JOURNAL OF
CROSS-DISCIPLINARY RESEARCH
IN COMPUTATIONAL LAW [CRCL]

Analogies and disanalogies between machine-driven and human-driven legal judgement

Reuben Binns *

Abstract

Are there certain desirable properties from text-driven law, which have parallels in code and data-driven law? As a preliminary exercise, this article explores a range of analogies and disanalogies between text-driven normativity (understood as patterns of acting) and its code and data-driven counterparts. Ultimately, the conclusion is that the analogies are weaker than the disanalogies. But the hope is that, in the process of drawing them, we learn something more about the comparison between text and code-/data-driven normativities and the (im?)possibility of code-/data-driven law.

Keywords: Algorithmic decision-making, normativity, rule of law, discretion, feedback loops

Replier: Emily M. Bender, Professor of Linguistics, University of Washington
ebender@uw.edu

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

*Associate Professor, University of Oxford reuben.binns@cs.ox.ac.uk

Introduction

Statistical models derived from machine learning (ML) techniques are increasingly involved in the exercise of decision-making power by state and private actors, prompting concerns about justice and the rule of law. In some scenarios, these systems may be applied to help guide human decision-making, and in others, to automatically make decisions relating to people. In the most contentious cases, the decision being predicted, pre-empted or influenced by an algorithm may be a judgement in a court of law, for instance, a judge deciding upon a criminal sentence. In such cases, decisions have both legal effects on the parties involved, as well as potentially making (case-)law in so far as the decision serves as an authority for future decisions. In such cases, there may be concerns that the algorithmic system will not or cannot satisfy certain properties inherent in human decision-making which are necessary for it to be both lawful and constitutive of law. Similar concerns also arise with respect to decision-making by public servants, whose decisions are not (and do not make) law, but do need to be lawful and may have legal effect.[1] The same may also apply to decision-making in the context of enforcing the law by private actors where the law has horizontal effect.

Sometimes, these concerns appear to turn on claims about whether or not algorithms can replicate human legal reasoning. What it would mean for an algorithm to replicate legal reasoning and how that could be measured, are themselves ambiguous. A simple (but altogether too simple) definition of successful replication would consist in an algorithm reproducing human decisions to a sufficiently accurate degree. On this interpretation, what constitutes legal decision-making is reducible to what computer scientists would understand as a decision problem, where the problem is defined by a fixed range of inputs and outputs, and is solved by finding an algorithm which successfully maps inputs to outputs.

But successfully 'solving' the problem of legal decision-making using algorithms in this way may not be sufficient, because there may be other qualities which we consider necessary for lawfulness. In other words, it may not be enough just to get the right answer, but rather it may be necessary to get the right answer in the right way. And indeed, getting a single 'right' answer may not even be a necessary condition; because depending on one's view of law, there may not be a single right answer.[2] In which case, what matters is not getting to the *right* decision in *any* way, but getting to *any* decision *in the right way*. This reflects a procedural conception of justice in legal thought.

What this *right way* consists in has traditionally been closely aligned with the normativities of text-driven law. In this case, the concept of a normativity or a norm, is to be understood not in *moral* terms, but as a habit or pattern of acting that is 'neither necessarily the result of a conscious decision nor mere regularity of behaviour'.[3] Text-driven law generates certain normativities (or norms), among which are acts of textual interpretation. The possibility to scrutinise the written word opens up rich possibilities in relation to argumentation and justification. A single word may contain multiple equally valid meanings which can be argued over. Recording laws and legal reasoning through text opens up a trail of words which can serve as fresh interpretative ingredients for future legal decision-making. On this account, such acts of interpretation are partly constitutive of what it means to do legal reasoning in *right way*, and therefore lawfulness is wedded to text as a medium. By contrast, data-driven law does not appear to allow for the same forms of scrutiny and interpretation.

To summarise, concerns about algorithmic systems have sometimes been interpreted in terms of their inability to replicate human decision-making. In the context of legal decision-making, such 'replication' cannot be understood purely in terms of comparing the accuracy of algorithmic outputs against some human-produced ground truth. Concerns about machine-judgement differing from hu-

1 'Administration is not law but it does fall under the rule of law (legality principle).' See Mireille Hildebrandt, 'From Galatea 2.2 to Watson – And Back?' in Mireille Hildebrandt and Jeanne Gaakeer (eds), *Human Law and Computer Law: Comparative Perspectives* (Springer 2013).

2 Ronald Dworkin, 'No right answer?' (1978) 53(1) New York University Law Review 1; Douglas Litowitz, 'Dworkin and Critical Legal Studies on Right Answers and Conceptual Holism' (1994) 18(2) Legal Studies Forum 135.

3 Mireille Hildebrandt, 'The adaptive nature of text-driven law' (2020) 1(1) Journal of Cross-disciplinary Research in Computational Law; Peter Winch, *The Idea of a Social Science* (Routledge & Kegan Paul 1958).

man judgement are not merely (and should not merely be taken as) concerns regarding *what* decision is reached (the 'decision problem' theory of these concerns), but *how* such decisions are reached (getting to a decision in the right way). So the task is *not* to compare the *outputs* of data- and code-driven decision-making to those of text-driven law and determine if the former tracks the latter with sufficient accuracy. Instead, the task is to understand the extent to which either mode of decision-making allows us to engage in whatever this 'right way' of legal reasoning might be.

As a preliminary exercise in this vein, this article explores a range of analogies and disanalogies between text-driven normativity and its data-driven counterparts. It does not aim to provide fundamentally new insights into either domain, but rather, to compare them at an abstract level. These comparisons will be based on generalisations of two deep and diverse areas of research and practice, and so are necessarily limited in depth and nuance, painting broad brushstrokes and obscuring important divides within each domain. Furthermore, these comparisons are made at a descriptive, rather than normative level, leaving considerations of the desirability of either approach for subsequent discussion. Ultimately, the conclusion is that the analogies are weaker than the disanalogies. But the hope is that, in the process of drawing them, we learn something more about the comparison between text- and data-driven normativities and the (im?)possibility of code/data-driven legal decision making.

Exceptions and discretion

Humans can handle exceptions through discretion, but can machines?

In public administration and other contexts, the exercise of discretion 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 of certain contextual factors of the situation in question that render its application inappropriate.

Legal scholars commenting on the importance of human judgement over code- and data-driven decision-making have sometimes fleshed out such arguments with appeals to legal philosophy and jurisprudential theory.[4] They have typically drawn on distinctions between rules and discretion and related distinctions between e.g. rules and principles or standards;[5] between formalised application of law and its open-ended interpretation;[6] or between those figures most associated with differing positions, e.g. 'the Hart-Dworkin debate'.[7]

These arguments usually presume that while computational systems may be capable of implementing those aspects of law which can be characterised in terms of rules, rules are at most a constituent element of a more complex set of reasoning tools. Independent judgement is required

4 Danielle Keats Citron, 'Technological Due Process' (2008) 85(6) Washington University Law Review 1249; Jennifer Cobbe, 'Administrative law and the machines of government: judicial review of automated public-sector decision-making' (2019) 39(4) Legal Studies 636; Lilian Edwards, 'Modelling law using a feminist theoretical perspective' (1995) 4(1) Information and Communications Technology Law 95; James Grimmelmann, 'Regulation by software' (2005) 114(7) Yale Law Journal 1719; Mireille Hildebrandt, 'Algorithmic regulation and the rule of law' [2018] (376) Philosophical transactions of the Royal Society A 1; Andrew Le Sueur, 'Robot government: automated decision-making and its implications for Parliament' in Alexandre Horne and Andrew Le Sueur (eds), *Parliament: Legislation and Accountability* (Hart Publishing 2015); Ronald Leenes, 'Abort or Retry — A Role for Legal Knowledge Based Systems in Electronic Service Delivery?' (Springer 2003); Philip Leith, 'Fundamental errors in legal logic programming' (1986) 29(6) Computer Journal 545; Guido Noto La Diega, 'Against the Dehumanisation of Decision-Making — Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9(1) Journal of Intellectual Property, Information Technology and E-Commerce Law 3; Marion Oswald, 'Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power' (2018) 376(2128) Philosophical transactions of the Royal Society A.

5 Peter H Schuck, 'Legal complexity: some causes, consequences, and cures' (1992) 42(1) Duke Law Journal 1.

6 Brian Bix, 'Waismann, Wittgenstein, Hart, and Beyond: The Developing Idea of 'Open Texture' of Language and Law' in Dejan Makovec and Stewart Shapiro (eds), *Friedrich Waismann: The Open Texture of Analytic Philosophy* (Palgrave Macmillan, Cham 2019).

7 Scott J Shapiro, 'The Hart-Dworkin debate: A short guide for the perplexed' [2007] .

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 — 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'.[8] In other words, discretion is the preserve of human decision makers.

If discretion is just exception handling, both machines and humans can exercise it

Is it so clear that humans can 'do' discretion, but machines cannot? Without fleshing out further what is special about discretion, it may not be clear why computational processes could not at least model or emulate it.

On a simple account of discretion, it is the ability to recognise when a general rule should not apply in a particular case. To use a term from computing, it is a matter of exception handling. Whether an algorithmic system is 'rule-based' or statistical, it is possible to design it in such a way as to create exceptions to general rules. For an expert system, this is simply a matter of adding another rule to modify the more general one.

For a statistical model, it is a matter of including examples of the exceptions, which allow the system to distinguish those cases which fall under the general rule and those which fall under the exception. For instance, if one predictor variable normally positively correlates with a label, but the correlation is negative in the presence of a particular value for a second predictor variable, the model has the capacity to capture this relationship provided it is reflected in the training data.

Is discretion about soft versus hard edges?

If it is not about exception handling per se, defenders of human discretion may instead argue that its value lies in the human ability to navigate blurry distinctions which machines cannot. For Grimmelmann, the reason that software cannot capture human discretion is because while 'software rules can become almost unimaginably complex', they do so 'without their hard edges blurring'.[9]

This critique may appear to be aimed at rule-based systems which involve absolute if-then rules.[10] But from a computational perspective, other available approaches would appear to sidestep this problem. For instance, some forms of rule-based systems incorporate fuzzy statements or predicates.[11] These allow for the traditionally binary *modus ponens* inference to be modified so that the truth values of the premises and conclusions can be represented by values between 0 and 1. This enables the representation of degrees of truth associated with fuzzy predicates. The conclusions follow only to the extent that the premise(s) are (partially) true.

Note that there are two distinct kinds of non-binary claims that can be modelled in this way. One concerns the degree of truth of a predicate as applied to a subject, while the other concerns our uncertainty about a statement in terms of probabilities. To say 'I am 80% confident the bottle is full' (a statement about probability of the bottle being full) means something quite different to 'the bottle is 80% full' (a statement about the degree to which a fuzzy predicate is true of the bottle).[12] Both fuzzy predicates and uncertainty modelling allow the 'hard edges' of absolute if-then rules to be softened.

Similarly, ML systems do not deal with binary rules, but rather in 'feature spaces' (or 'vector spaces'). Individual cases, represented by an array of values for a set of features (or a 'vector'), are placed in relation to one another.

8 Citron (n 4).

9 Grimmelmann (n 4).

10 Frederick Hayes-Roth, 'Rule-based systems' (1985) 28(9) Communications of the ACM 921.

11 Lotfi Asker Zadeh, 'Fuzzy sets' (1965) 8(3) Information and Control 338; Lotfi Asker Zadeh, George J Klir, and Bo Yuan, *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers* (vol 6, World Scientific 1996).

12 Didier Dubois and Henri Prade, 'Possibility Theory, Probability Theory and Multiple-Valued Logics: A Clarification' [2001] (32) Annals of Mathematics and Artificial Intelligence 35.

A distance function can measure the distance between any two cases in the feature space of a model. These distances are used to distinguish between cases for the purpose of prediction and classification.

It is true that if the model is used to generate a discrete classification rather than a continuous output, it is necessary to use functions which sort cases into precisely one or another class; so in these cases, the decision boundaries will be 'hard'. But even here, the underlying structure is not one of hard edges; it is one of smooth inclines which can represent infinitely gradual differences between cases. Hence why such systems often provide multiple possible classifications with degrees of confidence (e.g. a fraud detection model can give a confidence score indicating how likely the transaction is to be fraudulent or legitimate; or a dog breed classifier whose output is '[Labrador: 81%], [Poodle: 11%]', etc.).

Soft edges as indeterminacy

Despite the various ways in which the apparently hard edges of software can be softened and exceptions to general rules can be handled, there remain different kinds of 'softness' in text-driven systems which cannot be built into data- or code-driven ones. The kinds of softness involved in fuzzy logic, or in the smooth gradations of ML model feature spaces, are still determinate. A predicate which comes in degrees still needs to be given a particular value between 0 and 1; distances in feature spaces are similarly fixed.

But the kind of softness involved in text-driven law is qualitatively different; it is indeterminate. In some cases, there are no rules at all, and therefore no exceptions, because the decision relies on concepts such as reasonableness. There may also even be multiple equally valid ways to apply a facially clear rule, which would therefore be subject to judgement and discretion. Under a legal realist conception of law,[13] these discretionary elements of judgement are inevitable, even when applying an apparently bright-line, rule-based law.

Learning from new cases

Both human legal judgements and machine decision-making processes can evolve through consideration of new cases

Legal decision-making develops partly as a result of decisions made about new cases. As new cases are brought before the courts and decided upon, new precedent is set which influences or even binds future decisions which are similar in relevant respects. This allows novel issues and nuances of a particular context to be drawn out and reflected in future decisions.

Precedent exists as a distinct source of law in common law systems (alongside statute), as well as a doctrine of jurisprudence in civil law. But it also exists in 'softer' forms, for instance, where administrative decision-makers in practice bind themselves to be consistent with previous positions or else face the charge of unjustified differential treatment (e.g. an administrative agency which interprets a legal requirement or policy in a particular way in response to one case is effectively bound to that interpretation in similar cases). In this way, new cases introduce new considerations and distinctions into the system which influence subsequent decisions.

In some cases, the precedent which is set might have knock on effects on other areas of law or administrative practice. For instance, a case about a shipping dispute might end up setting precedent on the standard of proof for demonstrating the jurisdiction of a court which affects a wide variety of subsequent international cases. Or an administrator who decides to interpret an identity check as requiring a passport in relation to one case may be effectively bound to apply a similar standard elsewhere. In this way, a local case might have wider repercussions for the global system of decision-making.

Similarly, ML systems, once trained and deployed, may be subsequently updated as a result of training on fresh labelled examples from the domain of application. For instance, new types of previously unidentified fraudulent

[13] See e.g. Wesley Newcomb Hohfeld, 'Some fundamental legal conceptions as applied in judicial reasoning' (1913) 23(1) Yale Law Journal 16.

transactions can be incorporated into a fraud detection model, or new and different images of a certain dog breed can be included into the dog classifier. As with the incorporation of new cases into precedent, the incorporation of these new examples allows for novel cases and nuances to be captured in the next iteration of the model.

Similar to the way in which new case-law can have far-reaching implications, new cases used to retrain a model may have repercussions beyond their locale. Any changes that result from this fresh training data might not only affect nearby portions of the model (i.e. the way the model operates on a select portion of the feature space) but have ripple effects on the global model. For instance, by introducing new examples of a particular dog breed, the dog breed classifier might not only change the portion of the model which deals with the 'nearby' cases, but affect the model at a global level; for instance, the way the model detects different kinds of ears might change for all types of dogs, not just those that are similar to the new examples.

Disanalogies in the processes of learning from new cases

Despite these surface similarities, there are some important differences between the ways in which human-driven and machine-driven decision-making develops in response to new cases.

In case-law, when a new case is applied, the result applies to all similar cases in future, no matter how many previous similar cases had a different result. Old precedents can be contradicted and superseded by new ones. In other words, there is a temporal dimension according to which the latest case is (typically) what matters. In machine learning, one new data point may not be enough to effect the same change in all similar cases in future. Typically, each data point in the training data has equal weight; if there are more previous cases in the training data which have a different label, they will have a heavier weight on future decisions than the new one. In this way, fresh data adds to, but does not overrule the model globally. If the new case contradicts the old cases, the ML algorithm will attempt to find a model with the best fit between all the cases it has seen in training.

Furthermore, while case-law allows for similar cases to be distinguished according to material differences between them (and thereby determine if the precedent applies or not), in the machine learning context, there is no such mechanism. It may be that the material and non-material differences that make a (legal) difference are not captured by the distance between two points in feature space. Or if they are, they may be accompanied by many other irrelevant differences, between which the ML algorithm cannot distinguish.

In case-law (and its softer equivalents), there is a set of rules and hierarchy (e.g. between higher and lower courts) to address the way that new cases should be factored in, and how precedent can develop. Machine learning has no strongly analogous equivalent to allow certain data points to be given precedence over other ones. The typical ways to change how an ML model behaves are to provide it with different data or to change the loss function so that certain kinds of errors are designated as more costly than others. It is also possible for ML engineers to assign different weights to specific individual samples in the training data, but the default is to give all samples equal weight.

Finally, there are obvious differences in the reach or remit of each system. Legal jurisdictions typically concern whole nations, while international and transnational private law have even wider global scope. While code- and data-driven systems sometimes have global reach, they are generally substantially isolated from one another. There are exceptions, such as in federated learning, where multiple local models are combined into global models,[14] or in domain adaptation and transfer learning, where a general purpose model is adapted for more specific contexts.[15] However, when one model learns from new cases, unlike law, such learning rarely has ramifications for other models.

So, while some analogies can be drawn between the ways in which both legal and machine learning driven systems

[14] Jakub Konečný, Brendan McMahan, and Daniel Ramage, 'Federated Optimization: Distributed Optimization Beyond the Datacenter' (1913) 23(1) arXiv preprint arXiv:1511.03575 16.

[15] Koby Crammer, Michael Kearns, and Jennifer Wortman, 'Learning from multiple sources' [2008] (45) Journal of Machine Learning Research 1757.

both learn from new cases, their methods, purposes, and implications of doing so differ significantly.

Ground truth

Both human legal judgement and machine learning have complicated relationships with the notion of correctness or ground truth. There is no single agreed upon account of what it is that legal or machine judgements are aiming at, beyond vague or question-begging claims, such as that human legal judgement aims to apply the law 'correctly' to cases, or that statistical models aim to generate labels for unseen data points which reflect their 'true' labels.

In machine learning, the term 'ground truth' relates to the facts of the matter regarding the correct label for a given data point about which a prediction or classification is being made. This is evaluated by reference to labelled 'test data' which is drawn from the domain of interest but which the machine learning model has not yet 'seen' during its training. Of course, in many cases, the data available is only a proxy for the ground truth. Sometimes, the dog breed may be misidentified, or a financial transaction is wrongly logged as fraudulent.

But the notion of a ground truth against which the model's predictions can be hypothetically, even if not perfectly, tested is very often a slippery one. Machine learning is often deployed to deal with matters which are subjective, or whose truth is only established by social convention. For instance, natural language classifiers for offensive or abusive messages are trained on datasets which have been labelled by multiple human labellers. The reason for employing multiple labellers rather than just one, is that concepts like offense and abuse are subjective. Multiple human judgements are aggregated and the average score is taken as 'ground truth' for the classifier, but this ground truth is understood to be an inter-subjective average. Even the examples above, of dog breeds and fraudulent transactions, are matters about which reasonable people may disagree. Many machine learning researchers and practitioners will therefore use the term 'ground truth' with such caveats in mind.

There is a long history of debate in jurisprudence about whether or not law can be conceived as aiming towards 'legal' truth or any other kind of ideal. On some accounts, legal judgements should alight on a single correct answer, which could perhaps be described as analogous to the 'ground truths' provided by the data against which machine learning models are evaluated. For Dicey, the correct answer was to be determined by deterministic application of the rules.[16] For Dworkin, judges have more flexibility in their interpretation of rules and principles, but nonetheless ought to alight on a single judgement reflecting the legal truth of the case.[17] The implication is that given two cases in which the features are identical, and assuming an identical legal system with the same set of statutes, case-law and institutional history, judges ought to come to the same conclusion. In this sense, proponents of such accounts of law could argue there is a kind of 'ground truth' which legal decisions aim to capture.

However, many do not share Dworkin's faith in there always being a right answer; instead, there may be multiple different decisions which a court could reach; it is the form of legal reasoning that is deployed which matters. On this view, there are multiple right answers, even if only certain ways of reaching them are permitted.[18] Many legal theorists and practitioners, particularly those associated with the American Realist tradition, would reject the notion that legal reasoning is guided by any kind of ideal.[19] For them, the idea that legal judgements aim to correspond to some ideal form of correctness or 'legal truth' (even if that is understood in terms of the rather modest ideal of the *lawful* outcome) misunderstands law. Instead, they conceptualise law as operating in a relativist or pluralistic mode, a kind of 'rough morality' that is indeterminate; contingent

16 Albert Venn Dicey, *The Law of the Constitution* (Oxford University Press 2013).

17 Dworkin (n 2).

18 Bix (n 6).

19 Henry Hart and Albert Sacks, *The Legal Process: Basic Problems in the Making and Application of Law* (Foundation Press 1958); Oliver Wendell Holmes Jr, *The Path of the Law* (The Floating Press 2009); Hohfeld (n 13); Karl N Llewellyn, *Jurisprudence: Realism in Theory and Practice* (vol 1, Transaction publishers 2011).

on and expressed by the particular facts of the case before them.

Both domains contain spectra with analogous poles. On one pole, the idea that the outputs of the process (whether legal judgement or ML models) track some singular external ideal (e.g. lawfully correct judgements and ground truth in test data). On the other pole, scepticism that those outputs reflect anything other than pluralistic or relativistic set of context-specific values. In which case, the extent to which ML and law are analogous in their relationship to any notion of ground truth will depend on the position taken with respect to each. If they took the time to consider it, a Dworkinian about law might be a sceptic about the 'ground truths' of ML, while a believer in the latter might also be a legal realist. Allegiances might be split in different ways between theoreticians and practitioners. The aim here is not to take up particular points on either spectrum. The salient point for our purposes is that both domains have similar debates about the nature of the decisions/outputs reached and their relationship to truth or correctness.

Treating cases individually[20]

One of the objections to algorithmic decision-making is that, unlike human judgement, it treats individual cases on the basis of generalisations made from similar cases in the past and is therefore individually unjust. In other words, it fails to treat people as individuals. As articulated by a research participant 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'.[21] The inability to treat people as individuals — even if they appear identical to previous cases — is thus an endemic feature of machine learning systems.[22]

As a counterpoint to this, it is worth noting that machine learning models can in theory capture a near-infinite number of minute differences between cases. With a high-dimensional model, specific points in feature space may uniquely identify individuals. For instance, in a feature space consisting of just three features, each of which takes a number between 0 and 100, there are 1 million possible combinations. If the number of cases is small enough, the input values which represent individual cases could potentially be unique to each case. Machine learning models can treat people as individuals, in the sense of being able to distinguish them from each other. And wherever there is a difference (or, more precisely, a distance that is not zero) between them, there is the possibility of distinguishing cases and thus to come to a different decision.

However, in practice, there is no guarantee that the legally important differences will be captured by such differences in vector space. If we need human judgement to assess whether and how differences between individual cases matter, then only humans are capable of truly case-by-case judgement. Such arguments have both an epistemic and a normative component.

The epistemic argument for case-by-case judgement

The epistemic component of the argument for case-by-case judgement is based on the claim that we are unable to pre-specify how to reason appropriately about new cases without human intervention. 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.[23] Both were 'particularist' about new individual cases, according to which it is impossible to specify in advance how previous reasoning might apply in a particular case.

[20] Parts of the following section are adapted from another paper by the author — see Reuben Binns, 'Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making' Regulation & Governance (forthcoming).

[21] Reuben Binns and others, '"It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions' [2018] (377) Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems 1.

[22] ibid.

[23] Shapiro (n 7).

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.[24] 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) does or does not apply. Waismann argued that language was not like this. Certain concepts are 'porous'; we do not always know what to say about whether they still apply to a new case. An example would be the term 'arms' in the Second Amendment to the United States constitution. At one time this referred to a manually reloading musket, but we might now reasonably debate whether it ought to be taken to refer also to automatically reloading assault rifles.

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 comparable 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. 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[25]); 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 cannot be captured systematically enough to feature in most cases, or features which are irrelevant in all previous cases but are unexpectedly relevant in a new one. The way that any principles affecting the mapping from features to classifications interact and are balanced is similarly unspecifiable 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.[26] 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 cannot 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,[27] and a problem space may be just too complex to capture all cases in a parsimonious and generalisable way.

24 Friedrich Waismann, 'Verifiability' in Rom Harré (ed), *How I See Philosophy* (Palgrave Macmillan 1968); Bix (n 6).

25 Jonathan Dancy, 'Moral Particularism' in Edward N Zalta (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2017, Metaphysics Research Lab, Stanford University 2017).

26 The point that rule-based expert systems and machine learning models are not fundamentally different in this respect has also been made in Monika Zalnieriute, Lyria Bennett Moses, and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82(3) Modern Law Review 425.

27 Cristian S Calude and Giuseppe Longo, 'The Deluge of Spurious Correlations in Big Data' (2017) 22(3) Foundations of Science 595; Ronald L Graham, Bruce L Rothshild, and Joel H Spencer, *Ramsey theory* (John Wiley & Sons 1990).

The normative argument 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: '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.'[28] 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,[29] and in German jurisprudence as *Einzelfallgerechtigkeit* ('justice in each particular case').[30] 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.[31] As US 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'. 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.[32]

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 or function to use from scratch in every case.

Conclusions

The previous sections have summarised some broad analogies and disanalogies between text-driven and code- and data-driven decision-making. Broad statements about fundamental differences often have easy counterexamples. For instance, the legally minded may characterise code as always binary, not allowing soft or blurry edges. To which the computer scientists may retort: 'But look: we have fuzzy logic! We can model uncertainty!'. The legally minded may say: 'a human decision maker can treat each case individually', to which the machine learning engineer may say: 'but I can model individuals as unique points in high-dimensional feature space!'. Ultimately, such conversations may be talking at cross-purposes. Whatever functional similarities exist between these different normativities, they exist for different reasons and it would be

28 Aristotle and Hugh Tredennick, *The Ethics of Aristotle: The Nicomachean Ethics* (Penguin 1976).

29 Frederick Schauer, *Profiles, probabilities, and stereotypes* (Harvard University Press 2009).

30 Gabriele Britz, *Einzelfallgerechtigkeit versus Generalisierung: verfassungsrechtliche Grenzen statistischer Diskriminierung* (Mohr Siebeck 2008).

31 Benjamin Eidelson, 'Treating People as Individuals' in Deborah Hellman and Sophia Moreau (eds), *Philosophical Foundations of Discrimination Law* (Oxford University Press 2013); Kasper Lippert-Rasmussen, '"We are all Different": Statistical Discrimination and the Right to be Treated as an Individual' (2011) 15(1) Journal of Ethics 47.

32 Schauer (n 29).

naïve to expect them to be able to serve the same purposes.

At the same time, both are in search of analogies and disanalogies through which to understand themselves and their relationship to each other. In this context, it is perhaps instructive that in the most recent presidential address to the conference on Artificial Intelligence and Law, those working in the field were asked to consider both the metaphor of 'AI as law', and metaphors of law in terms of constituent disciplines of AI (mathematics, technology, psychology).[33]

Finally, the domains of law and computer science/AI are both pluralistic, containing various conflicting schools of thought and practice. For instance, as briefly alluded to above, both contain different views as to the 'ground truths' of the test data against which machine learning algorithms are evaluated, and the analogous notion of a single right answer against which legal judgements can be assessed as correct. It may be that the differences within each domain are greater than the differences between them.

References

Aristotle and Tredennick H, *The Ethics of Aristotle: The Nicomachean Ethics* (Penguin 1976).

Binns R, 'Human Judgement in Algorithmic Loops; Individual Justice and Automated Decision-Making' Regulation & Governance (forthcoming).

Binns R and others, '"It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions' [2018] (377) Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems 1.

Bix B, 'Waismann, Wittgenstein, Hart, and Beyond: The Developing Idea of 'Open Texture' of Language and Law' in Makovec D and Shapiro S (eds), *Friedrich Waismann: The Open Texture of Analytic Philosophy* (Palgrave Macmillan, Cham 2019).

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

Calude CS and Longo G, 'The Deluge of Spurious Correlations in Big Data' (2017) 22(3) Foundations of Science 595.

Citron DK, 'Technological Due Process' (2008) 85(6) Washington University Law Review 1249.

Cobbe J, 'Administrative law and the machines of government: judicial review of automated public-sector decision-making' (2019) 39(4) Legal Studies 636.

Crammer K, Kearns M, and Wortman J, 'Learning from multiple sources' [2008] (45) Journal of Machine Learning Research 1757.

Dancy J, 'Moral Particularism' in Zalta EN (ed), *The Stanford Encyclopedia of Philosophy* (Winter 2017, Metaphysics Research Lab, Stanford University 2017).

Dicey AV, *The Law of the Constitution* (Oxford University Press 2013).

Dubois D and Prade H, 'Possibility Theory, Probability Theory and Multiple-Valued Logics: A Clarification' [2001] (32) Annals of Mathematics and Artificial Intelligence 35.

Dworkin R, 'No right answer?' (1978) 53(1) New York University Law Review 1.

Edwards L, 'Modelling law using a feminist theoretical perspective' (1995) 4(1) Information and Communications Technology Law 95.

Eidelson B, 'Treating People as Individuals' in Hellman D and Moreau S (eds), *Philosophical Foundations of Discrimination Law* (Oxford University Press 2013).

Graham RL, Rothshild BL, and Spencer JH, *Ramsey theory* (John Wiley & Sons 1990).

Grimmelmann J, 'Regulation by software' (2005) 114(7) Yale Law Journal 1719.

Hart H and Sacks A, *The Legal Process: Basic Problems in the Making and Application of Law* (Foundation Press 1958).

Hayes-Roth F, 'Rule-based systems' (1985) 28(9) Communications of the ACM 921.

Hildebrandt M, 'From Galatea 2.2 to Watson – And Back?' in Hildebrandt M and Gaakeer J (eds), *Human Law and Computer Law: Comparative Perspectives* (Springer 2013).

— 'Algorithmic regulation and the rule of law' [2018] (376) Philosophical transactions of the Royal Society A 1.

[33] Bart Verheij, 'Artificial intelligence as law' (2020) 28 Artificial Intelligence and Law 181.

Hildebrandt M, 'The adaptive nature of text-driven law' (2020) 1(1) Journal of Cross-disciplinary Research in Computational Law.

Hohfeld WN, 'Some fundamental legal conceptions as applied in judicial reasoning' (1913) 23(1) Yale Law Journal 16.

Konečný J, McMahan B, and Ramage D, 'Federated Optimization: Distributed Optimization Beyond the Datacenter' (1913) 23(1) arXiv preprint arXiv:1511.03575 16.

Le Sueur A, 'Robot government: automated decision-making and its implications for Parliament' in Horne A and Le Sueur A (eds), *Parliament: Legislation and Accountability* (Hart Publishing 2015).

Leenes R, 'Abort or Retry — A Role for Legal Knowledge Based Systems in Electronic Service Delivery?' (Springer 2003).

Leith P, 'Fundamental errors in legal logic programming' (1986) 29(6) Computer Journal 545.

Lippert-Rasmussen K, '"We are all Different": Statistical Discrimination and the Right to be Treated as an Individual' (2011) 15(1) Journal of Ethics 47.

Litowitz D, 'Dworkin and Critical Legal Studies on Right Answers and Conceptual Holism' (1994) 18(2) Legal Studies Forum 135.

Llewellyn KN, *Jurisprudence: Realism in Theory and Practice* (vol 1, Transaction publishers 2011).

Noto La Diega G, 'Against the Dehumanisation of Decision-Making — Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9(1) Journal of Intellectual Property, Information Technology and E-Commerce Law 3.

Oswald M, 'Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power' (2018) 376(2128) Philosophical transactions of the Royal Society A.

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

Schuck PH, 'Legal complexity: some causes, consequences, and cures' (1992) 42(1) Duke Law Journal 1.

Shapiro SJ, 'The Hart-Dworkin debate: A short guide for the perplexed' [2007].

Verheij B, 'Artificial intelligence as law' (2020) 28 Artificial Intelligence and Law 181.

Waismann F, 'Verifiability' in Harré R (ed), *How I See Philosophy* (Palgrave Macmillan 1968).

Wendell Holmes Jr O, *The Path of the Law* (The Floating Press 2009).

Winch P, *The Idea of a Social Science* (Routledge & Kegan Paul 1958).

Zadeh LA, 'Fuzzy sets' (1965) 8(3) Information and Control 338.

Zadeh LA, Klir GJ, and Yuan B, *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers* (vol 6, World Scientific 1996).

Zalnieriute M, Bennett Moses L, and Williams G, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82(3) Modern Law Review 425.

A reply: On the differences between human and machine processing of legal language

Emily M. Bender • University of Washington ebender@uw.edu

Binns explores analogies and disanologies between decision making in legal contexts by humans and machines. Writing from the perspective of computational linguistics, I explore a key difference between the two: what it means for humans and machines to process language. Humans rely heavily on joint attention and modeling each other's possible intents. Machines, conversely, lack not only a model of their interlocutor's world but also the ability to relate language input to that model. I argue that a clear understanding of humans' language use and the actual capacities of machines to mimic that usage is important when applying natural language processing (NLP) in sensitive contexts such as the law.

English speakers talk about language as if words and sentences carry meaning: speakers use language to encode their intent, which is then in the words, from which listeners decode it. However, this metaphor is misleading [11]. Instead, we use words and phrases which have conventional 'standing' meanings within our speech communities [10, 9], not to communicate precisely the content of those meanings, but rather as rich cues to what we intend. Clark describes spoken communication as a *joint activity* wherein speaker and listener manage to coordinate on the messages being communicated because they are mutually aware of the joint activity and their roles therein [4].[1] With asynchronous written language, we rely much more heavily on the linguistic signal, being unable to enlist joint attention to shared immediate surroundings. Nonetheless, the activity of understanding remains fundamentally the same: readers take the arrangement of words into sentences and paragraphs as constructed by the writer, reconstruct their conventional meaning, and then reason from there about what the author intended to convey.

Legal language, including laws and the textual records of legal reasoning, presents an interesting special case. It is not meant to describe a state of affairs that is current as of the writing, nor to produce fictional worlds, but rather to shape the world of the present and in particular of the future. It is meant to be read, interpreted and acted upon by future interlocutors who may inhabit surroundings that are unfamiliar to the authors.[2] As a linguist, I imagine that it is for this reason that law is 'a linguistic practice' and legal professionals devote so much effort to the interpretation and careful drafting of texts.

Under the 'data-driven' paradigm that Binns describes, it is therefore important to ask what the machines have access to as their representation of the data. Binns does not specify, but one can assume that the inputs to machine learning systems include descriptions of the cases to be decided and relevant legal texts (statutes, previous court decisions, etc.), as well as possibly many irrelevant legal texts, leaving it to the machine to discover which to attend to.

When presented with texts in languages we are competent in, people tend to think of the information as being 'in' the text. But to the extent this is true (see above), it is only because we are in command of the linguistic system corresponding to the text. Modern NLP systems that rely on machine learning generally are not. Current state-of-the-

[1] Clark illustrates this with a commercial transaction, where the clerk said '$12.77' and Clark replied '$12.77'. The clerk's utterance in context meant roughly 'Those items together cost $12.77' and Clark's reply of exactly the same string meant 'I confirm I understand the amount to be $12.77' [4, p. 66]. There was no ambiguity because they had a shared understanding of the communicative situation.

[2] See also White: 'My aim [is] to try to transform our sense of law by putting it together with something else: to try to see it as compositional art, as a set of activities by which minds use language to make meaning and establish relations with others' [13, p.17].

art systems like BERT [6] or GPT-3 [2] are primarily language models. This means that their training tasks consist of predicting linguistic form given other linguistic forms as context. But language is a system of signs, i.e. pairings of form and meaning [12].[3] As we have recently argued [1], if the training data includes only form, a machine learning system has in principle no means of learning from it either meaning or a mapping between form and meaning. More linguistically sophisticated systems, such as (for English) Boxer [5] or the English Resource Grammar (ERG) [7, 8] model linguistic systems including morphology, syntax and compositional semantics, assigning semantic representations to input strings.[4] Nonetheless, these semantic representations at best can only be seen to capture standing meaning.

Thus, when Binns describes machine learning systems as representing cases in terms of 'features', it matters quite a bit what those features are. With current technology, they are most likely to be vector-space representations of words or word sequences, that is, representations of words in terms of which other words they co-occur with. A slightly more sophisticated approach would build those features in terms of semantic graphs extracted from representations of standing meaning. To get to an imagined future where the features could instead be in terms of precise representations of the actual situations and legal intents described in the textual input would require massive advances in NLP and AI, to support the reasoning required to go from standing meaning to an understanding of communicative intent in context. Currently, NLP systems can do that only in domains with limited sets of possible intents (such as the commands that voice assistants understand). But legal systems are domain-independent and must work with an open-ended set of circumstances. When Binns writes, 'a case-by-case assessment is needed because no two cases can be identified as exactly alike ahead of time without examining each', we must remember that machines that can actually turn text into the kind of representations required for the type of reasoning Binns assumes are far off, if possible to construct at all.

References

[1] Emily M. Bender and Alexander Koller. 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198.

[2] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: `2005.14165 [cs.CL]`.

[3] Yufei Chen et al. 'Neural Maximum Subgraph Parsing for Cross-Domain Semantic Dependency Analysis'. In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 562–572.

[4] Herbert H. Clark. *Using Language*. Cambridge: Cambridge University Press, 1996.

[5] James Curran, Stephen Clark, and Johan Bos. 'Linguistically Motivated Large-Scale NLP with C&C and Boxer'. In: *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*. Prague, Czech Republic: Association for Computational Linguistics, June 2007, pp. 33–36.

[6] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[7] Dan Flickinger. 'On Building a More Efficient Grammar by Exploiting Types'. *Natural Language Engineering* 6.1 Special Issue on Efficient Processing with HPSG (2000). Ed. by Dan Flickinger et al., pp. 15–28.

[8] Dan Flickinger. 'Accuracy v. Robustness in Grammar Engineering'. In: *Language from a Cognitive Perspective: Grammar, Usage and Processing*. Ed. by Emily

[3] Specifically, conventional pairings of form and meaning, the latter used as cues to understand communicative intent.

[4] Similarly machine learning systems trained on ERG output [e.g. 3].

M. Bender and Jennifer E. Arnold. Stanford, CA: CSLI Publications, 2011, pp. 31–50.

[9] H. Paul Grice. 'Utterer's Meaning, Sentence-Meaning, and Word-Meaning'. *Foundations of Language* 4.3 (1968), pp. 225–242.

[10] Willard V. Quine. *Word and Object*. Cambridge MA: MIT Press, 1960.

[11] Michael J. Reddy. 'The conduit metaphor: A case of frame conflict in our language about language'. In: *Metaphor and Thought*. 1979, pp. 164–201.

[12] Ferdinand de Saussure. *Course in General Linguistics*. Translated by Wade Baskin. New York: The Philosophical Society, 1959.

[13] James Boyd White. *Justice as Translation: An Essay in Cultural and Legal Criticism*. University of Chicago Press, 1994.

Author's response: Why even inferring legal intent would not be enough

Reuben Binns

I am very grateful for Bender's thoughtful response, which takes us beyond generalities of my article by focusing on a specific challenge of a technical sub-domain.

In my introduction, I argued that where algorithms are used to make or support decisions, and such decisions are expected to be lawful, the aim should not (only) be to get the 'right' decision. Rather, it should (also) be to get to a decision in the right way. I see a similar concern reflected in Bender's compelling case for doubting that computational models of language are anywhere near 'understanding'. Natural language understanding requires more than just performing well, in the sense of (e.g.) frequently producing convincing conversational responses. Rather, as Bender has argued elsewhere, language models need to 'perform well for the right reasons'.[1]

For Bender, a precondition for such 'right reasons' is correctly grasping the intent behind the expressions in a conversation, since inferring the intent behind an expression is necessary for grasping meaning. But intent is not encoded in the data used to train these systems, either explicitly or implicitly (absent grounding in the physical and mental worlds of interlocuters). If legal decision making relies on grasping legal intent (and therefore, is a special case of the more general language understanding problem), then the absence of representations of legal intent prevent such systems from being able to learn to make or guide decisions in the right way. I concur with Bender's conclusion that the current failure (and potential impossibility) of turning text into such representations prevents the kind of case by case assessment that I earlier argued was necessary for individual justice.

Awareness of what another person is attending to and intending to communicate are clearly important aspects of general natural language understanding, and certainly, interpreting the intent behind statutes and case law often play a part in legal reasoning. However, whether that is sufficient or even necessary may differ depending on different schools of jurisprudence and legal philosophy. In the U.S. constitutional context, for example, originalists place emphasis on modelling intent at the time of legislative drafting (although may differ regarding whose intent — the drafters' or society's — is relevant). Pragmatists, by contrast, take the view that original intent should not prevent novel textual interpretations which reflect changes in language and circumstance. For them, the aim should not be to infer the intention behind the original expression, but rather, to create new meaning; legal text is thus inert, waiting to have life breathed into it through fresh interpretation, open ended exploration and judgement. Such divides are consequential — as argued above, the differences within legal theory and computational approaches may be more important than those between them.

A consequence is that, compared to modelling intent between interlocuters, modelling legal intent is not straightforward, even for humans. Furthermore, the role intent ought to play in legal reasoning is essentially contested in both theory and practice. As such, algorithmic systems in legal decision-making are problematic not only because of more general limitations of natural language understanding outlined by Bender, but also due to additional and qualitatively different factors; to borrow loosely from mathematical terminology, they are a degenerate case rather than a (merely) special case.

[1] Emily M Bender and Alexander Koller, 'Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data' (Association for Computational Linguistics July 2020) p. 5192.