Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making

Reuben Binns

University of Oxford, Department of Computer Science

Reuben.binns@cs.ox.ac.uk

DRAFT 01/09/2019. UNDER REVIEW.

Abstract:

There are various arguments in favour of tempering algorithmic decision-making with human judgement. One common family of arguments appeal to concepts and criteria derived from legal philosophy about the nature of law and legal reasoning, and argue that algorithmic systems cannot satisfy them (but humans can). This paper argues that among the latter family of arguments, there is often an implicit appeal to the notion that each case needs to be assessed on its own merits, without comparison to or generalisation from previous cases. This notion of 'individual justice' has featured in jurisprudential debates about the granularity of rules and tests, and the (in)justice of discrimination, but has not yet been explicitly imported into debates about justice and algorithmic decision-making.

This paper has several aims. The first is to provide an overview of the concept of individual justice and distinguish it from related but distinct arguments about the value of human discretion. Equipped with this account of human judgement as a guarantor of individual justice, the second aim is to consider its place within and beside algorithmic decision-making. It argues that in so far as individual justice is valuable, it can only be meaningfully served through human judgement, because it antithetical to the kind of pre-determined reasoning that characterises algorithmic systems. This suggests that – to the extent that individual justice is deemed important – a requirement for human intervention and oversight over algorithmic decisions is necessary. The third aim is to consider how individual justice relates to other dimensions of justice, namely consistency and fairness or non-discrimination. Finally, the article discusses two challenges that are raised by this account. The first challenge concerns how individual justice can be accommodated alongside other dimensions of justice in the socio-technical contexts in

which humans-in-the-loop are situated. The second concerns the potential inequities in individual justice that might result from an uneven application of human judgement in algorithmic settings.

Keywords: algorithmic regulation, data protection, discretion, human-in-the-loop, justice

1. Introduction

Algorithmic systems are increasingly involved in the exercise of power by state and private actors, prompting concerns about justice. Referred to variously by terms including predictive analytics, machine learning, and artificial intelligence (AI), they are used to determine, solely or in combination with human input, the allocation of loans and jobs, the treatment of incarcerated people, the investigation of taxpayers for fraud, or the prioritisation of health treatments. Taking a broad view of regulation as the *management of risk* or *altering of behaviour*, such technologies have been described as enabling a form of 'algorithmic regulation' (Yeung, 2018).

Despite its newfound attention, algorithmic regulation has arguably been around for much longer, in the form of automated business processes such as credit scoring, and automated assessment of citizens by the state. A common safeguard found in various regulatory instruments and provisions, is the requirement to involve a human reviewer in the assessment of an individual case that would otherwise have been decided upon by an algorithmic decision-making system. The European Union's General Data Protection Regulation (GDPR) generally prohibits decisions based on personal data with legal or similar effects if they are solely automated, i.e. without human intervention (Article 22), outside of three specified legal bases; and where those legal bases are relied on, the right to obtain human intervention is explicitly called for as a safeguard in Article 22(3).

There are various reasons why regulators might be concerned to maintain human oversight over algorithmic systems. Some are based on the limitations of machines; the need to keep a 'human-in-the-loop' is frequently appealed to in human-computer interaction research and in discussions of the governance of artificial intelligence (Dautenhahn, 1998; Santoni de Sio and Van den Hoven, 2018; Weiner, 1950), primarily to avoid certain failure modes, or because a combination of human and machine collaboration is most cost-efficient (de Winter and Dodou, 2014). Others are based on a characterisation of relative strengths; while people do the creative work, machines can be used for administration (Applin and Fischer, 2015; Berners-Lee and Fischetti, 2001; Shadbolt et al., 2019). Another set of reasons for maintaining humans as the locus of decision-making power is that they are situated in social and institutional contexts which allow for liability and apportioning of responsibility (Bryson et al., 2017), or societal legitimation (Rahwan, 2018). According to this view, human involvement needs to be substantive to ensure that genuine thought has been applied, and that control can be exercised where necessary, otherwise these liability arrangements may risk humans being reduced to rubber-stamping *quasi-automated* decisions which they were not meaningfully involved in

(Wagner, 2019), or acting as a 'moral crumple zone' (Elish, 2016). Even if algorithms could effectively learn from observing human judgements, we may still want to keep human decision makers around to generate fresh ground truth and to avoid moral atrophy (Hildebrandt, 2013). Related arguments emphasise the importance of human judgement as something which can be justified with reasons (Oswald, 2018; Pasquale, 2018), which can be challenged and rebuked (Pasquale and Cashwell, 2018), and which respects the rule of law (Hildebrandt, 2018; Zalnieriute et al., 2019).

The focus of this paper is on those arguments which invoke legal philosophical accounts about the need for discretionary decision-making for justice. It begins with a review of the ways in which jurisprudential theory has figured into the debates about algorithmic decision-making in various waves since the late 1980's. I argue that a common element which implicitly underlies such arguments – but has not yet been made explicit – is the concept of individual justice. This refers to the notion that each case needs to be assessed on its own merits, without comparison to, or generalisation from, previous cases. Wherever individual justice is required, algorithmic decision-making systems cannot (entirely) replace human judgement. However, this conclusion presents several questions and some challenges. One question concerns how individual justice can be accommodated alongside other dimensions of justice, such as consistency, fairness, or non-discrimination. This may be especially tricky when we consider the ways in which human decision-makers may interact with socio-technical algorithmic systems in unexpected ways and bring their own commitments, pressures and norms to bear on the decision. A second challenge concerns the potential inequities in individual justice that might result from an uneven application of human judgement.

2. How the rules-versus-discretion debate enters into the automation-versus-human debate

The United Kingdom's 1998 Social Security Bill introduced new provisions for how decisions about welfare benefits may be made, including section 2 on 'the use of computers' (Le Sueur, 2015). Baroness Hollis of Heigham, speaking on behalf of the UK government, assured the House of Lords:

> 'decisions which require the *exercise of discretion or judgement* will continue to be made by the department's trained staff'.[1]

Similarly, a recent Council of Europe report argues that the 'discretion of decision-making processes cannot be automated' (Wagner, 2016). Exercising discretion in such contexts refers to the ability to deliberate about a case and come to a different decision than one which might otherwise be directly derived from a set of rules or protocols. This may involve weighing up conflicting rules and deciding which should take precedence in that particular case, or discounting a particular rule after consideration

[1] As quoted in (Le Sueur, 2015)

of certain contextual factors of the situation in question that render its application inappropriate. Algorithmic decision-making is usually regarded as incapable of exercising this kind of discretion, since there is no room for manoeuvre in a deterministic system. As Mireille Hildebrandt puts it: "Whoever determines 'this' as a condition of the 'that' decides the output of the system, which has no discretion whatsoever" (Hildebrandt 2018).[2]

Legal scholars commenting on the importance of human judgement over algorithmic decision-making have sometimes fleshed out such arguments with appeals to legal philosophy and jurisprudential theory (Citron, 2007; Cobbe, 2018; Edwards, 1995; Grimmelmann, 2004; Hildebrandt, 2018; Le Sueur, 2015; Leenes, 2003; Leith, 1986; Noto La Diega, 2018; Oswald, 2018). They have typically drawn on distinctions between *rules* versus *discretion*, and related distinctions between e.g. *rules* and *principles* or *standards* 00/00/0000 00:00:00; between formalised application of law and its open-ended interpretation (Bix, 1991); or between those figures most associated with differing positions, e.g. 'the Hart-Dworkin debate' (Shapiro, 2007).

While they have substantial differences and disagreements, these arguments generally involve some variation on the following claims. While law is partly about rules, and computational systems may be capable of implementing rule-based reasoning, rules are at most a constituent element of a more complex mixture of sources required for decision making to be legally compliant in particular cases. Independent judgement is required especially in cases that involve the application of possibly conflicting standards or principles, and is something that – so the argument goes – only humans can do. The implication is that while we might consider applying algorithms to the rule-based aspects of law – if they can be isolated[3] – we should leave humans in charge of the other aspects. "The rules-versus-standards literature", Danielle Citron argues, "can help guide an agency's initial decisions with regard to automation" (Citron, 2007).

Such arguments rest on two premises. First, that there is a clear distinction between two modes or sources of reasoning, namely, the *application of rules* on one hand, and *discretion* on the other. Second, that the differing nature of computational processes and human judgement means the former is naturally suited to the application of rules and the latter uniquely suited to discretion. Sceptics of these arguments might doubt these premises, either denying the distinction between rules and discretion, or denying that the exercise of discretion is a uniquely human faculty.

2.1 Are rules preferable to discretion? And can they be coded?

The claim that rules are preferable to discretion is a common theme in a variety of fields. Two which are particularly pertinent here are legal philosophy and public administration policy. As a philosophical

[2] Although Hildebrandt goes on to articulate how algorithmic systems in fact feature a different kind of discretion, situated in the design choices behind the system.
[3] Although the possibility of such isolation is questionable; see section XX below

position in its most extreme form, the claim is associated with A.V. Dicey, who argued that discretion is not only *outside* of law but in fact *opposed* to the rule of law. Drawing on Locke, who held that 'wherever law ends, tyranny begins', Dicey's position was that law is all and only rules, and therefore rules are the only alternative to tyranny (Dicey, 2013). There can be no legitimate exceptions to the rules outside of the rules themselves. This makes judicial decisions a matter of deducing a single correct output from a set of rules; if the same case were presented to two judges, they should always come to the same conclusion on the basis of the same rules. Rules apply in an all-or-nothing fashion, and must be consistent, in the sense that if two rules conflict, one of them must be invalid.

A second source of scepticism about the value of discretion comes from debates in public administration. For some, fears about the misuse of discretion by judges or public officials making arbitrary decisions, are an argument in favour of greater reliance on rule-making to confine and limit decision-makers (Davis, 1969). Attempts to curb administrative discretion can be seen in the U.S. in federal and state laws in the 1960s and 70s (Morgan, 1987), or in the 'new public management' (NPM) approaches to government and public services beginning in the 1980s (Lynn Jr, 1998). Similar trends are arguably observable in administrative law in favour of greater consistency and predictability rather than independent judgement.[4] With the introduction of algorithmic systems, administrators have become 'screen-level bureaucrats', whose discretion is potentially curtailed by the introduction of rules that are 'pre-programmed via algorithms and digital decision trees' (Bovens and Zouridis, 2002, p1).

If discretion can and should be replaced by more comprehensive and consistent rule-making, and if legal rules and logic for deriving valid inferences from those rules can be encoded into an algorithmic system, then one might be optimistic about the use of algorithmic regulation without human oversight. This optimism is reflected in early attempts to apply computation to legal reasoning using 'symbolic' artificial intelligence, which aimed to represent human knowledge as symbolic logic so that valid inferences can be derived. Such systems were first introduced by Layman Allen in 1956 (the very year the term 'artificial intelligence' was first introduced at the Dartmouth Summer Research Project) (Allen, 1956). Further work continued in the 1970's (e.g. (Buchanan and Headrick, 1970; McCarty, 1976)), but the topic gained prominence with the rise of 'expert systems' in the 1980's (Leith, 2016, 1987, 1986; Susskind, 1987). Had they been successful in validly deriving legal judgements from encoded rulesets, adherents of Dicey, NPM, and administrative rule-making might welcome such systems, and have no qualms about their inability replicate whatever is supposedly involved in human discretion.

However, legal expert systems, like expert systems in general, did not transpire to be as effective as hoped. It turned out to be too difficult to specify all the relevant rules for a given situation (McCarthy, 1980), and the legal domain was no exception (Leith, 2016). Furthermore, even if those systems had

[4] As Jennifer Cobbe notes, there is a 'trend towards preferring consistently applied policy', although recent Supreme Court decisions suggest that this is not 'a free-standing principle of administrative law in and of itself' (Cobbe, 2018).

worked, most do not share Dicey's confidence in legal certainty and NPM's hostility to discretion. As former Canadian Chief Justice Beverley McLachlin argues 'the law is not as certain as [Dicey] would have it, nor are administrators as arbitrary' (MccLachlin, 1992). Defenders of the first premise (that there is an important distinction between rules and discretion), tend to argue that legal judgements involve an irreducible element of contextual interpretation, which resists encoding into rules that can be derived for application in every context. To ignore this would risk a kind of tyrannical formalism, in which rules are applied regardless of contextual factors. Fictional treatments of such systems abound, from the unsympathetic, unthinking bureaucrats in Kafka's *The Trial*, to the rude bank clerk from the comedy series *Little Britain,* who repeatedly denies apparently qualified customers because the 'computer says no' (Wikipedia contributors, 2019). Typically, these examples are raised to illustrate the opacity of bureaucratic decision-making, but they are also fitting examples of their failure to address the individual case. Excessively rule-bound bureaucracies provide important case studies across the political spectrum (Arendt, 1973; Hayek, 1979; Weber, 2015). As these fictional and historical examples suggest, we should not only be concerned about constraining power through rules, but also about preventing the rules from being implemented in an automatic and tyrannical fashion.

2.2 Even if discretion is required, is it uniquely human?

Even if we accept that Dicey's rule-based conception of law is flawed, and concede that initial attempts to encode legal rules in symbolic AI were doomed to fail, is it nonetheless possible that whatever theorists believe to be good about discretionary human judgement could somehow be implemented in an algorithmic system? The argument would then shift from the familiar territory of 'rules-versus-discretion', to the more speculative question of whether there could be such a thing as machine discretion. Are human decision-makers truly exceptional in their ability to divine the relevant features and reasoning for a given case?

This argumentative shift challenges the defender of human judgement to articulate what exactly makes it so valuable, and to assess alternative, non-rule-based systems according to that account.

Such alternative systems have been pursued at various junctures over decades of work at the intersection of law and AI, which aims to model legal reasoning using a range of different computational techniques beyond the aforementioned rule-based and expert systems. In part this was prompted by a departure from early work on legal expert systems which focused on a rather limited set of legal theory – the 'Oxford school' of analytic jurisprudence of Dworkin and Hart, as in (Leith, 1987; Susskind, 1987). Other approaches, including critical and feminist perspectives, open up a larger space of models of legal reasoning against which different AI technologies could be compared (Edwards, 1995). For instance, inspired by the way pertinent historical *cases* are often marshalled in Socratic dialogues, 'case-based reasoning' systems aim to aid legal analysis by retrieving relevant similar cases from a knowledge base of prior cases (Rissland and Ashley, 1987).

A rich interdisciplinary body of work combining legal theory, philosophy, logic, informatics and computer science has addressed topics including the relationship between logic and legal argumentation, formalisation and justification, and more (see e.g. (Bankowski et al., 1995) for an overview). Other work, primarily by computer scientists, focuses not on theoretical questions about legal reasoning, but rather on practical methods for extracting legal information or predicting cases. 'Argumentation mining' based on techniques of information retrieval, attempts to automatically recognise the form and content of legal arguments from natural language text (Ashley and Walker, 2013; Lippi and Torroni, 2016). Machine learning techniques have been applied to providing automatic summaries of legal cases (Hachey and Grover, 2005), or to predict their outcomes based on historical data (Aletras et al., 2016).

Proponents of algorithmic decision-making might therefore be optimistic about the potential for new techniques such as machine learning, capable of representing hi-dimensional combinations of features, to overcome the problems faced by older systems which followed blunt rules (Russell and Norvig, 2016). If the problem with a system based on rules is that it can't infer when there should be exceptions to them, a model which considers many more factors ought to be better in so far as it can encode more exceptions.

Faced with this range of models of legal reasoning and proposed algorithmic systems for simulating them, the defender of human judgement must argue why the latter fail to genuinely embody the former. This could be done in a piecemeal way, focusing on particular permutations of models of legal reasoning and particular artificial intelligence approaches, and addressing their particular shortcomings in relation to human judgement. This has been the approach in many of the previous works cited above, and often to convincing effect. But such arguments require us to assess each algorithmic system in turn, and consider in each case whether that system is eluded by some special and distinctive form of human reasoning. Proponents of human judgement are thus always open to the charge that there is or will be some computational method that fully captures its allegedly unique qualities.

Alternatively, there may be more generic and fundamental arguments, which go beyond particular permutations of models of law and computation, and help establish the indispensability of human judgement regardless of the proposed algorithmic replacement. The next section outlines one such argument, what I call the argument from individual justice.

3. The argument from individual justice

In short, the argument is that *only humans* are capable of *truly* case-by-case judgement, and such judgement is necessary for justice, in particular the 'individual' dimension of justice. This argument does not aim to replace those arguments offered by previous proponents for human judgement, but rather to

draw out an implied element common to them, elucidate it, and thereby place them on stronger, more technologically-neutral ground. The argument has an epistemic and a normative component.

3.1 The epistemic case for case-by-case judgement

The epistemic component of the argument from individual justice is based on the claim that whatever logic should be involved in deciding a new case is in principle indeterminate. The consequence is that we are unable to pre-specify how to reason appropriately about new cases without human intervention. To illustrate the argumentation behind this rather strong claim regarding the indeterminacy of new cases, let us reprise some views of the above legal theorists about the epistemology of legal reasoning (remaining, for now, within the narrow analytical jurisprudence paradigm). Despite their differences, Dworkin and Hart both rejected the idea that a set of all-or-nothing rules could ever fully cover all the variety of different cases (Shapiro, 2007). Underlying this position is a kind of scepticism about the structure of rules (Dworkin) or language (Hart). Both subscribe to a form of `particularism' about new individual cases; both held it impossible to specify in advance how previous reasoning might apply in a particular case.

While Dworkin did not attribute his particularism to any particular philosophical view, Hart's position was explicitly influenced by Waismann's critique of the philosophical thesis of logical positivism (Waismann, 1968; Bix, 1991). The logical positivists believed that every meaningful sentence in a language could be treated as a hypothesis which could (in theory) be verified through scientific (i.e. inductive) or logical (i.e. deductive) methods. In so far as law is a linguistic practice, this would imply that it should in theory be possible to specify all the conditions under which a law (like any sentence) applies, or not. Waismann argued that language was not like this, citing the example of the concept 'cat'. We might feel confident in the scope of its application, until a particular cat grows to a gigantic size, or dies and returns to life. Such eventualities would stretch our concept of cat to the extent that the term would collapse. Concepts like this are 'porous'; we don't always know what to say about whether they still apply. For Hart, Waismann's arguments against the logical positivist approach to language applied to law, and made clear law is an *open-textured* domain. A domain having an open texture is distinct from its merely being *vague*, in the sense that vague concepts are *already* vague with respect to existing examples (e.g. the distinction between 'heap' and 'pile'). The open texture of language means that the possibility of future vagueness is always inherent, even for terms which are currently entirely clear but may later need to be adapted in the face of unknown examples.

While Hart's epistemic position draws on a particular dialectic within analytic philosophy of language, similar conclusions are reached by others to similar effect. The basic concern is that a case-by-case assessment is needed because no two cases can be identified as exactly alike ahead of time without examining each, whether that be due to the open-textured meaning of terms, the indeterminacy of conflicting rules, the appropriate weighing of standards or principles in particular cases, or some other

property of the process of assessing new cases (TODO: re-introduce citations for these). These arguments do not imply that no two cases are alike; only that it is not possible to specify the conditions under which any new case could be identified as exactly alike a previous one prior to examining it. Neither do they imply that *no* regularities exist between cases (which would be akin to the thesis of moral particularism (Dancy, 2001)); rather, they reflect that such regularities alone may not exhaust the possibly relevant criteria for any given new case.

In the context of algorithmic decision-making systems, these concerns affect both the *features* considered in cases, and the process of mapping *from* a set of features *to* a classification or decision. There may be features that can't be captured systematically enough, to feature in most cases, or features which are irrelevant in all previous cases, but unexpectedly relevant in a new one. The way that any principles affecting the mapping from features to classifications interact, and are balanced, is similarly un-specifiable in advance. Algorithmic decision-making systems can consider only those features which they have been trained to consider, and only the prediction or classification function they have been specified to use.[5] It is impossible to say ahead of time what all of the exceptional cases might be and why, and although high-dimensional models might allow for more complex functions, and more branches can always be added to a decision tree, ultimately the potential for a novel exception can't be considered in response to each new case, otherwise the process cannot be automated or in any meaningful sense be made independent from human judgement. The feature space that a model considers must also necessarily be constrained for practical reasons; too many features will likely lead to a overfitted model which contains features which are spurious (Calude and Longo, 2017; Graham et al., 1990), and a problem space may be just too complex to capture all cases in a parsimonious and generalisable way.

3.2 The normative case for case-by-case judgement

These epistemic limitations may lead us to a normative position according to which every case must be assessed on its own, even if it bears strong resemblances to previous cases. The features to be taken account of, the process of reasoning from those features, and the incorporation of existing rules, principles, and other criteria, are all open to question and must be considered afresh, even if the present case appears to be exactly the same as some previous case.

Frederik Schauer traces this idea back to Aristotle, who argued in *Nichomachean Ethics* and the *Rhetoric* that justice sometimes requires going against generalisations from previous cases:

[5] The point that rule-based expert systems and machine learning models are not fundamentally different in this respect, and the consequences for the justifiability of automated systems in government decision-making, has also been made by Monika Zalnieriute and co-authors (Zalnieriute et al., 2019)

> *'There are some things about which it is not possible to pronounce rightly in general terms … the raw material of human behaviour is of this kind'.*[6]

In so far as these epistemic limitations could lead us astray in making inferences about new cases, they may lead us to a position according to which each individual case must be assessed afresh as a matter of justice.

The notion that exceptions to generalisations need to be considered in each case, has been referred to by terms such as *individual* justice or *particularised* justice (F. F. Schauer, 2009), and in German jurisprudence as *Einzelfallgerechtigkeit* ('justice in each particular case') (Britz, 2008). A related concept is addressed in philosophical accounts of discrimination, and referred to as the duty to *treat people as individuals,* grounded in respect for individuality and autonomy (Eidelson, 2013; Lippert-Rasmussen, 2011). As U.S. Supreme Court Justice Kennedy argued, discrimination (in this case, on the basis of race) is wrong because it 'is not consistent with respect based on the unique personality each of us possesses'.[7] Note that the notion of treating people as individuals is not shared by all accounts of discrimination, some of which *do* allow generalisation as long as it avoids *protected* categories (F. F. Schauer, 2009).

Individual justice implies that even if the next case is apparently identical to a previous case, it might need to be treated differently (whether because of the unpredictability of the application of rules, the inability to generalise about human behaviour, or the uniqueness of each of our personalities). This is something that, by definition, an algorithmic model cannot do. Given the same set of inputs (i.e. features of the case), algorithmic systems will deterministically and consistently produce a single output. Of course, as new training data becomes available, the model may be updated and thereafter give contrary outputs, but this (again, by definition) cannot happen prior to each new case being processed.

Such systems are in this sense *incapable* of the case-by-case judgement required by individual justice, since they do not re-consider which features to include and the logic to use from scratch in every case. The point is well articulated in this quote from a research participant who took part in a study about algorithmic decisions: 'it's unfair to make the decision by just comparing him to other people and then looking at the statistics; he isn't the same person' (Binns et al., 2018). The inability to treat people as individuals - even if they appear identical to previous cases – is thus an endemic feature of machine learning systems, which are essentially based on generalisation (Binns, 2017).

To conclude this account of individual justice and its relevance to algorithmic decision-making, we can summarise the arguments thus far. Previous accounts of the importance of human judgement over algorithmic decision-making have appealed to the distinction made by legal philosophers between rules and discretion, and argued that particular algorithmic systems have only emulated the former, and that

---

[6] Aristotle, *Nichomachean Ethics*, (Aristotle et al., 1976, 1137a-b)
[7] In *Rice v. Cayetano , 528 U.S. 495, 517 (2000).*

the latter are the preserve of human decision-makers exercising discretion. Even if they are sound, such arguments leave open the possibility that some new approach to artificial intelligence might successfully capture whatever is supposedly involved in discretionary decision-making. The argument above presents a more specific and a priori defence of human discretion. Instead of attempting to assess particular models of legal reasoning and the deficiencies of their artificial pretenders, it grounds the case for human judgement in terms of an appeal to individual justice, which algorithmic systems cannot achieve by definition. This does not aim to replace but rather complement previous arguments. While this is in some senses more fundamental than an argument based on the traditional distinction between rules and discretion, it is not without its complications and challenges, to which the remaining sections of this article are devoted.

4. Conflicts between individual justice and other justice dimensions

While appealing to individual justice may give defenders of human judgement an additional line of argument, it also raises questions about how individual justice might cohere or conflict with other dimensions of justice, and how these different dimensions of justice might be affected by particular combinations of algorithmic decision-making systems and the discretion-exercising humans-in-their-loops.

4.1 Individual justice vs consistency

Even if individual justice is important, it is not absolute. In addition to merely being inefficient due to the necessity of human review, individual justice may conflict with other elements of justice too. In particular, it threatens what we might call 'consistency'. This is the notion that similar cases should be treated similarly, and unalike cases differently in their unalikeness; this maxim, also derived from Aristotle, leaves open the question of which aspects of cases are relevant to determining their 'similarity' (Barnes, 1984 V.3. 1131a10-b15). But in so far as algorithmic systems can account for the relevant similarities and differences, they satisfy Aristotle's maxim by default, since they always give the same output when given the same input (at least, until a model is re-trained with fresh data). In this respect, 'forms of automation hold out promise for legal certainty' (Le Sueur, 2015), a desirable property of any domain of legally bound decision-making. By contrast, human decision makers pursuing individual justice might lead to inconsistencies between cases, where two people who are alike in relevant respects are treated differently. Such inconsistency could simply be a result of the unpredictability inherent in human assessment on a case-by-case basis.

This tension, between individual justice and consistency, is already recognised and acknowledged in debates on jurisprudence in a non-algorithmic context. Neither value is an absolute, so a balance needs to be struck. As Sanne Taekema argues, 'the equal application of norms demanded by formal justice'

must be balanced against the 'equitable solution of the individual case' (Roughan and Halpin, 2017; Taekema, 2016). Similarly, Frederick Schauer contrasts the trade-offs involved thus:

> *'... Insofar as rules also bring the advantages of certainty, predictability, settlement, and stability for stability's sake, treating the rules as defeasible comes at the sacrifice of each of these values, even though of course it brings the potential advantages of fairness, equity, and, in theory, reaching the correct result in every instance.'* (F. Schauer, 2009)

In contexts where decision-making will be based on a combination of algorithmic and human elements, the result of this balancing act may help determine the right balance between the consistency-promoting algorithm and the individual-justice-serving human(s).

4.2 Individual justice vs algorithmic 'fairness'

However, consistency and individual justice are just two desirable criteria in a decision-making process, so one cannot necessarily determine the right 'mix' of human and algorithmic control by considering those two alone. Another element of justice concerns discrimination on the basis of protected characteristics: when comparing the outcomes between groups disaggregated by e.g. gender, race, sexuality or religious belief, one group might be disproportionately harmed or benefitted. The relationships between these three elements – consistency, individual justice, and non-discrimination – are non-obvious, which complicates further the problem of finding the right balance between human and algorithmic contributions to decision-making.

On the one hand, emphasising individual justice could come at the expense of non-discrimination. When decision-makers are given greater discretion, contextual factors can be dredged by a motivated decision-maker (whether consciously or not) to uncover an exculpatory or inculpatory factor, resulting in differential treatment. Such cases are individual *in*justices, and limiting discretion in favour of rules has historically been the proposed solution (Lacey, 1992). In so far as algorithmic systems replace the human decision-maker's ability to be inconsistent, discriminatory, or otherwise contrary to general justice, they may be seen as beneficial.[8] This line of thought would suggest that non-discrimination would be best achieved by limiting human discretion, and therefore the human-algorithm mix should involve a greater ratio of the algorithmic component.

But conversely, the fact that algorithmic systems avoid the potential inconsistencies of human adjudicators does not mean they will thereby avoid discrimination. Mere consistency is not enough to avoid disparate impacts where the factors used to assess the case are themselves a source of discrimination. So there is also a tension between generalisation (and hence, consistency) and anti-discrimination, as explored in work by Schauer and others (Britz, 2008; F. F. Schauer, 2009). This plays

[8] As Angele Christine argues, these systems could 'help limit prosecutorial discretion' in contexts where it is arguably abused, as in the U.S. criminal justice system where 'the vast majority of cases are settled by plea bargain rather than a trial' (Christin 2018).

out in the algorithmic setting, when variables used in machine learning models to make predictions about or classify people – such as qualifications, work history, or criminal record – are already shaped by structural discrimination (Barocas and Selbst, 2016; Custers et al., 2012; Gandy, 2016). Such variables may be distributed differently by e.g. gender or ethnicity in ways that reflect unjust structural discrimination, but a strict commitment to consistency would suggest neglecting these differences. These considerations have been the focus of recent research on 'fair' and 'discrimination-aware' machine learning (Dwork et al., 2012; Pedreshi et al., 2008), although they are pre-dated by older examples of testing and hiring systems which were found to have disparate impacts (Hutchinson and Mitchell, 2019; Lowry and Macpherson, 1988).

So, neither human nor algorithmic systems are inherently non-discriminatory; and neither seems to be overall better-positioned to encapsulate the three different dimensions of justice discussed above. While algorithmic systems arguably serve one dimension of justice very well (consistency), they fail to respect individual justice and are no guarantee of non-discrimination. Similarly, even though human judgement is necessary for the individual dimension of justice, it also risks conflicting with both consistency and non-discrimination.

How can these possibly conflicting dimensions of justice be resolved? Typically, the balancing between these dimensions has been undertaken by human decision-makers, who have been expected to pay due attention to the specifics of an individual case, whilst also ensuring consistency between decisions, as well as aiming to avoid disparate outcomes between protected groups. In cases where the application of existing law is indeterminate, discretionary decision making allows different dimensions to be weighed (Schauer, 1987), and for broader political, economic and moral considerations to come into play.[9] Similarly, judges are not free to come out however they want even when given discretionary powers; when they aren't constrained by particular *rules*, judges are still constrained by the broader context set by law, and can also be "questioned and rebuked for discriminatory behaviour" (Pasquale and Cashwell, 2018, p3). So as well as being the means to obtain individual justice, human judgement is also often the means by which conflicts *between* individual justice, consistency and other goals like antidiscrimination are typically addressed and given due weight.

But perhaps humans are not the only means of mediation between dimensions of justice; perhaps algorithmic systems could be designed to factor in more dimensions of justice than just consistency. In this way, while human discretion may need to be retained for purposes of individual justice, it may not need to play the exclusive role of *mediating between* different dimensions of justice, which could be partly taken on by algorithmic systems.

[9] Although the appropriate role(s) of the latter are subject to debate and critique between e.g. the law and economics approach and its detractors (Ash et al., 2018; Malloy, 1989)

For instance, they could be designed to incorporate anti-discrimination requirements *within the statistical model*, whilst maintaining consistency as much as possible. Indeed, this goal is pursued in much of the recent 'fairness in machine learning' (F-ML) research (Dwork et al., 2012; Hardt et al., 2016; Pedreshi et al., 2008; Ruggieri et al., 2010). Various methods propose that in order to ensure algorithmic systems do not generate 'unfair' results (where fairness may be formalised in various contestable ways), certain features may need to be treated differently by in the creation or operation of the model depending on the protected characteristics of the decision subject. The majority of approaches to quantifying how unfair a model is are based on 'group fairness'; these are based on comparisons between the performance of a model with respect to different protected groups. For instance, one type of group fairness measure defines a model as fair if any errors it makes are equally distributed between people from different protected groups. Methods are proposed to constrain a machine learning model to meet such fairness measures. The F-ML literature has even identified tensions between different mathematical measures of fairness, and attempted to find ways to strike an optimal balance between them.[10] In this way, such approaches attempt to gain the benefits of consistency that algorithmic systems bring (in addition to efficiency and accuracy), whilst also attempting to incorporate other important dimensions of justice.

If successful, increasingly sophisticated algorithmic approaches which incorporate multiple dimensions of justice could arguably reduce the need for human decision makers to play the role of mediator *between* those justice dimensions. It is beyond the scope of this article to address whether such technical approaches to algorithmic fairness are feasible, desirable, or genuinely reflective of the nature of justice at stake in real socio-technical systems (see, inter alia, (Binns, 2017; Hoffmann, 2019; Powles, 2018)). But even if they were, human decision-makers would still be necessary to assess individual justice and make their assessment commensurable with these other dimensions of justice. Ultimately, while algorithms may bring new means of formalising and mediating between e.g. consistency and discrimination, and may limit injustices deriving from biased human judgement, they cannot hope to provide the complete package.

---

[10] It is worth briefly addressing the tension identified in this literature between so called 'individual fairness' and 'group fairness' measures (Dwork 2012). Despite its lexical similarity, individual fairness bears little relationship to *individual justice* as I have defined it above. Individual fairness means giving 'similar predictions to similar individuals' (Dwork et al., 2012; Kusner et al., 2017). Two individuals are alike if their combinations of task-relevant attributes are nearby each other with respect to a geometric representation of their task-relevant features. A major limitation of individual fairness measures is the question of how to determine a task-relevant similarity metric, which arguably defers many of the normatively contestable questions to a policy-maker. Rather than reflecting individual justice, this is much closer to what I have called *consistency*, and what Schauer calls 'general' justice. As we have seen, this is often *in tension* with individual *justice*. Individual fairness measures involve pre-specifying the subset of features that define individual similarity in the context of the task. By contrast, the kind of individual justice addressed above starts from the assumption that even if the facts at hand suggest two cases are exactly alike, a fresh consideration is necessary, because there's a chance that different features, or alternative reasoning, might lead to a different decision.

5. Challenges for implementing individual justice in practice

In addition to the difficulties associated with integrating multiple dimensions of justice into human / algorithmic system, two further challenges are raised by this account. The first challenge concerns how individual justice can be accommodated alongside other dimensions of justice in the socio-technical contexts in which humans-in-the-loop are situated. The second concerns the potential inequities in individual justice that might result from an uneven application of human judgement in algorithmic settings.

5.1 Screen-level bureaucrats may have their own conceptions of justice

The argument above suggests that we need human decision makers in order to preserve individual justice. The implied model of collaboration between human and algorithmic elements is one in which human reviewers attend to the individual circumstances of the case; meanwhile, algorithms take care of inducing patterns across multiple cases to predict outputs (and perhaps meeting other formalised constraints, such as the algorithmic fairness measures mentioned above). This model of human involvement is arguably implied in regulation on automated decision making (e.g. Article 22 of the GDPR), where humans are required to either proactively interpret and challenge algorithmic outputs before making a decision, or to be on hand ready to intervene reactively when automated decisions are challenged.

But can this model survive contact with the messy realities of algorithm deployment on the ground? Can organisations using algorithms neatly apportion out the 'individual justice' element of a decision-making process to a human, whilst preserving any benefits that algorithmic systems might bring? What complications might arise when these abstract arrangements are designed with particular affordances and put into socio-technical contexts? The human usually imagined in the jurisprudential debate is a judge, but in practice the human-in-the-loop could be a police officer, welfare claims processor, human resources officer, or bank clerk; each is situated in widely varying social and structural conditions with different powers, pressures, motivations, and agendas affecting how they exercise judgement in ways which are unlikely to conform to expectations of the algorithm's designers.

In reality, human decision makers responsible for overseeing, interpreting and reviewing algorithmic systems are unlikely to confine themselves to focusing on the specifics of the case, or to do so in the ways that the designers of a system expect them to. The humans-in-the-loop will have their own ideas about the important matters of justice in their domain, as well as their own interpretations of the demands of consistency and non-discrimination. The use and abuse of discretion by such actors is a subject of study by sociologists, critical legal scholars, and others, who focus on non-judicial decision-makers in both public administration and (parts of) the private sector (Lipsky, 2010; Prottas, 1978). Such frontline workers effectively function as 'street-level bureaucrats' (Lipsky, 2010). These individuals use their discretion in line with their own convictions and visions of justice, often against the

prescriptions of management (Prottas, 1978). Various socio-legal scholarship charts how discretionary decision-making by frontline workers, while sometimes wielded in harmful ways, can equally serve transformative and progressive ends in relation to the needs of powerless groups (Pratt and Sossin, 2009; Sossin, 1994). For instance, Marisa Kelly analysed stories told by Californian schoolteachers which demonstrate how they use their discretionary powers to redress socio-economic inequalities between students in line with their own commitments to distributive justice norms (Kelly, 1994).

With the introduction of algorithmic systems, there is a new class of 'screen-level bureaucrats' (Bovens and Zouridis, 2002). Recent work on the use of machine learning in the public sector suggests that such systems are often significantly shaped by (and sometimes over-ridden in light of) the values of those designing, operating and responding to them on the ground in areas like taxation, law enforcement, and child welfare (Veale et al., 2018). In the same way that frontline decision-makers have sometimes fought against dictats from managers and disagreed with the substance or implementation of justice imposed from above, they may well do so in algorithmic contexts. In many cases they will be driven by their own commitments to broader political goals (whether they be distributive justice or, indeed, punishment of profligacy). A system which purports to be non-discriminatory according to one formal measurement of fairness may not seem fair to the human in the loop responsible for assessing individual cases.

This provides an added layer of complexity for the model of human intervention in algorithmic decision-making suggested above. In particular, it suggests that involving human judgement for the sole purpose of satisfying *individual* justice, while encoding other dimensions of justice such as non-discrimination into the algorithm, will run in to trouble because the human reviewers will bring other justice commitments to bear in their judgements. This makes it difficult to see how one could enforce system-level justice constraints by constraining the algorithm alone. Ideally, a human reviewer's conceptions of justice will be aligned with those encoded at the algorithmic level. But in reality, there is likely to be tension between the two.

This could be due to substantive differences between the individual and the institutionally-defined, algorithmically-encoded conceptions of justice (as in the misaligned commitments of street-level bureaucrats and their managers (Prottas, 1978)). Historically, where street-level bureaucrats have been regarded as wielding discretion in ways that are misaligned with managerial or institutional policies (for better or worse), the answer has been to reduce discretion in favour of stricter rules; this constraining role may increasingly be played by algorithmic systems with formalisations of justice encoded into them. Alternatively, even if individual and institutional / algorithmic justice commitments are aligned, they may lack the ability to effectively collaborate due to mismatches in information and approach to individual cases.

5.2 Individual justice may be unequally distributed

Setting aside the complications arising from the human in the loop bringing their own possibly conflicting justice-related commitments to the decision-making process, there are a set of problems that may arise as a result of design choices about *when* human judgement is inserted into a decision process, and how intensive it should be in different cases. These issues arise if we grant that, within a single decision-making context, individual justice is sufficiently important to merit human review in some cases but not others.

It would of course be convenient if it were possible to automatically determine in advance which cases are which. Algorithmic systems can be designed to defer to human judgements according to certain triggers, e.g. when the confidence value associated with a model's classification is below a certain threshold, where the effect of the output class is more likely to be particularly significant for the decision subject, or where deference to a human is predicted to increase some statistical fairness measure (see e.g. (Madras et al., 2018)).

If the sifting process were to be undertaken by humans, then it would be redundant; but if it is undertaken by an algorithm, then individual justice concerns may arise again at a secondary level. This could be seen as a version of the common bootstrapping problem in legal automation (Hildebrandt, 2013; Leenes, 2003); namely, that determining which cases require more or less human judgement, may itself require human judgement. This was one of the sticking points in the 1980's debates about legal automation between Susskind and Leith (Leith, 1987; Susskind, 1987). The contention concerned whether there were any areas where the cases were 'clear enough', such that rules for them could be encoded in an expert system (as Susskind hoped), or whether even in those cases, rules were dogged by a penumbra of doubt (as Leith argued), such that you can never know when a difficult case might crop up.

If individual justice is required only in *some* cases and not others, but the identification of those cases itself requires human judgement, then this undermines the extent to which partial integration of humans into algorithmic loops can meaningfully serve individual justice. While these triggers might capture some of the cases in which greater consideration of individual justice is required, they would do so imperfectly, and any imperfections will have repercussions for one's chances down the line. Unless the system is perfect, among those individuals whose case is automatically classified as *not* requiring individual justice, at least some will be incorrect, thereby unjustly denying them individual justice. If we take a stringent approach to individual justice, we might hold that the determination of whether individual justice is important in any given case should itself be subject to individual justice. This is separate to the question of whether the right decision was made in any particular case. This implies that *any* kind of algorithmic selection of cases for human review simply does not respect its essence, in so far

as it implies that *every* case deserves individual assessment, no matter how similar it may be to a pre-existing case.

Another way to allocate human review is on the basis of contestation by those who are subject to an automated decision. This is the approach envisioned under the safeguard included in Article 22(3) for data subjects to contest and request human intervention in a solely automated decision with legal or similarly significant effects. While an important safeguard, this may undermine the equitable application of individual justice in two ways. First, it puts the onus on decision subjects to mount a successful challenge, which may require resources and privileges that are not distributed equally among the population, compounding disadvantage by failing to give individual justice to all those to whom it is due. Second, it means that decision-makers are likely to only review false positives (where people have been incorrectly been denied a benefit), and ignore false negatives (where people have incorrectly been granted a benefit), because the latter have no incentive to challenge a positive decision. Relying on contestation to trigger human review will therefore likely overlook cases of undue lenience.

Similar considerations may apply to cases in which an algorithmic decision is only applied automatically if it is a positive outcome for the individual, and negative outcomes must be reviewed by humans (e.g., releasing defendants predicted 'low risk' on bail, and holding others in order to wait for a judge to review their case). This may sound reassuring. But in addition to the problem that those who are not automatically approved potentially face costs associated with waiting for a human decision (e.g. more time spent waiting to be released from jail), it also could mean an uneven distribution of individual justice. Reserving human judgement for negative decisions neglects that one's chances of obtaining a deserved positive decision will depend on whether one makes it through the initial automated prioritisation process. The purpose of human review is to overturn any false negatives, but unless this process is perfect, then arguably harm is being done to those who fail to obtain a true positive at the algorithmic sorting stage.[11] So while it might appear that as long as human discretion is incorporated before any negative decisions are made, such a process could still be unjust. Such inequalities be especially wrong if those who do not benefit automatically are already otherwise disadvantaged or marginalised.

[11] This is especially problematic in decision-making contexts where the decision maker has limited resources or a limited risk appetite, because the threshold for automatically receiving a benefit may depend on the total proportion of beneficial decisions. Every beneficial automated decision that would have been overturned by a human reviewer, will have contributed to an increase in the threshold, thereby excluding decision-subjects who are just below the threshold and would have otherwise automatically qualified.

6. Conclusions

It is reassuring to assert, as Baroness Hollinghurst did in front of the UK parliament, that decisions which have traditionally required discretion rather than rules will continue to do so. But it is another matter to determine when discretion is required, and when it is meaningfully applied in the right way. Legal philosophy and jurisprudence provide ample theoretical debate about when discretion is desirable rather than rules, which can and have been adapted to form analogous arguments against algorithmic alternatives. This article has taken a narrow focus on one particular line of argument; that human judgement remains necessary for individual justice. Individual justice involves the case in question being judged on its own, with a fresh consideration of the relevant factors and the reasoning involved, no matter how similar it may appear to previous cases. This is just one of several arguments explicating the value of human judgement (with others focusing, amongst other things, on maintaining human liability and responsibility (Bryson et al., 2017; Wagner, 2019), or the need to preserve the rule of law as an argumentative and text-driven system (Hildebrandt, 2018; Zalnieriute et al., 2019)).

Algorithmic systems which are intended to only 'augment and enrich' human decisions, or sift cases for human review, appear to combine the benefits of algorithmic decisions, such as consistency, with human discretion which serves individual justice. But as argued above, it is not obvious how these different elements of justice, and the humans and machines which are supposed to serve them, can be combined without undermining each other. Individual justice will often conflict with algorithm-driven consistency and fairness, but conversely algorithmic systems are incapable of respecting individual justice. To complicate matters further, the screen-level bureaucrats dealing with algorithmic systems on the ground will exercise discretion according to their own commitments in ways that may be at odds with the goals of the organisation deploying them. The potential for unequal application of individual justice raises additional challenges for how we assess the role, value, and scope of discretion. Finally, having articulated the notion of individual justice, it remains to be seen how important the concept is in specific contexts, and what the social expectations might be of those who will enjoy or suffer its consequences. These issues will have to be worked out as algorithmic systems are deployed in context; if individual justice is worth protecting, we cannot assume that it will be secured by simply putting a human in the algorithmic loop.

References:

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V., 2016. Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. PeerJ Comput. Sci. 2, e93.

Allen, L.E., 1956. Symbolic logic: A razor-edged tool for drafting and interpreting legal documents. Yale LJ 66, 833.

Applin, S.A., Fischer, M.D., 2015. New technologies and mixed-use convergence: How humans and algorithms are adapting to each other, in: 2015 IEEE International Symposium on Technology and Society (ISTAS). IEEE, pp. 1–6.

Arendt, H., 1973. The origins of totalitarianism. Houghton Mifflin Harcourt.

Ash, E., Chen, D.L., Naidu, S., 2018. Ideas have consequences: The impact of law and economics on american justice. working paper.

Ashley, K.D., Walker, V.R., 2013. From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study., in: JURIX. pp. 29–38.

Bankowski, Z., White, I., Hahn, U. (Eds.), 1995. Informatics and the Foundations of Legal Reasoning, Law and Philosophy Library. Springer Netherlands.

Barnes, J., 1984. Complete works of Aristotle, volume 1: The revised Oxford translation. Princeton University Press.

Barocas, S., Selbst, A.D., 2016. Big data's disparate impact. Calif Rev 104, 671.

Berners-Lee, T., Fischetti, M., 2001. Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor. DIANE Publishing Company.

Binns, R., 2017. Fairness in Machine Learning: Lessons from Political Philosophy. Proc. Mach. Learn. Res. 81, 1–11.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., Shadbolt, N., 2018. "It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, p. 377.

Bix, B., 1991. HLA Hart and the "open texture" of language. Law Philos. 10, 51–72.

Bovens, M., Zouridis, S., 2002. From street-level to system-level bureaucracies: How information and communication technology is transforming administrative discretion and constitutional control. Public Adm. Rev. 62, 174–184.

Britz, G., 2008. Einzelfallgerechtigkeit versus Generalisierung: verfassungsrechtliche Grenzen statistischer Diskriminierung. Mohr Siebeck.

Bryson, J.J., Diamantis, M.E., Grant, T.D., 2017. Of, for, and by the people: the legal lacuna of synthetic persons. Artif. Intell. Law 25, 273–291. https://doi.org/10.1007/s10506-017-9214-9

Buchanan, B.G., Headrick, T.E., 1970. Some speculation about artificial intelligence and legal reasoning. Stan Rev 23, 40.

Calude, C.S., Longo, G., 2017. The Deluge of Spurious Correlations in Big Data. Found. Sci. 22, 595–612. https://doi.org/10.1007/s10699-016-9489-4

Citron, D.K., 2007. Technological due process. Wash UL Rev 85, 1249.

Cobbe, J., 2018. Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making.

Custers, B., Calders, T., Schermer, B., Zarsky, T., 2012. Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases. Springer Science & Business Media.

Dancy, J., 2001. Moral particularism.

Dautenhahn, K., 1998. The art of designing socially intelligent agents: Science, fiction, and the human in the loop. Appl. Artif. Intell. 12, 573–617.

Davis, K.C., 1969. Discretionary justice: A preliminary inquiry. LSU Press.

de Winter, J.C., Dodou, D., 2014. Why the Fitts list has persisted throughout the history of function allocation. Cogn. Technol. Work 16, 1–11.

Dicey, A.V., 2013. The Law of the Constitution. OUP Oxford.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R., 2012. Fairness through awareness, in: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference. ACM, pp. 214–226.

Edwards, L., 1995. Modelling law using a feminist theoretical perspective. Inf. Commun. Technol. Law 4, 95–110.

Eidelson, B., 2013. Treating People as Individuals, in: Hellman, D., Moreau, S. (Eds.), Philosophical Foundations of Discrimination Law. Oxford University Press, pp. 203–227. https://doi.org/10.1093/acprof:oso/9780199664313.003.0011

Elish, M.C., 2016. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction (We Robot 2016).

Flew, A., 1953. Essays on logic and language.

Gandy, O.H., 2016. Coming to terms with chance: Engaging rational discrimination and cumulative disadvantage. Routledge.

Graham, Ronald L., Graham, Ronald Lewis, Rothschild, B.L., Spencer, J.H., 1990. Ramsey theory. John Wiley & Sons.

Grimmelmann, J., 2004. Regulation by software. Yale LJ 114, 1719.

Hachey, B., Grover, C., 2005. Automatic legal text summarisation: experiments with summary structuring, in: Proceedings of the 10th International Conference on Artificial Intelligence and Law. ACM, pp. 75–84.

Hardt, M., Price, E., Srebro, N., 2016. Equality of opportunity in supervised learning, in: Advances in Neural Information Processing Systems. pp. 3315–3323.

Hayek, F.A., 1979. The political order of a free people.

Hildebrandt, M., 2018. Algorithmic regulation and the rule of law. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 376, 20170355.

Hildebrandt, M., 2013. From Galatea 2.2 to Watson–And Back?, in: Human Law and Computer Law: Comparative Perspectives. Springer, pp. 23–45.

Hoffmann, A.L., 2019. Where fairness fails: On data, algorithms, and the limits of antidiscrimination discourse. Rev. Inf. Commun. Soc.

Hutchinson, B., Mitchell, M., 2019. 50 Years of Test (Un)fairness: Lessons for Machine Learning. Proc. Conf. Fairness Account. Transpar. - FAT 19 49–58. https://doi.org/10.1145/3287560.3287600

Kelly, M., 1994. Theories of justice and street-level discretion. J. Public Adm. Res. Theory 4, 119–140.

Kusner, M.J., Loftus, J., Russell, C., Silva, R., 2017. Counterfactual fairness, in: Advances in Neural Information Processing Systems. pp. 4066–4076.

Lacey, N., 1992. The jurisprudence of discretion: escaping the legal paradigm. Uses Discret. 361–88.

Le Sueur, A., 2015. Robot government: automated decision-making and its implications for parliament.

Leenes, R., 2003. Abort or Retry—A Role for Legal Knowledge Based Systems in Electronic Service Delivery?, in: IFIP International Working Conference on Knowledge Management in Electronic Government. Springer, pp. 60–69.

Leith, P., 2016. The rise and fall of the legal expert system. Int. Rev. Law Comput. Technol. 30, 94–106.

Leith, P., 1987. The Emperor's New Expert System. Mod. Law Rev. 50, 128–132.

Leith, P., 1986. Fundamental errors in legal logic programming. Comput. J. 29, 545–552.

Lippert-Rasmussen, K., 2011. "We are all Different": Statistical Discrimination and the Right to be Treated as an Individual. J. Ethics 15, 47–59.

Lippi, M., Torroni, P., 2016. Argumentation mining: State of the art and emerging trends. ACM Trans. Internet Technol. TOIT 16, 10.

Lipsky, M., 2010. Street-level bureaucracy: Dilemmas of the individual in public service. Russell Sage Foundation.

Lowry, S., Macpherson, G., 1988. A blot on the profession. Br. Med. J. Clin. Res. Ed 296, 657.

Lynn Jr, L.E., 1998. A critical analysis of the new public management. Int. Public Manag. J. 1, 107–123.

Madras, D., Pitassi, T., Zemel, R., 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer, in: Advances in Neural Information Processing Systems. pp. 6147–6157.

Malloy, R.P., 1989. Is Law and Economics Moral - Humanistic Economics and a Classical Liberal Critique of Posner's Economic Analysis Debate: Is Law and Economics Moral. Valpso. Univ. Law Rev. 24, 147–162.

McCarthy, J., 1980. Circumscription—a form of non-monotonic reasoning. Artif. Intell. 13, 27–39.

McCarty, L.T., 1976. Reflections on TAXMAN: An experiment in artificial intelligence and legal reasoning. Harv Rev 90, 837.

MccLachlin, T.H.M.J.B., 1992. Mess and Discretion in the Governance of Canada. Sask. Law Rev. 1, 168.

Morgan, D.F., 1987. Varieties of administrative abuse: Some reflections on ethics and discretion. Adm. Soc. 19, 267–284.

Noto La Diega, G., 2018. Against the Dehumanisation of Decision-Making–Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information.

Oswald, M., 2018. Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 376, 20170359.

Pasquale, F., Cashwell, G., 2018. Prediction, persuasion, and the jurisprudence of behaviourism. Univ. Tor. Law J. 68, 63–81.

Pasquale, F.A., 2018. A Rule of Persons, Not Machines: The Limits of Legal Automation (SSRN Scholarly Paper No. ID 3135549). Social Science Research Network, Rochester, NY.

Pedreshi, D., Ruggieri, S., Turini, F., 2008. Discrimination-aware data mining, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 560–568.

Powles, J., 2018. The Seductive Diversion of 'Solving' Bias in Artificial Intelligence [WWW Document]. Medium. URL https://medium.com/s/story/the-seductive-diversion-of-solving-bias-in-artificial-intelligence-890df5e5ef53 (accessed 3.29.19).

Pratt, A., Sossin, L., 2009. A brief introduction of the puzzle of discretion. Can. J. Law Soc. Rev. Can. Droit Société 24, 301–312.

Prottas, J.M., 1978. The power of the street-level bureaucrat in public service bureaucracies. Urban Aff. Q. 13, 285–312.

Rahwan, I., 2018. Society-in-the-loop: programming the algorithmic social contract. Ethics Inf. Technol. 20, 5–14.

Rissland, E.L., Ashley, K.D., 1987. A case-based system for trade secrets law, in: Proceedings of the 1st International Conference on Artificial Intelligence and Law. ACM, pp. 60–66.

Roughan, N., Halpin, A., 2017. In Pursuit of Pluralist Jurisprudence. Cambridge University Press.

Ruggieri, S., Pedreschi, D., Turini, F., 2010. Data mining for discrimination discovery. ACM Trans. Knowl. Discov. Data TKDD 4, 9.

Russell, S.J., Norvig, P., 2016. Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited,.

Santoni de Sio, F., Van den Hoven, J., 2018. Meaningful human control over autonomous systems: a philosophical account. Front. Robot. AI 5, 15.

Schauer, F., 2009. Is Defeasibility An Essential Property of Law? Univ. Va. Leg. Work. Pap. Ser. 130.

Schauer, F., 1987. Judging in a Corner of the Law. Cal Rev 61, 1717.

Schauer, F.F., 2009. Profiles, probabilities, and stereotypes. Harvard University Press.

Shadbolt, N., O'Hara, K., De Roure, D., Hall, W., 2019. The Theory and Practice of Social Machines. Springer Nature Switzerland AG.

Shapiro, S.J., 2007. The Hart-Dworkin debate: A short guide for the perplexed. Available SSRN 968657.

Sossin, L., 1994. Redistributing democracy: An inquiry into authority, discretion and the possibility of engagement in the welfare state. Ott. Rev 26, 1.

Susskind, R.E., 1987. Expert systems in law: a jurisprudential inquiry. Clarendon.

Taekema, S., 2016. The Many Uses of Law. Interactional Law as a Bridge between Instrumentalism and Law's Values.

Thomson, J.A.K., 1976. The ethics of Aristotle: the Nicomachean ethics. Penguin, Harmondsworth; New York.

Veale, M., Van Kleek, M., Binns, R., 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM, p. 440.

Wagner, B., 2019. Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. Policy Internet.

Wagner, B., 2016. Study on the human rights dimensions of automated data processing techniques (in particular algorithms) and possible regulatory implications.

Waismann, F., 1968. Verifiability, in: How I See Philosophy. Springer, pp. 39–66.

Weber, M., 2015. Bureaucracy, in: Working in America. Routledge, pp. 29–34.

Weiner, N., 1950. The human use of human beings. Cybern. Soc. Boston Houghton Mifflin Co 71.

Wikipedia contributors, 2019. Computer says no. Wikipedia Free Encycl.

Yeung, K., 2018. Algorithmic regulation: a critical interrogation. Regul. Gov. 12, 505–523.

Zalnieriute, M., Moses, L.B., Williams, G., 2019. The Rule of Law and Automation of Government Decision-Making. Mod. Law Rev. 0. https://doi.org/10.1111/1468-2230.12412