machine learning in expert systems attempts to make algorithmic logics more transparent [75, 81, 54, 66], or to better understand how users work with inductive systems and build mental models of them [60, 48, 77]. These tools have attracted attention from governmental bodies, such as the European Commission [22].

Yet these tools are primarily being built in isolation both from specific users and use contexts. While these are early days for fairness/transparency methods, which are mostly under a decade old [64], the predominant mode of development often involves characterising a problem in a way that might often be at odds with the real world context—such as assuming you have access to the sensitive characteristics you are trying to remove the proxy influence of in a system [79]. Yet practitioners are *already* deploying machine learning systems in the public sector, both those designed in-house and those bought



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI 2018, April 21–26, 2018, Montréal, QC, Canada.
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5620-6/18/04.
DOI: https://doi.org/10.1145/3173574.3174014

from vendors, and are facing immediate, value laden challenges [14]. As observed in studies of clinical decision tools, which often fail to be adopted [83], organisational conditions are often missed in design contexts. Indeed fields studying technologies in practice have long struggled to reconcile experimental traditions with the understanding that results from many methodologies fall apart when applied to political, noisy, stressful, complex and contested deployed settings [27]. In this paper, we seek to examine existing real-world contexts to assess whether FAT/ML tools might be airdropped from laboratory settings into the messy world of public sector IT with limited adjustment, and if not, in what direction change might be needed.

To open up questions of practitioner context in relation to fairness and accountability of data-driven systems in the public sector, we interviewed 27 actors working in and with 5 countries' public sectors. These ranged from in-house machine learning modellers to those managing projects, procurement, or contractors delivering and maintaining models. A diverse set of application areas including taxation, child protection, policing, justice, emergency response and interior security were consciously selected. This was in order to sample widely across practices and challenges emerging in this space rather than try to aim at representative data, which might result in closing-down the issue before it has been sufficiently prised open. The aim was to ascertain how concerns and challenges appear from the perspective of those commissioning and designing these systems on-the-ground. We are only beginning to see this reflective mode in published literature [62], as opposed to attempts to simulate or reverse engineer systems from the outside [4, 21] (which despite being a highly limited method [69], is one of few remaining choices where systems are resolutely closed).

Following further background and our methodology, we highlight and elaborate on core themes from undertaken interviews. Themes are grouped into those relating to internal users and stakeholders of the decision support systems and those relating to external users, stakeholders and decision subjects. We then discuss some central interdisciplinary challenges that emerge from these findings that we believe to be of importance to both design communities and to broader sets of scholars, policy-makers and other practitioners.

BACKGROUND AND MOTIVATION

Significant prior research interest exists concerning high-stakes, operational decision support. Decision-making in clinical contexts was an early focus in HCI and human factors due to its intrinsic importance to lives and livelihoods, the possibility of codification in highly specialised areas, and the development (and funding) of a range of relevant, high-profile expert systems. Canonical domains in human factors have also focussed on even more dramatic life-and-death situations, such as issues around decision support in airplane cockpits [57] and air traffic control [37].

While these studies are important to reflect and build upon, much algorithmically-informed public sector decision-making is not best conceived as simply a direct continuation of these specific challenges. The reason for this stems from an assumption of most work on decision-support systems: that it is relatively straight-forward to measure whether a design intervention would 'reliably improve human decision-making, particularly in real-world contexts' [84]. Does the patient improve, or at least receive the same treatment that would have been recommended by a top specialist? Does the plane glide or plummet? These situations, many of which can be characterised as 'safety', represent problems structured so that there is high consensus on both the *ends* (goals such as remaining in flight) and the *means* (well-understood steps that achieve those goals).

Those managing public policy problems rarely have the luxury of the settled consensus on ends and means some engineers are used to [38]. If a police department turns to a machine learned predictive model to anticipate crime risk in different parts of a city, they face a range of debates. A desired end might be to treat all crime equally. But does that imply police should focus resources on areas of high crime at the expense of those with low crime, to maximise total arrests? Or does it mean that a crime in a low-risk area is just as likely to be intervened in as a crime in a high risk area? Areas conceived of as 'high risk' are rarely distributed at random, coupled instead to communities with different demographic or vulnerability distributions. The means are also unclear. Should models be used to increase preventative measures, such as community policing, or to heighten response capacity after crimes have been reported? Of course, decisions are rarely this binary, but that does not mean they are settled. At first glance, a problem might seem like a clinical decision. Scratching the surface, myriad subjective choices quickly arise.

Compounding this, public sector IT projects are heavily resource constrained, path-dependent given existing infrastructures, and prone to failure from poor initial scoping [16]. Their information systems are notorious for crossing scales and chains of accountability, engendering many practical challenges relating to performance and maintenance [51]. Uptake of models in government is as much a social process as it is a technical one [50]. Furthermore, many of the important dynamics are somewhat hidden from public view. As noted in the public administration literature, values in public service are primarily exercised in the discretionary spaces between the rules rather than in the high level political rule-making processes themselves [44, 52]. Many ethical issues are only likely to surface far down the road of creation, procurement or deployment of a system, after many choices are baked-in [25]. This chimes with a common observation from clinical decision support—that designed tools, while promising in laboratories, often fail in practice in a variety of ways [47]. Only recently have HCI researchers in these settings begun to investigate *institutional* contexts behind these failures [83], noting primarily that airdropped decision support systems often fail to embrace the richness of the clinical context.

A parallel concern motivates this work: that nascent FAT/ML tools, such as debiasing or transparency systems, will also fail unless contextual challenges are understood early on. We are not aware of other research considering how value-laden

concerns manifest and are coped with in real, public sector decision support settings—something that speaks more to the recent increase in these types of tools than to the novelty of our broad approach. Indeed, while there has been a surge in researchers considering the social implications of 'algorithms', we are wary that issues are being framed from afar, rather than in collaboration with those facing them and perhaps understanding aspects that may not always be immediately apparent. As a result, we do not seek to populate or validate any of the young and primarily unverified theoretical frameworks, but lay empirical foundations for a more grounded approach that might enable more positively impactful work in the future. In particular, in summarising findings for this work we sought to highlight aspects infrequently discussed in or omitted entirely from relevant contemporary research discussions. The section that follows explains how we sought to do that.

METHOD

Twenty-seven individuals agreed to be interviewed in late 2016—predominantly public servants and attached contractors either in modelling or project management. Each individual was interviewed only once, either in person (seventeen interviews) or on the telephone (ten interviews). They were all undertaken with one interviewer, and each lasted between forty and sixty minutes. Just over one-fifth of informants were female. Interviewees worked in one of five OECD countries located over three continents. It was decided in the initial ethical approval for this work not to publicly name the countries in order to reduce the risk of informant identification, but we will note that the US was not one of the countries included.

Informants were identified with an *ad hoc* sampling approach. This was chosen for several reasons. Firstly, at this relatively early stage of deployment, projects are emerging without central mandates—no coordinating body was identified to have a reliable compiled register of activities. Indeed central agencies occasionally shared registers that turned out to be a poor representation of on-the-ground activities. Secondly, we sought as many perspectives as possible from within public sector organisations deploying machine learning for decision-support today, and felt this was best achieved by looking across sectors to very different types of agencies and bodies. To recruit participants, projects and contacts were assembled from grey literature, freedom of information requests (both actively made and through platforms such as *WhatDoTheyKnow*), snowball sampling, direct inquiries with organisation contacts, and the use of news databases including *Factiva* and *LexisNexis*. Terms including *predictive modelling*, *entity-level prediction*, *predictive analytics* and *machine learning* were entered into these databases and public document repositories. Participants additionally played an important role in sampling themselves, and were usually willing and often even eager to flag colleagues in other domestic or foreign agencies working on projects they felt would benefit the study. Similarly to challenges arranging interviews with societal 'elites', candidacy for interviews 'often cannot be planned for adequately in advance of the project; rather, it emerges as part of the fieldwork' [61].

Because of the open-ended nature of the sampling, the varied nature of the roles (particularly across sectors), and the many different systems concerned, it was neither possible nor helpful to stick to a rigid script. Instead, the approach taken was similar to other open-ended work in policy research, involving prompting the participant to not only outline their role but explain the process behind the development and maintenance of the project.[1] First, the purpose of the study was explained to participants, at which point any ambiguities could be resolved. Following that, participants were asked about their role (and history of roles) in this area, then to give a high level outline of relevant project(s) and a more detailed view on their position within them. They were then steered at opportune moments in the discussion towards topics of fairness and accountability, effectiveness and complexity/robustness (mirroring the public sector values framework introduced by [36]). At times, this steering was unnecessary and avoided, particularly as the nature of the study was made clear to participants: many already had considered these issues during their job, albeit often under different names. The other main prompt used to elicit relevant insights, particularly where participants had not considered their job in the above framing before, was to ask whether 'anything surprising or unexpected' had happened to them in relation to their work, such as a deployed model. This was especially useful in eliciting institutional events, or novel incidences of model failure.

Conversations were not taped. While that might have been desirable, recording audio of individuals discussing sensitive public sector work is extremely difficult. Bodies commonly disallow it when individuals are not spokespersons for the organisation, precluding it as an interview approach, more so where new technologies are involved, and questions asked are likely to be totally new. Even where taping is permitted, it can risk inhibiting openness and frankness in discussions. These politically-charged contexts pose methodological restrictions infrequently seen in HCI, but frequently encountered by public administration researchers, and we follow methodological practices developed in this field [63]. These are further exacerbated here by fear of negative media coverage—both journalists and academics in this field have exhibited a recent taste for algorithmic 'shock stories'. Instead, verbose notes were continuously taken with the aim of authentically capturing both interviewees' tone, phrasing and terminology, as well as the core points they explained. Where longer continuous answers were given, interviewees kindly paused for note-taking purposes. Notes were typed up by the interviewer, always on the same day as the interview took place and often immediately following. Some highly context-specific terminology, such as geographic subunits or revealing was substituted with equivalent generic alternatives to increase the difficulty of project re-identification. Handwritten notes were then destroyed in line with data protection and the study's ethical approval.

To analyse the interviews, open coding was used (*NVivo 11 for Mac*), with codes iteratively generated and grouped concerning the challenges and coping mechanisms observed. These were then iteratively grouped according to a public sector values

[1] See [63], who conducted 128 interviews in the UK civil service to understand the nature of policy work, asking only 'what do you do?' and 'how do you come to be in this job?'.

framework from the public administration literature [43] as a thematic organisational principle.

FINDINGS

In this section, we summarise some of the key themes from the interviews undertaken. They are split into two broad sections: those concerning internal actors and their relation to the algorithmic systems, such as other departments, and those concerning external actors, such as decision subjects.

Internal actors and machine learning–driven decisions

The first category of themes relate to discussions of how the deployed systems are connected to and perceived by a range of internal actors. Many of the issues around algorithmic transparency so far have focussed on external algorithmic accountability and transparency-based rights, such as a 'right to an explanation' [20], although broad reasons to make systems transparent and interpretable exist [53]. Yet within organisations there are a wide array of reasons for understanding data and models.

Getting individual and organisational buy-in

Informants reported a need to use different approaches to clarify the workings of or process behind machine learning powered decision-support systems for internal actors. Some of these were strategic actors in management positions, either the clients of external contractors or customers of internal modelling teams.

Several interviewed practitioners noted that this organisational pressure led them to make more 'transparent' machine learning systems. Detection systems for fraudulent tax returns illustrated this. The analytics lead at one tax agency [X1] noted that they "*have better buy-in*" when they provide the logic of their machine learning systems to internal customers, while their counterpart in another tax agency [X2] described a need to "*explain what was done to the business user*". Both these individuals and modellers around them emphasised they had in-house capability for more complex machine learning systems, such as support vector machines or neural networks, but often chose against them for these reasons. Instead, many of the systems that ended up being deployed were logistic regression or random forest based.

Some saw transparency in relation to input variables more than model family. One contractor that constructed a random-forest based risk score for gang members around knife crime on behalf of a global city's police department [X3] described an "*Occam's razor*" process, where they started with 18,000 variables, working down to 200, then 20, then 8—"*because it's important to see how it works, we believe*". To steer this, they established a target percentage of accuracy with the police department *before* modelling—around 75%—which they argued helped them avoid trading off transparency. When users of analytics are not "*confident they know what a model is doing*", they "*get wary of picking up protected characteristics*", noted the modelling lead at tax agency [X4]. To make this more transparent, the police contractor above [X3] would "*make a model with and without the sensitive variables and see what lift you get in comparison*", presenting those options to the client to decide what was appropriate.

Another issue raised by several modellers was the difficulty in communicating the performance of designed systems. One modeller in a regional police department [X5] was designing a collaborative system with neighbouring police departments to anticipate the location of car accidents. They noted that

> *We have a huge accuracy in our collision risk, but that's also because we have 40 million records and thankfully very few of them crash, so it looks like we have 100% accuracy—which to the senior managers looks great, but really we only have 20% precision. The only kind of communication I think people really want or get is if you say there is a 1/5 chance of an accident here tomorrow—that, they understand.*

An analytics lead at a tax department [X2] faced parallel issues. When discussing the effectiveness of a model with clients, he would often find that "*people tend to lose faith if their personally preferred risk indicators aren't in a model, even without looking at performance of results.*"

Performance was often judged by the commissioning departments or users based on the additional insight it was thought to provide, compared to what they thought to be known or easily knowable. There was a tension between those who were seeking insight beyond existing processes, and those seeking efficiency/partial automation of current processes. One contracted modeller for a police department [X3] noted that during modelling, they "*focussed on additionality. The core challenge from [the police department] was to ascertain whether the information we could provide would tell them things they did not already know. How would it complement the current way of doing things?*" Yet another case, an in-house police modeller [X6] noted that a focus on additionality by the users of the system often clouded the intended purpose of the software in the first place.

> *What we noticed is that the maps were often disappointing to those involved. They often looked at them and thought they looked similar to the maps that they were drawing up before with analysts. However, that's also not quite the point—the maps we were making were automatic, so we were saving several days of time.*

Over-reliance, under-reliance and discretion

Over and under-reliance on decision support, extensively highlighted in the literature on *automation bias* [71, 18], featured considerably in informants' responses. A lead machine learning modeller in a national justice ministry [X7], whose work allocates resources such as courses within prisons, described how linking systems with professional judgement "*can also mean that [the model output is] only used when it aligns with the intuition of the user of the system*". To avoid this, some informants considered more explicitly how to bring discretion into decision-support design. A lead of a geospatial predictive policing project in a world city [X8] noted that they designed a user interface

> *to actively hedge against [officers resenting being told what to do by models] by letting them look at the predictions and use their own intuition. They might see the top 3 and think 'I think the third is the most likely' and that's*

okay, that's good. We want to give them options and empower them to review them, the uptake will hopefully then be better than when us propellorheads and academics tell them what to do...

Model outputs were not treated similarly as decision support in all areas. The former lead of a national predictive policing strategy [X9] explained how they saw discretion vary by domain.

We [use machine learning to] give guidance to helicopter pilots, best position them to to optimise revenue—which means they need to follow directions. They lose a lot of flexibility, which made them reluctant to use this system, as they're used to deciding themselves whether to go left or right, not to be told 'go left'! But it's different every time. There were cases where agents were happy to follow directions. Our police on motorcycles provide an example of this. They were presented with sequential high risk areas where criminals should be and would go and apprehend one after another—and said "yes, this is why we joined, this is what we like to be doing!" The helicopters on the other hand did not like this as much.

Also faced with a list of sequential high risk activities, this time relating to vulnerability of victims, the analytics lead at one regional police department [X10], sought advice from their internal ethics committee on how to use the prioritised lists their model outputted.

We had guidance from the ethics committee on [how to ethically use rank-ordered lists to inform decision-making]. We were to work down the list, allocating resources in that order, and that's the way they told us would be the most ethical way to use them... It's also important to make clear that the professional judgement always overrides the system. It is just another tool that they can use to help them come to decisions.

Augmenting models with additional knowledge

Informants often recognised the limitations of modelling, and were concerned with improving the decisions that were being made with external or qualitative information. A lead of a national geospatial predictive policing project [X11] discussed transparency in more social terms, surrounding how the intelligence officers, who used to spend their time making patrol maps, now spent their time augmenting them.

We ask local intelligence officers, the people who read all the local news, reports made and other sources of information, to look at the regions of the [predictive project name] maps which have high predictions of crimes. They might say they know something about the offender for a string of burglaries, or that building is no longer at such high risk of burglary because they local government just arranged all the locks to be changed. [...] We also have weekly meeting with all the officers, leadership, management, patrol and so on, with the intelligence officers at the core. There, he or she presents what they think is going on, and what should be done about it.

Other types of knowledge that modellers wished to integrate were not always fully external to the data being used. In particular, information needs also arose linked to the primary collectors of training data. One in-house modeller in a regional police department [X5], building several machine learning models including one to predict human trafficking hotspots, described how without better communication of the ways the models deployed worked, they risked large failure.

Thankfully we barely have any reports of human trafficking. But someone at intel got a tip-off and looked into cases at car washes, because we hadn't really investigated those much.[2] *But now when we try to model human trafficking we only see human trafficking being predicted at car washes, which suddenly seem very high risk. So because of increased intel we've essentially produced models that tell us where car washes are. This kind of loop is hard to explain to those higher up.*

Similarly, external factors such as legal changes can present challenges to robust modelling. A modeller in a justice ministry building recidivism prediction systems noted that while changes in the justice system were slow, they were still "*susceptible to changes in sentencing, which create influxes of different sorts of people into the prison systems.*" These kinds of rule change are unavoidable in a democratic society, but awareness of them and adequate communication and preparation for them is far from straightforward.

Gaming by decision-support users

'Gaming' or manipulation of data-driven systems, and the concern of this occurring if greater transparency is introduced, is often raised as an issue in relation to the targets of algorithmic decisions. This will be discussed in a following section. Yet types of *internal gaming* within organisations have received considerably less treatment by those concerned about value-laden challenges around algorithmically informed decisions. This is despite how internal gaming is extensively highlighted in the public administration literature in relation to targets and the rise of New Public Management [7], a broad movement towards 'rationalisation' in the public sector that clearly affected informants around the world.

One tax analytics lead [X2] worried that releasing the input variables and their weightings in a model could make their own auditors investigate according to their perception of the model structure, rather than the actual model outputs—where they believed that bias, through fairness analysis, could ostensibly be controlled.

To explain these models we talk about the target parameter and the population, rather than the explanation of individuals. The target parameter is what we are trying to find—the development of debts, bankruptcy in six months. The target population is what we are looking for: for example, businesses with minor problems. We only give the auditors [these], not an individual risk profile or risk indicators [...] in case they investigate according to them.

[2] Modern slavery is a problem in the car wash industry [59].

Additionally, some tax auditors are tasked with using the decision-support from machine learning systems to inform their fraud investigations. Yet at the same time, the fraud they discover feeds future modelling; they are both decision arbiter and data collector. The effect these conflicting incentives might have on a model were highlighted by a different tax agency [X2], as when auditors accumulate their own wages, "*[i]f I found an initial [case of fraud], I might want to wait for found individuals to accumulate it, which would create perverse incentives for action*".

External actors and machine learning–driven decisions

The second theme focusses on when informants reflected upon value concerns that related to both institutional actors that were outside their immediate projects, or that were at a distance, such as subjects of algorithmically informed decisions.

Sharing models and pushing practices

Scaling-up is an important part of experimentation. This is particularly the case in public sector organisations replicated by region—while some of them, particularly those in the richest or densest areas, can afford to try new, risky ideas with the hope of significant performance or efficiency payoffs to outweigh their investment, for smaller or poorer organisations that economic logic does not balance. The latter set of organisations are more reliant on the import and adaptation of ideas and practices from more well-resourced sister organisations (which could also be abroad) or from firms. Yet in practice, this is challenging, as machine learning systems also come imbued with very context specific assumptions, both in terms of the problem they are attempting to model, and the expertise that surrounds the decision-making process each day it is used. A modeller and software developer in a spatiotemporal predictive policing project [X6] emphasised the challenges in scaling up these social practices, as they were not as mobile as the software itself.

> *If you want to roll out to more precincts, they have to actually invest in the working process to transform the models into police patrols. To get more complete deployment advice... it takes a lot of effort to get people to do that. What you see is that other precincts usually—well, sometimes—set up some process but sometimes it is too pragmatic. What I mean by this is that the role of those looking at the maps before passing them to the planner might be fulfilled by someone not quite qualified enough to do that.*

Similar sentiments were also echoed by individuals in national tax offices, particularly around the 'trading' of models by large vendors. One tax analytics lead [X2] in a European country expressed concerns that another less resourced European country was being sold models pre-trained in other jurisdictions by a large predictive analytics supplier, and that they would not only transpose badly onto unique national problems, but that the country interested in purchasing this model seemed unprepared to invest in the in-house modelling capacity needed to understand the model or to change or augment it for appropriate use.

Accountability to decision subjects

Interpretable models were seen as useful in relation to citizens. One lead tax analyst [X2] described how transparency provided "*value-add, particularly where an administrative decision needs explaining to a customer, or goes to tribunal*". They noted that "*sometimes [they] justif[ied] things by saying here are the inputs, here are the outputs*" but they were "*not really happy with that as an ongoing strategy.*" Yet on occasion, more detailed explanations were needed. The same informant recalled an incident where a new model, in line with the law, was flagging tax deductions to refuse that were often erroneously allowed to some individuals in previous years. Naturally, many people called in to complain that their returns were not processed as expected—so the tax agency had to build a tool to provide call centre operators with client-specific explanations.[3]

Other organisations focussed on providing knowledge of the system to other interested parties, such as media organsiations. One national predictive policing lead [X11] explained how they found it difficult to have discussions around equity and accountability with police officers themselves, who are often narrowly focussed on "*where they think they can catch someone*", and have less capacity or incentive to devote time and energy to frame broader questions. Instead, this police force would invite journalists over twice a year to see what the predictive teams "*do, how [the algorithms] work, and what we are doing*". Several public sector organisations using machine learning systems already publish information about the weights within their model, the variable importance scores, or record ethical experiences and challenges in the modelling process [76, 55, 62].

Discriminating between decision-subjects

Discrimination has taken centre-stage as the algorithmic issue that perhaps most concerns the media and the public. Direct use of illegal-to-use protected characteristics was unsurprisingly not found, and interviewees were broadly wary of directly using protected characteristics in their models. Input data was seen as a key, if not the only point of control, but the reasons and the logics behind this varied. A lead of analytics at a national tax agency [X2] noted that "*if someone wanted to use gender, or age, or ethnicity or sexual preference into a model, [they] would not allow that—it's grounded in constitutional law.*" In one case guidance was to be released clarifying forbidden variables, but made no difference as the tax agency was already compliant [X1]. Even when characteristics were found to be legally permitted after consultation with lawyers (characteristics are not protected in all contexts), they might still have been avoided. Informant [X4], a lead modeller in a tax agency, noted that they have an informal list "*of variables that [they] don't feed into models*", which included age and location, both of which were legally permissible in their context. Location was avoided by this informant because even though different cities have different tax fraud risks, they "*don't usually want to investigate on those grounds.*" In other cases, home location was avoided as it was a "*proxy for social depri-*

[3] It was unclear if this system was machine learning or rule-based.

vation", in the words of the lead modelling a justice ministry [X7] .

Occasionally, there would be pressure to use protected characteristics to increase predictive power. The same justice modeller [X7], noted that "*we had feedback from a senior [foreign nationality, omitted] academic in this space on our [criminal justice] model, noting that 'if you've got something as predictive as race is, why aren't you using it?' Many of [this experts' deployed] models do, but it's an ethical decision in my mind and this is the route we've taken.*" Relatedly, they were also concerned by how the proxy outcome that could be measured (conviction) related to sensitive variables, rather than the outcome variable of real interest (offending).

> *Race is very predictive of re-offending, [but] we don't include race in our predictive models [...] we are aware that we are using conviction as the proxy variable for offending, and if you do this then you can get into cycles looking at certain races which might have a higher chance of being convicted, and train models on this data instead. That would mean you're building systems and catching people not based on the outcome, but on proxy outcomes.*

Gaming by decision-subjects

It is commonly expressed that extensive transparency of algorithms to the public might encourage system gaming [19], and this is brought to bear as a justification for opacity. Correspondingly, external gaming was raised as an issue by some informants. One contractor developing predictive policing software for a world city [X3] noted that concerns in his sector concerned "*criminal gangs that might send nine guinea pigs through the application process looking for loopholes to get arrested, just to find a tenth that highlights a way they can reliably get passports from under the noses of the authorities.*". An analyst from a large NGO working in collaboration with the police on developing a predictive system to detect child abuse [X12] noted that "*it's much harder to game when you've linked up lots of different aspects, education and the like.*", although their colleague [X13] warned that they were concerned about many of the usual sophisticated practices being used to game ML-supported systems, such as "*turning professionals against each other*" or the "*strategic withholding of consent at opportune moments*". The analytics lead at one tax agency [X1] explained that while they would publicly share the areas they were interested in modelling tax fraud for, such as sectors or size, they were "*primarily concerned that if the model weights were public, their usefulness might diminish*".

Other incidents resembled gaming—and could feasibly be interpreted as such—but served more to demonstrate the current fragility of models towards concerted attempts to change them. A modeller at a police department [X5] noted, in relation to a model they had built to pre-empt when the force should ensure they had the most staff available to deal with missing persons, that

> *There's one woman who calls in whenever her kid is out after 10pm. She then calls back about 30 minutes or so later to say that everything is fine, or we follow up with her. But then it looks like in the model that kids always go missing at 10pm, which obviously is a bit misleading. In the end I had to manually remove her from the model to remove the spurious pattern.*

While in this case, the model failed—resembling an *availability attack*, to draw on the adversarial machine learning literature—this might not always be the case. Indeed, models might not fail in obvious ways, or might even be subject to attacks designed to change them in targeted ways [41]. Even where an attack is not planned, simply responding to decisions informed by the model—such as patrol patterns—might look like gaming, or at least a game of cat-and-mouse. The police lead on a geospatial predictive policing project for a world city [X8] noted this in their own system. While it wasn't clear whether they were just removing the lowest hanging fruit or criminals were responding, in response, they linked a further feedback effect to try to compensate for the performance loss.

> *The highest probability assessments are on the mark, but actual deployment causes displacement, dispersion and diffusion, and that throws the algorithm into a loop. You have to remodel, though typical patterns of unresponded-to crime are predicted well [...] we decided to re-evaluate learning every 2–3 weeks, pull in all sorts of other variables, such as feeding it with what police were deployed, what they did—I've never seen this in other similar systems. In the first four weeks of trialling it out, the probability of being correct just tanked [...] in the 3rd update, it started to figure shit out.*

ISSUES AHEAD

In this section we draw upon the responses from informants to point to several 'grand challenges' for high stakes, algorithmically-informed public decisions. This is not a comprehensive list or typology. Instead we are seeking to emphasise areas where we believe vigorous discussion in the FAT/ML, HCI, human factors, critical data studies, information systems and public administration communities, among others, is lacking and needed. While these are not intended as direct implications for design, we see opportunities for design within each of them, as well as opportunities for many other fields and types of knowledge to be brought to bear.

'The probability of being correct tanked': Data changes

Data in the public sector is usually collected, categorised and cleaned for primarily operational reasons, such as recording who has been arrested, calculating tax bills, or delivering mail—not for modelling. While increasing emphasis is now put on secondary uses of data [2], primary needs remain primary. Upstream changes in the logic of collection—as was the case above when investigative patterns led to a human trafficking risk model becoming a car-wash detector [59]—can have significant downstream effect. Particularly where models are quietly being developed or piloted, or are located in a different part of the organisation from data collection efforts, it is easy for changes in practices to occur without those responsible for model performance to be aware of them. Accountability becomes difficult to trace in these situations. As Nick Seaver puts it, these systems are 'not standalone little

boxes, but massive, networked ones with hundreds of hands reaching into them' [68]. Accountable systems should to be internally accountable, else it would appear to be difficult for external accountability to either make sense or be sustained.

Data can also change *because* of the model rather than only in spite of it. Where models allocate the same resources that collect data then they are directly influencing the future sampling of their training data [67]. Sending police officers to areas of high predicted crime is an example of this. In the worst cases, the model can have a polarising effect: directing police resources disproportionately to areas with slightly higher crime risk will, without corrections, skew future training data collection in those area, which might be demographically or socioeconomically disproportionate [21]. In other cases, effects might be more subtle but of equal importance, and might cause particular failures to occur in unforeseen ways. If individuals react to try and influence or game a system—as the example stories above indicate is certainly possible—then the future population distribution becomes a function of past model decisions or structure. Little work has focussed on this so far, particularly on the impacts of these on the research into statistical fairness and non-discrimination properties, which broadly implicitly assume stationarity in their problem set-up. This is also a topic not substantively covered in existing literature, which is largely founded on data collected online, such as in the process of optimising advertising revenue. The adverts you are delivered might slowly change your behaviour, but each one can hardly be thought to have a significant impact. This is not the case in the public sector. As [X8] recalled above when discussing crime dispersion, feedback effects in practice can be so strong that they make models rapidly fail. The effect of this property on fairness and accountability in systems has yet to be properly unpacked and explored in context.

How to respond to concerns around shifting data? To some extent, the problem is highly interpersonal. The notion of a *visibility debt* has received some attention from engineers of both machine learning and traditional software engineering [56, 67]. To the upstream data collectors, there are *undeclared users* of their data streams. To the downstream users, there are individuals exerting influence over their system that that might not even be aware that such a system exists, particularly when data is collected in a decentralised manner by, say, auditors or police patrol officers. This problem is only going to be exacerbated as more models are made using the same upstream data sources, and bilateral communication becomes more and more challenging. Better communication might help, but must overcome difficult hurdles of explaining to upstream actors the kind of changes that matter downstream, and the kind that don't, in ways that they not only understand (as they might be relatively statistical) but that they can identify and act on within their roles. This is all compounded by how changing upstream data collection might not be an explicit act at all, but one emerging from cultural change or use of discretion. This is emphasised in the importance of so-called 'street-level ministers' in the public administration literature, which points out how formal rules are only part of the picture, and that many day-to-day choices in the grey zones are made by bureaucrats at the frontlines of public service [52]. Where change does occur, managers might not notice it, as in their day-to-day roles or through their monitoring and evaluation tools, they only see part of the picture [10].

A second approach would assume communication failure is inevitable, pushing instead a focus on the changing data itself. This would involve concept drift detection, sets of techniques designed to automatically detect shifts in distributions potentially relevant to a modelling task. Concept drift detection, particularly in complex real-world contexts, is difficult and daunting theoretically, let alone practically [65, 28]. Some of the more recent reviews in the field call for the integration of domain knowledge in order to discern relevant drift [28], yet there are few, if any well-explored methods for doing this.

'Always a person involved': Augmenting outputs

While we hear horror stories of the results of algorithms unquestioningly replacing swathes of existing analytical practice and institutional knowledge, our informants' experiences do not reflect that. Many organisations interviewed here have well-developed routines for augmenting algorithmic outputs, such as crime maps, with contextual data using manual analysts. As one informant described, their 'predictive policing' system was not supposed to bring in shocking new insights, but relieve analysts from the slog of generating maps so that they could get on with more advanced work. How algorithmic systems are examined day-to-day and how humans enter 'the loop' of decision-making at different stages is an important area for future design focus. There are many points for intervention in a decision support system outside of the modelling process—for example, in the training data (many systems attempting to make fairer machine learning system intervene at this point [45, 23]) or after the model has been generated [46], such as the stage between model output and map dissemination [12]. Particularly in this latter stage, design interventions are likely to be key. If a statistical definition of fairness is reached, it may be possible to make a 'fair' model, for example by introducing fairness constraints to optimisation. This provides no guarantees about decision-support being interpreted fairly. Designers should not just consider how to design artifacts such as maps to promote fairness, but should also do so in contexts imagining that models have been 'scrubbed' of certain types of bias, to understand if this introduces any additional effects. In the messy outside world, these efforts may interact, and it is not guaranteed that the sum of two good efforts is also effective.

Taking this areas forward will likely require building upon and rethinking traditional knowledge elicitation techniques. Effective knowledge elicitation, as part of the hot topic of knowledge acquisition in heady days of expert systems, was thought to be a foundational building block of AI [15, 34]. With the inductive, data-driven turn, we may need to rediscover it as something which constrains and augments patterns learned from data, less around tough or rare cases [35] as much as around contentious, value-laden ones. This will require very different sorts of prioritisation and elicitation methods than developed so far, and seems a promising and urgent avenue for future research.

'When it aligns with intuition': Understanding discretion

It is commonly claimed that people over-rely on algorithmic systems, or increasingly consider them neutral or authoritative [9]. We do not claim this is not an issue—but according to the informants in this project, this framing is one-dimensional. In particular, if and how individuals trust and rely on decision-support systems seems highly contextual in nature. The design strategies used to improve uptake of these systems, such as presenting prioritised lists or options, are understudied in relation to how these affect the mental models constructed by those using these systems day-to-day.

The way that different tasks, stakes or contexts mediate these effects is even less studied. We might expect there to be a difference in the perception of 'neutrality' of algorithms between those that direct police helicopters and those that flag children at risk of abuse; two very different tasks. We might not expect however, as informants reported, there to be a significant difference in the way algorithms were considered by helicopter pilots versus by police motorcyclists. Research in risk perception by helicopter pilots has found additional disparities between experienced and inexperienced users which is also worth unpacking in this context [74]. Ultimately, to make blanket and somewhat alarmist statements about how algorithms are or are not being questioned is likely to alienate practitioners who recognise a much more nuanced picture on the ground, and hinder co-operation in this space between researchers and those who would benefit from research uptake.

As well as the demographics and contexts of when algorithms are trusted more or less on aggregate, we might be interested in patterns of over- or under-reliance *within* individuals or use settings. If users of decision-support choose to ignore or to follow advice at random, we may not be wholly concerned with this, or at least our concern might centre on the dimension of increasing adherence. Yet if there are *systematic* biases in the way that advice is or is not used—particularly if they result in individuals holding different protected characteristics being treated differently—then this may create cause for alarm, or at least merit further study. Assuming a system can be 'scrubbed' of bias and then forced onto users to obey is clearly not what will happen in real world deployments.

Lastly, researchers considering human factors in computer security have emphasised 'shadow security practices', which consist of 'workarounds employees devise to ensure primary business goals are achieved' and 'reflect the working compromise staff find between security and "getting the job done"' [49]. Similarly, studies of fairness and accountability in socio-technical systems must incorporate an assumption that there will be a mixture of technological resistance and ad-hoc efforts, which, similarly to the findings in human factors of security, will surely be 'sometimes not as secure as employees think.' You can't engineer ethics, and you can't expect some individuals not to try, rigorously or not, to uphold it in ways they see fit. It is a useful heuristic to assume systems are trained on 'pure' streams of data and then must be cleaned of bias downstream, but in real data collection environments, even upstream actors in the data collection process attempt to work in the discretionary places computer systems allow (and create) to inject fairness where they see fit [44].

'I'm called the single point of failure': Moving practices

Most of the work in discrimination-aware data mining involves statistical assurance of fairer systems, or the installation of interfaces to make them more transparent. Most of the experiences of informants in this study were the opposite—social detection of challenges and social solutions to those challenges, none of which were mathematically demonstrated to work, but which organisationally at least were perceived to be somehow effective. Managing these challenges will require a balance between the two that has seldom been effectively struck. It seems unlikely that statistical practices could exist without the social practices, or the other way around.

This means that how the social practices are developed, maintained and transferred across contexts or over time is important to consider. Public sector bodies are under the constant shadow of their core quantitatively trained staff being poached, moving agencies, or leaving the sector entirely. Several interviewees had recently entered their job from another part of government where they pioneered analytics, or were about to leave from their current post. One modeller described how their manager called them "*the single point of failure for the entire force*" [X14]. As discussed above, there is significant concern within the sector that less resourced sister organisations will import the models without the hard-won practices to understand and mitigate issues such as bias and discrimination. Some of the informal practices that are established might be able to be documented, at least for inspiration if not for reproduction—employee handover notes are of course commonplace in these organisations. Yet other practices, particularly any critical skills that led to the establishment of practices in the first place, will likely be more challenging to codify.

Encoding social practices that surround software systems has always been challenging. The stakes are now higher than ever. Relevant efforts might involve the creation of informal and dynamic knowledgebases and virtual communities to share ethical issues and quandaries in relation to algorithmic support in practice [79], but expecting this to arise organically in competitive or resource-scarce fields is risky. Considering what collaboration in these domains could and should look like is of immediate importance to practitioners today.

'Looks like we've 100% accuracy': Talking performance

Some of the most value laden aspects of machine learned models relate to loss functions and performance metrics. Yet, beyond accuracy, false positives or negatives, it fast becomes difficult to explain performance effectively to those lacking technical background, but whose vertical accountability for the project or necessary, extensive domain knowledge makes it necessary. As recalled above, some informants complained of challenges explaining performance when accuracy was not the appropriate task-specific metric, such as in heavily imbalanced datasets (where you can get a high accuracy by using a dumb classifier that always predicts one class). There are a range of performance metrics suitable for imbalanced data [42], but these mostly lack clear analogies for laypeople. Moving away

from binary classification, explaining performance metrics for continuous regression tasks or multiple classification tasks is arguably more challenging still.

In other cases described above, performance was judged in other ways: models were not trusted or thought valuable if they did not contain individuals' "*preferred risk indicators*" [X2]; were too similar to analysis that existed before [X6]; or even if they were *more* accurate than was initially planned for, as the commissioners would rather the rest of that performance be substituted for interpretability [X3]. Other informants emphasised the importance of talking to users before determining performance metrics [X15], as in some cases only actionable knowledge is worth optimising for (see also [1]). This broadly chimed with many respondents' conception of the most important performance metric of all—for contractors, whether a client bought a model, and for public servants or in-house modellers, whether their department actually used it.

Given that performance metrics are one of the most value-laden parts of the machine learning process [5, 14], it will be key to discuss them both with statistical rigour and with practical relevance. This intuitively seems to present domain-specific challenges in training, visualisation, user interfaces, statistics and metrics, problem structuring and knowledge elicitation, among other fields.

CONCLUDING REMARKS

Researchers should be wary of assuming, as seems often the case in current discourse, that those involved in the procurement and deployment of these systems are necessarily naïve about challenges such as fairness and accountability in the public sector's use of algorithmic decision support. This assumption sits particularly uncomfortably with the value attributed to participatory design and action research in HCI and information systems [33, 6]. While those involved in acquiring these technologies for the public sector might not be prime candidates for developing new statistical technologies for understanding bias and outputs in complex models, this does not mean that they do not care or do not try to tackle ethical issues that they perceive. Indeed, as well as the individual perspectives in this paper, some public agencies are already developing their own in-house ethical codes for data science activities [30]. Yet issues like fairness have been shown to come with technically difficult to reconcile, or even irreconcilable trade-offs—something well-demonstrated by Alexandra Chouldechova's impossibility theorem illustrating that independently plausible formal definitions of fairness can be statistically incompatible with one another [13], or concerns raised that explanation facilities might work better for some outputs than for others [20]. Reconciling these harder boundaries and issues within messy organisational contexts will present a major challenge to research uptake in this field in the coming years.

Where to go from here? We believe that the challenges we outlined above—dealing with changing data, better understanding discretion and the augmentation of model outputs, better transmission of social practices and improved communication of nuanced aspects of performance—sit amongst a range of promising areas for future interdisciplinary collaboration. The implicit and explicit assumptions of proposed solutions to both these challenges and to the broader issues must be stress-tested in real situations. This presents important questions of methodology. Domain-specific, organisational and contextual factors are crucial to closely consider in the context of interventions intended to improve the fairness and accountability of algorithmic decision-support. The institutional constraints, high stakes and crossed lines of accountability in the public sector arguably presents even more reason to do so. Only so much can be learned from studying systems *in vitro*, even with access to impressive quantities of relevant, quality data with which to experiment. Those interested in transformative impact in the area of fair and accountable machine learning must move towards studying these processes *in vivo*, in the messy, socio-technical contexts in which they inevitably exist. Interventions will have to cope with institutional factors, political winds, technical lock-in and ancient, withering infrastructure head on, as they would have to in the real world. Researchers will have to facilitate the navigation of contested values, and will not always have the freedom of seeking the types of accountability or fairness that they feel most comfortable with. Such challenges should be embraced. To enable this, trust will need to be built between public bodies and researchers; trust that is currently being endangered by 'gotcha!'–style research that seeks to identify problematic aspects of algorithmic systems from afar without working collaboratively to understand the processes by which they came about and might be practically remedied. Action research is a core methodology that would support these aims [6], but the combination of high stakes and a wariness that researchers might be spending more effort looking for algorithmic harms than offering help to fix it might make public agencies reluctant to open up to research interventions.

Rarely have the issues HCI concerns itself with been as directly involved in steering choices related to the use of governmental power as much as they are today. As we involve more advanced decision-support, and even decision-making, systems in the workings of the state, this field might even be the 'difference that makes a difference' to the rights and freedoms of vulnerable societal groups. We believe that making this difference is possible, but only in close collaboration with different disciplines, practitioners and affected stakeholders. Future research must engage with not only with the new questions and avenues of exploration such research brings, but also the practical constraints that come with studying politically charged settings and developing workable social and technical improvements within them.

ACKNOWLEDGMENTS

Funded by the Engineering & Physical Sciences Research Council (EPSRC): EP/M507970/1 [MV], EP/J017728/2 [MVK, RB]. This study was approved by UCL's Research Ethics Committee (7617/001) and is distributed CC-BY 3.0. Participants in several workshops—FAT/ML '17 (Halifax, CA); Politiques de modélisation algorithmique (ULB, BE); TRILCon (Winchester, UK); Big Data: New Challenges for Law & Ethics (Ljubljana, SI); and The Human Use of Machine Learning (Ca'Foscari, Venice, IT)—provided helpful comments on presentations of early forms of this work.

REFERENCES

1. Monsuru Adepeju, Gabriel Rosser, and Tao Cheng. 2016. Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions—a crime case study. *International Journal of Geographical Information Science* 30, 11 (2016), 2133–2154. DOI: http://dx.doi.org/10.1080/13658816.2016.1159684
2. Administrative Data Taskforce. 2012. *The UK Administrative Data Research Network: Improving access for research and policy*. Economic and Social Research Council. http://www.esrc.ac.uk/files/research/administrative-data-taskforce-adt/improving-access-for-research-and-policy/
3. AI Now. 2016. *The AI Now Report: The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*. https://artificialintelligencenow.com/
4. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
5. Solon Barocas and Andrew D Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104 (2016), 671–732. DOI:http://dx.doi.org/10.15779/Z38BG31
6. Richard L Baskerville and A Trevor Wood-Harper. 1996. A critical perspective on action research as a method for information systems research. *Journal of Information Technology* 11, 3 (1996), 235–246. DOI: http://dx.doi.org/10.1080/026839696345289
7. Gwyn Bevan and Christopher Hood. 2006. What's measured is what matters: Targets and gaming in the English public health care system. *Public Administration* 84, 3 (2006), 517–538. DOI: http://dx.doi.org/10.1111/j.1467-9299.2006.00600.x
8. Julia Black. 2005. The emergence of risk-based regulation and the new public risk management in the United Kingdom. *Public Law* (2005), 512–549. Issue Autumn. https://perma.cc/Z8AU-4VNN
9. danah boyd. 2016. Undoing the neutrality of Big Data. *Florida Law Review Forum* 16 (2016), 226–232.
10. Aurélien Buffat. 2015. Street-level bureaucracy and e-government. *Public Management Review* 17, 1 (2015), 149–161. DOI: http://dx.doi.org/10.1080/14719037.2013.771699
11. Matthew Chalmers and Ian MacColl. 2003. Seamful and seamless design in ubiquitous computing. In *Workshop at the crossroads: The interaction of HCI and systems issues in UbiComp*, Vol. 8. https://perma.cc/2A3D-NMJP
12. Hsinchun Chen, Homa Atabakhsh, Chunju Tseng, Byron Marshall, Siddharth Kaza, Shauna Eggers, Hemanth Gowda, Ankit Shah, Tim Petersen, and Chuck Violette. 2005. Visualization in law enforcement. In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*. 1268–1271.
13. Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163. DOI: http://dx.doi.org/10.1089/big.2016.0047
14. Cary Coglianese and David Lehr. 2016. Regulating by Robot: Administrative Decision Making in the Machine-Learning Era. *Geo. LJ* 105 (2016), 1147. https://ssrn.com/abstract=2928293
15. Nancy J. Cooke. 1994. Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies* 41, 6 (1994), 801–849. DOI: http://dx.doi.org/10.1006/ijhc.1994.1083
16. Patrick Dunleavy, Helen Margetts, Simon Bastow, and Jane Tinkler. 2006. *Digital Era Governance: IT Corporations, the State and e-Government*. Oxford University Press, Oxford. DOI:http://dx.doi.org/10.1093/acprof:oso/9780199296194.001.0001
17. Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. 214–226. DOI: http://dx.doi.org/10.1145/2090236.2090255
18. Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718. DOI: http://dx.doi.org/10.1016/S1071-5819(03)00038-7
19. Editor. 2016. More accountability for big-data algorithms. *Nature* 537, 7621 (2016), 449. DOI: http://dx.doi.org/10.1038/537449a
20. Lilian Edwards and Michael Veale. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You are Looking For. *Duke Law & Technology Review* 16, 1 (2017), 18–84. DOI: http://dx.doi.org/10.2139/ssrn.2972855
21. Danielle Ensign, Sorelle A. Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. *Presented as a talk at the 4th Workshop on Fairness, Accountability and Transparency in Machine Learning (FAT/ML 2017), Halifax, Canada* (2017). https://arxiv.org/abs/1706.09847
22. European Commission. 2017. *Tender specifications: Study on Algorithmic Awareness Building, SMART 2017/0055*. https://etendering.ted.europa.eu/cft/cft-document.html?docId=28267
23. Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. 259–268. DOI: http://dx.doi.org/10.1145/2783258.2783311

24. Gerhard Fischer. 1991. The importance of models in making complex systems comprehensible. In *Mental Models and Human-Computer Interaction*, MJ Tauber and D Ackermann (Eds.). Elsevier, Noord Holland.
25. Diana E Forsythe. 1995. Using ethnography in the design of an explanation system. *Expert Systems with Applications* 8, 4 (1995), 403–417. DOI: http://dx.doi.org/10.1016/0957-4174(94)E0032-P
26. Batya Friedman and Helen Nissenbaum. 1996. Bias in Computer Systems. *ACM Trans. Inf. Syst.* 14, 3 (July 1996), 330–347. DOI: http://dx.doi.org/10.1145/230538.230561
27. Robert D Galliers and Frank F Land. 1987. Choosing appropriate information systems research methodologies. *Commun. ACM* 30, 11 (1987), 901–902. DOI: http://dx.doi.org/10.1145/32206.315753
28. J Gama, Indre Žliobaitė, A Bifet, M Pechenizkiy, and A Bouchachia. 2013. A survey on concept drift adaptation. *Comput. Surveys* 1, 1 (2013). DOI: http://dx.doi.org/10.1145/2523813
29. Raphaël Gellert, Katja de Vries, Paul de Hert, and Serge Gutwirth. 2013. A Comparative Analysis of Anti-Discrimination and Data Protection Legislations. In *Discrimination and privacy in the information society*, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.). Springer, Heidelberg. DOI: http://dx.doi.org/10.1007/978-3-642-30487-3_4
30. Government Digital Service. 2015. *Data science ethical framework*. HM Government, London. https://www.gov.uk/government/publications/data-science-ethical-framework
31. Government Office for Science. 2016. *Artificial intelligence: Opportunities and implications for the future of decision making*. HM Government, London. https://www.gov.uk/government/publications/artificial-intelligence-an-overview-for-policy-makers
32. Sara Hajian and Josep Domingo-Ferrer. 2012. Direct and indirect discrimination prevention methods. In *Discrimination and privacy in the information society*, Bart Custers, Toon Calders, Bart Schermer, and Tal Zarsky (Eds.). Springer, Berlin, Heidelberg, 241–254.
33. Gillian R Hayes. 2011. The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 3 (2011), 15. DOI: http://dx.doi.org/10.1145/1993060.1993065
34. Robert R Hoffman. 2008. Human factors contributions to knowledge elicitation. *Human factors* 50, 3 (2008), 481–488. DOI: http://dx.doi.org/10.1518/001872008X288475
35. Robert R Hoffman, Beth Crandall, and Nigel Shadbolt. 1998. Use of the critical decision method to elicit expert knowledge: A case study in the methodology of cognitive task analysis. *Human Factors* 40, 2 (1998), 254–276. DOI:http://dx.doi.org/10.1518/001872098779480442
36. Christopher Hood. 1991. A public management for all seasons? *Public Administration* 69 (1991), 3–19. DOI: http://dx.doi.org/10.1111/j.1467-9299.1991.tb00779.x
37. V David Hopkin. 1995. *Human factors in air traffic control*. CRC Press, London.
38. Robert Hoppe. 2011. *The governance of problems: Puzzling, powering and participation*. Policy Press.
39. House of Common Science and Technology Committee. 2016. *Robotics and artificial intelligence (HC 145)*. The House of Commons, London. http://www.publications.parliament.uk/pa/cm201617/cmselect/cmsctech/145/145.pdf
40. House of Commons Science and Technology Committee. 2016. *The big data dilemma (HC 468)*. House of Commons, London. http://www.publications.parliament.uk/pa/cm201516/cmselect/cmsctech/468/468.pdf
41. Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin I P Rubinstein, and J D Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*. 43–58. DOI:http://dx.doi.org/10.1145/2046684.2046692
42. Nathalie Japkowicz and Mohak Shah. 2011. *Evaluating learning algorithms: A classification perspective*. Cambridge University Press, Cambridge, UK.
43. Torben Beck Jørgensen and Barry Bozeman. 2007. Public values: An inventory. *Administration & Society* 39, 3 (2007), 354–381. DOI: http://dx.doi.org/10.1177/0095399707300703
44. Frans Jorna and Pieter Wagenaar. 2007. The 'iron cage' strengthened? Discretion and digital discipline. *Public Administration* 85, 1 (2007), 189–214. DOI: http://dx.doi.org/10.1111/j.1467-9299.2007.00640.x
45. Faisal Kamiran and Toon Calders. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33, 1 (2012), 1–33. DOI: http://dx.doi.org/10.1007/s10115-011-0463-8
46. Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. 2010. Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining*. 869–874. DOI:http://dx.doi.org/10.1109/ICDM.2010.50
47. Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas, and David F Lobach. 2005. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *BMJ* 330, 7494 (2005), 765. DOI: http://dx.doi.org/10.1136/bmj.38398.500764.8F
48. Sara Kiesler and Jennifer Goetz. 2002. Mental Models of Robotic Assistants. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. 576–577. DOI:http://dx.doi.org/10.1145/506443.506491

49. Iacovos Kirlappos, Simon Parkin, and M. Angela Sasse. 2015. "Shadow Security" As a Tool for the Learning Organization. *SIGCAS Comput. Soc.* 45, 1 (2015), 29–37. DOI:http://dx.doi.org/10.1145/2738210.2738216

50. Daniel Antony Kolkman, Paolo Campo, Tina Balke-Visser, and Nigel Gilbert. 2016. How to build models for government: Criteria driving model acceptance in policymaking. *Policy Sciences* 49, 4 (2016), 489–504. DOI: http://dx.doi.org/10.1007/s11077-016-9250-4

51. Christopher A Le Dantec and W Keith Edwards. 2010. Across boundaries of influence and accountability: The multiple scales of public sector information systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'10*. ACM, 113–122. DOI:http://dx.doi.org/10.1145/1753326.1753345

52. Michael Lipsky. 2010. *Street-level bureaucracy: Dilemmas of the individual in public services*. Russell Sage Foundation, New York.

53. Zachary C Lipton. 2016. The Mythos of Model Interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York. https://arxiv.org/abs/1606.03490

54. Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition* 65 (2017), 211 – 222. DOI:http://dx.doi.org/https://doi.org/10.1016/j.patcog.2016.11.008

55. Robin Moore (Ed.). 2015. *A compendium of research and analysis on the Offender Assessment System*. Ministry of Justice Analytical Series, London. DOI: http://dx.doi.org/https://perma.cc/W2FT-NFWZ

56. J. David Morgenthaler, Misha Gridnev, Raluca Sauciuc, and Sanjay Bhansali. 2012. Searching for Build Debt: Experiences Managing Technical Debt at Google. In *Proceedings of the Third International Workshop on Managing Technical Debt, MTD'12, Zurich, Switzerland — June 05, 2012*. 1–6. DOI: http://dx.doi.org/10.1109/MTD.2012.6225994

57. Kathleen L. Mosier, Linda J. Skitka, Susan Heers, and Mark Burdick. 1998. Automation Bias: Decision Making and Performance in High-Tech Cockpits. *The International Journal of Aviation Psychology* 8, 1 (1998), 47–63. DOI: http://dx.doi.org/10.1207/s15327108ijap0801_3

58. Nesta. 2015. *Machines that learn in the wild: Machine learning capabilities, limitations and implications*. Nesta, London. https://perma.cc/A6AM-GV6X

59. BBC News. 2016. Kent slavery raids 'uncover 21 victims'. *BBC News* (7 Dec. 2016). https://perma.cc/AM4S-RMHR

60. Donald A Norman. 1983. Some observations on mental models. In *Mental Models*, Dedre Gentner and Albert L Stevens (Eds.). Psychology Press, New York City, NY, 7–14.

61. Teresa Odendahl and Aileen M Shaw. 2002. Interviewing elites. *Handbook of Interview Research* (2002), 299–316. DOI:http://dx.doi.org/10.4135/9781412973588.n19

62. Marion Oswald, Jamie Grace, Sheena Urwin, and Geoffrey C. Barnes. forthcoming. Algorithmic Risk Assessment Policing Models: Lessons from the Durham Hart Model and 'Experimental' Proportionality. *Information & Communications Technology Laws* (forthcoming). https://ssrn.com/abstract=3029345

63. Edward C Page and Bill Jenkins. 2005. *Policy bureaucracy: Government with a cast of thousands*. Oxford University Press, Oxford. DOI:http://dx.doi.org/10.1093/acprof:oso/9780199280414.001.0001

64. Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 560–568. DOI: http://dx.doi.org/10.1145/1401890.1401959

65. Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press, Cambridge, MA. DOI:http://dx.doi.org/10.7551/mitpress/9780262170055.001.0001

66. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. DOI: http://dx.doi.org/10.1145/2939672.2939778

67. D Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden Technical Debt in Machine Learning Systems. In *Proceedings of the 28th International Conference on Neural Information Processing Systems, Montréal, Canada — December 07–12, 2015*. MIT Press, Cambridge, MA, 2503–2511. https://perma.cc/G6VN-9B86

68. Nick Seaver. 2013. Knowing algorithms. *Media in Transition* 8 (2013). https://perma.cc/8USJ-VTWS

69. Nick Seaver. 2014. On reverse engineering: Looking for the cultural work of engineers [blog. *Medium* (2014). https://medium.com/anthropology-and-algorithms/on-reverse-engineering-d9f5bae87812

70. Andrew Selbst. forthcoming. Disparate Impact in Big Data Policing. *Georgia Law Review* (forthcoming). DOI: http://dx.doi.org/10.2139/ssrn.2819182

71. Linda J Skitka, Kathleen L Mosier, and Mark Burdick. 1999. Does automation bias decision-making? *International Journal of Human-Computer Studies* 51 (1999), 991–1006. DOI: http://dx.doi.org/10.1006/ijhc.1999.0252

72. The Royal Society. 2017. *Machine learning: The power and promise of computers that learn by example*. The Royal Society, London. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/machine-learning-report.pdf

73. The Royal Society and the British Academy. 2017. *Data management and use: Governance in the 21st Century*. The Royal Society and the British Academy, London. https://royalsociety.org/~/media/policy/projects/data-governance/data-management-governance.pdf

74. Mary E Thomson, Dilek Önkal, Ali Avcioğlu, and Paul Goodwin. 2004. Aviation risk perception: A comparison between experts and novices. *Risk Analysis* 24, 6 (2004), 1585–1595. DOI: http://dx.doi.org/10.1111/j.0272-4332.2004.00552.x

75. Alan B Tickle, Robert Andrews, Mostefa Golea, and Joachim Diederich. 1998. The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9, 6 (1998), 1057–1068. DOI:http://dx.doi.org/10.1109/72.728352

76. Nikolaj Tollenaar, B. S. J. Wartna, P.G.M Van Der Heijden, and Stefan Bogaerts. 2016. StatRec — Performance, validation and preservability of a static risk prediction instrument. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 129, 1 (2016), 25–44. DOI: http://dx.doi.org/10.1177/0759106315615504

77. Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How It Works: A Field Study of Non-technical Users Interacting with an Intelligent System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. New York, NY, USA, 31–40. DOI: http://dx.doi.org/10.1145/1240624.1240630

78. Berk Ustun and Cynthia Rudin. 2016. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning* 102, 3 (2016), 349–391. DOI: http://dx.doi.org/10.1007/s10994-015-5528-6

79. Michael Veale and Reuben Binns. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4, 2 (2017). DOI: http://dx.doi.org/10.1177/2053951717743530

80. Wetenschappelijke Raad voor het Regeringsbeleid. 2016. *Big Data in een vrije en veilige samenleving (WRR-Rapport 95)*. WRR, Den Haag. http://www.wrr.nl/publicaties/publicatie/article/big-data-in-een-vrije-en-veilige-samenleving/

81. Michael R Wick and William B Thompson. 1992. Reconstructive expert system explanation. *Artificial Intelligence* 54, 1-2 (1992), 33–70. DOI: http://dx.doi.org/10.1016/0004-3702(92)90087-E

82. Langdon Winner. 1980. Do Artifacts Have Politics? *Dædelus* 109, 1 (1980), 121–136. http://www.jstor.org/stable/20024652

83. Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the Heart Pump Implant Decision Process: Opportunities for Decision Support Tools to Help. In *Proceedings of the 2016 SIGCHI Conference on Human Factors in Computing Systems, CHI'16*. 4477–4488. DOI: http://dx.doi.org/10.1145/2858036.2858373

84. Yunfeng Zhang, Rachel KE Bellamy, and Wendy A Kellogg. 2015. Designing information for remediating cognitive biases in decision-making. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI'15*. 2211–2220. DOI: http://dx.doi.org/10.1145/2702123.2702239