Diachronic interpretability and machine learning systems

Delacroix, Sylvie

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

Diachronic interpretability and machine learning systems

Sylvie Delacroix *

Abstract

If a system is interpretable today, why would it not be as interpretable in five or ten years time? Years of societal transformations can negatively impact the interpretability of some machine learning (ML) systems for two types of reasons. These two types of reasons are rooted in a truism: interpretability requires both an interpretable object and a subject capable of interpretation. This object *versus* subject perspective ties in with distinct rationales for interpretable systems: generalisability and contestability. On the generalisability front, when it comes to ascertaining whether the accuracy of some ML model holds beyond the training data, a variety of transparency and explainability strategies have been put forward. These strategies can make us blind to the fact that what an ML system has learned may produce helpful insights when deployed in real-life contexts this year yet become useless faced with next year's socially transformed cohort. On the contestability front, ethically and legally significant practices presuppose the continuous, uncertain (re)articulation of conflicting values. Without our continued drive to call for better ways of doing things, these discursive practices would wither away. To retain such a collective ability calls for a change in the way we articulate interpretability requirements for systems deployed in ethically and legally significant contexts: we need to build systems whose outputs we are capable of contesting today, as well as in five years' time. This calls for what I call 'ensemble contestability' features.

Keywords: interpretability, ethical agency, concept drift, contestability, fairness

Replier: Zachary C. Lipton, Carnegie Mellon University. zlipton@cmu.edu.

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

* Professor in Law & Ethics, University of Birmingham. s.delacroix@bham.ac.uk.

Introduction

The title of this paper may sound puzzling. That is not only because the term 'diachronic' is infrequently used. As an adjective, it is defined as 'concerned with the way in which something (...) has developed and evolved through time (often contrasted with 'synchronic')'.[1] At a more fundamental level, the idea that the interpretability of machine learning (ML) systems may 'develop and evolve through time' will appear foreign at best, unnecessarily distracting at worst; it is difficult enough to delineate with precision what the interpretability of such systems entails. Depending on both the objectives and the kind of system at stake, efforts to design 'interpretable' ML systems will translate into very different strategies.

One major driver behind demands for interpretable ML systems is the need to be able to assess the extent to which a trained model is likely to yield insights whose validity perdures beyond the training data. For real-world tasks, this means checking whether a model's prediction accuracy not only holds on validation data (which may be taken from the same distribution as the training set) but also on testing data. The relative unpredictability of real-life, incoming data flows[2] makes such generalisability assessments both vital and thorny, especially since the very reliance upon the trained model can skew the incoming, testing data.[3] The tools deemed suitable for such generalisability assessments will vary, depending on the ML system's degree of complexity.[4] For systems that can be decomposed into a series of meaningful steps leading from inputs to outputs, a variety of transparency strategies have been put forward.[5] When a model's degree of complexity makes transparency strategies less likely to be fruitful, a range of *post hoc* explanation tools have been proposed. These tools rely on a statistical analysis of the way in which certain input characteristics relate to the outputs produced by the system.[6]

The development of 'interpretable' ML systems also proceeds from an altogether different rationale. When I am — or my fellow citizens are — at the receiving end of a decision based on the insights yielded by an ML algorithm, the fact that the trained model has been deemed to produce dependable predictions in a dynamic, real-world setting may be deemed welcome, yet not enough. Here the concern(s) at stake proceed from the nature of the practices within which the trained model is deployed. Agents taking part in practices that fulfil ethically or legally significant functions are expected to be able to give and ask for reasons underlying their decisions. The demand for and provision of such reasons is only occasionally driven by a need for explanation (and even less frequently by an endeavour to make such practices 'transparent'). What is most often at stake is the ability to contest discursively — rather than through force — each other's stances or decisions. Why should this ability disappear when the 'other' we are confronted with is an ML agent?

The logic behind this particular rationale for interpretable ML systems leads to different requirements from those that proceed from the need to ascertain the generalisability of trained models. Whereas transparency and/or explainability strategies sometimes prove adequate[7] when the aim is to ascertain generalisability, they are inadequate when the aim is to develop systems that are sufficiently contestable, such that they do not end up compromising the very practices in which they play a part.

To illustrate this contrast, the section *Case studies* below outlines various ways of understanding interpretability requirements within ethically or legally significant con-

1 Definition taken from Angus Stevenson, *Oxford Dictionary of English* (Oxford University Press 2010).

2 As an example, the characteristics of the cohort of college applicants of 2021 may have changed in unpredictable ways compared to the characteristics of earlier cohorts used as training data.

3 Zachary C Lipton, 'The Mythos of Interpretability: In machine learning, the concept of interpretability is both important and slippery' (2018) 16(3) ACM Queue 1.

4 Lipton analyses these tools' relative suitability for different types of ML systems in ibid.

5 Though increasingly these transparency strategies are pursued in the context of less decomposable systems too, see for instance Sarah Tan and others, 'Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation' (AIES '18, Association for Computing Machinery 2018).

6 This statistical analysis can be 'global' or focus on one input only i.e. 'local'.

7 Depending on the nature of the ML model, see Lipton (n 3).

texts through concrete examples. Each example encapsulates an increasing degree of concern for the fostering of collective contestability. On the back of these examples, the next section argues that what poor — or overly 'synchronic' — takes on interpretability threaten is not so much the freedom but rather the *power* to question a system's embedded values.[8] That power is at the heart of our capacity for normative agency. Without that capacity there would be no ethically or legally significant practices to speak of. The section *Mistaking the tree for the forest* traces the roots of this dominant, synchronic take on interpretability to an enduring fascination with the 'fairness' of ML systems.

It is in part because of this over-emphasis on one particular value (which is often artificially 'frozen' to yield a seemingly univocal definition) that relatively little attention has been paid to the challenges that stem from the passage of time — and the societal changes that often result. The section *Societal change: a passive and active challenge for ethically and legally significant ML* distinguishes between two kinds of challenges, which pertain to each of the two rationales for ML interpretability. On the generalisability front, we know that large societal changes can mean that what an ML model has learned becomes inadequate, yet poor societal and regulatory awareness mean monitoring methods are hardly debated. Worryingly, this issue is also ignored in the otherwise relevant provisions of the proposed EU 'Artificial Intelligence Act'.[9] On the contestability front, as a distinct rationale for ML interpretability, it is all about ensuring we retain our own capacity to effect changes in the practices within which ML systems are deployed. This capacity can be fostered, rather than hindered, by relying on what I call 'ensemble contestability features'. These features allow end-users to compare counterfactual or 'shadow' models. This comparison in turn facilitates a critical grasp of the pertinent parameters (and objectives). The final section outlines this ensemble contestability proposal and draws a parallel with the resolution of disagreements among human experts, when both the objectives and ways of fulfilling these objectives are uncertain and contested.

Case studies

This section compares four different ways of understanding interpretability requirements, for two different kinds of ML systems. One is not yet operational; it is trained to predict educational needs and personalise remote content delivery and assignments of high school students.[10] This example is chosen because in this instance desired outcomes are not necessarily known. The other is trained to predict the future performance of a job applicant.[11] Those two systems are relied on by a course coordinator and by a recruitment team respectively. The four different ways of understanding interpretability requirements (from 1 to 4) reflect different degrees of concern for normative agency and a progressive shift in emphasis. While (1) and (2) could only ever facilitate synchronic interpretability, (3) and (4) may be seen as driven by a concern to facilitate diachronic interpretability.

1a. Sophie is not offered the job she applied for. Her rejection letter includes a link to a website that can generate a 'simple' approximation of the decision-making algorithm that informed the recruitment team's decision (just how much that algorithm was relied on is left unsaid). To produce such a 'local' approximation of the overall model, one needs to narrow down the domain (and extent) of the variables deemed relevant. This narrowing-down process is

[8] On such embedded values, see Helen Nissenbaum, 'How computer systems embody values' 34(3) Computer 120; Mary Flanagan, Daniel C Howe, and Helen Nissenbaum, 'Embodying Values in Technology: Theory and Practice' in Jeroen van den Hoven and John Weckert (eds), *Information Technology and Moral Philosophy* (Cambridge University Press 2008).

[9] European Commission Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence act) and amending certain Union legislative acts 2021.

[10] For a survey of the potential (and pitfalls) inherent in data-intensive technologies deployed within an education context, see Mireille Hildebrandt, 'Learning as a Machine. Crossovers Between Humans and Machines' (2017) 4(1) Journal of Learning Analytics 6.

[11] 'The terminology here can be confusing since there are actually two algorithms: one algorithm (the 'screener') that for every potential applicant produces an evaluative score (such as an estimate of future performance); and another algorithm (the 'trainer') that uses data to produce the screener that best optimises some objective function'. Jon Kleinberg and others, 'Discrimination in the Age of Algorithms' (2018) 10 Journal of Legal Analysis 113.

judgment-based and will significantly affect the substance, accuracy[12] and clarity of the explanation.[13]

1b. Because of some medical condition that makes school attendance difficult, Sophie's twin children, Alexa and Paul, are following a remote high school learning program that claims to deliver superior results compared to traditional remote schooling. It does so by optimising the timing and selection of educational content based on Alexa and Paul's respective profiles. When Alexa asks why she gets far less challenging science lessons than Paul, the course coordinator sends her a link that is very similar to the one Sophie received in relation to her job application.

Assessment: Neither Sophie nor Alexa know what to make of the 'explanation' they have been given. Neither of these explanations empower them to effect change. Sophie, in particular, is increasingly wary and resolves to search for jobs whose recruitment process does not lean on automated profiling tools.

2a. Sophie's rejection letter includes a reference to factors that had a significant influence on the recruitment algorithm's performance prediction. The letter suggests that had these factors been different, the outcome would probably have been different. These factors include her history of frequent absences at school, as well as her psychological tests results, which suggest she is particularly risk-averse. While she cannot change her history of absences, she can contact the school to ask for those recorded absences to be accompanied by an explanation. She can also ponder which of her answers led to the 'risk-averse' label, as she does not feel that is a particularly accurate trait attribution. 2b. In response to her query about science lessons, the course coordinator explains to Alexa that had her recent psychological test results been different, she would probably have been given harder materials. As it stands, her psychological profile suggests that she does not respond well to very challenging content or tasks. As such, the content and tasks she is assigned are designed to be just marginally harder than what she has successfully achieved previously. The course coordinator also includes an (anonymised) reference to other past pupils who were given very similar science content, so that she may compare herself and possibly reach out to them.

Assessment: While the counterfactual explanation provided in 2a is meant to facilitate some degree of agency, in practice it is unclear how helpful it will be for Sophie. The explanation provided suggests that none of her protected characteristics (such as race and gender) played a role, yet she may still deem the decision unfair. If the system's giving weight to her school absences is unjustified (due to some pressing circumstances, for instance), Sophie has no way of expressing her disagreement or giving feedback to that effect. The counterfactual explanation given in 2b is problematic not only because it does nothing to improve Alexa's degree of agency within her education program but also because it may become a harmful, self-fulfilling prophecy.

3a. Sophie's rejection letter contains a link to a webpage that gives her a snapshot of four different systems. Each system has the same core learning algorithm, but the latter is either trained on a different dataset (system W) or constrained in the factors which it can take into account in different ways. These key differences are documented and the link enables her to see the different 'performance predictions' generated by each system. While system X rules out any reference to data prior to the award of A-levels — hence school absences cannot make any difference — system Y rules out reliance on psychological tests of any kind. The letter explains that the recruitment team favoured the prediction produced by system Z (system Z takes both school absences and psychological tests into account) which has been shown to lead to reliable and long-term hires for that company.

3b. The course coordinator refers Alexa to a set of three different education personalisation algorithms. In contrast to the system (Z) favoured by the course coordinator, system

[12] This accuracy issue is sometimes referred to as the 'fidelity' of an explanatory approximation of a machine learning system. The degree of fidelity depends on how well it mimics the system it is meant to explain. See notably Alan B Tickle and others, 'The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks' (1998) 9(6) IEEE Transactions on Neural Networks 1057.

[13] The difficulties and limitations inherent in the production of such 'local' models are outlined in detail in Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31(2) Harvard Journal of Law & Technology 841.

X does not allow psychological tests results to influence the selection of content and tasks. System Y comes in two versions: one trained on data generated by girls-only schools and one trained on data generated by boys-only schools. Alexa is struck by the very different content recommendations issued by each system and starts questioning the extent to which she is well served by the 'not too challenging' approach.

Assessment: Sophie may argue that reliance on pre-A levels data is inappropriate, since it fails to respect the fact that minors need to be able to make mistakes (and learn from them) without long-term repercussions for their life choices. She may write a letter to the employer to that effect and possibly start a campaign to outlaw reliance on such data in the context of job recruitment. Alexa may ask her course coordinator to switch to the system trained on data generated by girls-only schools for a while, to see how she fares, as she suspects she is not especially averse to challenging content.

4a. Sophie's rejection letter includes a link to the same snapshot of four different systems as in 3a, but with a twist: two out of the four ML systems are built 'interactively', in that they demand regular input on the part of end-users. Recruitment teams can feed the performance of new hires back to the learning algorithm. In addition to this, all job applicants (whether successful or not) are asked to comment on whether they feel the three most 'weighty' parameters (both for the prediction which the recruitment team relied on and for another, alternative prediction produced by system W) are fairly taken into account, with a score of 1 to 10. Aside from potentially improving the system's learning performance (assuming adequate monitoring), such a method, sometimes referred to as 'interactive machine learning' or IML,[14] has the advantage of endeavouring to address the 'normative holiday' risk that will be unpacked later in this paper.

4b. Not only is Alexa's questioning of the adequacy of the 'not too challenging' science content fed back into system Z, but the students are also regularly 'switched' from one personalisation system to another. Every time this switch takes place, students are notified and asked to reflect upon the extent to which they felt adequately challenged, motivated etc. They then provide such feedback in both a formalised (scale of 1 to 10) and non-formalised way (describing their experience in their own words). Similar (but separate) feedback is open to both parents and course coordinators. Students, teachers, parents and the wider community are encouraged to discuss their views on the criteria and objectives that should drive education in discussion boards, online fora etc., which soon feed into wider societal debates.

Assessment: The rationale behind (4) rather than (3) lies in the acknowledgment that sooner or later normative fatigue will creep in. When the emotional de-sensitisation or 'loafing'[15] effect sets in, the extent to which end-users continue to gain from 'both the information embedded within explanations given by the system and the information provided by the system's transparency level'[16] diminishes.

The cause of this 'diminishing return' from information that would otherwise be useful cannot be grasped unless one pays attention to the nature of the practices within which the ML algorithm is deployed. In both of the chosen examples, the practices at stake are not only ethically and legally significant, they are also structured around conflicting values. This conflict of values is not an unfortunate and ideally resolvable characteristic of such practices. The next section outlines why.

[14] 'Although humans are an integral part of the learning process (they provide labels, rankings etc.), traditional machine learning systems used in these applications are agnostic to the fact that inputs/outputs are from/for humans. In contrast, interactive machine learning places end-users in the learning loop (end users are an integral part of the learning process), observing the result of learning and providing input meant to improve the learning outcome. Canonical applications of IML include scenarios involving humans interacting with robots to teach them to perform certain tasks, humans helping virtual agents play computer games by giving them feedback on their performance'. Wendell Wallach and Colin Allen, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press 2009).

[15] See n 21 and associated text *infra*.

[16] This is taken from a definition of interpretability as 'the level to which an agent gains, and can make use of, both the information embedded within explanations given by the system and the information provided by the system's transparency level'. This definition is put forward in Richard Tomsett and others, 'Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems' .

Freedom *versus* power to question embedded values: normative agency as a capability

A scenario in which we would be able to claim certain and undisputed knowledge of the ideal educational trajectory for a given student profile is a scenario where both our own nature and the nature of education would be unrecognisable.[17] It is because human nature is a never-ending 'work in progress' that we have a propensity to question the way things are and see how they could be made different. That we often reach divergent conclusions in the process is a problem only for those scientists in need of neatly optimisable objectives that are not open to discussion. The number of such scientists is growing as ML algorithms are deployed within an increasing number of ethically and legally significant practices. Few among such scientists worry about the extent to which the process of ironing out value conflicts for the purpose of building implementable utility functions can end up compromising the very source of those value conflicts. Fewer still will consider the extent to which this source — our capacity for normative agency — is of inherent value.

Most philosophers treat normative agency as a given. Our ability to see how the world and the way we live in it could be different and organise ways of living together accordingly, is often deemed central to our humanity. It is a cornerstone of the 'noumenal self' that structures Kant's moral philosophy. On the latter, Kantian account, while we may not always exercise our capacity for normative agency, or indeed do so rarely, we cannot lose that capacity, central as it is to our human nature. This dominant view is coming under pressure.

On the theoretical front, the growing influence of so-called 'capability approaches' to autonomy have brought renewed attention to the social, cultural and economic conditions that can compromise what Sen refers to as 'critical agency'. Aware of the variety of meanings associated with the term 'agency', Sen uses it 'in its older — and "grander" — sense as someone who acts and brings about change, and whose achievements can be judged in terms of her own values and objectives'.[18] To foster such normative agency, '[w]hat is needed is not merely freedom and power to act, but also freedom and power to question and reassess the prevailing norms and values.'[19]

On a more applied front, human-computer interaction studies have emphasised for some time the dangers inherent in our increased, uncritical reliance on systems designed to simplify our practical reasoning. Among the factors that Skitka and others hypothesise might contribute to sub-optimal decisions associated with automated decision aids, so-called 'cognitive miserliness' features prominently: 'most people will take the road of least cognitive effort, and rather than systematically analyse each decision, will use decision rules of thumb or heuristics'.[20] Automated systems will act as the latter. Skitka and others also refer to what they call 'social loafing, diffusion of responsibility and possible belief in the relative authority of computers and automated decision aids'.[21]

Reliance upon non-ambiguous systems — whose opaque, multi-objective optimisation process makes any effort of critical engagement redundant — will affect the extent to which we are made to flex our 'normative muscles' in the longer term. What if we enjoy the comforts of automated, simplified practical reasoning a bit too much, a bit too long? What was born out of efficiency and accessibility

[17] The same applies to recruitment decisions.

[18] Amartya Sen, *Development as Freedom* (Oxford University Press 2001).

[19] Jean Dreze and Amartya Sen, *India: Development and participation* (Oxford University Press 2002).

[20] Linda J Skitka, Kathleen L Mosier, and Mark Burdick, 'Does automation bias decision-making?' (1999) 51(5) International Journal of Human-Computer Studies 991.

[21] ibid.

concerns may become a 'normative holiday'[22] which we are unable[23] to bring to an end[24] for want of being able to mobilise normative muscles that have become atrophied through lack of exercise.[25]

The next section critically analyses the dangers inherent in allowing one particular value — fairness — to structure the shape of our efforts when it comes to building interpretable systems meant for ethically and legally significant practices.

Mistaking the tree for the forest: the dangers inherent in our algorithmic fairness obsession

Many of the calls to build interpretability features in systems deployed within ethically or legally significant practices proceed from concerns about the 'fairness' of such systems. The typical line of reasoning goes like this: it is because such ML systems can either perpetuate or create unfairness – in various guises[26] that we, end-users, need to be able to make sense of the outputs generated by such systems. In this case, 'making sense' entails being able to assess the 'fairness' of the outputs generated by those systems. Depending on the nature of the system at stake, this assessment is often facilitated by explainability and/or transparency features. The latter need not be high-tech. Systematically documenting the 'human' choices made by system designers can sometimes facilitate fairness assessments that may prove more reliable than those available for human decisions. Along this line, Kleinberg and others emphasise the contrast between the degree of *post hoc* scrutiny and experimentation available for such algorithms versus that available for obfuscating and unconsciously biased humans.[27] Armed with access to the training data, delineation of the decision space and choice of observable features, discrimination will be easier to prove in the case of algorithmic decisions.

Yet the excitement at the prospect of 'doing better' than humans when it comes to producing 'verifiably fair' outcomes is not without its dangers. One of them is the propensity to forget that far from being some longed-for, univocal and disambiguated ethical yardstick, fairness lends itself to varied and incompatible translations.[28] The choice of one translation over another reflects value-laden judgments that are central to ongoing political disputes. In this context, the fact that Kleinberg et al. mathematically proved that the concept of fairness gives rise to multiple, irrec-

22 This concept of 'normative holidays' is borrowed from William James, *Pragmatism: A New Name for Some Old Ways of thinking* (Pragmatism and other writings, Penguin Classics 2000) referring to 'moral holidays'. The following passage highlights its relationship to what James calls 'absolutism': '[The world of pluralism] is always vulnerable, for some part may go astray; and having no 'eternal' edition of it to draw comfort from, its partisans must always feel to some degree insecure. If, as pluralists, we grant ourselves moral holidays, they can only be provisional breathing-spells, intended to refresh us for the morrow's fight. This forms one permanent inferiority of pluralism from the pragmatic point of view. It has no saving message for incurably sick souls. Absolutism, among its other messages, has that message [...] That constitutes its chief superiority and is the source of its religious power. That is why, desiring to do it full justice, I valued its aptitude for moral-holiday giving so highly.' William James, 'The Absolute and the Strenuous Life' in *The Meaning of Truth* (Longman Green and Co 1911).

23 A commitment to avoid precisely such a 'state of dependence' was at the heart of Wilhelm von Humboldt's critique of State interventionism, which may be read as an early — perhaps the earliest — 'capability account of autonomy'.

24 In a related vein, see Hildebrandt emphasising that '[t]he elasticity, ex-centricity and ecological nature of the inner mind are what makes us human, but thereby also vulnerable to being hacked by an environment that is conducive to cognitive automation.' Mireille Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20(1) Theoretical Inquiries in Law.

25 To picture the utter state of dependence that would result from such never-ending normative holidays, the 'Wall-E' cartoon is particularly evocative: due to lack of exercise while in low gravity, ballooned humans each sipping their smoothie are simply unable to stand up (https://www.pixar.com/feature-films/walle).

26 Aside from the fundamental distinction between individual and group fairness, the latter can be translated into starkly different and often incompatible requirements. See Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian, 'The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making' (2021) 64(4) Communications of the ACM 136.

27 Kleinberg and others (n 11).

28 The disputed nature of fairness as a concept has nothing to do with its 'abstractness'. This bears emphasising, given those who argue that the problem is that 'fairness is too abstract to be completely encoded into the system'. Finale Doshi-Velez and Been Kim, 'Towards A Rigorous Science of Interpretable Machine Learning' [2017] .

oncilable requirements[29] is both good and bad news. It is good news to the extent that it hopefully puts an end to naïve and potentially dangerous efforts to distil fairness to some computable, supposedly crystallised form. It is bad news in that a mathematical proof was needed in the first place.

The other sort of danger is aptly phrased in terms of 'mistaking the tree for the forest': as a tree, formal algorithmic fairness assessments have distracted us from a range of wider problems. At a substantive level, and sitting fairly close to the tree, are questions about the role which algorithmic prediction tools should play within ethically and legally significant practices. In the criminal justice context, Mayson astutely calls for our being more 'thoughtful about what we want to learn from the past, and more honest about what we can learn from it':[30]

> If the risk that really matters is the risk of serious crime, but we have no access to data that fairly represent the incidence of it, then there is no basis for predicting serious crime at all. Nor is it acceptable to resort to predicting some other event, like "any arrest," that happens to be easier to measure. (...) If the data fairly represent the incidence of serious crime, however, the place to redress racial disparity is not in the measurement of risk, but in the response to it. Risk assessment must reflect the past; it need not dictate the future.[31]

The last sentence is key. Within human affairs, predictions are necessarily based upon past experience. Given the limits inherent in human, finite and biased experience, ML systems can be trained to offer better prediction accuracy, provided the characteristics of those whose behaviour needs to be predicted do not change too much compared to the cohorts at the heart of the training data.

Societal change: a passive and active challenge for ethically and legally significant ML

How might the characteristics of future cohorts change? Significant societal change can be brought about by large-scale events such as a pandemic. It can also be brought about by human interventions, such as those that can be yielded by social, medical care and education frameworks, as well as the criminal justice system itself. In a recent study, Neil and Sampson analyse 'inter- and intracohort variations in becoming arrested as individuals came of age during some of the largest social changes of recent times'[32] — including the rise of mass incarceration and proactive policing:

> Societal changes have been so large that they rendered socioeconomically disadvantaged and low self-control individuals of recent cohorts nearly indistinguishable from socio-economically advantaged and high self-control individuals of cohorts born just one decade earlier.[33]

The deployment of ML systems within practices that are not only capable but tasked with effecting such momentous change — such as the criminal justice system or education — thus yields two challenges. They pertain to each of the two distinct rationales for interpretable ML that were highlighted in the introduction.

On the generalisability front, how does one anticipate — or proactively monitor — drops in predictive accuracy that are rooted in momentous societal changes such as those highlighted above? After all, it takes years to understand the nature of changes often less explicit than a pandemic. Is there a risk that the very features that are meant to enable us to grasp some of what the system has learned make us

29 Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores' . See also Alexandra Chouldechova, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' 5(2) Big data 153.

30 Sandra G Mayson, 'Bias In, Bias Out' (2019) 128(8) Yale Law Journal 2122.

31 ibid.

32 Roland Neil and Robert J Sampson, 'The Birth Lottery of History: Arrest over the Life Course of Multiple Cohorts Coming of Age, 1995–2018' (2021) 126(5) American Journal of Sociology 1127.

33 ibid.

inattentive to the extent to which what has been learned may become wholly inadequate? This is what I refer to as a 'passive' challenge, since one available — and dominant — option, when faced with the possibility of what is sometimes referred to as 'concept drift', is to do nothing. Other options include a variety of more or less sophisticated monitoring strategies.[34] The more sensitive the domain of application — such as education or criminal justice — the greater the need for such monitoring strategies, given the potential for momentous changes. Yet today there is remarkably little societal debate about the relative adequacy of these 'concept drift monitoring' strategies (or absence thereof). Worse, in its proposal for an EU 'Artificial Intelligence Act', the European Commission seems content to ignore the challenges raised by the dynamic contexts within which many ML systems of ethical or legal significance are deployed. Three provisions are of particular relevance:

> Article 42: '[t]aking into account their intended purpose, high-risk AI systems that have been trained and tested on data concerning the specific geographical, behavioural and functional setting within which they are intended to be used shall be presumed to be in compliance with the requirement set out in Article 10(4).
>
> Article 10(4): 'Training, validation and testing data sets shall take into account, to the extent required by the intended purpose, the characteristics or elements that are particular to the specific geographical, behavioural or functional setting within which the high-risk AI system is intended to be used'.
>
> Article 10(3): Training, validation and testing data sets shall be relevant, representative, free of errors and complete. They shall have the appropriate statistical properties, including, where applicable, as regards the persons or groups of persons on which the high-risk AI system is intended to be used. These characteristics of the data sets may be met at the level of individual data sets or a combination thereof'.[35]

Given that these provisions are animated by an endeavour to avoid scenarios where an 'AI system' is deployed in a setting whose characteristics are too distant from those of the training, validation and testing datasets, it is peculiar that none of these provisions consider the impact of the passage of time. In contexts like education or criminal justice, massive shifts (on the scale documented in Neil and Sampson's study)[36] can and do take place within 'geographical, behavioural and functional settings' that are otherwise in line with the characteristics of the training, validation and testing datasets. In that context, to speak of 'appropriate statistical properties' (Art. 10(3)) as if these properties did not change, exemplifies the amplitude of the 'blind spot' when it comes to societal and regulatory awareness of the challenges that stem from the passage of time (in other words, diachronic challenges).

The other, 'active' challenge that stems from the passage of time — and the societal changes that come with it — is less familiar to the ML community and rarely discussed. It stems from the fact that once deployed, an ML system becomes an agent capable of effecting – or affecting — change on a significant scale. Behind the vast literature on different ways of formalising fairness stand two fundamental questions. There is the question highlighted by Mayson: what — and how — do we want to learn from our (necessarily imperfect) past? And there is a related, yet distinct question: how much do we care about retaining some ability to 'dictate the future'? The section *Freedom versus power to question embedded values: normative agency as capability* highlighted the extent to which our capacity for normative agency can be compromised. Unless we develop robustly contestable ML systems, the answers to these questions about learning from our past and shaping our future risk being 'algorithmically set' for us. We would end up 'rote learning' from our past to such an extent as to

[34] Aside from periodically 're-fitting' or updating the static model that has become out of date, another option is to build a separate model that learns to correct the predictions from the static model based on the characteristics of the incoming, more recent data: choosing between these different options is not value-neutral and should be documented and debated for ethically and legally significant ML algorithms. For a useful overview, see Indrė Žliobaitė, 'Learning under Concept Dift: an Overview' [2010] (arXiv preprint arXiv:10104784).

[35] European Commission Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence act) and amending certain Union legislative acts 21 April 2021.

[36] Neil and Sampson (n 32).

become incapable of questioning — let alone dictating — the future.

Individual counterfactual explanations versus 'ensemble contestability' features

The limits inherent in individual counterfactual explanations

The dominance of concerns with the 'fairness' of ML systems has contributed to an overly synchronic[37] take on interpretability. This focus on building interpretable systems in order to facilitate the assessment of ML systems' fairness 'here and now' tends to yield insights that are:

(1) *Unlikely to support any endeavour to effect societal change* (as illustrated in case studies 1a, 1b, 2a and 2b). Because it bypasses the need to 'convey the internal state or logic of an algorithm',[38] the counterfactual type of explanation provided in 2a is often deemed attractive. It is meant to 'help a data subject *act* rather than merely understand',[39] since it points at what would need 'to be changed to receive a desired result in the future, based on the current decision making model'.[40] Yet the type of agency facilitated by such a counterfactual explanation is highly specific, dependent as it is on the data subject having formulated a 'desired result'.[41] Outside the realm of mortgage, insurance or job applications, there are many instances where the data subject will not know what their 'desired result' might be (from education to social networking tools). Even when they do, the punctual nature of such counterfactual explanations makes them unlikely to support the need for broader, society-wide questioning of the assumptions and parameters that inform a given system.

(2) *Individualistic*: That one negatively affected individual is in a position to usefully interpret an automated output says nothing about our retaining a collective ability to interpret the vast arrays of automated outputs whose effects are not tied to one individual in particular. To retain such a collective ability is not only a matter of preserving crucial 'interpretive resources',[42] it is also a matter of preserving our ability to discursively (re)articulate conflicting values in light of changing aspirations (as seen in the section *Freedom versus power to question embedded values: normative agency as a capability*).

Overcoming the above two limitations calls for a change in the way we articulate interpretability requirements for ML systems meant to be deployed within ethically or legally significant practices. The next section emphasises the contrast between fallibility strategies suited to contexts that are structured around a well-established objective and theoretical framework, versus contexts that are not. Ethically and legally significant practices rarely fit within the former category.

Contrasting fallibility strategies

We humans get things wrong all the time, for all sorts of reasons. We often do not notice, until others (or circumstances) set us straight. Machine learning systems are likely to get things wrong too, for all sorts of reasons. When it comes to addressing the possibility of mistakes in different contexts, we learned to develop distinct fallibility strategies well before ML systems were ever developed. De-

[37] This synchronic ambition is corroborated (or possibly made worse) by so-called 'fairness by design' aspirations. The latter tend to proceed from the idea that some systems may be designed in such a way as to 'settle' fairness concerns, thereby liberating us from the clutches of arbitrary, biased human decisions. The 'finality' implicit in 'settling fairness' is important. The logic behind 'fair by design' systems has no room for the idea that something inherently valuable may be lost were the process of system (re)-evaluation and contestation be deemed 'concluded'.

[38] Wachter, Mittelstadt, and Russell (n 13) p.5.

[39] ibid p.4.

[40] ibid p.4.

[41] Yash Goyal and others, 'Counterfactual visual explanations' discuss 'visual' counterfactual explanations; this is of particular relevance for surveillance systems relying on facial recognition algorithms.

[42] Van den Hoven for instance draws attention to the extent to which extensive reliance on automation in the justice system might end up depriving us of crucial interpretive resources, in a striking parallel with what Fricker refers to as situations of 'hermeneutical injustice'. Emilie van den Hoven, 'Hermeneutical injustice and the computational turn in law' (2021) 1(1) Journal of Cross-disciplinary Research in Computational Law.

manding explanations[43] will not always be helpful. What follows compares different strategies when it comes to resolving disagreements among human experts.

In the first instance, the objective that structures the expert's task is both clear and uncontested, and there is a well-established theoretical background. In the second instance, there is uncertainty about the underlying method, which is theoretically opaque. In a third instance, both the objectives and ways of fulfilling these objectives are not only uncertain. They are also highly contested. Asking for an 'explanation'[44] in either the second or third instance is a poor way of enabling a discussion that can improve upon that decision. Counterfactual enquiries can help in the second instance. Yet when the objectives that structure a task are contested, as in the third instance, enlarging the circle of expertise is often the only constructive option.

Clear, objective and well-established theoretical framework

It is difficult to find any example of ethically or legally significant practices that fit within this category, save perhaps for the most trivial parking-ticket-like scenario. In that case demands for explanation can and do help. At the other end of the spectrum — least trivial — 'lives at stake' applications such as the one below raise a different kind of challenge (even if it is disputable whether the latter applications really count as ethically or legally significant practices).

Take systems like Airborne aircraft collision avoidance systems (ACAS).[45] These systems are in the process of being made to rely on neural networks to make the relevant, high dimensional data easily retrievable and hence make the systems operational on aircraft. They had better not get things wrong, since they require pilots to override air traffic control instructions (unless doing so would put the plane at risk). In this case, the demand for robust safety analysis *prior* to deployment has led to the ongoing development of increasingly sophisticated mathematical verification methods.[46] Reliance on such methods is only conceivable because the task is so clearly structured around an undisputed, 'avoid collisions' objective. The fact that we have a solid theoretical grasp of aerodynamics helps too. Yet such verification methods are of little use when considering domains that either lack such solid theoretical foundations or whose task is a complex fudge of disparate and often unarticulated concerns. In short, verification methods are of little use in most ethically or legally relevant contexts.

Clear, objective and poor or absent theoretical framework

Domains of human expertise that draw upon intuitive skills rather than abstract, model-based understanding have long been the focus of the so-called 'naturalistic decision making' (NDM) tradition.[47] This tradition developed from an attempt to analyse the way fireground commanders make decisions under conditions of uncertainty and time pressure. Works within NDM studies tend to show that reliance on intuition to detect patterns of similarity between past and present situations can enable experts to perform much better than if they had systematically sought to analyse and evaluate all feasible options.[48] The following example is put forward in a bid to draw the par-

43 '[T]rying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable'. Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' [2019] Nature Machine Intelligence 206.

44 'Explanation' is helpfully defined as 'the information provided by a system to outline the cause and reason for a decision or output for a performed task'. Tomsett and others (n 16).

45 Kyle D Julian, Mykel J Kochenderfer, and Michael P Owen, 'Deep Neural Network Compression for Aircraft Collision Avoidance Systems' 42(3) Journal of Guidance, Control, and Dynamics 585.

46 The complexity stems from the need to consider the aircrafts' likely trajectories within the vast number of relevant geometric configurations: 'ACAS X uses a large machine-optimised score table for its decisions. This setting leads to a disciplined way of reaching optimal compromises between safety and operational suitability, but makes verification more difficult'. The latter uses 'hybrid systems modelling and theorem proving to formally assess the safety of ACAS X'. Jean-Baptiste Jeannin and others, 'Formal verification of ACAS X, an industrial airborne collision avoidance system' (2015).

47 Gary Klein, 'Naturalistic Decision Making' (2008) 50(3) The Journal of the Human Factors and Ergonomics Society 456; Caroline E Zsambok and Gary Klein, *Naturalistic decision making* (Psychology Press 2014).

48 '[S]imple heuristics that ignore information can be better — faster, more frugal, and more accurate — than complex strategies that use all available information'. Gerd Gigerenzer and Henry Brighton, 'Can hunches be rational' (2007) 4 Journal of Law, Economics, and Policy.

allel between such scenarios and applications such as ML-based loan triage systems. In both cases, there is no independent, given model that one may consult to explain how given inputs give rise to given outputs: the model is constructed 'on the go', as the system (or, in the case of the example below, the nurse) learns from exposure to a variety of inputs.

Crandall and Getchell-Reiter[49] have studied the intuitions that allow nurses in a neonatal intensive care unit to detect life-threatening infections even before blood tests come back positive. These intuitions draw upon tacit[50] rather than explicit knowledge. The nurses' remarkable ability is acquired through a process of gradual habituation, rather than any formalised training based on a set of rules or principles. Were these nurses to be asked to explain how or why they fear a life-threatening infection, their answer is unlikely to be helpful. Yet counterfactual explanations — 'if the skin had looked brighter, I would not have been worried' — may go some way towards teasing out key factors influencing a nurse's judgement and offer a striking parallel with the type of counterfactual explanation envisaged by Wachter and others for an ML-based loan triage tool (in both cases, the objective is clear and well defined).

Contested objectives and poor or absent theoretical framework

The vast majority of ethically or legally relevant practices are structured around a range of conflicting values. These values, and their relative prevalence, are constantly in the process of being rearticulated, both in the private, ethical sphere and in the political sphere. Education and criminal justice are two areas where ML systems are increasingly relied on as triage tools, in a bid to anticipate chances of success or recidivism. The very definition of those objectives is contested: should you rank college applicants according to anticipated college grades, according to some notion of merit or according to anticipated 'transformative potential' in a given field? Controversy also affects the delineation of the training data or the choice of observed features (such as family background) considered in the process of reaching the system's predictive score. Each of these choices will reflect value judgments. Since there is no established 'model' that could lend the admissions process some scientific credentials, none of these choices can be ruled out 'a priori' and objectives are often 'fudged' together.[51]

In practice, by the time those impacted by a decision (or indeed, those who must implement it[52]) are confronted with those systems, the choices have been made. These choices will have a large impact on decisions that will shape not just the future of those affected by the decision, but our collective future too. Case study 4 is built around an endeavour to create 'built-in' opportunities for collective feedback and debate. This debate would remain very abstract without an ability to compare the outcomes of models trained differently. This is where 'ensemble contestability' features (so-called to flag their borrowing from parts of 'ensemble models' techniques) come in.

These techniques rely on running one learning algorithm (or 'base learner') on different data subsets in parallel. Their degree of rigour depends in large part on the way in which the data subsets are selected (and subsequent outcome differences resolved): when combined with 'bootstrap sampling' methodologies,[53] these ensemble techniques can help reduce the risk of overfitting. For our distinct purposes, such ensemble techniques could be just as helpful as those relying on multiple, slightly different[54] learning algorithms which may have different constraints imposed on the optimisation process. What matters is that the resolution process is taken out: rather than combining the results of each 'base learner' (whether through 'voting', 'averaging' or otherwise), emphasis would be placed

[49] Beth Crandall and Karen Getchell-Reiter, 'Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses' (1993) 16(1) Advances in Nursing Science 42.

[50] For a study delving into the characteristics of such tacit knowledge see Nicky Priaulx, Martin Weinel, and Anthony Wrigley, 'Rethinking moral expertise' (2016) 24 Health Care Analysis 393.

[51] Diane Coyle, The tensions between explainable AI and good public policy (15 September 2020).

[52] In this case admission officers. See Tomsett and others (n 16).

[53] There are various ways of extracting those data subsets from the larger data set. 'Bootstrap sampling' or 'bagging' (which randomly draws data subsets, thus allowing one data point to potentially re-appear in several subsets) is frequently relied on.

[54] Such techniques are sometimes referred to as multiple classifier systems.

on documenting the differences/factors that lead to each of the base learners' outcomes, in an 'agonistic machine learning' spirit, to borrow Hildebrandt's phrase.[55]

By facilitating the comparison of counterfactual or 'shadow'[56] systems, such 'ensemble contestability' features would put end-users in a position where they may appreciate concretely the impact of different training datasets and/or different optimising constraints. This 'ensemble contestability' aspect would ideally be accompanied (as in case study 4) by interactive features allowing decision-subjects and those implementing decisions to 'interrogate, investigate, scrutinize the system'.[57] Again, the importance of this interactive dimension stems from the nature of the practices within which the ML agent is deployed. The (re)articulation of the conflicting values at the heart of education or criminal justice practices does not proceed ex-nihilo: it is nurtured by the 'imperfect rationalisations'[58] characteristic of our intuitive, ethical grasp of a situation. To convey what is at stake in fostering interactive contestability, the following passage from Williams' classic 'Conflict of values' is worth quoting in full:

> '[T]he public order, if it is to carry conviction, and also not to flatten human experience, has to find ways in which it can be adequately related to private sentiment, which remains more "intuitive" and open to conflict than public rules can be. For the intuitive condition is not only a state which private understanding can live with, but a state which it must have as part of its life, if that life is going to have any density or conviction and succeed in being that worthwhile kind of life which human beings lack unless they feel more than they can say, and grasp more than they can explain'.[59]

To design ML systems meant for ethically or legally significant contexts that are equipped with such interactive, 'ensemble contestability' features may sound like a tall order. As Miller and others put it: 'AI researchers [are used to] building explanatory agents for [them]selves, rather than for the intended users. But explainable AI is more likely to succeed if researchers and practitioners understand, adopt, implement, and improve models'.[60]

Hopefully this paper has shown the extent to which such design choices are not just a matter of instrumental 'success'. They are also a matter of preserving what is distinctive and inherently valuable, about those ethically and legally significant practices: at their heart are our ongoing, collective efforts to (re)articulate the kind of lives we wish to live.

Conclusion

The drive to build 'interpretable' ML systems was largely born out of what could be characterised as 'due diligence' concerns. When one writes a piece of code with intent, as a set of explicit instructions to achieve a given objective, the source of potential errors or 'mishaps' is relatively easy to trace back to the code itself (or the delineation of objectives). When, by contrast, a system is made to infer traits or predict outcomes based on what it has 'learned' from training data, system designers will rightly insist on

55 Hildebrandt, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (n 24) suggests that 'one way of protecting our privacy is to require what I call "agonistic machine learning", i.e., demanding that companies or governments that base decisions on machine learning must explore and enable alternative ways of datafying and modelling the same event, person or action'.

56 I am grateful to Eric Meissner for the suggested terminology.

57 '[U]nlike simple contestation in which disagreement or attempts to shape the decision-making process may be asynchronous, pursued through outside channels, or otherwise externalised, contestability is built into the system to support iteration on the decision-making process. This makes contestability a deep system property: the ability to interrogate, investigate, scrutinise the system throughout the process of coming to a joint decision between human and algorithm. It must surface information to the user but also support interaction with and co-construction of the decision making process'. Kristen Vaccaro and others, 'Contestability in Algorithmic Systems' (CSCW '19, Association for Computing Machinery 2019).

58 Bernard Williams, *Moral luck: Philosophical papers* (Cambridge University Press 1981).

59 ibid p. 82.

60 Tim Miller, Piers Howe, and Liz Sonenberg, 'Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences' [2017] .

having some way of grasping what the system has learned. This concern is especially pressing when those systems are meant to be deployed in real-life settings, since what was an accurate prediction tool can become useless under the impact of large-scale events and human interventions.

The impact of societal change upon ML systems deployed in real-life contexts is a well-known (if under-debated) challenge for the purpose of assessing the generalisability of such systems. Yet the enduring possibility of such societal change also brings to light an often forgotten, distinct rationale for ML interpretability in ethically or legally significant contexts. This distinct rationale has to do with our retaining the capacity to trigger change within the practices where these ML systems have been deployed. The gradual restructuring of conflicting values that typically gives rise to change within legally or ethically significant practices does not happen by 'fiat', as if we had suddenly decided that some 'value spring cleaning' was needed. This gradual restructuring is an effortful process. It depends on our continued drive to question both the way we do things as well as agents responsible for doing things that way. Once deployed in ethically or legally significant practices, ML systems become such agents; what they have learned, they have learned based on our past practices.

Being in a position to continually reassess just how much we want this past to inform our future is key to our enduring capacity to trigger societal change. To preserve this enduring capacity demands a particular kind of contestability: it is not just a matter of our being able to contest a system's outputs today. A bigger challenge is to design ways of interacting with the system that foster – rather than discourage – a vigilant perspective on the value-choices that inform the design of that system.

We have heard an awful lot about the relative merits of transparency versus *post hoc* explainability strategies when it comes to assessing the generalisability of ML systems. It is time for cross-disciplinary research to consider the relative merits of concrete ways of building collectively contestable ML systems. As a distinct rationale for ML interpretability, contestability is of relatively little significance if it is only ever envisaged in a synchronic, individualistic way ('is individual X in a position to contest this system's outputs today?'). Once it is considered from a collective and diachronic perspective, contestability demands ways of building ML systems that incentivise continuous, critical feedback over time. By facilitating the comparison of counterfactual or 'shadow' systems, the 'ensemble contestability' features put forward in this paper offer one concrete avenue for future research in this domain.

References

Chouldechova A, 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments' 5(2) Big data 153.

Coyle D, The tensions between explainable AI and good public policy (15 September 2020).

Crandall B and Getchell-Reiter K, 'Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses' (1993) 16(1) Advances in Nursing Science 42.

Doshi-Velez F and Kim B, 'Towards A Rigorous Science of Interpretable Machine Learning' [2017].

Dreze J and Sen A, *India: Development and participation* (Oxford University Press 2002).

European Commission Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial intelligence act) and amending certain Union legislative acts 2021.

Flanagan M, Howe DC, and Nissenbaum H, 'Embodying Values in Technology: Theory and Practice' in Hoven J van den and Weckert J (eds), *Information Technology and Moral Philosophy* (Cambridge University Press 2008).

Friedler SA, Scheidegger C, and Venkatasubramanian S, 'The (Im)possibility of Fairness: Different Value Systems Require Different Mechanisms For Fair Decision Making' (2021) 64(4) Communications of the ACM 136.

Gigerenzer G and Brighton H, 'Can hunches be rational' (2007) 4 Journal of Law, Economics, and Policy.

Goyal Y and others, 'Counterfactual visual explanations'.

Hildebrandt M, 'Learning as a Machine. Crossovers Between Humans and Machines' (2017) 4(1) Journal of Learning Analytics 6.

Hildebrandt M, 'Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning' (2019) 20(1) Theoretical Inquiries in Law.

Hoven E van den, 'Hermeneutical injustice and the computational turn in law' (2021) 1(1) Journal of Cross-disciplinary Research in Computational Law.

James W, 'The Absolute and the Strenuous Life' in *The Meaning of Truth* (Longman Green and Co 1911).

— *Pragmatism: A New Name for Some Old Ways of thinking* (Pragmatism and other writings, Penguin Classics 2000).

Jeannin J.-B and others, 'Formal verification of ACAS X, an industrial airborne collision avoidance system' (2015).

Julian KD, Kochenderfer MJ, and Owen MP, 'Deep Neural Network Compression for Aircraft Collision Avoidance Systems' 42(3) Journal of Guidance, Control, and Dynamics 585.

Klein G, 'Naturalistic Decision Making' (2008) 50(3) The Journal of the Human Factors and Ergonomics Society 456.

Kleinberg J, Mullainathan S, and Raghavan M, 'Inherent Trade-Offs in the Fair Determination of Risk Scores'.

Kleinberg J and others, 'Discrimination in the Age of Algorithms' (2018) 10 Journal of Legal Analysis 113.

Lipton ZC, 'The Mythos of Interpretability: In machine learning, the concept of interpretability is both important and slippery' (2018) 16(3) ACM Queue 1.

Mayson SG, 'Bias In, Bias Out' (2019) 128(8) Yale Law Journal 2122.

Miller T, Howe P, and Sonenberg L, 'Explainable AI: Beware of Inmates Running the Asylum or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences' [2017].

Neil R and Sampson RJ, 'The Birth Lottery of History: Arrest over the Life Course of Multiple Cohorts Coming of Age, 1995–2018' (2021) 126(5) American Journal of Sociology 1127.

Nissenbaum H, 'How computer systems embody values' 34(3) Computer 120.

Priaulx N, Weinel M, and Wrigley A, 'Rethinking moral expertise' (2016) 24 Health Care Analysis 393.

Rudin C, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead' [2019] Nature Machine Intelligence 206.

Sen A, *Development as Freedom* (Oxford University Press 2001).

Skitka LJ, Mosier KL, and Burdick M, 'Does automation bias decision-making?' (1999) 51(5) International Journal of Human-Computer Studies 991.

Stevenson A, *Oxford Dictionary of English* (Oxford University Press 2010).

Tan S and others, 'Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation' (AIES '18, Association for Computing Machinery 2018).

Tickle AB and others, 'The truth will come to light: directions and challenges in extracting the knowledge embedded within trained artificial neural networks' (1998) 9(6) IEEE Transactions on Neural Networks 1057.

Tomsett R and others, 'Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems'.

Vaccaro K and others, 'Contestability in Algorithmic Systems' (CSCW '19, Association for Computing Machinery 2019).

Wachter S, Mittelstadt B, and Russell C, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2017) 31(2) Harvard Journal of Law & Technology 841.

Wallach W and Allen C, *Moral Machines: Teaching Robots Right from Wrong* (Oxford University Press 2009).

Williams B, *Moral luck: Philosophical papers* (Cambridge University Press 1981).

Žliobaitė I, 'Learning under Concept Dift: an Overview' [2010] (arXiv preprint arXiv:10104784).

Zsambok CE and Klein G, *Naturalistic decision making* (Psychology Press 2014).

A reply: Diachronicity raises new questions. Interpretability offers few answers.

Zachary C. Lipton • Carnegie Mellon University, zlipton@cmu.edu.

Diachronic Interpretability and Machine Learning Systems [4] spells out several arguments concerning (i) the desiderata, nature, robustness, and adaptability of interpretable machine learning (ML) systems; (ii) factors influencing the agency that such systems might afford; and (iii) a proposal for *ensemble contestability features*. The theme of *diachronicity* (roughly, temporal dynamics) is woven throughout, emphasizing the complications that arise in real-world ML deployments as environments change, often in response to the introduction of ML systems. Delacroix argues that the dominant take on interpretable ML is overly synchronic (static), which she attributes to a misguided fixation on the fairness of ML systems.

I share Delacroix's central concern: that developing ML systems suitable for real-world deployment, whether vis-a-vis traditional performance measures or other notions of impact on societal systems, requires that we account for the dynamics by which environments evolve. Moreover, I agree with Delacroix's argument that automation threatens normative agency and the discursive practices through which values evolve. However, I would like to push back on three finer points: (a) the extent to which interpretability research has addressed the generalizability of ML systems; (b) the attributability of the synchronic view to a fixation on 'ML fairness'; and (c) the usefulness of the proposed *ensemble contestability features*.

First, some context: for four years, my lab has focused squarely on temporal dynamics problems (Delacroix's *diachronicity*) from both theoretical, empirical, and philosophical perspectives. Concerning fairness and explainability, we demonstrated that static fairness interventions can suggest policies that prove counterproductive when the slightest attention is paid to temporal dynamics [12, 3]. In two recent philosophy papers [6, 5], we argued that coherent approaches to algorithmic fairness must account for the dynamics by which proposed interventions (actually) influence the allocation of benefits and harms in society, rather than analyzing naïve idealizations of these problems from a local (single decision maker) and static (single slice of time) perspective. Moreover, our lab is not alone in recognizing the centrality of temporal dynamics, either in the societal computing [15, 9, 16] or the broader ML literature. Further on the technical side, we focus on developing ML methods resilient to distribution shift [14, 17, 19, 11, 7, 10, 8]. These problems are vexing for several reasons: (i) there are no general solutions — progress typically requires structural assumptions on how the world can change; (ii) compared to the case of static environments, empirical evaluation cannot carry us so far.

Now to our disagreements: First, Delacroix echoes claims from the interpretability literature that *post hoc* explanations hold relevance for out-of-domain generalizability. While I note the prevalence of such claims in [13], they remain unsubstantiated. In fact, I know of no concrete problem for which such methods offer any solution. Delacroix takes the erroneous claim that *post hoc* explanations address generalizability as a starting point to advocate elevating such interpretations themselves to respond to diachronicity. However this framing misses (1) that out-of-domain generalization remains a fundamentally open problem, and (2) while interpretability research has been voluminous, it has contributed little to this (or any other) problem. While I echo Delacroix's calls for a diachronic focus, I worry that *interpretability* here (as usual) serves as a device to lump together diverse problems, giving false hope of a common elixir.

Second, I disagree with Delacroix's suggestion that an enduring fascination with fairness accounts for static takes on interpretability. Arguments about the relevant merits of black box predictors (often more accurate) versus

more mechanistic models (justified by their putative *interpretability*) have raged for many decades, long predating the current dialogue about fairness in ML, which emerged, largely, in the 2010s. Most problems in statistics and ML more broadly have been framed in primarily static terms, and arguments over interpretability have long roots withing this (static) framing (consider Vladimir Vapnik's *instrumentalist* approach to predictive modeling, which diverged from the model-based approach dominant among his contemporaries in statistics [18, 2]). The discourse and scholarship on *fairness in ML* are not the root cause. Rather, the interpretable ML and fair ML literatures both inherit the static view that dominate statistics and machine learning. Moreover, while the static view is dissatisfying, and while my research aims to transcend it, academics adopting a static focus should not be dismissed so lightly. For most technical problems, we could hardly articulate the dynamic setting without understanding simplified, static versions of these problems. Additionally, any statement about non-stationary settings rests precariously on unverifiable assumptions, making them complementarily unsatisfying.

Finally, I would like to counter the suggestion that ensemble contestable features offer a solution. To summarize, the idea of ensemble contestability features seems to be that the subject of a decision might be told how their decision would be different among a set of counterfactual models, each trained either on different data or in a different fashion. The broad framing is not new. For example, others have investigated classifier stability, comparing models trained on slightly modified datasets [1]. Crucially, the important details are missing: what are the relevant subsets of the data for training counterfactual models? What other modelling decisions should be modified and how? Precisely what questions can be answered in such a fashion? And what does any of this have to do with the central problem animating the paper? Like current (static) interpretable ML, this proposal aims too wide while offering too little. Delacroix argues that *ensemble contestability features* may reduce overfitting. First, the relationship between ensembling (e.g., bootstrap aggregation and boosting) and generalization have been studied with some seriousness for decades; there's nothing new to the suggestion. Second (and crucially), the overfitting at play here concerns generalization from finite samples to the underlying (static!) population. In short, this ostensibly key proposal (i) does not appear to address diachronicity and (ii) is hardly a proposal at all absent guidance for how to choose the relevant counterfactuals.

The failure of ML systems to account for dynamics presents a significant challenge and represents a risk to all stakeholders. This shortcoming is old and enduring and owes little to a fixation on fairness. However, these are real problems and coherent solutions require actually modeling these aspects of the world. Addressing these problems requires normative principles outlining the ends technology should serve, and guidance for when and how one can justify deploying technologies ill-suited to a changing world. It also requires technical progress towards coherent dynamics-aware methods. Delacroix is right to focus on diachronicity. At the same time, the paper reminds us that these problem are easier to recognize than to solve and that before attempting to take interpretable ML into the future, one ought to recognize its present failings.

References

[1] Emily Black and Matt Fredrikson. 'Leave-one-out Unfairness'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 285–295.

[2] Leo Breiman. 'Statistical modeling: The two cultures (with comments and a rejoinder by the author)'. *Statistical science* 16.3 (2001), pp. 199–231.

[3] Jessica Dai, Sina Fazelpour, and Zachary Lipton. 'Fair machine learning under partial compliance'. In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 55–65.

[4] Sylvie Delacroix. 'Diachronic Interpretability & Machine Learning Systems'. *Journal of Cross-Disciplinary Research in Computational Law* (2021).

[5] Sina Fazelpour and Zachary C Lipton. 'Algorithmic fairness from a non-ideal perspective'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 57–63.

[6] Sina Fazelpour, Zachary C Lipton, and David Danks. 'Algorithmic Fairness & the Situated Dynamics of Justice'. *Canadian Journal of Philosophy* (2021).

[7] Saurabh Garg et al. 'A Unified View of Label Shift Estimation'. *arXiv preprint arXiv:2003.07554* (2020).

[8] Saurabh Garg et al. 'Mixture Proportion Estimation and PU Learning: A Modern Approach'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2021.

[9] Lily Hu and Yiling Chen. 'A short-term intervention for long-term fairness in the labor market'. In: *Proceedings of the 2018 World Wide Web Conference*. 2018, pp. 1389–1398.

[10] Audrey Huang et al. 'Off-Policy Risk Assessment in Contextual Bandits'. *arXiv preprint arXiv:2104.08977* (2021).

[11] Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 'Learning the difference that makes a difference with counterfactually-augmented data'. *International Conference on Learning Representations (ICLR)* (2020).

[12] Zachary Lipton, Julian McAuley, and Alexandra Chouldechova. 'Does mitigating ML's impact disparity require treatment disparity?' In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2018.

[13] Zachary C Lipton. 'The mythos of model interpretability'. *Communications of the ACM (CACM)* (2018).

[14] Zachary C Lipton, Yu-Xiang Wang, and Alex Smola. 'Detecting and Correcting for Label Shift with Black Box Predictors'. In: *International Conference on Machine Learning (ICML)*. 2018.

[15] Lydia T Liu et al. 'Delayed impact of fair machine learning'. In: *International Conference on Machine Learning (ICML)*. 2018.

[16] Smitha Milli et al. 'The Social Cost of Strategic Classification'. In: *Conference on Fairness Accountability and Transparency (FAT*)*. 2018.

[17] Stephan Rabanser, Stephan Günnemann, and Zachary C Lipton. 'Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.

[18] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[19] Yifan Wu et al. 'Domain Adaptation with Asymmetrically-Relaxed Distribution Alignment'. In: *International Conference on Machine Learning (ICML)*. 2019.

Author's response: Why preserving 'a subject capable of interpreting' might be a challenge too

Sylvie Delacroix

Lipton's response is helpful on more than one level. It throws light on both the challenges and benefits of cross-disciplinary research.

My paper emphasises the difficulties inherent in preserving the interpretability of machine learning (ML) systems over time, particularly when these systems are deployed in ethically significant contexts. It also points at the potential inherent in what I call 'ensemble contestability features'. To understand the roots of Lipton's scepticism on the latter front, one must start by reiterating the following: interpretability requires both an interpretable object and a subject capable of interpretation.

On the 'interpretable object' front, Lipton and his team have produced seminal work on the challenges inherent in the fact that the real-world environments within which ML systems are deployed change over time. These temporal dynamics can mean that a system whose predictions were mostly accurate at the time of deployment become inaccurate five or ten years later. This is a known problem in computer sciences. Yet today there is little societal awareness when it comes to ways of addressing such 'concept drift' problems. In Europe, the 'Artificial Intelligence Act' contains provisions meant to avoid deployment in a setting whose characteristics are too distant from those of the training, validation and testing datasets. Yet none consider the impact of the passage of time.

'Ensemble contestability features' have nothing to do with the above, 'interpretable object' side, and everything to do with the 'subject capable of interpretation' side. Most take it for granted. Yet when an ML system is deployed within practices that presuppose our (re)-articulating the values that preside over such practices (such as justice or education), a concern for that system's interpretability over time must consider the 'subject side' too. That entails a concern for the extent to which these systems foster our ability to flex the muscles necessary to what are value-loaded interpretations. Ensemble contestability features are envisaged to do just that. As a 'normative workout affordance', they are certainly not designed to reduce overfitting (in contrast with the techniques from which they borrow). Nor are they meant to be laid out in any technical detail.

This concrete ensemble contestability proposal is meant to improve on philosophers' tendency to discuss the need for things like collective contestability without bothering to look into concrete ways of going about it. This paper's attempt to do so by borrowing from existing tools and vocabulary shows just how difficult it is. One ambiguous turn of phrase is all it takes to generate impatience from across the disciplinary divide. Sometimes this impatience is more warranted than others. I do not think Lipton's queries regarding the relevant data subsets or modification of modelling decisions are on point. Aside from the fact that answers to these implementation questions are context-dependent, they are also precisely the questions meant to be answered by a cross-disciplinary team.

On fairness and the roots of the 'static' takes on interpretability, however, I plead guilty to a cross-disciplinary research sin: I let my own disciplinary background shape an abrupt problem formulation. As per Lipton, making do with a static version of the problem which an ML algorithm is trying to tackle is often a necessity. It is therefore hardly surprising if systems deployed within ethically significant practices often neglect the effect of temporal dynamics. Given how momentous the impact of such dynamics can be on both accuracy and fairness, I hope Lipton's work continues to flourish. I also hope that this journal's impact is such that values other than fairness might become as salient.