Reverse engineering the "normativity" in machine learning: A rule-based modelling of data-driven decisions for contestation[*]

➢ *Normativity: the key to theorize transparency*

Theorising transparency with a view to see automated decision systems "at work" is a territory ever expanding as we attempt to map it.[1] The opacities, (in)transparencies and informational asymmetries inherent in machine learning (ML), resulting with a "mental invisibility" on the side of the individuals, may only be counteracted through a visibility of different type—namely, an *actionable transparency*[2] as an instrument to enforce rights. Based on this, what could follow Ruben Binns's premise, "*algorithmic decision-making necessarily embodies contestable epistemic and normative assumptions"*[3] is that, a systemisation of what transparency[4] can offer for the contestation of automated decisions, rather requires an understanding of the *system* as a regulatory process—containing *normativity* in different forms, constructs and disguises.[5]

As decision-making systems are goal-oriented, their behaviour may be attributed to the inherent values and assumptions guiding their response to a given input.[6] This allows us to infer certain normativity from the system's output as aiming to achieve some pre-set goals. Hence with normativity, we not only refer to the capacity to control and guide conduct but also to a claim, or contention, to do so which is ultimately reducible to some moral ground—say, a right to rule.[7] Since, by themselves, facts

---

[*] Emre Bayamlıoğlu, Tilburg University / TILT. The author is deeply thankful to Mireille Hildebrandt and Ronald Leenes for their invaluable comments, academic guidance, and for all the *good manners* they teach.

[1] As Introna notes 'the algorithm is a black box, which when opened simply introduce more black boxes, which when subsequently opened simply introduce more black boxes, and so forth'. Lucas Introna "Algorithms, performativity and governability" in '*Governing Algorithms: A Conference On Computation, Automation, and Control*' New York University, 16–17 May 2013 cited in Matthias Leese, 'The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union' (2014) 45 (5) *Security Dialogue* 498; Jenna Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms" *Big Data & Society*, 1-12 (2016).

[2] The transparency approach summarised in this piece is rooted in Hildebrandt's concept of "agonistic machine learning". Mireille Hildebrandt, 'Privacy As Protection of the Incomputable Self: Agonistic Machine Learning' (December 3, 2017, https://ssrn.com/abstract=3081776).

[3] Reuben Binns, "Algorithmic Accountability and Public Reason." *Philosophy & Technology*, 2017. For a general view, see Vries MJ, Hansson SO and Meijers (eds.), *Norms in Technology* (Springer: Netherlands 2013).

[4] C. Hood, 'Transparency in Historical Perspective,' in C. Hood and D. Heald (eds), *Transparency: The Key to Better Government?* Oxford, Oxford University Press 2006; A. Buijze, *The Principle of Transparency in EU Law*, 2013, Utrecht ; Roberts, J. (2009) 'No One is Perfect: The Limits of Transparency and an Ethic for "Intelligent" Accountability' *Accounting, Organization and Society* 34: 957–70; Andrea Brighenti, 'Visibility: A Category for the Social Sciences' *Current Sociology* 2007; Mike Ananny and Kate Crawford "Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability" *New Media & Society* 20, no. 3 (2016): 973-89.

[5] For a general account of normativity, see Sylvie Delacroix, *Legal Norms and Normativity: An Essay in Genealogy (*Hart Publishing Limited, 2006).

[6] Debora Hammond, *Science of Synthesis: Exploring the Social Implications of General Systems Theory*. (University Press of Colorado, 2010)

[7] Michael Giudice *Understanding the Nature of Law: A Case for Constructive Conceptual Explanation* (2015, Edward Elgar Pub); Stefano Bertea, *The Normative Claim of Law*, 11-12; Rouse, J. "Social practices and normativity" *Philosophy of the Social Sciences,* (2007) *37*(1), 46–56; Franssen, M. Artefacts and normativity, in A. Meijers (Ed.), *Handbook of the philosophy of science: Vol. 9: Philosophy of technology and engineering sciences (*2009, Oxford: Elsevier), 923–952; George Pavlakos et. al., *Reasons and Intentions in Law and Practical Agency*, Cambridge University Press (2015);

(data) cannot provide "reasons for action"[8], looking from the lens of normativity informs us about the motives, assumptions and the further decisional criteria underlying the systems, and thus, opens the way to a normative evaluation of the observed behaviour/action.[9]

Accordingly challenging the truth claim or the accuracy of a decision, prescribing of "what ought to be" in a given situation, will initially require a conceptualisation of the outcome as a process based on ***facts, norms*** and ***decisions/effects*** in the most abstract sense. In the context of automated decisions based on personal data processing, this would simply imply how and why a person, event, or situation is classified in certain ways, and what consequences follow from that. Such modelling, which maps input/data with the effects/consequences within a contemplated normative framework, provides us with a rule-based "explanation" of the system which helps contextualise the decision at the appropriate level of generality for the purposes of contestation.[10]

➢ *A rule-based modelling of transparency: reverse engineering the implicit normativity*

Following from above, a rule-based explanation will mean that given certain factual input, the result could be verified, justified, or alternatively contested with reference to a certain set of rules (normative framework) inherent in the system. Concrete transparency requirements of the rule-based modelling as an operable scheme enabling effective contestation of data-driven automated decisions will require clarities, verifications and justifications with regard to the below "informational components" of the system as a regulatory process—the *transparency desiderata*.

- *Provided, observed and inferred dat;, data types; data structures; together with all the derived representations and inferences: factual input*[11]
  The core decisional features of the system as a regulatory process: decisional criteria (norms)
- *The impact and context of the decision in a regulatory perspective*
- *The responsible actors (agency) behind the decisions or the benefits accrued*

As a normative construct, the rule based-modelling and the ensuing informational components do not aim to analyse the system by the mechanisms of its operation, but rather by the normativity embedded in its behaviour/action.[12] Such modelling entails a more abstract and multi-layered conception of

---

[8] Joseph Raz, *The Authority of Law* (Oxford: Clarendon Press, 1979).

[9] As Coglianese and Lehr put it: *[…]in the rulemaking context machine learning would need to be nested within a larger decision-making model in order to support automated regulatory decisions. Machine learning predictions would, within an agent-based simulation, inform agents' actions, which in turn would generate predicted outcomes from different regulatory permutations*. Cary Coglianese and David Lehr, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era" (2017).

[10] "Explanation" here is not limited to the concepts of technical analysis of AI-systems (e.g. *local explanation* or *local counterfactual faithfulness*) but rather used in the broader sense to refer to the efforts to render decisions of data-driven systems *reviewable on normative grounds*. For more on the concept of explanation, see Doshi-Velez, Finale, et. al. "Accountability of AI Under the Law: The Role of Explanation".

[11] Within this perspective, the concept of "data" is regarded not as a tool of insight, but simply as informational or factual input similar to the facts in a legal case. This is where the observations and the feedback in the form of data are transformed into factual input (constructed as representations of "reality") for the system. What we intend to encapsulate by the concept of "factual input" is all the inferences and representations ("data derivatives") which relate to the world/reality, and serve as the basis for the operation of decisional norms. The abstract concept of factual input is the product of the effort to differentiate between "rules of fact-making" and "rules of decision-making". In the legal domain, decisions or judgments are reached, first, by the establishment of facts (in light of the relevant rules—akin to "constitutive rules"), and second, through the application of the norm (in light of the relevant facts). However, such distinction is difficult and questionable in the ML context. Will address this issue in the final part.

[12] As Leenes clearly notes *[...] in the case of automated decision making about individuals on the basis of profiles, transparency is required with respect to the relevant data and the rules (heuristics) used to draw the inferences. This allows the validity of the inferences to be checked by the individual concerned, in order to*

transparency which equally takes into consideration both the outcome and the process itself.[13] Rather than reflecting on the underlying algorithmic processes, it reverse engineers the decisional process for a reconstruction on the basis of facts, norms and the resulting effects; and by doing so, it employs *a synthetic* method aiming to acquire an understanding of the realit or the phenomenon by means of model-building.[14]

The informational components (*transparency desiderata)* are intended as model-agnostic formulations which may not be seen as independent items of check but rather need to be implemented in a systemic way— treating each *desideratum* as an indispensable constituent of a framework which eventually aims to render automated decisions contestable on normative grounds.[15]

To some extent, the idea here, is not to interpret the domain of ML through legal knowledge but to define legal requirements which would render the data-driven systems more responsive, communicative and engageable from the legal or regulatory perspective. Rule based modelling is not a top-down initiative ordering system owners and engineers how they should design their systems but rather a bottom-up call from the view of the informed data subject simply formulating what the totality of the data-driven activities entail for review and contestation.[16]

➢ *Impediments between facts and norms*

As a theoretical construct, the rule-based modelling and the ensuing informational components draw the horizon of the desirable (but not necessarily the possible or the optimal) without any regard to the feasibility or technical, legal, or epistemological permissibility of these components. A viable implementation of the rule-based model requires a balancing of the trade-offs arising out of the impediments inherent to data-driven decisions, namely i) the legal limits: security, integrity and commercial secrecy; ii) the physical limits due to computational complexity; and iii) the economic feasibility in consideration of the risks. As such, each component needs implementation at various levels through different tools in a manner reconciling a balance among the risks, computational difficulties and the economic constraints—while also taking into account the legal and the systemic integrity concerns (e.g., to prevent competitors from reverse-engineering a particular scorer's model or customers from gaming the smart grid). Since the ingredients and conditions of realization for each *informational component* may vary depending on the nature of the analysis together with its scope,

---

notice and possibly remedy unjust judgements. Ronald Leenes, "Reply: Addressing the Obscurity of Data Clouds" in Mireille Hildebrandt and Serge Gutwirth (eds.) *Profiling the European Citizen: Cross-disciplinary Perspectives, 293-300*. An approach closer to this modelling may be found in the paper by Sorelle A. Friedler and others, that focuses on transparency and fairness issues in embedded value systems with a view to explore the mapping from the *observation* and *construct space* to *decision space*. Sorelle A. Friedler et. al, "On the (im)possibility of fairness" 2016 arXiv:1609.07236v1.. For a more practical but similar formulation, see Mikella Hurley and Julius Adebayo, "Credit Scoring In The Age of Big Data," *Yale Journal of Law and Technology*: Vol. 18 : Iss.1, 196. Also, see Bert – Jaap Koops, "Reply: Some Reflections on Profiling, Power Shifts and Protection Paradigms" in Mireille Hildebrandt and Serge Gutwirth (eds.) *Profiling the European Citizen: Cross-disciplinary Perspective,* 326-337

[13] Bert-Jaap Koops, 'On Decision Transparency, or How to Enhance Data Protection after the Computational Turn' 196-220.

[14] Alcibiades Malapi-Nelson, *The Nature of the Machine and the Collapse of Cybernetics: A Transhumanist Lesson for Emerging Technologies* (Cham, Switzerland: Palgrave Macmillan, 2017), 134.

[15] "*The purpose of identifying operating parts and determining their organization is to go beyond describing the phenomenon to showing how the working entities cause and constitute the phenomenon*." Marta Halina, "Mechanistic Explanation and Its Limits" in Stuart Glennan and Phyllis Illari (eds.*) The Routledge Handbook of Mechanisms and Mechanical Philosophy*. (London: Routledge, Taylor & Francis Group, 2017) 213-225, 217.

[16] This approach inevitably prioritises procedural regularity and intelligibility over efficiency. "*Procedural regularity is a core idea behind due process: the state cannot single out an individual for a different procedure.*" Kroll et. al. "Accountable Algorithms" (2017) *University of Pennsylvania Law Review*, 679.

intensity, and duration; the appropriate tools and the necessary forms of transparency (e.g., notification/disclosure, audit and design principles) cannot not be detailed in the abstract but require further refinement in light of the specifities of the domain together with the context of the data operations in hand. Nevertheless, for the purposes of provocation, we may identify some preliminary theoretical gaps that are yet to be bridged.

* * *

As mentioned above, the rule-based modelling, which is intended as a normative reading[17] of the totality of computational expression put forward by the system, treats data as input (stimuli) triggering certain operational processes followed by the effects in the form of classification/decision. Any regulator, whether it is in the realm of the law or within other normative/regulatory frameworks, will weigh various factors, and decide what norm should be applicable in case of a certain constellation of facts. Thus, our contestation model starting with the facts, secondly requires some normative understanding of the reasons giving rise to a particular decision. Take the example of a data-driven health insurance system which is constructed, among others, on the premise that eating deep-fried foods is an indicator of bad health. Based on the assumption that a deep-fried diet is the major cause of cardiovascular problems, the data analysis may decide that those searching for deep fryers through online retailing websites are in a risky category. Seen from the normative perspective, in automated decisions based on personal data processing, we can identify normativity primarily at two levels: first, for the determination of facts through inference rules/mechanisms (e.g. the assumed relation between the search for deep fryers and eating deep-fried), and the second, for the determination of the consequent effects (being classified as risky) based on the decisional norms embodied in the system. For instance, speech analysis in a micro-targeting campaign can detect one's dialect and, irrespective of legal or ethical admissibility of such inquiry, dialect is a factual input accuracy or validity of which may be empirically challenged on the basis of first-order experience or other conventional verification methods. On the other hand, the selection of the suitable online political content based on this "factual" finding is the result of the decisional norm which should be regarded as distinct from the fact-generating mechanisms(rules) used to infer one's dialect.

Although theoretically every decision regarded as "rational" can be, albeit in varying abstraction, decomposed to infer which rules have been followed in what order; in case of automated decisions, facts and rules do not part or differentiate as easily as the way conventional lawyers are accustomed to. The formulation of the factual input may be so complex and unstructured that it may conflate the fact-generating inference process(rules) and the decisional norms (criteria). An example may be the detection of social relationship between persons based solely on acoustic and conversational characteristics. In a given speech analysis task, the degree of intimacy between the parties of a phone conversation may be the target variable to be predicted through a set of selected features. Accordingly, the relevant inference rule may provide that the length of silent pauses in a phone conversation is a predictor of intimacy between the parties (longer silent gaps during the conversation means more intimacy), resulting with the application of a certain decisional norm (e.g. discard calls with an intimacy score of "X" for surveillance purposes). Apparently, the intimacy between persons is not a kind of factual input like eating deep-fried food but rather more judgmental and value-laden characteristic, contestation of which would require a different argumentation.[18] Similarly, we can think

---

[17] In this context, Stiegler's very concept of *grammatization (as a theoretical framework for orienting rhetorical inquiry)* may be a useful avenue for further inquiry. John Tinnell, "Grammatization: Bernard Stieglers Theory of Writing and Technology" *Computers and Composition* 37 (2015): 132-46.

[18] *"[w]hile reasoning about the facts can (at least in principle) still be regarded as probabilistic, reasoning about normative issues clearly is of a different nature. Moreover, even in matters of evidence reliable numbers are usually not available so that the reasoning has to be qualitative."* Henry Prakken, 'Logics of

of a college admissions process which, for example, take personal grit together with high school scores as the important factors for entry. Since grit cannot be measured or verified as the high-school grades, taking grit as input (calculating the *incalculable* with far too remote inferences) could be a way to conceal decisional norms.

The epistemological effort to keep decisional norms and rules behind factual inferences distinct is a key challenge in terms of articulating the embedded normativity inherent in the system. Leaving aside the inaccuracy of calculation, or the problem of passing spurious correlations as causation; refuting of a factual inference is one thing, and challenging of the decisional criteria which underlie a specific result is quite another. In ML context, even if we could identify certain rule-based factors affecting the decision, the problem lies in determining what it takes for a rule to be a decisional norm, and when we are faced with a rule of fact-making. This stands as a major difficulty in terms of differentiating between the inference rules (mechanisms) generating *factual input* and the *decisional norms (criteria)* interpreting the results according to the assumptions and legitimations relating to the wider objectives of the system in use.[19] IN sum, ML is fraught with the problems of distinguishing between facts and norms—a case of *normativities* within *normativities*.

* * *

So, the question remains: Is rule-based modelling a viable approach in that whether the contemplation and construction of automated decisions on the basis of *facts, norms* and *effects* could be enforced as a design choice? Whatever the chances of *ex-ante* implementation[20], there are always instances and situations where the normativity implicit in the system could not be articulated at a general level by a review of the system in the abstract. This is primarily because adaptive systems operate on dynamic correlation patterns where the decisional rule itself emerges autonomously from the streaming data. The "norm" is no longer predetermined, but constantly adjusted. Such *fluid hypotheses* [21] make any challenge on normative credentials of the system hard to formulate; thus, the decisional *criteria* remain vague and cannot be pinned down in sufficient precision. Rather than being based on factual identifications, categorizing through data could be seen as social procedures that initially create the groups they aim to define.[22] The so-called neutrality of data somehow naturalises this segmentation, and falsely renders its own construction—or say, normativity— invisible as a regulatory process.

Against this fuzzy entanglement of facts and norms in ML context, there are various efforts to develop methodologies and algorithmic tools explaining black-box models. An example of such efforts is the LIME[23] project which aims to disclose the implicit rules behind predictions, while taking into account

---

Argumentation and the Law' in H. Patrick Glenn and Lionel D. Smith (eds.), *Law and the New Logics* (Cambridge University Press 2017) 3-32, 4.

[19] So far, some theoretical hints for this conundrum is found in Sandra Dewitz's work which speaks of four kinds of predicates in legal reasoning: *purely descriptive predicates, descriptive-interpretive predicates, evaluative predicates*, and *normative predicate*." Sandra Dewitz, "Using Information Technology as a Determiner of Legal Facts" in Z. Bankowski, et al. (eds.), *Informatics and the Foundations of Legal Reasoning,* (1995 Kluwer Academic Publishers) 357-369.

[20] *[…] we should acknowledge that to a large extent the methodological integrity of the machine learning requires advance specification of the purpose as this will inform the solidity and productivity of the relevant research design*. Mireille Hildebrandt, "Privacy as protection of the incomputable self: Agonistic machine learning", 13.

[21] Matthias Leese, 'The new profiling: Algorithms, black boxes" 505-506.

[22] Karoline Krenn, "Markets and Classifications - Constructing Market Orders in the Digital Age. An Introduction" *Historical Social Research*. 2017, Vol. 42 Issue 1, 14.

[23] LIME (Local Interpretable Model-Agnostic Explanations) is a "model induction" technique that experiments with any given machine learning model—as a black box—to infer an approximate explainable model. It primarily works on text classifiers.

the human limitations (e.g. the explanations should not be too long). The idea is to design an interpretable model by taking on the predictions of a supposedly uninterpretable (black-box) model.[24] The tools for this purpose generally focus on the importance-measuring methods that operate on the individual level explaining what most important variables were for a specific result.[25]

As a final remark, it is important to acknowledge that machine learning and the sphere of automated decisions are not monolith concepts, and they have bifurcated implications resulting with diversely harmful effects. The rule-based modelling deals with the type of harms which may be individually contested on normative grounds such as unfair treatment or due process violations. There also exist other methods for detecting and ameliorating various different types of harms—e.g., invasiveness, group-level harms, harms from economic manipulation or exclusionary practices, and last but not the least, societal harms such as disrespect to human dignity— which cannot be effectively challenged or remedied under an individual contestation scheme but require novel contemplations.

[24] Marco Tulio Ribeiro, et. al., "Why should I trust you?: Explaining the predictions of any classifier" in *Knowledge Discovery and Data Mining (KDD)*, 2016; Ross, Andrew Slavin, Michael C. Hughes, and Finale Doshi-Velez. "Right for the Right Reasons: Training Differentiable Models by Constraining Their Explanations" *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, 2017. For a similar "norm inference" approach, see Daniel Kasenberg et. al., "Norms, Rewards, and the Intentional Stance: Comparing Machine Learning Approaches to Ethical Training". Also, see Marco Tulio Ribeiro's blog. https://homes.cs.washington.edu/~marcotcr/blog/lime/

[25]S. Datta, and Y. Zick, "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems" in 2016 I.E. Symposium on Security and Privacy (SP) 598– 617.