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

Rules, judgment and mechanisation

Mazviita Chirimuuta

Abstract

This paper is a philosophical exploration of the notion of judgment, a mode of reasoning that has a central role in legal practice as it currently stands. The first part considers the distinction proposed by Kant, and recently explored historically by Lorraine Daston, between the capacity to follow and execute rules and the capacity to determine whether a general rule applies to a particular situation (that is, judgment). This characterisation of judgment is compared with one proposed by Brian Cantwell Smith, as part of an argument that current AI technologies do not have judgment. The second part of the paper asks whether digital computers could in principle have judgment and concludes with a negative answer.

Keywords: philosophy of AI, cognitivism, legal judgment, automation, history of AI

Replier: William Lucy, Law School, Durham University • w.n.lucy@durham.ac.uk

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

* Senior Lecturer, University of Edinburgh. m.chirimuuta@ed.ac.uk.

Introduction

The notion of *judgment* is central to the law, not only as embodied in the figure of the judge, but as the mode of cognition that mediates between the often conflicting formal, moral and instrumental demands that animate the legal system. In recent times there have been extreme predictions made about the overturning of law as we know it through the roll out of automated tools based on artificial intelligence technologies such as deep learning, designed to replicate (and improve on) human legal reasoning. These scenarios have been depicted in both a positive and negative light,[1] but reasonable expectations for the impact of automation in law must be informed by consideration of what judgment is, the circumstances in which it is required, and whether or not it can be automated (that is, carried out by digital computers). Hildebrandt criticises Alarie's forecast of a 'legal singularity' for confusing 'the mathematical simulation of legal judgment for legal judgment itself,'[2] and yet, the computational theory of human cognition (*cognitivism*), which dominates cognitive science and is foundational to AI, tends to dismiss suggestions that there is daylight between the human enactment of a thought process, and its functionally equivalent, silicon counterpart. Cognitivism has of course been contested, and the roots of the disagreement go deep into the history of philosophy, as we will see.

This paper offers an anti-computationalist, philosophical perspective on the debate. What follows is an exploration of judgment — where it is called for, and how the capacity stands in relation to conduct and decision making governed by explicit, codified rules. The first part will ask what judgment is and the second part will ask if digital computers could ever have judgment. I draw on sources and arguments outside of law, whose relevance should be clear to those concerned with legal judgment. The argument of the second part will largely be illustrated with historical examples rather than the machines of today. Examination of the pre-history of computation and first accounts of digital computers are helpful when addressing our topic because it makes clear how computers fit into a bigger story concerning mechanistic theories of human behaviour and processes of industrial automation. These foundational issues are far clearer than the opaque workings of the artificial intelligence (AI) algorithms of today. But they are still relevant because however impressive today's technology may be, its basic operations are not categorically different from those carried out by the first digital computing machines.

What is judgment?

I will be offering two answers to the question. Judgment is first defined as the sense of how a universal principle should be applied to a particular instance, how an abstract rule fits to a concrete case. The second answer takes judgement to be an essentially normative mode of cognition, in both an ethical and epistemic sense, one resting on ontological commitment to a world to which judgments must defer. I conclude

[1] Benjamin Alarie, 'The path of the law: Towards legal singularity' (2016) 66 University of Toronto Law Journal 443; William Lucy, 'The Death of Law: Another Obituary' (2022) 81 Cambridge Law Journal 109.

[2] Mireille Hildebrandt, 'Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics' (2018) 68 University of Toronto Law Journal 12, p. 23.

the section by considering how these two characterisations of judgment relate to one another.

Daston and Kant

Historian of science Lorraine Daston discusses how the concept of a rule as something that can be followed to the letter, without need for intelligent interpretation and adaptation to context, is quite a recent innovation. 'Rule', she explains, originally meant something like a model case or paradigm — a standard for emulation.[3] The monastic Rule of St. Benedict was a case in point. This body of regulation foresaw the need for exceptions, depending on circumstances, and hence discernment about application of the rules. Authority to make discriminations in the deployment of codes of conduct was vested in the abbot, himself a rule, in the sense of model, of the ideal monk and Christian. Even in what we now think of as the formal disciplines of arithmetic and geometry, rules and even 'algorithms' were not always conceived as foolproof, explicit, step-by-step instructions. Instead, procedures for calculation in the Middle Ages were taught via sketched out generalisations which made no sense without the examples accompanying them. The apprentice calculator was expected to absorb these cases and use them as the basis for sensible analogical transfer when new problems arose. The mathematics of the past, with its resemblance to case law, was very different from the clear, precise, and inexorable system of deductive closure that we have in mind today.[4]

The current notion of rules as, ideally, exceptionless and infallible codes for conduct, is epitomised by modern arithmetical algorithms, like one for long division. These are fail-safe recipes for calculation that require no mathematical intuition or discernment for their execution and they are not context-dependent, which means they demand no sensitivity to attenuating circumstances on the part of their user.[5] As long as a small set of basic operations can be carried out and the instructions followed to the letter, the person who executes the rule will always be able to reach the correct answer (that is, the output corresponding to a given input), granted sufficient time and paper to carry out all the steps.[6] The development of such

[3] Cognates are '*règle*' (French) and '*regula*' (Latin). To give one of Daston's examples, the first definition in the entry for *Règle* in the *Encyclopédie* of 1780 (eds. Denis Diderot and Jean d'Alembert), is the following: 'the life of Our Savior is the *rule* or the *model* for Christians.' See Daston, *Rules: A Short History of What We Live By* (Princeton University Press 2022), p. 22.

[4] Daston's distinction between pre-modern and modern rules helps illuminate the contrast made by Alarie between 'legal standards' and law as a 'completed' system of rules. Legal standards are rules of thumb that do not come with explicit instructions for their application. Their existence, according to Alarie, indicates that there are gaps in what the law specifies, which means that law is not complete. He writes that 'big data and machine learning can help us to complete the law by substituting extremely complex, fact-sensitive, and query-able systems of rules for existing legal standards' (Alarie, n 1, p. 454).

[5] Onora O'Neill nicely captures the modern notion of algorithmic rules: 'they must determine answers for all cases that fall under them and that particular decisions must be given by, and so deducible from, rules.' See Onora O'Neill, 'Abstraction, Idealization and Ideology in Ethics' (1987) 22 Royal Institute of Philosophy Supplements 55, p. 58.

[6] A confusion may arise at this point. Computer science is founded on the concept of the algorithm (in the modern sense), which is a fail-safe recipe for computing a function by following the explicitly coded procedure. This means that a computing machine will infallibly generate the output determined by its input and its algorithm (program), as long as it is in working order. This does *not* mean that a computer will always be right (in an informal sense) or produce the output intended by its programmer. Computer codes are of course not fail-safe means to achieve the answers or performance people want from them. But that does not make their basic operation any less algorithmic.

algorithms is the result of a long process of automation of calculation, which preceded the invention of computing machines.[7] A point emphasised by Daston is that the old woolly and mushy kinds of rules — to use pejorative terms that express a common contemporary reaction to cognitive modes that are not clear, distinct and formalisable — were bound up with an implicit scepticism about the idea that any universal formulation of a rule could anticipate all the particulars it would be expected to meet in practice. Context-dependency of rules was assumed, not denied. Rules needed to be soft so that they could be adapted to an open-ended series of novel situations; and because of this flexibility, this lack of full pre-specification about how they should operate, knowledge of rules needed supplementation with a capacity to see when and how to best apply them. Let us call this judgment.[8]

The supplementary capacity of judgment appears nicely characterised in the following passage from Immanuel Kant's *Critique of Pure Reason*:[9]

> If the understanding in general is explained as the faculty of rules [Regeln], then the power of judgment is the faculty of subsuming under rules, i.e., of determining whether something stands under a given rule (*casus datae legis*) or not. General logic contains no precepts at all for the power of judgment, and moreover cannot contain them. For since it abstracts from all content of cognition, nothing remains to it but the business of analytically dividing the mere form of cognition into concepts, judgments, and inferences, and thereby achieving formal rules for all use of the understanding. Now if it wanted to show generally how one ought to subsume under these rules, i.e., distinguish whether something stands under them or not, this could not happen except once again through a rule. But just because this is a rule, it would demand another instruction for the power of judgment, and so it becomes clear that although the understanding is certainly capable of being instructed and equipped through rules, the power of judgment is a special talent that cannot be taught but only practiced. Thus this is also what is specific to so-called mother-wit, the lack of which cannot be made good by any school; for, although such a school can provide a limited understanding with plenty of rules borrowed from the insight of others and as it were graft these onto it, nevertheless the faculty for making use of them correctly must belong to the student himself, and in the absence of such a natural gift no rule that one might prescribe to him for this aim is safe from misuse.* A physician therefore, a judge, or a statesman, can have many fine pathological, juridical, or political rules in his head, of which he can even be a thorough teacher, and yet can easily stumble in their application, either because he is lacking in

[7] Lorraine Daston, 'Enlightenment Calculations' (1994) 21 Critical Inquiry 182; Lorraine Daston, 'Calculation and the Division of Labor, 1750-1950' (2018) 62 Bulletin of the German Historical Institute 9.

[8] You might also be reminded here of the term *casuistry,* which has a negative connotation — the over-subtle kind of argumentation that bends general principles to get the outcome that suits the reasoner in the particular instance. Indeed, it is when rules are not algorithmic but indeterminate and do not spell out how they will apply in each case, that this further work is needed, leaving the rules open to biased and self-interested application.

[9] Immanuel Kant, *Critique of Pure Reason* (Cambridge Edition of the Works of Immanuel Kant, Cambridge University Press 1781/1787/1998). This passage is quoted because of its affinity with Daston's account. Of course, Kant wrote plenty more about judgment, but in relation to his 'transcendental logic' which concerns the possibility of *a priori* knowledge (A12/B26), not the 'general logic' relevant to modern computing. See n 12.

> natural power of judgment (though not in understanding), and to be sure understands the universal *in abstracto* but cannot distinguish whether a case *in concreto* belongs under it, or also because he has not received adequate training for this judgment through examples and actual business. This is also the sole and great utility of examples: that they sharpen the power of judgment. For as far as the correctness and precision of the insight of the understanding is concerned, examples more usually do it some damage, since they only seldom adequately fulfil the condition of the rule (as *casus in terminis*) and beyond this often weaken the effort of the understanding to gain sufficient insight into rules in the universal and independently of the particular circumstances of experience...[.]
>
> * The lack of the power of judgment is that which is properly called stupidity, and such a failing is not to be helped... (A133/B72-A34/B73)

Note the points of connection with the pre-modern rules characterised by Daston: there can be general, abstract rules, but the rules alone do not pre-specify their application to concrete particulars. For this, one needs judgment, a 'power' that cannot be taught because it cannot be explicated and formulated as a set of guidelines for an apprentice to follow. You either have it or you do not, and if you have it the only way you can improve it is through practice, like muscles strengthened by lifting weights. O'Neill summarises this passage in a way quite helpful for our discussion: 'Kant insisted that we can have no algorithms for judgment, since every application of a rule would itself need supplementing with further rules.'[10]

In the light of Daston's long historical narrative about the shift towards satisfaction with a 'thin,' context-independent, algorithmic notion of rules, Kant appears as a curious figure with one foot in the past and the other far in the future. For his depiction of understanding as a logical processing plant for sensory representations, all rule governed, Kant is heralded as a forerunner of contemporary *cognitivism*, the view that thought is no more than the manipulation of symbolic structures in the mind, according to rules that ensure maximisation of a pre-specified desideratum, such as truth or well-being.[11] Thought is a kind of

[10] O'Neill, 'Abstraction, Idealization and Ideology in Ethics' (n 5), p. 58.

[11] Some philosophers who have made this connection between Kant and cognitivism are Daniel C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology* (MIT Press: Cambridge, Massachusetts 1981), ch. 7, Andrew Brook, *Kant and the Mind* (Cambridge University Press 1994), p. 12, and John Haugeland, 'Two Dogmas of Rationalism' in Zed Adams and J. Browning (eds), *Giving a damn: Essays in dialogue with John Haugeland* (MIT Press: Cambridge, Massachusetts 2016), p. 301. The following passage from the start of Kant's *Logic* shows this side of him:

> The exercise of our own powers also takes place according to certain rules which we first follow without being conscious of them, until we gradually come to cognize them through experiments and long use of our powers.
>
> Like all our powers, *the understanding* in particular is bound in its acts to rules we can investigate. Indeed, the understanding is to be regarded as the source and faculty of thinking rules generatim. (*Logic* (Dover Publications 1800/1988), p. 13)

The project of cognitive science since the mid-twentieth century — symbolic AI or 'GOFAI' being one important strand of this — has precisely been to explicate the rules of thought operational in human thinkers, and to replicate those rules in computer models.

computation, according to cognitivism, so that the ability to perform certain kinds of computation is sufficient for intelligence. On the other hand, we have here a Kant who points out the insufficiency of rule-determined reasoning (at least for 'general logic'),[12] who sees rules as generalities that do not meet particular cases unless brought to ground by this inexplicable extra something, judgment. An implication that might be drawn out here is that a computer simulation of a mind, a machine built just to execute rules of thinking, would be stupid. Insofar as a computer can only execute rules,[13] it will be a stupid machine because it has no power to see how to use the rules properly, this not being codifiable.[14]

Some people, taken with the achievements of current AI, would of course dispute this pre-emptive dismissal of computerised intelligence. Either the premise that computers can only execute coded rules might be challenged, or the appeal to a non-codifiable extra something will be rejected. On the first point, the thought here is that the paradigm of current AI is deep *learning*. These machines do not solve problems with the rules they are coded with, but are coded to train themselves to perform certain tasks. As such, they are both rule-executing and rule-creating devices.[15] I find the argument unpersuasive since all of the outputs of deep learning systems are still only ever the result of procedures carried out following coded instructions (e.g. according to the back propagation algorithm used to train artificial neural networks). Whereas the point at issue here is whether the machine can do something, namely judgment, that is radically different from the execution of rules. This would require a postulation of judgement as a strongly emergent capacity of sufficiently complex artificial neural networks. Such claims should be met with skepticism because strong emergence is controversial in metaphysics, and is, in any case, inconsistent with the physicalist ontology which is presupposed by cognitivism.[16] On the latter response, defenders of the possibility of human-like AI have always been keen to burst the bubble of pretension that says that human thought involves anything other than the codifiable processes of

[12] *Nota bene*: the stated limitation is on 'general logic', but not of Kant's proposed 'transcendental logic' which in fact can give 'precepts to the power of judgment' because 'in addition to the rule … it can at the same time indicate *a priori* the case to which the rules ought to be applied' (*Critique of Pure Reason* (B175)). Since the notion of transcendental logic is peculiar to Kant's theory of the synthetic *a priori*, but general logic is a predecessor of logic in the sense used today (e.g. the logic executed in computing machines), Kant's remarks on the deficit of general logic are quite relevant.

[13] Strictly speaking a digital computer cannot follow rules but only execute them. Rule following involves a deliberate intention to conform to some norms, whereas a computer has no such intention but is physically determined, 'to behave exactly as if it were following rules' (see John Searle, *The Rediscovery of the Mind* (MIT Press: Cambridge, Massachusetts 1992), p. 216). Shanker's interpretation of Wittgenstein's account of rule following in mathematics makes precisely this point (see Stuart Shanker, *Wittgenstein's Remarks on the Foundations of AI* (Routledge 1998)).

[14] I use the term 'codifiable' to mean 'explicable in a finite list of rules.' Pippin does draw out the anti-cognitivist implication, in response to philosophers such as Dennett (n 7, ch. 7) who find in Kant a founder of the computational theory of mind (see Robert B. Pippin, 'Kant on the Spontaneity of Mind' (1987) 17 Canadian Journal of Philosophy 449). Kant's insistence that thought is *spontaneous* stands against the proposal that thought could be replicated in a causal mechanism that executes logical rules.

[15] As Shanker (n 13, pp. 22-23) points out, the transition from fixed to self-modifying algorithms has long been taken as the transition point from ordinary computing to artificial intelligence.

[16] Timothy O'Connor, 'Emergent Properties' (2020) *The Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/archives/win2021/entries/properties-emergent.

mechanisms (see section 'Argument 1: Restriction to fixed rules'). But one piece of evidence that something equivalent to Kant's judgment is lacking in current AI systems is their failure in analogical reasoning.[17] Sound judgment of how to apply the rules to novel cases involves seeing relevant similarities from one case to another, where the form the comparison should take is left indeterminate in the particulars. Melanie Mitchell sees the lack of the ability to form analogies as a significant barrier to the invention of artificial general intelligence because of the importance of this process in concept formation.[18] This supports my claim that judgment shows no signs of emerging out of the aggregate activity of many millions of rule executing, electronic components, and that human cognition does involve some additional process, not as yet subject to mechanisation, and which may not be mechanisable.

Cantwell Smith and back to Mach

A different characterisation of judgment appears in a recent book by Brian Cantwell Smith, a philosophically minded computer scientist. Judgment stands in contrast to 'reckoning', the calculative capacity at which digital computers excel. According to Cantwell Smith, all the kinds of AI invented to date have only been reckoning devices. He warns of the danger of relying on reckoning machines for decisions that require judgment and argues that there is no way to scale up the current generation of technology (deep learning) to achieve judgment. Judgment here involves ontological commitment to a world containing objects transcending the self, a normative epistemic commitment to having one's claims answerable to states of the world, and an ethical stance:

> I reserve the term '*judgment*,' [...] for the sort of understanding I have been talking about — the understanding that is capable of taking objects to be objects, that knows the difference between appearance and reality, that is existentially committed to its own existence and to the integrity of the world as world, that is beholden to objects.
>
> Judgment is something like *phronesis*, that is, involving wisdom, prudence, even virtue.[19]

Cantwell Smith argues that a future generation of AI, if acculturated in the right way, *could* achieve judgment — a conjecture we will return to in the second part. Here it is instructive to explore more deeply the reasons why deep learning devices lack judgment. My examples will centre on automation in science and pattern recognition using this technology, but the parallels with issues around automation in other decision-making contexts, like law, should be clear.

According to Cantwell Smith, judgment requires positing that there is something beyond the 'registrations',

[17] Melanie Mitchell, 'On Crashing the Barrier of Meaning in AI' (2020) 41 AI Magazine 86.

[18] Another important connection with Kant on judgment comes via the issue concept formation. In the third *Critique,* Kant introduced a distinction between 'determinative judgment' (the sort described above), which is the ability to move appropriately from the general rule to the particular case, and 'reflective judgment', which goes in the opposite direction from particulars to general rules and concepts applicable to them. See Immanuel Kant, *The Critique of Judgement* (Oxford University Press 1790/1952), p. 18. It is significant that both forms of judgment seem to require the ability to form analogies, needed for successful navigation between the concrete/particular and abstract/general.

[19] Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press: Cambridge, Massachusetts 2019), pp. 110-11.

the surface data that a human gets from their sensory organs or are fed into an AI system:

> Yet no matter their necessity, our *registrations* are not what matters — what matters is *that which we register*. To be accountable, to hold things to account, is to know the difference between the two — and to be committed to the latter, not the former.[20]

It is revealing to compare Cantwell Smith's insistence on the importance of ontological commitment with the scientific empiricism set out by Ernst Mach over 100 years ago. According to Mach, the task of science is to order and represent the data economically, so as to make accurate predictions of new data, with the ultimate aim of instrumental control.[21] There is no question of the existence of a stable object 'beyond' or 'behind' the data stream, for entertaining this question brings unnecessary metaphysical difficulties into science. As such, a deep neural network, limited to reckoning and not judgment, is a perfect instantiation of Mach's empiricist norms.[22]

However, what is incomplete about this data-adhering 'mindset' is revealed when one considers the vulnerability of artificial neural networks to adversarial attacks. These occur when, for example, a deep convolutional neural network (DCNN) trained to classify everyday objects is presented with photographs with small perturbations (or in some cases, photographs of objects with a few specially designed stickers placed on them), which would not lead a human perceiver to alter their categorisation, but result in dramatic changes in classification for the DCNN. Initial hypotheses about adversarial vulnerability assumed that the networks were succumbing to some noisiness in their

[20] Ibid p.112.

[21] Ernst Mach, 'The Economical Nature of Physical Inquiry' in Thomas McCormack (ed), *Popular Scientific Lectures* (Open Court 1882/1895); E. C. Banks in 'The Philosophical Roots of Ernst Mach's Economy of Thought' (2004) 139 Synthese 23 summarises at p. 23:

> As it is usually understood, that doctrine [of the economy of thought] holds that scientific laws and abstract class terms are tools for compiling and organizing experience by means of the fewest possible concepts, a mastery that is useful for the prediction and control of events.

It Is striking that Mach himself defined the task of science as an optimisation problem:

> [s]cience itself, therefore, may be regarded as a minimal problem, consisting of the completest possible presentment of facts with the least possible expenditure of thought. (Ernst Mach, *The Science of Mechanics* (Open Court 1883/1919), p. 490)

and also envisaged the eventual automation of at least the mathematical procedures of science:

> Even a total disburdening of the mind can be effected in mathematical operations. This happens where operations of counting hitherto performed are symbolised by mechanical operations with signs, and our brain energy, instead of being wasted on the repetition of old operations, is spared for more important tasks. ... The drudgery of computation may even be relegated to a machine. Several different types of calculating machines are actually in practical use. The earliest of these (of any complexity) was the difference-engine of Babbage, who was familiar with the ideas here presented. (ibid p. 488).

[22] See Cameron Buckner, 'Empiricism without magic: transformational abstraction in deep convolutional neural networks' (2018) 195 Synthese 5339, for the case that deep learning is a working model of empiricist theories of the mind, more generally.

systems, that is, that the adversarial misclassifications were not connected with the learned data structures that enable successful classification. However, more recent work has shown, surprisingly, that the features in the data that cause the networks to make adversarial misclassifications, are also ones relied on in successful cases.

As Ilyas and colleagues summarise their finding, '[a]dversarial vulnerability is a direct result of our models' sensitivity to well generalizing features in the data.'[23] To appreciate this point intuitively, examine Figure 1. It shows how an adversarial image can be generated by taking an ordinary photograph of an object and superimposing a very low contrast 'noise' texture which, to the network, is diagnostic of another kind of object. When the typical texture and shape cues of different objects are combined in one figure, human perceivers overwhelmingly make the object identification on the basis of shape whereas a DCNN's identification will be determined by texture.[24] Texture is a 'well generalizing feature' in datasets comprising images of everyday objects: when the DCNN under supervised learning comes to associate each name label with the texture of those objects when presented in the training data, it can reliably use those learned texture-name associations to classify objects in images not previously presented in the training set. The texture bias of DCNNs accounts for some cases of adversarial vulnerability[25] and illustrates Ilyas et al.'s general point that adversarial vulnerability is due to the DCNN learning features of image data that are actually diagnostic for classification but would not be relied upon by a human perceiver.

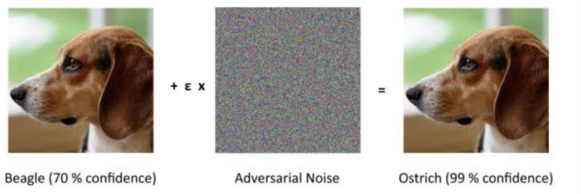

Figure 1. Illustration of object classification by DCNN altered by superposition of a very low contrast noise pattern (centre) to image of a beagle (left), resulting in the high confidence classification of the photo as an ostrich, even though it looks identical to a human observer.
(Credit: Joshua Clymer, CC-BY-SA 4.0: https://commons.wikimedia.org/w/index.php?curid=126027330)

In their commentary on Ilyas et al.'s findings, Gilmer and Hendrycks write that the problem of adversarial vulnerability is due to the tendency that a DCNN, 'latches onto superficial statistics in the data.'[26] This is a very telling remark. From the human perspective, we encounter the data as being relatable to *things,* objects which have core properties and surface properties, essential and inessential features. A furry texture should not be diagnostic for the classification of a cat because a cat can lose its fur and still *be* a cat; the presence of glasses should not determine the identification of a person, because people take glasses on and off, and this never changes who they *are.* Our thinking is constrained in this way regardless of the statistical regularities we have experienced. In my lifetime of accumulated cat data, I have never looked at a shaved cat, and yet if I did, I would still visually classify it with

[23] Andrew Ilyas and others, 'Adversarial Examples Are Not Bugs, They Are Features' (2019), available at https://arxiv.org/abs/1905.02175v4.

[24] Robert Geirhos and others, 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness' (2019), available at https://arxiv.org/abs/1811.12231.

[25] Thomas Serre, 'Deep Learning: The Good, the Bad, and the Ugly' (2019) 5 Annual Review of Vision Science 399, p. 412.

[26] Justin Gilmer and Dan Hendrycks, 'Adversarial Example Researchers Need to Expand What is Meant by "Robustness"' (2019) Distill, available at https://distill.pub/2019/advex-bugs-discussion/response-1.

the furry ones whose image statistics are so different. But one cannot make this distinction between 'superficial statistics' and deep ones unless one has an ontology of things. And so, from the perspective of the DCNN, this distinction cannot be made, for it is entirely bound to its data stream and cannot discriminate between essential and spurious features of objects; it has no dealings with objects as independently existing things, and no notions of such things.

It remains to be seen whether adversarial vulnerability presents a serious barrier to the roll-out of automated science, if restricted to the instrumental aims of prediction and control of natural phenomena. But there are already some worries that arise about the ethical peril that comes with reliance on a cognitive system restricted to reckoning and lacking judgment, commitment to a world beyond what appears in the data. In fact, long before automated science was dreamed of, Max Horkheimer raised these concerns about a purely empiricist social science.[27] In a tale about exemplary empiricists working under a dictatorship, he makes the point that optimal efficiency in data processing leads inexorably away from knowing the truth about peoples' lives, because those truths lie not in the surface data, no matter how readily those data may be deployed for prediction and control. Much has been written recently about algorithmic bias, and the danger that automated decision making will do no more than reinforce injustices pre-existing in society. But the essential problem was already outlined in Horkheimer's polemic against empiricism, which was, by the way, a response to Carnap's polemic against metaphysics[28] — the essential point being that a system of thought that is limited to the data as they stand and has no imagination to envisage a world beyond the data stream, let alone a world more just than the current one, cannot do other than reinforce a status quo. In sum, the notion of judging thinkers as ones with ontological commitment to a world that has epistemic and normative significance — where judgment stands apart from the value of accurate prediction of the data stream — can bring clarity to what is deficient about taking calculating machines to be emulators of human cognition, however super-humanly impressive those feats of calculation and prediction may be.

Points of convergence

I have presented two very different characterisations of judgment, which come with different suggestions about what AI can be said to lack. I would now like to consider what these proposals have in common. On a historical note, I mention that Horkheimer's criticism of empiricism is continuous with some points made by Kantians of his time.[29] A dimension of Kant's own project was to demonstrate the inadequacy of a purely empiricist account of cognition, showing why empirical knowledge of the natural world could not be

[27] M. Horkheimer, 'The Latest Attack on Metaphysics', *Critical Theory: Selected Essays* (Continuum 1937/2002), pp. 159-60.

[28] As discussed by H. J. Dahms, *Positivismusstreit. Die Auseinandersetzungen der Frankfurter Schuke mit dem logischen Positivismus, dem amerikanischen Pragmatismus und dem kritischen Rationalismus* (Suhrkamp 1994) and John O'Neill and Thomas Uebel, 'Horkheimer and Neurath: Restarting a Disrupted Debate' (2004) 12 European Journal of Philosophy 75, though their accounts of the dispute are quite partisan towards the logical empiricist side and indicate a lack of comprehension of the issues at stake for Horkheimer.

[29] See Horkheimer (n 27, p. 158) on his agreement with Neo-Kantianism on the cognitive 'activity which produces and organizes the facts', and more friendly criticism of Mach from Ernst Cassirer, *Substance and Function, and Einstein's Theory of Relativity* (Open Court 1910/1923), p. 261.

attributed merely to the ordered accumulation of sensory data, be they Lockean *ideas* or Hume's *impressions.* According to Kant, the passive process of receiving sensations is complemented by, and conjoined with, the active, 'spontaneous,' ordering brought by the understanding, to result in knowledge or 'cognitions.' Pure empiricism, of the sort modelled in an artificial neural network, never arrives at the concepts of substance or cause, which are indispensable for genuine cognitions of the empirical world.

Another connection between the two accounts of judgement is that they both highlight the importance of the particular and concrete cases. They both insist on the judge's sense of there being a concrete world, that must be related in some way to the general and abstract. A DCNN with its learned data structures, or any 'stupid' system with only a catalogue of rules (like the old symbolic AIs), will never get outside of the abstract formulations and relate them to concrete instances. Nor do such devices have any sense of there being something lacking. Because of the absence of a sense of a world beyond it, a DCNN does not learn from particulars in the way that people, and probably other animals do. Despite a superficial resemblance between the training process of deep nets and the case-by-case experience described by Daston as the essential complement to pre-modern rules, we should not think that the networks are ever the least bit in contact with the concrete world, for they only ever deal with input data, such as digital photographs, abstract representations of things, which are then transformed into even more abstract representational spaces. As Cantwell Smith points out, none of these representations are ever taken, by the machine, to stand in for anything beyond the representations. If they were people, we would say that these systems were trapped in the net of their own abstractions. Judgment is what gives the cognising agent free passage between the general and particular, the abstract and concrete. The ability to range in an appropriate way between these is prerequisite for good decision making in the actual world.

Could a digital computer have judgment?

Returning to Daston's narrative about the slow change in the meaning of 'rule' towards something more algorithmic — a set of instructions that pre-specifies how it should be applied to cases and unambiguous about the decisions it mandates — we can wonder here about some of the historical drivers of this shift. Perhaps it was because the Jesuits gave casuistry such a bad name, that the idea of flexible rules requiring judicious application lost its appeal in comparison with the promise of iron-clad rules that left no room for subjective manipulation. As will be familiar to you, the roll-out of automated, algorithmic decision-making in recent years came with the promise that objective machines would be fairer than subjective human beings; and of course, this promise was soon undercut with the recognition of algorithmic bias. As was concluded at the end of the section titled 'Cantwell Smith and back to Mach,' a reasoning system that has no capacity to treat new cases in somewhat unprecedented ways will only ever be a reinforcer of existing injustice. Until we have arrived at Kant's 'kingdom of ends' — a perfectly just society — reasoning systems without judgment should not be left unsupervised to make normatively charged decisions. The tautologous conclusion is that justice cannot be served without judgment. Therefore, the question to ask now is whether machines could have the capacity for judgment. Cantwell Smith answers yes, but sees no sign of it yet in existing technologies.[30] I will argue that we should not expect to see this development, at least for any AI programs running on digital computers. My argument will bring into consideration some drivers of the historical shift towards the algorithmic notion of rules (a little more credible than blaming the Jesuits), to do with the development of stable environments for industrial work, ones in which machines can flourish in spite of their inflexibility and hyper-specialisation.

Argument 1: restriction to fixed rules

My arguments for the impossibility of automated judgment only cover digital computers. All current AI runs on digital computers and there is no reason to think that AI software could somehow transcend the limitations of the hardware that implements it, developing systems with capacities that are strongly emergent in the sense of having novel causal powers. Therefore, we can ground the arguments on considerations of what kinds of machines digital computers are, getting back to the bare bones specifications that came out at their inception, from inventors such as Alan Turing, and not having to concern ourselves too much with the intricacies of machine learning as it has developed since then. A quick answer to the question, *could a digital computer have judgment?*, is to say, *no, because such machines can do no more than execute fixed rules.*

The reasons for attributing this limitation to digital computers comes with looking at what they are basically designed to do. Turing's paper in which he characterised the functionality of what is now known as the Turing Machine (which is an idealised mathematical model of computation that actual computers approximate), is a formal definition of the prior notion of 'mechanical' or 'effective' methods in mathematics.[31] Before Turing , mathematicians had a sense that some functions could be calculated by these methods that were, in other words, algorithmic — where a correct answer could always be produced by following step by step recipes employing only the basic arithmetical operations. The important point here is that Turing got to the idea of his machine by modelling the operations of the *human* computer — the kind of low skilled clerical worker whose job it was to execute algorithms. As Turing later described:

> The idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer. The human computer is supposed to be following fixed rules; he has no authority to deviate from them in any detail.[32]

Of course, a human computer *could* deviate from the fixed rules or interpret them in strange ways, but that would lead to them failing to execute the algorithm.

[30] Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment*, p.115.

[31] Alan Turing, 'On Computable Numbers, with an Application to the Entscheidungsproblem' (1936–37) 42 Proceedings of the London Mathematical Society (Series 2) 230.

[32] Alan Turing, 'Computing Machinery and Intelligence' (1950) 59 Mind 433, p. 436. Cf. '[t]hese machines are *humans* who calculate' (Wittgenstein, quoted in Shanker, *Wittgenstein's Remarks on the Foundations of AI* (n 13 p.2)). What we might gather from this puzzling remark is the thought that it is only humans who follow rules, and therefore calculate, strictly speaking, since their behaviour can be constrained by normativity, not only physical causation. Digital machines are only ever physical, causal systems. Turing's mistake was to equate the rule following behaviour of a human computer with the state transition processes instantiated in physical machines.

The digital computer is an executor of algorithms more perfect than the human computer, precisely because the only thing it can do is carry out basic procedures in a step-by-step fashion. It is a mechanism causally determined to do so, until there is some malfunction in its hardware. Since it has no other cognitive powers, it cannot get bored, lose focus, or behave eccentrically. But being so narrowly specialised, it is also condemned to be without judgment in that sense of having a view to the fit between abstract rule and concrete case.

As it happens, Turing anticipates this dismissal of the potential of computer intelligence, in the paper where he presents the 'Imitation Game,' now known as the Turing Test. This is what he calls the 'Argument from Informality of Behaviour':

> It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances. One might for instance have a rule that one is to stop when one sees a red traffic light, and to go if one sees a green one, but what if by some fault both appear together? One may perhaps decide that it is safest to stop. But some further difficulty may well arise from this decision later. To attempt to provide rules of conduct to cover every eventuality even those arising from traffic lights, appears to be impossible.[33]

Turing articulates the line of thought traced above via Daston and Kant, that in the real world of human activity, there can be no formal, codifiable rules that cover all cases and determine how should one act in every conceivable situation. Therefore, the argument goes, human cognition must encompass more than the capacity to execute algorithms, the capacity for which computers were designed to have functional equivalence with us. In response, Turing maintains that this objection to computerised intelligence rests on a conflation between the notion of an explicit rule of conduct (which cannot cover all cases) and the notion of a rule of behaviour, more like a law of nature that governs human behaviour. He points out the weak justification for the claim that rules of behaviour do not always exist. In essence, he tells us that human actions are more mechanical and rule-bound than we actually realise, because the laws of nature governing our behaviour are not transparent to us.[34] And so, the argument from restriction to fixed rules comes down to a disagreement about whether physical reality, and therefore everything in it, including behaving humans, is ultimately governed by exact, fixed laws of nature. This is a deep metaphysical issue, which cannot be settled here, and so this first line of argument is inconclusive. That said, it is important to appreciate that arguments for all cognition being computational, on which prospects for general artificial intelligence depend, rest in turn on very general metaphysical

[33] Turing, 'Computing Machinery and Intelligence', p. 452.

[34] See Shanker (n 13, pp. 45-47) for a nice account of how Turing's argument for computerised thought depends on a disputable assumption about human cognition: that at its most elementary level it involves merely mechanical processes equivalent to the basic operations of the Turing machine. It is disputable for reasons given in Section 1, namely, the conjecture that human cognition involves judgment, a capacity fundamentally different from this. Turing's idea is that we call people (and potentially computers) intelligent when they adapt their rules for operation in ways not pre-specified by an external agent. The critical point is that this form of 'learning' does not involve anything more than rearranging and complexifying a sequence of the elementary operations.

beliefs about the adequacy of exact (i.e. mathematical) description for all of natural reality.[35]

Argument 2: automatisation

A different form of argument, one that does not arrive at a metaphysical disagreement too large to be entertained in this essay, comes from widening the scope and attending to the material and historical context in which computing technology came into being. This better justifies the claim that there is a meaningful distinction between algorithmic reasoning which is simulable in a digital computer, and more context dependent, flexible forms of thought that involve capacities that go beyond rule following. In answer to the question, *could a digital computer have judgment?*, the response is now, *no, because a digital computer is a model of automatised human cognition, and so its capacities are limited to those left behind after the process of automatisation; judgment is not one of these.* The crucial concept here is 'automatisation', so I will begin with that.

It is a common observation that human beings often do things automatically. You can drive home 'on autopilot', getting from A to B without ever thinking about the route, or the process of steering, or about anything really. Once a piano piece has been learned well, the fingers just play it by themselves. Trained athletes do not have to think about the moves that they make with such expert precision. The other common observation is that when a skill, like driving, or knowing a route, or musical piece, or sports technique are new, then focussed attention is indispensable for their correct execution. But once the behaviour is established, the need to think about what you are doing floats away — that is what is meant by saying it has been automatised.

In his *Principles of Psychology*, William James wrote a chapter on habit, which describes many instances of this phenomenon. This publication came at the tail end of an intense debate about the significance of automatised behaviour, along with other curiosities such as the intelligent performances of decerebrated frogs and sleep-walking army lieutenants and recent discoveries of the nerves for reflex arcs, which together suggested, to some commentators, that automatism was the model for *all* behaviour.[36] This was the view put forward by Thomas Henry Huxley: that human beings are *conscious* automata.[37] In other words, that there is no difference in kind between actions that are consciously directed, and movements that occur automatically, without consciousness. All intelligent behaviour, Huxley argued, is just the outcome of the workings of the reflex machine that is the nervous system, and conscious thought is the ineffectual

[35] It is to be noted that Phenomenology, the 20th century philosophical school most opposed to the computational theory of mind, also rejected scientific realism and therefore Turing's premise that the natural laws described in the sciences are built into reality. Similarly, Simon Deakin and Christopher Markou, 'From Rule of Law to Legal Singularity' in Simon Deakin and Christopher Markou (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing, 2020) make the point at p. 16 that those expecting that legal reasoning can be implemented in computers are committed to a general metaphysical view of all of nature *being* mathematical. In my view, the weaker thesis that all of the phenomena in nature are in principle pre-statable in mathematical terms (that is, the fixed quantitative laws of a hypothetical, completed natural science) is all that needs to be presupposed.

[36] Roger Smith, *Free Will & the Human Sciences in Britain 1870-1910* (University of Pittsburgh Press 2016), ch. 2.

[37] T. H. Huxley, 'On the hypothesis that animals are automata, and its history' (1875) 21 The Eclectic Magazine of Foreign Literature (1844-1898) 1.

whistling that gets generated by this physical engine. James offered many good reasons to resist the conflation of these two modes of human action.[38] I will take these arguments against Huxley to be decisive, without reiterating them here. The issue I will dwell on instead is the similarity between the automatised behaviour of humans and the automatic movements of actual machines. It cannot be coincidental that the expansion of the scope of thoughts and actions deemed attributable to mechanical processes was attendant on waves of industrialisation in which machines replaced human labourers and, moreover, in which some humans had to learn to regularise and restrict their own activity in order to cohere with the operation of machines in factory settings.[39]

Collins and Kusch make a useful distinction between *mimeomorphic* and *polimorphic* actions.[40] The former are best exemplified by the kinds of behaviours that people aim to perform in exactly the same way, again and again, like a golf swing, or spray painting a chair on a Taylorist production line. The best examples of polimorphic actions are ones where exact repetition is not sought after and is even undesirable, like greeting a person, or feeding a baby. The connection to the argument of this section is that mimeomorphic actions are the ones that can profitably be automatised by a person, or replicated by a robotic machine,[41] whereas one runs a practical or social risk in 'going on autopilot', or attempting robot replacement, for polimorphic activities where flexibility and sensitivity to context (e.g. shaking this person's hand, but kissing this person on the cheek, feeding the baby just enough, but not too much) are expected. As Collins and Kusch explain, polimorphic actions take place in the human life-world, which is complex and open-ended in the kinds of events and novelties that may occur, and to which behaviour may need to be adapted. Mimeomorphic actions, in contrast, have their home in simple, delineated 'micro-worlds' — like a production line or a game of golf — in which there is a pre-defined number of things that can occur within the parameters of the task, and so all of the necessary actions can be pre-specified and matched to the circumstances which should elicit them. The important point about automatisation in humans is that it is the process of removal of the requirement that the action be thought about. This may seem paradoxical in our discussion, when the replication of thought is what is at issue. But the point is very apt: the kinds of mental operations that can become so routinised that a person can do them *unthinkingly* are the ones that

[38] William James, *Principles of Psychology*, vol. 1 (Henry Holt & Co. 1890), ch. 5; William James, 'Are We Automata?' (1879) 4 Mind 1.

[39] Of course, the automaton as a model for living body pre-dates industrialisation, a famous example being Descartes' comparison between the human body and a hydraulic statue in the *Traité de l'homme*. See Jessica Riskin, *The Restless Clock* (University of Chicago Press 2016). The critical difference is that Descartes excluded the human mind and willed behaviour from this form of explanation.

[40] Harry Collins and Martin Kusch, 'Two Kinds of Actions: A Phenomenological Study' (1995) 55 Philosophy and Phenomenological Research 799.

[41] I am not talking about sophisticated AI robots here. Industrial robots can be built to mimic particular actions without use of AI.

can be taken over by unthinking machines, which is what digital computers are.[42]

Collins and Kusch refer to Dreyfus' critique of symbolic AI in their use of the term micro-world,[43] and it is here that we can also see the connection to the issue of automated thought. My contention is that the kinds of human mental operations that can be successfully replicated in digital computers are the equivalents of human physical movements that can be automatised. In both cases they are mimeomorphic in that they involve doing the same thing again, and again, without continual adjustment to circumstance. In the case of symbolic AI this means that rule execution without judgment is sufficient because all of the potential cases have been pre-specified by the programmer, who is also the architect of the microworld. With deep learning, the situation is superficially different because they are not programmed to make decisions within a pre-specified micro-world but learn from enormous amounts of data sampled from the life-world — the giant trawls of digitised image, text or speech that go into training a network. Here we need to attend to the constraints that also apply to these systems. It is an assumption of supervised learning that the training data and test data (the data about which the network will be required to make decisions, predictions or classifications) are *independently and identically distributed* (IID). What this means is that the network learns the features of an arena larger than a micro-world — let us call it a *macro-world* — but in order for the network to perform adequately, this macro-world cannot change; its statistics must remain the same as they did during the training phase. Another way of putting this, is that unlike humans and some animals, the network can only *interpolate* within its macro-world; it cannot reliably *extrapolate,* adapting what it has learned about the statistics of one context to make its representations applicable to diverse, novel situations.[44]

All this does not imply that these machines are lacking in practical applications, or that they do not transcend human ability in other respects. The appeal and fascination of automatisation of human behaviour, and automation (i.e. machine replacement) in the labour

[42] This notion of there being an 'unthinking' use of reason is there in Kant's *Logic*, which he describes in terms of people reasoning by imitation, and 'merely historically.' This '*mechanical*' use of reason is not appropriate for jurisprudence or philosophy, Kant maintains:

> It is harmful to know some rational cognitions merely historically; this does not matter with others. For example, the navigator knows the rules of navigation historically from his tables, and that is enough for him. But if a lawyer knows jurisprudence merely historically, he is completely ruined for being truly a judge, let alone a legislator..... [I]t becomes clear that, in a certain way, one can learn philosophy without being able to philosophize. He who truly wants to become a philosopher must practice free use of reason and not merely imitative and, so to speak, mechanical use. (1800/1988, p. 26)

[43] See also Harry Collins, 'Embedded or embodied? A review of Hubert Dreyfus' *What Computers Still Can't Do*' (1996) 80 Artificial Intelligence 99.

[44] Mitchell, 'On Crashing the Barrier of Meaning in AI' (n 17). See Ali Alkhatib and Michael Bernstein, 'Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions' (2019) Paper No. 530 CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems 1, on how the performance of a deep learning network cannot be improved by use of more training data, when the statistics of the system it is learning to represent are themselves non-stationary, or when reliable decisions are sought for novel and marginal cases. I highly recommend this paper for its exposition of the limitations of deep learning in real-life (people-facing) decision making, due to lack of judgment.

force is that it can lead to tremendous gains in productivity. Think of Adam Smith's rhapsody of the pin factory, which as Daston and others have shown,[45] played an inspirational role in the history of computation: a team of specialised workers, each one repeating a step in the pin-making process all day long, can produce an order of magnitude more pins than any single labourer doing all the tasks himself, having to switch from one role to another. Machine production in factories follows on from the automatisation of human handiwork, and computation is the application of this same principle of division of labour to mental work, as was made clear by Charles Babbage, the Victorian political economist and inventor of mechanical computers.[46]

As with mimeomorphic actions, automated reasoning processes are reliant on the prior specification of a micro/macro-world in which success is possible via adherence to a set pattern of processing.[47] The critical point for Collins and Kusch, and it holds as much for deep nets confined to macro-worlds as for the symbolic AIs they discuss, is that success within a delimited world cannot translate into competence within a life-world, because its complexity is so far beyond that of a micro-world (or macro-world) — the life-world is unstable, changing, and therefore inherently uncertain.[48] We already see this limitation following the roll out of deep learning to real world situations involving human beings. Alkhatib and Bernstein describe how the task of the 'street level bureaucrat', like the police officer or judge, is to interpret the rules of the regulatory system so that appropriate decisions are made on a case-by-case basis in real life. This involves 'discretion' or 'reflexivity' — agents 'thinking about their roles as observers ... and decision–makers in a given setting, and about the impact that their decision will have'[49] — which may be compared to what I called judgment in the section titled 'Daston and Kant'. When a 'street level algorithm' faces a novel case for which its current policy (based on the training data's distribution, to which this new case does not belong), all it can do is make a decision based on the current policy, and update its policy after-the-fact *if* feedback is given. It cannot anticipate that its policy will be inadequate, and modification of policy after-the-fact is not indicative of judgment.

Conclusion

To conclude, it is the automatisation of action and cognition, the stripping away of the requirement of thought — attention, awareness, self-evaluation, and judgment — that prepares the ground for automation. It is the fact that people, whether through choice or

[45] Daston, 'Enlightenment Calculations' and 'Calculation and the Division of Labor, 1750-1950' (n 7).

[46] Charles Babbage, *On the Economy of Machinery and Manufactures* (3rd edn, Charles Knight 1835); Simon Schaffer, 'Babbage's Intelligence: Calculating Engines and the Factory System' (1994) 21 Critical Inquiry 203.

[47] I say 'pattern of processing', noting that deep learning algorithms are probabilistic and so do not give exactly the same output on repeated presentations of the same input.

[48] Similarly, Mireille Hildebrandt describes how the radical uncertainty of the future puts a limitation on code-driven law: see 'Code-driven Law: Freezing the Future and Scaling the Past' in Simon Deakin and Christopher Markou (eds) (n 35) at p. 78. One kind of uncertainty is due to the fact that human behaviour is responsive to predictions made about it. The behaviour of free agents is not fully predictable in the long run because the very act of making predictions has the potential to change people's behaviour.

[49] Alkhatib and Bernstein, 'Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions' (n 44), p. 4.

coercion, sometimes behave automatically, like machines, that makes their behaviour simulable in machines. It was observed by A. N. Whitehead over 100 years ago that, '[c]ivilization advances by extending the number of important operations which we can perform without thinking about them.'[50] This is a telling remark about the world that we have inherited from the Victorian industrialists, including Babbage. We have a form of social and economic organisation that relentlessly extends its reach by scaling up information processing at a rate that far outstrips the growth of available human cognitive resources, the actual thinking that can be employed to help conduct these tasks. This growth is only possible by regularisation and simplification of the worlds of agriculture, manufacture, education, healthcare and social interaction, so that they become more like micro/macro-worlds and less like life-worlds.

I have argued that digital computers cannot have judgment, but there is a further question that can be asked in conclusion: *do they even need judgment?* At present, automated tools cannot overcome the challenge of novelty, outliers and human unpredictability, that dealing with a life-world presents. However, if context can be fixed, through the delimitation of stable micro/macro-worlds, rules can then be explicated, neural networks trained up, and thought can be automated — which is to say, performed by a person not using their judgment or an unthinking machine. If the society of the future is a eutopia/dystopia (depending on how you look at it) in which events are never totally unprecedented, and all actions and outcomes are themselves constrained, then the bulk of regulation could be delegated to rule bound machines.[51] In which case judgment, and judges, would be obsolete.

Acknowledgments

I am most grateful to participants at the COHUBICOL seminar on Legal Effect of Code-Driven Law, at the meeting of the University of Edinburgh Philosophy, Psychology, and Informatics Group, and at the Histories of AI symposium at the University of Cambridge, for their generous engagement with the paper. I also thank Brian Cantwell Smith for his comments, the editors of *CRCL*, and three anonymous reviewers for numerous helpful suggestions. The original version of this paper began with the words, 'I know nothing about the law…'. I like to think that I have learned a little in the process of discussing this work with so many experts from different domains.

References

Alarie B, 'The path of the law: Towards legal singularity' (2016) 66 University of Toronto Law Journal 443

Alkhatib A and Bernstein M, 'Street–Level Algorithms: A Theory at the Gaps Between Policy and Decisions' (2019) Paper No. 530 CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems 1

[50] A.N. Whitehead, *An Introduction to Mathematics* (Oxford university Press 1911/1948), pp. 41-42. Compare the quotations from Mach in n 21.

[51] This conclusion chimes with the argument of Lucy, 'The Death of Law: Another Obituary' that a new regime of automated law would have to be bound together with wider changes in governance and society, involving a hollowing out of the notion of human agency in which rule of law is grounded.

Babbage C, *On the Economy of Machinery and Manufactures* (3rd edn, Charles Knight 1835)

Banks EC, 'The Philosophical Roots of Ernst Mach's Economy of Thought' (2004) 139 Synthese 23

Brook A, *Kant and the Mind* (Cambridge University Press 1994)

Buckner C, 'Empiricism without magic: transformational abstraction in deep convolutional neural networks' (2018) 195 Synthese 5339

Buckner C, 'Understanding adversarial examples requires a theory of artefacts for deep learning' (2020) 2 Nature Machine Intelligence 731

Cantwell Smith B, *The Promise of Artificial Intelligence: Reckoning and Judgment* (MIT Press: Cambridge, Massachusetts 2019)

Cassirer E, *Substance and Function, and Einstein's Theory of Relativity* (Open Court 1910/1923)

Church A, 'An Unsolvable Problem of Elementary Number Theory' (1936) 58 American Journal of Mathematics 345

Collins H and Kusch M, 'Two Kinds of Actions: A Phenomenological Study' (1995) 55 Philosophy and Phenomenological Research 799

Collins H, 'Embedded or embodied? A review of Hubert Dreyfus' *What Computers Still Can't Do'* (1996) 80 Artificial Intelligence 99

Dahms HJ, *Positivismusstreit. Die Auseinandersetzungen der Frankfurter Schuke mit dem logischen Positivismus, dem amerikanischen Pragmatismus und dem kritischen Rationalismus.* (Suhrkamp 1994)

Daston L, 'Calculation and the Division of Labor, 1750-1950' (2018) 62 Bulletin of the German Historical Institute 9

—, 'Enlightenment Calculations' (1994) 21 Critical Inquiry 182

—, *Rules: A Short History of What We Live By* (Princeton University Press 2022)

Deakin S and Markou C, 'From Rule of Law to Legal Singularity' in Deakin S and Markou C (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing, 2020)

Dennett DC, *Brainstorms: Philosophical Essays on Mind and Psychology* (MIT Press: Cambridge, Massachusetts 1981)

Geirhos R and others, 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness' (2019) https://arxiv.org/abs/1811.12231

Gilmer J and Hendrycks D, 'Adversarial Example Researchers Need to Expand What is Meant by "Robustness"' (2019) Distill

Haugeland J, 'Two Dogmas of Rationalism' in Adams Z and Browning J (eds), *Giving a damn: Essays in dialogue with John Haugeland* (MIT Press: Cambridge, Massachusetts 2016)

Hildebrandt M, 'Code-driven Law: Freezing the Future and Scaling the Past' in Deakin S and Markou C (eds), *Is Law Computable?: Critical Perspectives on Law and Artificial Intelligence* (Hart Publishing, 2020)

—, 'Law as computation in the era of artificial legal intelligence: Speaking law to the power of statistics' (2018) 68 University of Toronto Law Journal 12

Horkheimer M, 'The Latest Attack on Metaphysics,' *Critical Theory: Selected Essays* (Continuum 1937/2002)

Huxley TH, 'On the hypothesis that animals are automata, and its history' (1875) 21 The Eclectic Magazine of Foreign Literature (1844-1898) 1

Ilyas A and others, 'Adversarial Examples Are Not Bugs, They Are Features' (2019) https://arxiv.org/abs/1905.02175v4

James W, 'Are We Automata?' (1879) 4 Mind 1

—, *Principles of Psychology*, vol 1 (Henry Holt & Co. 1890)

Kant I, *Critique of Pure Reason* (Cambridge Edition of the Works of Immanuel Kant, Cambridge University Press 1781/1787/1998)

—, *Logic* (Dover Publications 1800/1988)

—, *The Critique of Judgement* (Oxford University Press 1790/1952)

Lucy W, 'The Death of Law: Another Obituary' (2022) 81 Cambridge Law Journal 109

Mach E, 'The Economical Nature of Physical Inquiry' in McCormack T (ed), *Popular Scientific Lectures* (Open Court 1882/1895)

Mach E, *The Science of Mechanics* (Open Court 1883/1919)

Mitchell M, 'Abstraction and Analogy-Making in Artificial Intelligence' (2021) https://arxiv.org/abs/2102.10717'

—, 'On Crashing the Barrier of Meaning in AI' (2020) 41 AI Magazine 86

O'Connor T, 'Emergent Properties' (2020) *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/archives/win2021/entries/properties-emergent/

O'Neill J and Uebel T, 'Horkheimer and Neurath: Restarting a Disrupted Debate' (2004) 12 European Journal of Philosophy 75

O'Neill O, 'Abstraction, Idealization and Ideology in Ethics' (1987) 22 Moral Philosophy and Contemporary Problems Royal Institute of Philosophy Supplements 55

Pippin RB, 'Kant on the Spontaneity of Mind' (1987) 17 Canadian Journal of Philosophy 449

Riskin J, *The Restless Clock* (University of Chicago Press 2016)

Schaffer S, 'Babbage's Intelligence: Calculating Engines and the Factory System' (1994) 21 Critical Inquiry 203

Searle J, *The Rediscovery of the Mind* (MIT Press: Cambridge, Massachusetts 1992)

Serre T, 'Deep Learning: The Good, the Bad, and the Ugly' (2019) 5 Annual Review of Vision Science 399

Shanker S, *Wittgenstein's Remarks on the Foundations of AI* (Routledge 1998)

Smith R, *Free Will & the Human Sciences in Britain 1870-1910* (University of Pittsburgh Press 2016)

Turing A, 'Computing Machinery and Intelligence' (1950) 59 Mind 433

Turing A, 'On Computable Numbers, with an Application to the Entscheidungsproblem' (1936–37) 42 Proceedings of the London Mathematical Society (Series 2) 230

Whitehead AN, *An Introduction to Mathematics* (Oxford university Press 1911/1948)

A reply: Judgment: between reckonability and ineffability (and an observation about rules)

William Lucy • Law School, Durham University, w.n.lucy@durham.ac.uk

Mazviita Chirimuuta's admirable paper on *Rules, judgment and mechanisation* occupies the middle ground — possibly also a no man's land — between two alternative views about the nature of judgment. On one side we have cognitivism, which 'dominates cognitive science and is foundational to AI' (p. 2). It holds that '[t]hought is a kind of computation . . . so that the ability to perform certain kinds of computation is sufficient for intelligence' (pp. 5-6). Judgment, on this view, is undeniably a matter of rule following and is therefore formalisable or codifiable. On the other side there is a roughly Kantian view, holding that judgment can be neither completely determined nor explained by 'rule-determined reasoning' (p. 6). Rather, it 'sees rules as generalities that do not meet particular cases unless brought to ground by [an] (...) inexplicable extra something' (ibid). That something is judgment and it is 'non-codifiable' (p. 6).

Taking the middle ground — walking into no man's land — risks attracting fire from both sides. I am sure that cognitivists, always keen to 'burst the bubble of pretension that says human thought involves anything other than the codifiable processes of mechanisms' (ibid), will have their sights set on Chirimuuta's *via media*, and I do not intend to join them. Indeed, my hunch is the same as that unpacked and defended by Chirimuuta, that the centre ground is intellectually the right place. But there is a risk in that space that goes beyond having to defend oneself on both flanks, namely, that we make judgment seem ineffable and thus inflate the bubble of pretension. To the tough-minded, claims like Kant's, that '[g]eneral logic contains no precepts at all for the power of judgment, and moreover cannot contain them' (p. 4), can look mystifying, seemingly obfuscatory. They resemble those alleged defences of mercy (or equity) in which we are told that, in making a merciful judgment, we enter into 'the unanswering void of particularity, the realm of love, about which only mystical, poetic things can be said.'[1] If this is the best we can do in defence of merciful judgment, then we might wonder whether or not the defence is actually an indictment.[2] Similarly, if our only defence of non-calculable, non-reckonable judgment makes it ineffable and mysterious, then we might worry that that defence dishonours rather than burnishes the notion.

What to do? A careful reading of Kant on judgment, which is a key background refrain of Chirimuuta's paper, is an obvious corrective. So, too, is engagement with the rich tradition of relatively recent English-language work about practical reason and reasoning

[1] Michael J Detmold, 'Law as Practical Reason' (1989) 48 Cambridge Law Journal 436-71, p. 457.

[2] For discussion, see William Lucy, *Law's Judgement* (Hart/Bloomsbury: Oxford 2017), pp. 116-121.

initiated by Stephen Toulmin.[3] Although some have sought to reduce Toulmin's account of the nature of practical arguments and reasoning to a replicable schema that might inform automated decision procedures, I invoke it here instead as marking the modern 'rediscovery' of practical reason: the more or less implicit impetus behind the tidal wave of contemporary work on the topic.[4] That work, in the main, reminds us not just of the intricacies involved in, but also of the quotidian nature of, non-reckonable, non-calculable judgment. Once we see that such judgment is ubiquitous, it will surely lose whatever mystery or ineffability it might otherwise have.

Judgment is, of course, ubiquitous in the law and not just within the context of adjudication. And Chirimuuta, by utilising a distinction illuminated by Lorraine Daston's marvellous book,[5] reminds us of what might be an interesting peculiarity of common law adjudication and judgment. In Chirimuuta's characterisation, Daston tells us that 'rule' 'originally meant something like a model case or a paradigm — a standard for emulation' (p. 3). By contrast, '[t]he current notion of rules [regards them] as, ideally, exceptionless and infallible codes for conduct, (...) epitomised by modern arithmetical algorithms, like the one for long division' (p. 3). The old conception of rule looks a lot like the way in which lawyers and judges in common law jurisdictions regarded — and still regard — precedents. Certainly, precedents have never functioned in those jurisdictions like rules on Daston's current conception: they have always been regarded as being in principle open to revision as the common law practice of 'distinguishing,' combined with the obvious leeway in the process of determining what is *ratio* and what is *obiter* in judicial decisions, attest.[6]

The most they can be is therefore a kind of model or paradigm whose application is by no means automatic but, as jurists are often wont to say, a matter of judgment. Furthermore, judgments as to the weight and standing of a precedent can change, sometimes quite radically. A not particularly unusual instance is the judicial *volte-face*, in English law, as to the correct answer to the question of whether or not the maker of an inaccurate statement owes a non-contractual duty of care to a party who relied upon the statement and suffered economic loss as a result. This question was answered unambiguously in the negative by the majority judgment of the England and Wales Court of Appeal in *Candler* v. *Crane, Christmas & Co.* in 1951.[7] However, in 1963 the House of Lords in *Hedley Byrne*

[3] See Toulmin's *An Examination of the Place of Reason in Ethics* (Cambridge University Press 1952) and *The Uses of Argument* (Cambridge University Press 1958).

[4] See, for example: Robert Alexy, A Theory of Legal Argumentation (Clarendon Press: Oxford 1989); Garrett Cullity and Berys Gaut (eds), Ethics and Practical Reason (Clarendon Press: Oxford 1997); Neil MacCormick, Practical Reason in Law and Morality (Clarendon Press: Oxford 2008); Elijah Millgram (ed), Varieties of Practical Reasoning (MIT Press: Cambridge, Massachusetts 2001); T M Scanlon, Being Realistic About Reasons (Clarendon Press: Oxford 2014); J David Velleman, The Possibility of Practical Reason (Clarendon Press: Oxford 2000).

[5] Lorraine Daston, *Rules: A Short History of What We Live By* (Princeton University Press 2022).

[6] See Neil MacCormick, 'Why Cases Have Rationes and What These Are' in Goldstein, L (ed) *Precedent in Law* (Clarendon Press: Oxford 1987), pp. 155-182 (also published as ch. 8 of Neil MacCormick, *Rhetoric and the Rule of Law* (Clarendon Press: Oxford 2005)), and Grant Lamond, 'Precedent and Analogy in Legal Reasoning' (2006) *The Stanford Encyclopedia of Philosophy*, available at https://plato.stanford.edu/entries/legal-reas-prec.

[7] [1951] 2 KB 164.

& Co., Ltd. v. Heller & Partners, Ltd.[8] gave an affirmative answer to that question. At some point in the intervening decade (and not least by the time the House of Lords granted leave to appeal from the Court of Appeal decision in *Hedley*[9]), the standing and weight of *Candler* had become problematic.

Of course, the reasons used by the court in *Hedley* to set aside *Candler* might be formalisable and thus converted into a rule-script that all, including AI decision-makers, could follow. But, if we accept, as I do, Chirimuuta's arguments that that seems unlikely with regard to judgment in general, then it is also equally unlikely with regard to changes in judgments. There is perhaps something going on here which, while neither mysterious nor ineffable, is not as straightforward as following a recipe. Nor does it seem likely that the process can be captured or understood by those of our existing AI technologies which purport to predict adjudicative outcomes (for an overview, see the Typology of Legal Technologies at https://publications.cohubicol.com/typology/).

References

Alexy R, *A Theory of Legal Argumentation* (Clarendon Press: Oxford 1989)

Cullity G and Gaut B (eds), *Ethics and Practical Reason* (Clarendon Press: Oxford 1997)

Daston L, *Rules: A Short History of What We Live By* (Princeton University Press 2022)

Detmold M J, (1989) 48 'Law as Practical Reason,' Cambridge Law Journal 436-71

Diver L and others, 'Typology of Legal Technologies' (Counting as a Human Being in the Era of Computational Law (COHUBICOL) 2022) https://publications.cohubicol.com/typology/

Lamond G, 'Precedent and Analogy in Legal Reasoning' (2006) *The Stanford Encyclopedia of Philosophy* https://plato.stanford.edu/entries/legal-reas-prec/

Lucy W, *Law's Judgement* (Hart/Bloomsbury: Oxford 2017)

MacCormick N, *Practical Reason in Law and Morality* (Clarendon Press: Oxford 2008)

—, 'Why Cases Have Rationes and What These Are' in Goldstein, L (ed) *Precedent in Law* (Clarendon Press: Oxford 1987) 155-182 (also published as ch. 8 of MacCormick, N, *Rhetoric and the Rule of Law* (Clarendon Press: Oxford 2005))

Millgram E (ed) *Varieties of Practical Reasoning* (MIT Press: Cambridge, Massachusetts 2001)

Scanlon T M, *Being Realistic About Reasons* (Clarendon Press: Oxford 2014)

Toulmin S E, *An Examination of the Place of Reason in Ethics* (Cambridge University Press 1952)

—, *The Uses of Argument* (Cambridge University Press 1958)

Velleman J D, *The Possibility of Practical Reason* (Clarendon Press: Oxford 2000)

[8] [1963] 2 All ER 575 (judgment delivered 28 May 1963).

[9] [1961] 3 All ER 891 (judgment delivered 18 October 1961).

Author's response: A response to William Lucy

Mazviita Chirimuuta

Firstly, I thank William Lucy for his appreciative remarks on my essay. He characterises my account of judgment as a *via media* between cognitivists and those who take judgment to be mysterious and irrational. An incautious reader of Kant might put him in that latter camp but, as Lucy points out, that is not a good representation of Kant's discussions of the discernment that only comes with effort and talent, exemplified by professionals, such as lawyers and judges, who must think hard about how novel, particular cases relate to established norms and precedents.

Instead of the image of Kantian judgment standing in between the extremes of cognitivist and anti-rationalist characterisations of thought, I prefer to invoke Max Horkheimer's metaphor of the *Eclipse of Reason*. In the book bearing that title in English, but called in the German edition, *Critique of Instrumental Reason* [Zur Kritik der instrumentellen Vernunft], Horkheimer takes issue with the late modern conception of rationality that encompasses only instrumental, means-end reasoning. The result is that the very idea that we can rationally assess and debate our purposes is put in the shade. The process of choosing values and ends looks only like opting for some subjective preferences, for which no rational scrutiny is appropriate.

We are in a parallel situation with respect to the eclipse of judgment by the form of reasoning that, as it happens, can be mechanised. Thinking that does not conform to this conception of reasoning, is either ignored or classified as too obscure and arbitrary to be bothered with. Daston's history shows how the 'demotion of judgment from exercise in reason to indulgence in darkling subjectivity.... [is] part of the modern history of rationality'.[1] A pertinent question is the relationship between the reduction of reason to instrumental rationality, described by Horkheimer, and its reduction to mechanisable operations delegated to computers. I suspect these are alternative descriptions of the same phenomenon. We should appreciate here why the connection is important to the issue of computational law.

In Kant's theory of practical reason, a key notion is *autonomy* — self-legislation. This is the ability not only to by bound by laws or norms, but to devise norms that do not merely express subjective preferences. For Kant, autonomy is integral to our status as humans *qua* rational beings, and as Lucy describes, the form of reasoning employed in case law, with its possibility of *volte face,* depends on a certain agency with respect to the formation of norms. Situations occur which are literally unprecedented. The open texture of law, as traditionally practiced, reflects a philosophical appreciation of the open texture of human lives and cultures, and our autonomy. Legal systems are mirrors of ourselves. What legal scholars and practitioners must appreciate is that a transition to legal reasoning on the model of computation, which in my view excludes

[1] Lorraine Daston, *Rules: A Short History of What We Live By* (Princeton University Press 2022), p. 15.

judgment, must come with a shift in our self-conception as human beings. This is not any old discussion about the pros and cons of automation.

References

Daston L, *Rules: A Short History of What We Live By* (Princeton University Press 2022)

Horkheimer M, *Eclipse of Reason* (Martino Publishing 1947/2013)