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

The Interpretability Problem and the Critique of Abstraction(s)

Patrick Allo *

Abstract

Classifiers implement a level of abstraction: they classify entities by taking into account certain features and ignoring other features. Explanations and interpretations of algorithmic decisions require us to identify and to critically assess these abstractions. But what does it mean to critically assess an abstraction? While computer scientists see abstraction as something desirable, many legal scholars would see it as a cause for concern. Each of these views entails a different kind of critique of abstraction. This paper argues that a relative critique of abstraction is more appropriate in the context of the interpretability problem. It proposes a model, inspired by Floridi's Method of Abstraction, of a relative critique of abstractions that can be used to reason about the explanation, justification and contestation of classifiers.

Keywords: abstraction, method of levels of abstraction, proxies

Replier: Sandra Wachter, Professor of Technology and Regulation, University of Oxford • sandra.wachter@oii.ox.ac.uk

Journal of Cross-disciplinary Research in Computational Law

DOI: pending

www.journalcrcl.org

* Vrije Universiteit Brussel, patrick.allo@vub.be.

Introduction

For computer scientists, abstraction is something desirable. Abstraction allows them to construct and reason about computational processes,[1] and drives technologies that classify entities, such as persons, through supervised machine learning. Both uses of abstraction have been criticized by legal scholars.[2] While I agree with those critics that abstractions should be treated with caution, I do not agree that abstraction should entirely be avoided. To show that a different kind of critique of abstraction is possible, I develop a way of exposing the limits of a given abstraction that does not refer to an unabstracted reality (as most critics of abstraction would do), but only requires a comparison with a different, typically more refined, level of abstraction.

Classifiers (as well as other learned functions) implement a level of abstraction: they classify entities by taking into account certain features and ignoring other features.[3] When we want to interpret the workings of a classifier, the initial task is always to make explicit the level of abstraction it implements: we enumerate the features that influenced a specific decision and/or enumerate the features that, in a more general sense, are taken into account by the classifier. Highly simplified, if we want to make sense of an algorithmic system, we try to identify the features (such as income, education-level, age, ...) that lead a classifier to predict that a given entity belongs to a certain class (e.g. low financial risk).

This description captures the minimum we expect from *explanations* or *interpretations* of the workings or outcomes of algorithms. It is a bare account that does not depend on how such explanations are obtained (directly from the classifier that makes the prediction, or by means of a new model that approximates the original model) and does not depend on the purpose of such explanations.[4] As such, this minimal account remains consistent with the view that useful explanations require more information than just an enumeration of relevant features.[5]

Explanations have intrinsic value (providing understanding), but also instrumental value. The engineer might use explanations to improve the system, while the data-subject might use them to influence a future outcome. When it comes to *post-hoc* explanations, the instrumental value of explanations is naturally related to the further goals of (epistemic) justification and

1 Timothy Colburn and Gary Shute, 'Abstraction in Computer Science' (2007) 17(2) Minds and Machines 169.

2 Andrew D Selbst and others, 'Fairness and Abstraction in Sociotechnical Systems' (FAT* '19, Association for Computing Machinery 2019); Guido Noto La Diega, 'Against the Dehumanisation of Decision-Making — Algorithmic Decisions at the Crossroads of Intellectual Property, Data Protection, and Freedom of Information' (2018) 9(1) JIPITEC 3.

3 See the analysis of profiling and grouping people based on the Method of Abstraction in Patrick Allo, 'The Epistemology of Non-distributive Profiles' (2020) 33(3) Philosophy & Technology 379, ss 2-3.

4 Zachary C Lipton, 'The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery' (2018) 16(3) Queue 31; Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI' (FAT* '19, Association for Computing Machinery 2019) section 3; Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31(2) Harvard Journal of Law & Technology 841, 843.

5 Biran and McKeown distinguish the *effect* and the *importance* of features. Or Biran and Kathleen McKeown, *Justification Narratives for Individual Classifications* (techspace rep, 2014). Selbst and Barocas argue that 'Explanations of technical systems are (...) not sufficient to achieve law and policy goals': Andrew D Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87(3) Fordham Law Review 1085, 1088.

contestation.[6] Each of these brings into focus the level of abstraction implemented by a classifier:

1. When we interpret or explain a classifier, we make the level of abstraction it implements explicit by enumerating (the) features that informed a decision.

2. When we justify a classifier, we show that the level of abstraction it implements is appropriate in view of its intended purpose; no irrelevant features play a role, no relevant feature is ignored, *etc.* Whereas an explanation is an answer to a 'why' question,[7] an (epistemic) justification is an answer to a 'why (likely) correct' question.[8]

3. When we contest the workings of a classifier (or some of its effects), we somehow challenge its justification (the reasons that support the claim that a given prediction or decision is correct or likely correct). As such, we may challenge the underlying level of abstraction (for instance, highlighting that a classifier takes into account features it should not take into account or ignores features it should have taken into account), or challenge claims about the (predicted) accuracy of the classifier.

When we consider the interpretability problem along these lines, it is intimately related to the identification and critical examination of abstractions and their justification. In the next section, I argue that despite the considerable attention that has been paid to the role of abstraction in new and old classificatory practices, the question of how we should look critically at the use and justification of specific abstractions appears under-theorised.

Criticising abstractions

Evaluating the workings of classifiers implies (amongst others) being critical of the abstractions they implement. Bowker and Star, for instance, note that

> Each standard and each category valorises some point of view and silences another. This is not inherently a bad thing — indeed it is inescapable. But it is an ethical choice, and as such it is dangerous — not bad, but dangerous.[9]

[6] Lipton (n 4); EML Moerel and Marijn Storm, 'Automated Decisions Based on Profiling: Information, Explanation or Justification — That Is The Question!' [2019] SSRN Electronic Journal.

[7] Bas C Van Fraassen, *The Scientific Image* (Oxford University Press 1980), ch 5.

[8] This notion of justification follows very closely the account provided in Biran and McKeown (n 5), but does not thereby exclude further aspects of justification such as reliance on meaningful and unbiased correlations (Moerel and Storm (n 6)). Smart and others correctly point out the close connection between reliabilist accounts of justification which require belief-generating processes that maximize true beliefs and minimize false beliefs, and the role of accuracy and expected accuracy (and other measures) in machine learning. Andrew Smart and others, 'Why Reliabilism Is Not Enough: Epistemic and Moral Justification in Machine Learning' (AIES '20, February 7–8, 2020, Association for Computing Machinery 7 February 2020). The conclusions Smart and others attach to this diagnosis are more doubtful. In particular, their assessment does not take into account that, even on a reliabilist account, we can only attribute knowledge of ϕ to an agent A if we know that A's true belief that ϕ was produced by a reliable process. Indeed, while we do not require A to be aware that this process is reliable, we (as knowledge ascribers) do require insight in the reliability of the belief-forming process. See William P Alston, 'Internalism and Externalism in Epistemology' (1986) 14(1) Philosophical Topics 179.

[9] Geoffrey C Bowker and Susan Leigh Star, *Sorting Things Out: Classification and its Consequences* (MIT Press 1999) 5-6.

Object and meta-level abstractions

The need to adopt a critical stance towards abstraction(s) arises twice in relation to the interpretability problem.

First, it arises at the object-level of the classifiers themselves. Classifiers implement levels of abstraction and create *models of data-subjects* (traditionally referred to as *profiles*, which are just individuals observed, classified, or grouped at a given level of abstraction).[10] Explanations build on the explicitation of these levels of abstraction. Traditional critiques of abstraction primarily focus on this object-level: abstraction, understood as the construction of models of individuals, can lead to alienation,[11] objectification,[12] dehumanisation,[13] deindividualisation[14] and stereotyping.[15]

Second, the issues of abstraction and its critique arise at the meta-level: how we represent and reason about object-level abstractions, relations between object-level abstractions and the decisions they inform. Meta-level abstractions mediate, for instance, the amount of information about the decision-making system that is used to generate explanations or the (amount of) information that is included in the explanation. By analogy to how classifiers create models of data-subjects, explanations as well as justifications rely on *models of classifiers*. Explanations are simplified accounts of why a decision was made (a selection of reasons, a selection of features, a selection of criteria, ...) and are often based on simplified versions of decision-making systems (approximations of the original models,[16] 'meaningful information about the logic involved,'[17] ...). Similarly, justifications often refer to the expected or predicted accuracy of a classifier, and the calculation of this accuracy is itself based on a model of the classifier whose accuracy is being assessed (see 'Meta-level abstractions' below).

Attention for these meta-level abstractions has recently accrued. Several authors refer to the different levels of abstraction at which explanations can be generated in terms of lower and higher levels of abstraction (the amount of detail that is taken into account).[18] Mittelstadt and others explicitly tie the process of generating explanations to the construction of

[10] Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen. Cross-disciplinary Perspectives* (Springer Netherlands 2008).

[11] Marion Fourcade and Kieran Healy, 'Classification situations: Life-chances in the neoliberal era' (2013) 38 Accounting, Organizations and Society 559.

[12] Theodore M Porter, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Publication Title: Princeton University Press, vol 22, 1995) 74ff.

[13] Noto La Diega (n 2).

[14] Antoinette Rouvroy and Bernard Stiegler, 'The digital regime of truth: from the algorithmic governmentality to a new rule of law' [2016] (3) La Deleuziana 6; Bart W Schermer, 'The limits of privacy in automated profiling and data mining' (2011) 27(1) Computer Law & Security Review 45.

[15] Mireille Hildebrandt, 'Defining profiling: A new type of knowledge?' in *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer Netherlands 2008) 23–25.

[16] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin, '"Why Should I Trust You?": Explaining the Predictions of Any Classifier' (KDD '16, ACM 2016).

[17] Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) International Data Privacy Law 76, 78.

[18] Michele Loi, Andrea Ferrario and Eleonora Viganò, 'Transparency as Design Publicity: Explaining and Justifying Inscrutable Algorithms' (2021) 23(3) Ethics and Information Technology 253; Lipton (n 4).

simplified models (of the decision or the decision-making system).[19] Barocas, Selbst and Raghavan add a critical counterpoint and expose the hidden assumptions of *feature-highlighting explanations* such as counterfactual explanations and principal reasons that only reveal a subset of features that matter to the individual decision *and* constitute a useful and actionable explanation for the subject involved.[20]

Selbst, boyd and others raise a similar, but more radical concern for the design of fair machine learning systems.[21] More than others involved in this debate, they develop an argument that pays close attention to the role of abstraction in computer science. Their initial focus is on the tension between what is necessary to make things or concepts computable — such as fixed definitions and reduced complexity — and the open-ended, shifting and contested nature of social concepts such as fairness.

> The most common abstractions in machine learning consist of choosing representations (of data), and labeling (of outcomes). Once these choices are made, they constitute the description of what we call the algorithmic frame. Within this frame, the efficacy of an algorithm is evaluated as properties of the output as related to the input.[22]

What they describe as the 'algorithmic frame' is an example of a meta-level abstraction. It is one level at which we may represent a classifier and reason about its functioning and performance.

Jacobs and Wallach approach the problem of fairness from the perspective of the operationalization of unobservable theoretical constructs.[23] The risks they identify extend well beyond the initial question of fairness, and their analysis of the role of algorithmic frames in machine learning is directly relevant to the interplay between abstraction(s) and the explanation/interpretation of classifications and decisions.

Abstractions as proxies and absolute versus relative critiques of abstraction

Through abstraction we hide information (properties, facts, distinctions, ...), and such hiding is not without consequences. A critique of abstraction is absolute whenever it rejects all forms of information hiding. Such absolute critiques are omnipresent in the disciplines that provide the theoretical foundations of critical data studies, such as (human) geography, ethnography and science and technology studies (STS). McCormack's account of 'abstraction's critique' in human geography quite aptly summarizes this position as (i) the problematisation of abstraction as 'an intellectual *withdrawal* from the world' and (ii) the adoption of 'a logic of distancing'; (iii) a charge of *reduction* aimed at simplification and generalisation that is (iv) driven by a desire for *universalism*; (v) and leads to *alienation*.[24]

[19] Mittelstadt, Russell and Wachter (n 4).

[20] Solon Barocas, Andrew D Selbst and Manish Raghavan, 'The hidden assumptions behind counterfactual explanations and principal reasons' (ACM January 2020) 81-82.

[21] Selbst and others (n 2).

[22] Ibid 60.

[23] Abigail Z Jacobs and Hanna Wallach, 'Measurement and Fairness' (FAccT '21, Association for Computing Machinery 2021) 375.

[24] Derek McCormack, 'Geography and abstraction: Towards an affirmative critique' (2012) 36(6) Progress in Human Geography 715.

A relative critique accepts that abstraction is inevitable and therefore only challenges the hiding of specific information when the omission of a specific piece of information leads to a mistaken inference. Abstraction, on this account, is problematic because hiding information influences outcomes, decisions or conclusions that are based on defeasible reasoning forms: inferences whose conclusions might have to be retracted when additional information is taken into account. Learned functions generated by supervised learning algorithms are defeasible in exactly this sense:[25] they are only intended to be correct in most cases and mistaken inferences can only be detected and corrected by taking into account additional information. If the difference between two outcomes of a defeasible inference process is not neutral, the ways in which we hide and reveal information is not neutral either — it may flip a significant decision.

The abstractions that inform such defeasible decisions do more than just grouping people (putting them together based on similarity). They act as *proxies* and are ways of accessing something else that might not be directly accessible (a given *target*). Proxies, in this sense, allow for what philosophers of science have called *surrogative reasoning*.[26] In the sciences, models are used as a more tractable alternative to the direct reasoning about reality itself. When applied to supervised learning, it is more appropriate to think of proxies as something we use to access something else that does not even have to be (part of) reality itself but can be another abstraction. This additional step is important. The targets used in the context of supervised learning are often only an approximation of the *true target*:[27] past decisions are used instead of correct decisions, 'performance reviews, sales numbers, and retention time' instead of true employee performance),[28] or student scores on standardised tests instead of actual teacher performance.[29]

Absolute and relative critiques understand such proxies along different lines. While absolute critiques of abstraction focus on the fact that abstractions are proxies of some unabstracted (hence, more authoritative) piece of reality, relative critiques of abstraction emphasise their inferential role and adopt a relational approach[30] to decide what counts as a proxy and what counts as a target. Something is a proxy if it is used to access a given target. A relative critique thus provides *a different explanation* of the problematic nature of abstraction (linked to the risk of mistakes when relying on defeasible inferences) as well as a *different method of assessment* of abstractions (shifting abstractions to identify the consequences of revealing previously hidden information). The intended method of assessment is comparative or relative: abstractions used as proxies are evaluated in terms of how good an access they provide to their intended target (how reliable the indirect reasoning will be) and not on the basis of what it takes away from a presumed unabstracted reality.

[25] Allo, 'The Epistemology of Non-distributive Profiles' (n 3).

[26] Chris Swoyer, 'Structural representation and surrogative reasoning' (1991) 87(3) Synthese 449; Mauricio Suárez, 'Scientific Representation' (2010) 5(1) Philosophy Compass 91.

[27] Jacobs and Wallach (n 23).

[28] Manish Raghavan and others, 'Mitigating bias in algorithmic hiring: evaluating claims and practices' (ACM January 2020) 472.

[29] Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016).

[30] Compare with L Floridi, *The Logic of Information* (Oxford University Press 2019) 40–6 and Sabina Leonelli, 'What Counts as Scientific Data? A Relational Framework' (2015) 82(5) Philosophy of Science 810.

The next section clarifies the distinction between absolute and relative critiques of abstraction(s) and explicitly connects it to the goal of supervised learning.

The reference-class problem

Imagine that Alice's credit-application is rejected because the features of Alice that are being taken into account place her in a group *G* whose members have a high risk of defaulting. One way to assess this situation focuses on the fact that Alice is treated as a member of a group (only based on what Alice has in common with other members of a group) rather than as an individual (based on what distinguishes Alice from any other individual). This approach would be in line with an absolute critique of abstraction. Another way to assess this situation would be to point out that if certain additional features of Alice would have been taken into account, she would have been situated in *G'* instead of *G*. As members of *G'* have a much lower risk of defaulting, the final decision about her credit-application would have been different.

The situation I just sketched is a version of the reference-class problem: how should we 'determine an appropriate single-case probability when an individual belongs to several reference classes for which data is available, and where estimates of chances differ from reference class to reference class'?[31] This is a notoriously recalcitrant problem.[32] It is also one of the problems that machine learning, and especially supervised learning, purports to solve.[33]

This reframing of the problem of how far we may abstract away Alice's features emphasises the comparative nature of the problem (why this, rather than that level of abstraction?), and can help us relate this comparative approach to what supervised learning is meant to achieve.

Let us, for the sake of illustration, assume that decisions about Alice's credit application rely on decision-trees. In such a scenario, the question whether Alice should be situated in *G* (using tree T) or in *G'* (using the alternative tree T' that makes some finer distinctions) would be equivalent to asking whether using T should be expected to lead to more accurate predictions than using T'. This is the comparative problem as it arises in the context of supervised learning. Given the set Λ of all decision-trees that can be built using the available data, the goal is to identify a tree within Λ that (on the basis of the available training data) can be expected to lead to more accurate predictions than other trees in the same set.[34]

Supervised learning deals with the comparative problem of figuring out which object-level abstraction is *best* for a given (predictive) purpose. The comparison of alternative abstractions is limited to trees that can be generated on the basis of the available data (a limited set of alternatives), based on a well-defined metric (a single criterion for what counts as a reliable

[31] Christian Wallmann and Jon Williamson, 'Four Approaches to the Reference Class Problem' in Gábor Hofer-Szabó and Leszek Wroński (eds), *Making it Formally Explicit* (Springer International Publishing 2017).

[32] Early versions of the problem are due to Venn and to Reichenbach, see Alan Hájek, 'The reference class problem is your problem too' (2007) 156(3) Synthese 563, 564.

[33] Gregory Wheeler, 'Machine Epistemology and Big Data' in Lee McIntyre and Alex Roseberg (eds), *The Routledge Companion to The Philosophy of Social Science* (Routledge 2016); Wallmann and Williamson (n 31).

[34] Compare with the more general version: 'The problem of learning is that of choosing from the given set of functions $f(x,\alpha), \alpha \in \Lambda$ the one which predicts the supervisor's response in the best possible way': Vladimir N Vapnik, 'An overview of statistical learning theory' (1999) (5) IEEE transactions on neural networks 988 (Publisher: IEEE) 988.

proxy) and expectations that are computed on the basis of a given training-set (a settled 'source of ground truth'). [35]

The way in which supervised learning deals with this comparative problem is, however, limited and cannot address the question which meta-level abstraction is best to reason about the initial comparative problem. As explained above, one version of this problem has been brought to attention by Selbst, boyd and others: supervised learning does not look beyond the algorithmic frame.[36] According to their analysis, this type of limitation can be avoided by adopting an STS-lens and recognizing that machine learning systems are always part of a broader sociotechnical system.[37] This narrative suggests that abstraction is inherently problematic,[38] and underscores the disciplinary and methodological discontinuity between the formal perspective on abstraction as we find it in the context of supervised learning and the social science perspective on abstraction that becomes relevant when we want to critically assess the use of abstractions in sociotechnical systems.

This paper develops an alternative narrative with a stronger emphasis on the continuity between the (object-level) comparative problem that supervised learning deals with, and the broader range of (object-level and meta-level) comparative problems that need to be considered to critically assess the abstractions that classifiers implement.

Critical formal methods?

My starting point is the *Method of Abstraction*, a core philosophical method of the *Philosophy of Information* (PI).[39] The two key tenets of the Method of Abstraction are (i) the view that our interactions with the world always occur at a level of abstraction (no unmediated access to the world) and (ii) the view that questions (about reality) can only be answered when the level of abstraction at which the question should be answered is made explicit (and kept fixed).[40] Formalisation is one way to make these levels of abstraction explicit.

Given the tendency in the literature to doubt the critical potential of formalisation, this use of formal methods needs a separate defence. Two positions from the literature are of particular interest.

Dourish on context

A first view is related to how scholars of sociotechnical systems conceptualise context and materiality. When Paul Dourish proposes to study the *materialities of information representation*,[41] he explicitly contrasts his perspective with Claude Shannon's conception of information as an abstract entity that remains independent of the matter in which it is represented. While Shannon would adhere to the commonplace that there is no information without physical

35 Jacobs and Wallach (n 23) 384.

36 Selbst and others (n 2) 60. See also the quote in 'Object and meta-level abstractions' above.

37 Ibid 60.

38 Although their suggestion to focus on a sociotechnical frame seems compatible with a relative critique of abstraction, this is still combined with a more radical critique of abstraction (see 'Critical formal methods!' below).

39 L Floridi, *The Philosophy of Information* (Oxford University Press 2011) ch 3.

40 Floridi, *The Logic of Information* (n 30) 21.

41 Paul Dourish, *The Stuff of Bits* (MIT Press 2017).

implementation[42] or material representation, he also believes that we can reason about information by making abstraction of this material representation.

In earlier work, Dourish zooms in on the epistemological divide between phenomenological and positivist accounts of context. He explains that positivist accounts will see the problem of defining context as a problem of representation. Context, then, is a form of *information* that is not only *delineable* and *stable*, but also something that can be *separated* from the activities that take place in these contexts. Such epistemologies inform most engineering practices. On a phenomenological account, however, context is a relational property that holds between objects and activities. It is not stable and cannot be defined or delineated if one ignores the particulars of the situation in which certain activities take place.

For present purposes, there is a useful analogy to be made between the way we approach context—deciding which contextual factors matter and which can be ignored—and our attitudes toward abstraction. When we delineate contextually relevant features, we do settle for a given level of abstraction. Doing so, and making the underlying level of abstraction explicit (as the Method of Abstraction recommends!), can then be seen as 'a misunderstanding of the nature and role of contextuality in actual everyday affairs' where 'context isn't something that describes a setting; it's something that people do.'[43] What the method of abstraction commends as the virtuous thing to do, is precisely what Dourish would want us to avoid.

The resulting tension touches upon the core of my proposal to only compare abstractions with other abstractions. First, because it correctly points to the initially limited critical potential of the method of abstraction. Fixing and making our abstractions explicit is simply not enough! Second, because Dourish's views about materiality reveal the limits of looking at technical artefacts such as classifiers as abstract objects,[44] or objects that can fully be understood in terms of the level of abstraction they implement. As such, well entrenched ways of studying technical artefacts such as algorithms call into question the feasibility of a critical evaluation of classifiers that is based on probing the abstractions they implement.

If we look at the Method of Abstraction from Dourish's perspective, we might just see a positivist account of information and context. This reading is at odds with Floridi's own relational understanding of the (epistemology behind the) Method of Abstraction.[45] The difference between both views is more subtle. The Method of Abstraction agrees that information is relative to the (epistemic) activities that take place in a given context, but it privileges a different class of epistemic agents. For Dourish, the users of a given technology are the relevant epistemic agents, but the Philosophy of Information privileges the knowing agent or modeller of a system or phenomenon (of interest to that agent). This gives rise to two different—but in my view compatible—relational accounts of information: one that rejects epistemological views that presuppose a *view from nowhere* and another that challenges the privileged perspective of the maker of a system.

[42] Luciano Floridi, 'Semantic Conceptions of Information' in Edward N Zalta (ed), *Stanford Encyclopedia of Information* (Stanford 2005).

[43] Paul Dourish, 'What We Talk About When We Talk About Context' (2004) 8(1) Personal and Ubiquitous Computing 19, 22.

[44] Paul Dourish, 'Algorithms and their others: Algorithmic culture in context' (2016) Big Data & Society 1; Rob Kitchin, 'Thinking critically about and researching algorithms' (2017) 20(1) Information Communication and Society 14.

[45] Floridi, *The Logic of Information* (n 30) 84.

When applied to the interpretability problem, the latter is essential. Explanations, justifications and contestations that privilege the perspective of the maker of the decision-making system are of limited value to the data subject.

Abstraction traps

A Second view comes to the fore in the recent work of Selbst, boyd and others that I referred to above.[46] This work specifically considers the relation between currently popular attempts to design ML-based systems that can achieve socially desirable and/or legal outcomes such as fairness and the role of abstraction and modular design in computer science. They identify an important problem in how 'fair ML-based systems' are currently being conceived and developed and point out that research in this area systematically 'bound[s] the system of interest narrowly. [And] abstract{s] away any context that surrounds the system.'[47] First, they draw attention to the gap between the existing practice of designing fair subsystems that maximise a given fairness metric and the socially desirable goal of designing fair global systems. Second, they contend that fairness and justice are social concepts whereas the current efforts in computer science only target computationally tractable proxies (metrics) of these social concepts.

In their words, 'Fairness and justice are properties of social and legal systems like employment and criminal justice, not properties of the technical tools within.'[48] This diagnosis is based on a more encompassing critique of abstraction that flags five different abstraction traps: (i) a framing trap, (ii) a portability trap, (iii) a formalism trap, (iv) a ripple effect trap and (v) a solutionism trap.

The framing trap is a '[f]ailure to model the entire system over which a social criterion, such as fairness, will be enforced.' Whereas the formalism trap is the '[f]ailure to account for the full meaning of social concepts such as fairness, which can be procedural, contextual, and contestable, and cannot be resolved through mathematical formalisms.' Each of these five traps is relevant in the present context. In the section 'Object and meta-level abstractions' above, I already considered the framing trap in relation to meta-level abstractions. In the remainder of this section, I take a closer look at the interplay between the framing trap and the formalism trap.

Critical formal methods!

My proposal to rely on formal methods for critical purposes and develop the critical potential of the Method of Abstraction is potentially affected by the framing trap. Indeed, in its striving towards tractable systems[49] or answerable questions,[50] the Method of Abstraction can entice one to bound a system of interest (too) narrowly by extremely simplifying a given problem or by abstracting away too much of the potentially relevant context in which a problem arises. This is a plausible, but not an inevitable risk.

To begin with, modelling a system within certain explicitly stated boundaries does not prevent one to

46 Selbst and others (n 2).

47 Ibid 59.

48 Ibid.

49 Floridi, *The Philosophy of Information* (n 39) s 3.4; Patrick Allo, 'A Constructionist Philosophy of Logic' (2017) 27(3) Minds and Machines 545, s 4.

50 L Floridi, 'Turing's three philosophical lessons and the philosophy of information' (2012) 370(1971) Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 3536.

extend these boundaries.[51] We can always start from a highly simplified version of a problem and then gradually increase its complexity. Such a shifting of boundaries (going beyond the algorithmic frame) or of levels of abstraction (considering alternative abstractions) is key to my proposal. Shifting boundaries means taking more into account and thus increasing the chance of finding a counterexample to a previously identified pattern.

Even without the possibility of shifting boundaries, the rigorous application of the Method of Abstraction has critical potential simply because it forces one to relativize (and thereby limit) the scope of one's claims. If we only solved a radically simplified version of a problem (and made these simplifications explicit), we should never claim that we solved the problem in all its original complexity. Properties that are attributed to a sub-system should not automatically be attributed to a more encompassing system. More generally, properties attributed to a system S modelled at a LoA l should not automatically by attributed to a larger system S' and/or model of this system at an alternative LoA l'.

For the formalism trap, Selbst, boyd and others rely on a slightly different argumentative strategy. Fairness cannot properly be assessed on the basis of a mathematically precise fairness metric because no such metric can ever capture the true (vague, open-ended, contextually sensitive, shifting, and contested) meaning of fairness. Any attempt to formalise fairness (and hence replace a concept with open texture with a concept with closed texture) will, on their account, be misguided. Here, the conclusion is not just that current formalisations or current metrics fail to capture crucial nuances, features or other fine-grained relevant distinctions that a more refined metric could capture. Rather, the implied conclusion is that no formalisation will ever get it right.[52]

Both arguments start out as a critical assessment of specific abstractions but end up as full-fledged critiques of abstraction. First, it is indicated that a given frame or a given formalisation does not include all the relevant features or capture all the relevant nuances, and shortly after this turns into the claim that one should look at social reality as a whole, or the claim that nothing short of the 'full concept' of fairness will be good enough. As I read it, the conclusion of this argument is not just that we need to be careful when we set the boundaries of a system or when we adopt a formal definition. Instead, the conclusion is that by delineating a system of interest or adopting a formal definition we are not just selective (and I fully agree that this is a reason to be careful), but fundamentally mistaken. And this sets the bar for what is good enough incredibly high.

This collapse is *prima facie* worrisome because it resembles the collapse of fallibilism (we might be wrong) into radical scepticism (knowledge is impossible),[53] which is problematic because it turns a productive form of criticism into a degenerate form of criticism.[54] I do not think that this collapse is desirable, and I am not sure at all that this collapse is intended

[51] Floridi, *The Philosophy of Information* (n 39) 73.

[52] Their argument is more elaborate than my reconstruction suggests. I do not, for instance, consider the issue they raise about procedural versus outcome-based notions of fairness. Such omissions notwithstanding, I do not think I misrepresent their sceptical attitude towards fairness metrics.

[53] Recall that Smart and others (n 8) similarly struggle with the fallibility of reliable methods that, according to reliabilist accounts of justification, can form the basis of knowledge.

[54] Floridi, *The Logic of Information* (n 30) ch 6.

by these authors. At the very least, they do not explicitly rule it out: they first identify the risks associated with a given abstraction trap (framing, formalisation) and then go on to present the absence of abstraction, framing, or formalisation as an alternative.

When it comes to formalisation and the use of formal methods, the situation is somewhat different. As they see it, the process of formalisation is something that replaces a complex, open-ended and contested concept with a unique, precise and no longer contestable concept. This is a highly plausible reading of how formalisation currently affects the concept of fairness, but it is not the only conceivable use formal methods. There is a pluralistic and more explorative use that turns concepts with open texture into families of precise, distinct, and potentially conflicting concepts whose relative merits can be studied in ways that the original concept could not. In addition, there is also a more critical use[55] that is more in line with the tradition of proving limitative results (the mathematical study of the limitations of formal systems),[56] and the critical role of statistics and statistical literacy.[57]

Such critical uses of formal methods do indeed substantially diverge from the practices that authors such as Dourish or Selbst, boyd and others bring into focus. One particularly noteworthy example is the engineering practice of narrowly delineating a problem and sticking to one's chosen formalisms, and actively resist considerations based on alternative frames or alternative abstractions. As shown by Corinne Cath in her study of how IETF engineers deal with the social implications of the protocols and the standards they develop, engineers often restrict their responsibility (based on a logic of separating concerns) to the issues that can be made visible within their own frame (the module or layer they work on) and consider problems that fall outside that frame as 'Somebody Else's Problem'.[58]

Abstractions and networks of abstractions

A toy model based on Barwise and Seligman

Barwise and Seligman's formal theory of information flow can be used as a language for characterising abstractions and as a model to reason about relations between abstractions, such as proxy-target relations.[59] In its original formulation, it is a theory about the flow of information within distributed systems, where information flows between different components in virtue of the regularities within this system.[60] The same principle of information flow applies to abstractions as well: one abstraction can carry information about

55 Catarina Dutilh Novaes, *Formal Languages in Logic: A Philosophical and Cognitive Analysis* (Cambridge University Press 2012).

56 See the reply by Van Bendegem on e.g. Gödel's Incompleteness Theorems in Bernhard Anrig, Will Browne, and Mark Gasson, 'The Role of Algorithms in Profiling' in Mireille Hildebrandt and Serge Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer Netherlands 2008) s 4.7, or the references to Wolpert's No Free Lunch Theorem in Mireille Hildebrandt, 'Law as Information in the Era of Data-Driven Agency' (2016) 79(1) Modern Law Review 1.

57 Tim Harford, 'Big data: A big mistake?' (2014) 11(5) Significance 14; Chris J Wild, 'Statistical literacy as the earth moves' (2017) 16(1) Statistics Education Research Journal 31.

58 Corinne J N Cath, 'Changing Minds and Machines: A Case Study of Human Rights Advocacy in the Internet Engineering Task Force (IETF)' (DPhil in Information, Communication, and the Social Sciences, University of Oxford 2021) 37.

59 Jon Barwise and J Seligman, *Information Flow: The Logic of Distributed Systems* (Cambridge Tracts in Theoretical Computer Science, vol 44, Cambridge University Press 1997).

60 Ibid 8.

another abstraction in virtue of regularities (or patterns) within a broader system. Because Barwise and Seligman's theory can account for regularities that might have exceptions, it is particularly well-suited to reason about proxy-target relations in supervised learning.[61]

I start from the basic formal construct of a *classification*, and use this construct to study relations between *abstractions*. A classification, in Barwise and Seligman's theory,[62] is a triple consisting of a set of tokens (entities), a set of types (properties) and a binary relation that captures which tokens are classified as being of a given type. Because Barwise and Seligman's use of the term 'classification' does not match other uses of the same term,[63] I diverge from their terminology and refer instead to these triples as abstractions. This allows me to propose an important (but from a formal point of view entirely natural) generalisation of the relevant notion of abstraction.

Normally, we understand an abstraction as a selection of properties or features used to classify entities. An abstraction is then just a set of types. This captures the idea that our ability to distinguish between tokens depends on the types we have available to express those distinctions: no distinctions without the conceptual means to make these distinctions. We cannot just think of an abstraction as a selection of our means to classify (a set of the types), but also as a selection of the entities to which we can attribute these types (a set of the tokens). From a formal point of view this is an evident generalisation based on the so-called type/token duality according to which there is no fundamental difference between using types to classify tokens and using tokens to classify types.[64] This captures the idea that our ability to distinguish between types depends on the tokens we have available to detect those distinctions.

When applied to the classification of data-subjects, this generalisation reveals that omitting the entities under consideration is just as much an act of abstraction as the omission of the features we use to classify these entities: if we use tokens as the concrete examples that allow us to understand types and relations between types, our understanding of these types will be constrained by the (number of) tokens we can consider. A first, rather basic, effect of this type of abstraction is that fewer subjects (tokens) lead to less fine-grained distinctions between features (types) in exactly the same sense as having fewer features decreases our ability to make fine-grained distinctions between subjects. This is so for purely combinatorial reasons: a smaller number of tokens comes with fewer opportunities to instantiate differences between properties and combinations of properties. A more salient effect of this type of abstraction is that the actual selection of subjects (tokens) also affects our ability to detect fine distinctions and subtle relations between features that occur in the entire population. A large but homogeneous sample from a heterogeneous population inevitably hides (and hence misrepresent) the diversity as well as finer distinctions and patterns that are present in the entire population. If we moreover

61 Patrick Allo, 'Reasoning about Data and Information: Abstraction between States and Commodities' (2009) 167(2) Synthese 231, s 5.

62 Barwise and Seligman (n 59) 28.

63 First, it does not entirely match what in ML is understood as a classifier. Second, it does not correspond to how Bowker and Star characterize a classification as a segmentation of the world, wherein, amongst others, categories are mutually exclusive.

64 Barwise and Seligman (n 59) s 4.4.

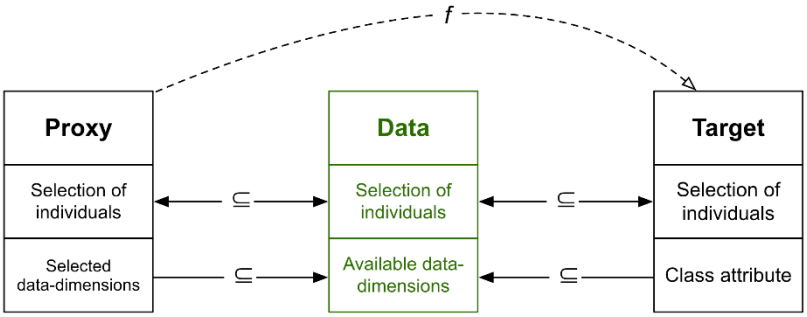

Figure 1. Algorithmic frame as relations between abstractions

consider the kinds of patterns that drive defeasible inference forms, such as inferences based on generalisations that might have exceptions (most *P*'s are *Q*), even a representative selection of subjects (tokens) can hide the finer structure underneath such generalisations. The concentration or uneven distribution of errors and error-types in specific (unaccounted for) subgroups is one example of such a hidden feature.[65]

To explain how information can move back and forth between classifications, Barwise and Seligman introduce *infomorphisms* as the primary relation between classifications. When we focus on information-flow between abstractions, we can restrict our attention to *refinement*-relations as a special case of infomorphisms.[66] We will say that one abstraction is a refinement of another if it takes a superset of types into account, while agreeing on the assignment of types to tokens for the types they have in common. By the type/token duality, a superset of tokens will then count as a refinement of the classification of types. On this account, the inverse of a refinement-relation is exactly the kind of abstraction-relation I described above; abstractions are just omissions of types (tokens). Patterns, such as regularities (generalisations) or similarities are not preserved under refinements, but they are preserved under abstractions. Adding types (tokens) affords additional distinctions between tokens (types); adding tokens introduces counterexamples to regularities. Conversely, counterexamples to regularities or dissimilarities are preserved under refinements, but they are not preserved under abstractions.

Abstraction and refinement relations can be used to reconstruct the algorithmic frame of supervised learning, but they can also be used to further extend this frame and take into account a wider range of proxy-target relations. To illustrate how such extensions fit a relative critique of abstractions, I take a closer look at (i) relations between sets of features within the algorithmic frame ('Features as proxies'), (ii) relations between sets of known and unknown individuals ('Individuals as proxies') and (iii) meta-level abstractions ('Meta-level abstractions').

[65] Compare with the findings in Julia Angwin and others, 'Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.' [2016] Pro Publica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>; Jeff Larson and others, 'How We Analyzed the COMPAS Recidivism Algorithm' [2016] Pro Publica <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>.

[66] Barwise and Seligman (n 59) 72, 76.

Features as proxies

As a first step, consider the representation of a proxy-target relation within the narrow confines of the algorithmic frame, as depicted in Figure 1.

Here, two types of relations between abstractions are included: the abstraction-relations between the available data and the proxy and target-abstractions (each based on a subset of the available data-dimensions) and the learned function *f* which maps combinations of selected data-dimensions to the class attribute. Because the abstraction-relations go in opposite directions, there is no guarantee that the proxy will allow us to reliably predict the target: the relation between the proxy and the target is itself not an abstraction-relation. Hence, the proxy will only contain information about the target if they are connected by a regularity within the available data. It is in virtue of this regularity that the learned function *f* can be used to predict the class attribute from the data-dimensions that are used to create the proxy.

Regularities could be supported by causal connections but could also just be correlations. For present purposes, regularities are just constraints that are linked to specific abstractions, and this allows us to verify their robustness by inquiring whether they are also present in other, more refined, abstractions (i.e. preserved under abstractions and refinements). This is done without considering regularities of a different kind (asking whether a correlation is also a causal connection): only extensional regularities need to be taken into consideration.[67]

This sets the stage for further explorations. It shows how the proxy-target relations supervised learning deals with can be described, but it can also be seen as a basic frame that can be further extended. Alternative selections of data-dimensions can easily be added, and attempts can be made to make the relation between the class attribute that is used as a target and the real target of the decisions it informs.

Individuals as proxies

A diagram that is surprisingly similar to that of Figure 1 can be used to show how data about known individuals (used for supervised learning) can serve as a proxy for data about unknown individuals (used for new predictions). In the context of supervised learning, a *known* individual is one with known proxy (used for the prediction) and target features (the feature that is being predicted, i.e. the class attribute); an *unknown* or *unseen* individual is one with only known proxy features. This situation is depicted in Figure 2, where we see a configuration that is the dual of the one in Figure 1, with the abstraction-relations between individuals (tokens) going in opposite directions, and the known data (proxy) and unknown data (target) as two different abstractions from a larger data set (the system).

Following a reasoning that is analogous to the one presented in 'Features as proxies' above, we note that the known data cannot be seen as an abstraction of the unknown data or vice-versa. Both are abstractions

[67] For the distinction between intensional and extensional definitions of tasks, see David Schlangen, 'Targeting the Benchmark: On Methodology in Current Natural Language Processing Research' in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers, Association for Computational Linguistics 2021) 671.

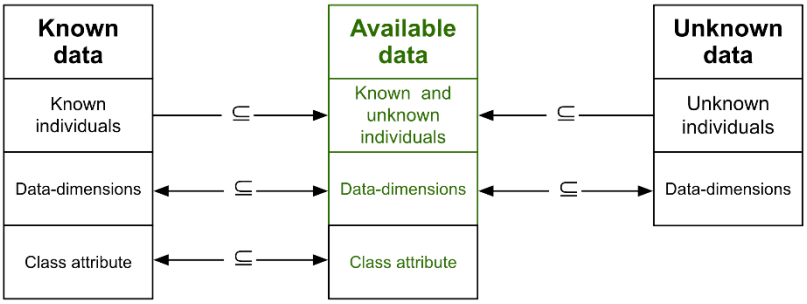

Figure 2. Abstraction-relations between sets of known and unknown individuals

(i.e. subsets or samples) from a larger dataset.[68] As we saw in the case of proxies and targets, we cannot directly conclude that what the known individuals (tokens) will tell us about regularities between features (i.e. the patterns that can be discovered through learning) will contain information about the regularities between the same features for a different set of individuals. Indeed, the known individuals will only contain information about the unknown individuals if they are connected by a regularity within the available data; intuitively and highly simplified, if both groups are sufficiently similar.

The resulting perspective allows us to characterise a second proxy-target relation that defines the practice of supervised learning, namely the use of knowledge extracted from one set of individuals to make claims (predictions) about other individuals. While itself not a new insight, this is a first departure from the initial setting of the algorithmic frame and an indication that at least two proxy-target relations (and two types of object-level abstractions) are relevant to the interpretability problem.

Meta-level abstractions

The previous two cases dealt with object-level abstractions: the selection of features to classify individuals and the selection of individuals to identify patterns (learn about relations between features) and make predictions. This approach can be extended to meta-level abstractions: the level of abstraction at which we model a proxy-target relation. As I will explain, claims about classifiers or about learned functions can be made at different levels of abstraction, and once we make this explicit, it is possible to ask whether these claims remain true if we adopt a different level of abstraction. To illustrate this, we take a closer look at the practice of attributing levels of accuracy to classifiers or learned functions.

Considered in isolation, a learned function f only maps features (selected data-dimensions) to features (the class attribute). It depends on the features that are available, but not on the individuals under consideration. This is precisely why f can be re-used to make predictions about *unseen* individuals. We can use f as soon as we have all the data that are needed as an input. When it comes to making claims about the accuracy of the predictions that f allows, we cannot ignore

[68] Do note that the fact that we do know more about the known individuals (the class attribute is known) than about the unknown individuals does not come into play here. The relevant abstraction-relation is at the level of the individuals under consideration, and we only focus on the fact that the set of known and unknown individuals are disjoint.

the individuals that were involved in the determination of its accuracy.

This insight is already implicit in what was said above with regard to the similarity of the individuals we learn from and those we make predictions about, and can immediately be linked to the standard practice of estimating the accuracy of a learned function.[69] At least three different such claims can be made: (i) about the accuracy of the predictions for the training and validation data that are used for the sake of model-selection (which *f* to choose from a given class of functions), (ii) about the accuracy of the predictions for the test data that are used for the sake of model-assessment (how we expect the chosen model to perform on unseen data) and (iii) about the actual accuracy of the predictions on unseen data. Here too, claims of one type are used as a proxy for claims about another type of claim. First, the accuracy measured on the training and validation data are used to select a model that one expects to perform equally well on the test data. Second, the accuracy measured on the test data is used to estimate the accuracy that really matters, namely with respect to unseen individuals.

These distinctions are entirely standard within the context of supervised learning, but informal claims about the accuracy of a model rarely make explicit relative to which dataset such claims are made. Recent work on the role of datasets in machine learning suggests that when widely used datasets are taken for granted and used as benchmarks to measure advancement, the relative nature the results remains hidden.[70] When we make the relative nature of claims about the accuracy of a model explicit, we do make the level of abstraction at which we reason about this model explicit or say which data-set we used as a proxy.

We can again use this as a basis for further explorations and alternative ways of assessing the accuracy of a model. One such alternative assessment is based on the decomposition of a single claim about the accuracy of a model relative to a given dataset into multiple such claims relative to different subsets of this dataset. This strategy is closely related to what researchers at Pro Publica did for the COMPAS Recidivism algorithm,[71], which we can now re-describe as a shifting from the level of abstraction adopted by the maker of a system to a level of abstraction that better suits the needs of a critical researcher: one coarse measure is replaced by multiple more fine-grained measures.

Concluding remarks

The model outlined in the previous section provides a formal language that can be used to characterise abstractions and relations between such abstractions. These include refinement and abstraction-relations, as well as relations between abstractions we use as proxies for other abstractions that are a target of interest. In the context of the interpretability problem, this model can be used to situate the proxy-target relations

[69] Trevor Hastie, Robert Tibshirani and Jerome Friedman, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd, Springer Verlag 2009) 222.

[70] Amandalynne Paullada and others, 'Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research' (2021) 2(11) Patterns 100336 <https://linkinghub.elsevier.com/retrieve/pii/S2666389921001847> accessed 10 July 2022; Bernard Koch and others, 'Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research' (35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia, arXiv 3 December 2021) <https://arxiv.org/abs/2112.01716> accessed 10 July 2022.

[71] Angwin and others (n 65).

supervised learning deals with within a broader network of abstractions.

As claimed at the outset of this paper, explanations or interpretations of the workings and outcomes of algorithms minimally require the explicitation of the level(s) of abstraction they implement. On a narrow reading, this means that the features that are taken into account for making a certain prediction should be enumerated. As soon as we think of an abstraction as a combined selection of types and tokens, it becomes clear that a simple list of features that are taken into account is not sufficient. The entities (data-subjects) that are classified with these features need to be enumerated as well. This requirement can apply to different sets of entities: entities used to identify patterns between different features (the learning-process), entities used to choose which classifier or learned function should be chosen (the validation-process), entities used to estimate the accuracy of the classifier (the test-process) and finally the entities actual predictions are made about. The relevance of these different sets implies that, even at the level of explanations, we cannot restrict our attention to a single proxy-target relation. Whereas Selbst, boyd and others argue that we should look beyond this algorithmic frame and pay more attention to the broader sociotechnical system, the model developed in the previous section only allows us to situate the algorithmic frame within a broader network of abstractions.

In 'Criticising abstractions' above, the resulting relative critique of abstractions was contrasted with absolute critiques. In 'Critical formal methods!' above, the choice to develop this relative critique within a formal setting was defended. The question that remains to be answered is how this can help us with explanation, justification, and contestation. I consider each in turn.

Even when understood as mere explicitations of the relevant levels of abstractions, *explanations* become much richer when they are placed in a broader network of abstractions. There are more than just two abstractions (the proxy and the target), the relevant abstraction occurs alongside two separate dimensions (types *and* tokens) and abstractions are nodes in a network rather than isolated islands. At the very least, this shows that explanations are not limited to the explicitation of a single set of features that a classifier takes into account. The algorithmic frame might single out a unique and well-defined set of features that are used to make predictions, but this set is only one of several selections of features and individuals that might have to be made explicit.

When it comes to *justifications*, the main benefit comes from the Method of Abstraction itself, namely from the requirement that each property attributed to a system should be relative to a specific and explicit level of abstraction (the meta-level abstractions). This includes that when a model or system is claimed to be accurate to a certain degree, one should make clear how narrowly or broadly the system was bounded, and which of its features were taken into account. Networks of abstraction can help us to characterise our perspective on a system. When we situate proxy-target relations within a broader network of abstractions, it is immediately clear that the reliability of a proxy is tied to a specific use and proxy-target relation. The reliability of a proxy is, as such, attributed with more precision and situated relative to several alternatives, which makes it harder to overstate its significance. This is especially the case when abstractions are understood as sets of types and tokens. Here too, we see that there is no unique and well-defined way of understanding the accuracy of the predictions made by a system (in general). The accuracy reported by the makers of that system is not only *just an estimate*, but

it is also a measure that is tied to a specific perspective. While the maker of a system might be interested in estimating the performance relative to a specific benchmark dataset, or might try to find out how it performs in a given population, a critical researcher might ask about its performance relative to a different data-set or a data-subject might inquire after the expected accuracy for a given subset of the intended population. All such additional questions arise from shifting from one Level of Abstraction to another; looking differently at a system or focusing on a different part of a network of abstractions.

The benefits for *contestation* should by now be obvious. If there is more than one way to explain and more than one way to justify, then surely there is plenty of space to contrast the explanations and justifications that are provided by the maker or operator of a system with alternatives that better reflect the needs of a data subject.

To conclude, I would like to clarify once more what the present proposal does and does not entail. The proposal to situate the algorithmic frame within a broader network is not intended as a way to identify better proxies or means to improve predictions by taking into account larger parts of the context (a familiar *better tools with more and better data* narrative!). It is a critical tool in the sense that it is a tool to uncover limitations of the algorithmic frame and the properties, such as the accuracy of predictions based on a given proxy, we attribute to it. Supervised learning can tell us to what extent the proxy-target relation that is taken into account by the algorithmic frame is reliable. Looking beyond this initial frame and considering a wider range of abstractions and proxy-target relations then reveals the limits of what a single accuracy-claim can tell us, and which other questions we might want to ask about the system. Unlike what critics of abstraction suggest, it does not refer to an unabstracted reality, but only seeks to expose the limits of a given abstraction by moving to a different or more refined abstraction. Unlike what Selbst, boyd and others propose, this approach does not turn to other disciplines or refer to a technical/social dichotomy to develop a critical assessment of abstractions. As such, it takes a step towards a better understanding of the critical potential of formal methods in the context of algorithmic decisions.

References

Allo P, 'Reasoning about Data and Information: Abstraction between States and Commodities' (2009) 167(2) Synthese 231

—, 'A Constructionist Philosophy of Logic' (2017) 27(3) Minds and Machines 545

—, 'The Epistemology of Non-distributive Profiles' (2020) 33(3) Philosophy & Technology 379

Alston WP, 'Internalism and Externalism in Epistemology:' (1986) 14(1) Philosophical Topics 179

Angwin J and others, 'Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.' [2016] Pro Publica <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Anrig B, Browne W and Gasson M, 'The Role of Algorithms in Profiling' in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008)

Barocas S, Selbst AD and Raghavan M, 'The hidden assumptions behind counterfactual explanations and principal reasons' (ACM January 2020)

Barwise J and Seligman J, *Information Flow: The logic of distributed systems* (Cambridge Tracts in Theoretical Computer Science, vol 44, Cambridge University Press 1997)

Biran O and McKeown K, *Justification Narratives for Individual Classifications* (techspace rep, 2014)

Bowker GC and Star SL, *Sorting Things Out: Classification and its Consequences* (MIT Press 1999)

Cath CJN, 'Changing Minds and Machines: A Case Study of Human Rights Advocacy in the Internet Engineering Task Force (IETF)' (DPhil in Information, Communication, and the Social Sciences, University of Oxford 2021)

Colburn T and Shute G, 'Abstraction in Computer Science' (2007) 17(2) Minds and Machines 169

Dourish P, 'Algorithms and their others: Algorithmic culture in context' (2016) Big Data & Society 1

—, *The Stuff of Bits* (MIT Press 2017)

—, 'What We Talk About When We Talk About Context' (2004) 8(1) Personal and Ubiquitous Computing 19

Dutilh Novaes C, *Formal Languages in Logic: A Philosophical and Cognitive Analysis* (Cambridge University Press 2012)

Floridi L, *The Logic of Information* (Oxford University Press 2019)

—, *The Philosophy of Information* (Oxford University Press 2011)

—, 'Semantic Conceptions of Information' in EN Zalta (ed), *Stanford Encyclopedia of Information* (2005)

—, 'Turing's three philosophical lessons and the philosophy of information' (2012) 370(1971) Philosophical Transactions of the Royal Society A Mathematical, Physical and Engineering Sciences 3536.

Fourcade M and Healy K, 'Classification Situations: Life-chances in the Neoliberal Era' (2013) 38 Accounting, Organizations and Society 559

Hájek A, 'The Reference Class Problem is Your Problem Too' (2007) 156(3) Synthese 563

Harford T, 'Big Data: A Big Mistake?' (2014) 11(5) Significance 14

Hastie T, Tibshirani R and Friedman J, *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd, Springer 2009)

Hildebrandt M, 'Defining profiling: A new type of knowledge?' in *Profiling the European Citizen: Cross-Disciplinary Perspectives* (Springer 2008)

—, 'Law as Information in the Era of Data-Driven Agency' (2016) 79(1) Modern Law Review 1.

Hildebrandt M and Gutwirth G (eds), *Profiling the European Citizen. Cross-Disciplinary Perspectives* (Springer 2008)

Jacobs AZ and Wallach H, 'Measurement and Fairness' (FAccT '21, Association for Computing Machinery 2021) <https://doi.org/10.1145/3442188. 3445901>

Kitchin R, 'Thinking critically about and researching algorithms' (2017) 20(1) Information Communication and Society 14

Koch B and others, 'Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research' (35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia, arXiv 3 December 2021) <http://arxiv.org/abs/2112.01716> accessed 10 July 2022

Larson J and others, 'How We Analyzed the COMPAS Recidivism Algorithm' [2016] Pro Publica hhttps://www.propublica.org/article/how-we-analyzedthe-compas-recidivism-algorithmi

Leonelli S, 'What Counts as Scientific Data? A Relational Framework' (2015) 82(5) Philosophy of Science 810

Lipton ZC, 'The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery.' (2018) 16(3) Queue 31

Loi M, Ferrario A and Viganò E, 'Transparency as Design Publicity: Explaining and Justifying

Inscrutable Algorithms' (2021) 23(3) Ethics and Information Technology 253

McCormack D, 'Geography and abstraction: Towards an affirmative critique' (2012) 36(6) Progress in Human Geography 715

Mittelstadt B, Russell C and Wachter S, 'Explaining Explanations in AI' (FAT* '19, Association for Computing Machinery 2019)

Moerel EL and Storm M, 'Automated Decisions Based on Profiling: Information, Explanation or Justification — That Is The Question!' [2019] SSRN Electronic Journal.

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

O'Neil C, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown 2016)

Paullada A and others, 'Data and Its (Dis)Contents: A Survey of Dataset Development and Use in Machine Learning Research' (2021) 2(11) Patterns 100336 <https://linkinghub.elsevier.com/retrieve/pii/S2666389921001847> accessed 10 July 2022

Porter TM, *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life* (Princeton University Press 1995)

Raghavan M and others, 'Mitigating bias in algorithmic hiring: evaluating claims and practices' (ACM January 2020)

Ribeiro MT, Singh S and Guestrin C, 'Why Should I Trust You?": Explaining the Predictions of Any Classifier' (KDD '16, ACM 2016)

Rouvroy A and Stiegler B, 'The digital regime of truth: from the algorithmic governmentality to a new rule of law' [2016] (3) La Deleuziana 6

Schermer BW, 'The limits of privacy in automated profiling and data mining' (2011) 27(1) Computer Law & Security Review 45

Schlangen D, 'Targeting the Benchmark: On Methodology in Current Natural Language Processing Research' in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers, Association for Computational Linguistics 2021)

Selbst AD and Barocas S, 'The Intuitive Appeal of Explainable Machines' (2018) 87(3) Fordham Law Review 1085

Selbst AD and others, 'Fairness and Abstraction in Sociotechnical Systems' (FAT* '19, Association for Computing Machinery 2019)

Smart A and others, 'Why Reliabilism Is Not Enough: Epistemic and Moral Justification in Machine Learning' (AIES '20, February 7–8, 2020, Association for Computing Machinery 7 February 2020)

Suárez M, 'Scientific Representation' (2010) 5(1) Philosophy Compass 91

Swoyer C, 'Structural representation and surrogative reasoning' (1991) 87(3) Synthese 449

Van Fraassen BC, *The Scientific Image* (Oxford University Press 1980)

Vapnik VN, 'An overview of statistical learning theory' (1999) 10(5) IEEE transactions on neural networks 988 (IEEE)

Wachter S, Mittelstadt B and Floridi L, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) International Data Privacy Law 76

Wachter S, Mittelstadt B and Russell C, 'Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR'

(2018) 31(2) Harvard Journal of Law & Technology 841

Wallmann C and Williamson J, 'Four Approaches to the Reference Class Problem' in G Hofer-Szabó and L Wroński (eds), *Making it Formally Explicit* (Springer 2017)

Wheeler G, 'Machine Epistemology and Big Data' in L McIntyre and A Roseberg (eds), *The Routledge Companion to The Philosophy of Social Science* (Routledge 2016)

Wild CJ, 'Statistical literacy as the earth moves' (2017) 16(1) Statistics Education Research Journal 31

A reply: Abstractions remain useful tools for algorithmic governance

Sandra Wachter • University of Oxford, sandra.wachter@oii.ox.ac.uk

In *The Interpretability Problem and the Critique of Abstraction(s)* Patrick Allo assesses criticisms of 'abstraction', a tool of computer science used 'to construct and reason about computational processes', as deployed in legal and social scholarship on machine learning. The paper describes a proposal for framing normative challenges and sociotechnical characteristics of supervised learning systems through the method of Levels of Abstraction. Allo focuses in particular on the analysis of Selbst, boyd and others who identify five 'abstraction traps' in work on fairness in machine learning (fairML) arising from 'the way in which the domain-specific aspects of the problem —broadly, the social context — are abstracted so that machine learning tools can be applied' across different domains.[1]

Allo's proposal to situate analysis of interpretability and related normative problems such as fairness within a 'network of abstractions' is best understood in relation to the insightful framing proposed by Selbst, boyd and others. When analysing social challenges arising from machine learning, they recognise three 'frames' or viewpoints from which such challenges can be conceptualised: the (1) algorithmic frame, (2) data frame and (3) sociotechnical frame.[2] These frames exist at increasingly higher Levels of Abstraction (LoA).[3] The algorithmic frame consists of 'the most common abstractions in machine learning… representations (of data), and labelling (of outcomes).' Within the algorithmic frame, 'the efficacy of an algorithm is evaluated as properties of the output as related to the input.' Above this is the data frame which, in addition to the algorithmic frame, also includes the algorithm's inputs and outputs, allowing for direct interrogation of underlying choices such as 'choices of representations and labels' and how they 'might affect the resulting model.' Finally, above these is the sociotechnical frame which locates a machine learning model within sociotechnical systems, allowing for broader elements of the context of use to be interrogated such as 'humans and human institutions within the abstraction boundary.'[4]

Selbst, boyd and others argue for a reorientation of fairML towards sociotechnical frames and a science and technology studies (STS) lens and away from the abstractions characterising the algorithmic and data frames. This transformation can help resolve seemingly intractable problems in the field, such as how to

1 Andrew D. Selbst and others, 'Fairness and Abstraction in Sociotechnical Systems' in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019) 59, 60.

2 Ibid.

3 Luciano Floridi, 'The Method of Levels of Abstraction' (2008) 18 Minds and Machines 303.

4 Selbst and others (n 1).

decide the 'right' measure of fairness among dozens of possibilities.[5]

Allo's proposal aims to expand the perceived utility of abstractions for problems such as fairness in machine learning while providing a method and framework to critically assess abstractions. He argues that critical analysis of abstractions can reveal greater limitations of models in relation to problems like fairness or interpretability without needing to adopt a different frame above the algorithmic frame. He proposes that situating 'the algorithmic frame within a broader network of abstractions' and 'proxy-target relations' reveals the utility of abstractions. This re-framing can 'uncover limitations of the algorithmic frame and the properties, such as the accuracy of predictions based on a given proxy, we attribute to it.' Bringing further abstraction and refinement relations into consideration can reveal limitations of the current algorithmic frame, such as the dependency of individual measures of model accuracy on a particular perspective or dataset.

From the perspective of law, this reframing of abstractions is informative for several domains. Allo explains that the granularity of classifications is constrained by the number of tokens (or subjects) and types (or features) available to the classifier. This observation is highly relevant when considering how to align fairML with equality law. In cases of EU non-discrimination law, for example, determining the appropriate composition of the harmed group and comparator group is a highly complex normative question. In practice, group compositions are often determined by the availability of information about individual members, group features, and population statistics to parties in the case. Setting appropriate group boundaries is of critical importance to identify intersectional discrimination and ensure inequality is not hidden by increasing the granularity of sub- groups.[6]

Whereas interpretability solutions often focus on disclosing partial or complete lists of model features, Allo's reframing reveals the importance of disclosing information about both tokens and types, or people as subjects and model features, to ensure claims about the fairness of models can be fully and appropriately assessed to live up to existing legal standards.[7] This finding is also relevant to broader standards for transparency in ML.[8] Allo's reframing reveals that manufacturer or developer claims about model performance (e.g. single-measure accuracy, fairness) can only be critically assessed if information about model features is accompanied by information about the subjects or tokens used. This requirement has clear implications for both individual and group privacy in relation to how training and test datasets are constructed, labelled and deployed.[9] Adding this requirement to

5 Selbst and others (n 1); Sandra Wachter, Brent Mittelstadt and Chris Russell, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review 105567.

6 Wachter, Mittelstadt and Russell (n 5).

7 Ibid.

8 Sandra Wachter, Brent Mittelstadt and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation', (2017) 7(2) International Data Privacy Law 76; Mike Ananny and Kate Crawford, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2018) 20(3) New Media & Society 973; Brent Mittelstadt, Chris Russell and Sandra Wachter, 'Explaining Explanations in AI' in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019) 279.

9 Brent Mittelstadt, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30 Philosophy & Technology 475; Sandra Wachter and Brent Mittelstadt, 'A right to reasonable inferences: re-thinking data protection law in the age of Big Data and

transparency requirements found, for example, in the General Data Protection Regulation, EU Artificial Intelligence Act or technical standards could prove highly disruptive by tightening pre-deployment certification and post-market surveillance requirements for ML. Further, as reflected in Allo's discussion of using 'knowledge extracted from one set of individuals to make claims (predictions) about other individuals,' it could empower data subjects who have little say over how their data is used to train ML systems and draw privacy invasive inferences about them.[10]

Overall, Allo's proposal to situate the algorithmic frame within a network of abstractions is highly insightful across regulatory frameworks addressing machine learning. His contribution can help operationalise transparency ideals and inform robust and appropriately critical evidence and testing requirements to contextualise manufacturer claims about ML systems. Recognising when a shift of LoA has occurred in answering normative questions is essential to the future of ML governance.

References

Ananny M and Crawford K, 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability' (2018) 20(3) New Media & Society 973

Floridi L, 'The Method of Levels of Abstraction' (2008) 18 Minds and Machines 303

Mittelstadt B, Russell C and Wachter S, 'Explaining Explanations in AI' in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019) 279

Mittelstadt B, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30 Philosophy & Technology 475

Selbst, AD and others, 'Fairness and Abstraction in Sociotechnical Systems' in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019) 59

Wachter S, Mittelstadt B and Floridi L, 'Why a Right to Explanation of Automated Decision-making Does Not Exist in the General Data Protection Regulation' (2017) 7(2) International Data Privacy Law 76

Wachter S, Mittelstadt B and Russell C, 'Why fairness cannot be automated: Bridging the gap between EU non-discrimination law and AI' (2021) 41 Computer Law & Security Review 105567

AI' (2019) Columbia Business Law Review 494; Linnet Taylor, Luciano Floridi and Bart van der Sloot (eds), *Group Privacy: New Challenges of Data Technologies* (Springer 2017).

[10] Wachter and Mittelstadt, ibid.

Author's response: The epistemological and normative limits of abstraction

Patrick Allo

What do we object to when we challenge or question the use of an abstraction? One possible answer is that we disagree about whether a good trade-off has been achieved between remaining faithful to the reality that is being modelled and reducing the complexity of that reality.[1] Abstraction is not inherently wrong, but some abstractions achieve a better balance between the exclusion and inclusion of features of what is being modelled. This is, in my view, the critical perspective on abstraction we should adopt to tackle the interpretability problem.

An alternative answer sees abstraction as something inherently problematic. This, I argue, is the answer that many critical scholars favour. Selbst, boyd and others develop this stance more systematically than many others. As they see it, the limitations of the algorithmic frame are best assessed by situating it within a broader sociotechnical frame. In doing so, they emphasise that the sociotechnical frame is of a different kind than the algorithmic frame — they oppose open-ended sociotechnical systems to delimited technical (formal, abstract) systems. As Wachter explains, I accord more importance to the fact that such different frames can be understood as different levels of abstraction.

This is the core of the theoretical claim I make in my contribution. Wachter's reply does not only address this more principled claim, but also asks what it might mean for the law. I would like to take a closer look at two aspects of this questioning.

From the outset, Wachter refers to the problems of interpretability, transparency and fairness as normative problems. This framing deserves more attention, as my contribution does not explicitly engage with this normative dimension. Instead, I approach abstraction from a purely informational (or epistemic) perspective, and even agree with the method of supervised learning that abstractions can be evaluated by asking how good one abstraction can serve as a proxy for another target abstraction. My proposal, however, aims at a more encompassing confrontation with alternative abstractions; including abstractions based on different sets of individuals. What I tried to show is that a purely informational understanding of the limits of abstractions has critical potential.

Whether this take on the problems of interpretability, transparency and fairness is sufficient to deal with all its normative implications, remains an open question. Wachter's exploration of how a critical analysis of abstractions could be useful from the perspective of the law reveals that normative assessments might indeed

[1] This formulation is inspired by and follows closely the account of formalisation given in Catarina Dutilh Novaes, *Formalizing Medieval Logical Theories: Suppositio, Consequentiae and Obligationes* (Logic, Epistemology, and the Unity of Science 7, Springer Netherlands 2007) 216 (I am indebted to Benedict Löwe for drawing my attention to this specific formulation). Abstraction and formalisation are distinct modelling practices, but they often go hand in hand and face similar issues.

benefit from a more fine-grained assessment of abstractions. The role of group-composition in EU non-discrimination law suggests that normative questions can depend on the availability of sufficiently fine-grained information about the composition of groups. Likewise, the limits of current interpretability solutions do seem to call for more fine-grained and more informative ways of reporting the predicted accuracy of a model for specific groups. In both cases, we see that the normative issues are closely related to the limits of specific abstractions. The abstractions that typically make up the algorithmic frame can be too coarse to allow for a good normative assessment but alternative, more fine-grained abstractions might effectively improve such assessments.

References

Dutilh Novaes C, *Formalizing Medieval Logical Theories: Suppositio, Consequentiae and Obligationes* (Logic, Epistemology, and the Unity of Science 7, Springer Netherlands 2007)