In: Emre Bayamlıoğlu, Irina Baraluic, Liisa Janssens and Mireille Hildebrandt (eds), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, 2018: Amsterdam University Press, 30-33. DOI 10.5117/9789463722124/CH4.

On the possibility of normative contestation of automated data-driven decisions

*Emre Bayamlıoğlu**

Abstract

This provocation intends to identify the possible requirements of a transparency model which aims to decompose and analyse automated decision-making systems not by the mechanisms of their operation but rather by the normativity embedded in their behaviour/action. For the effective contestation of automated decisions, essential components of a Rule-based Model (RbM) are briefly described as: i) the data as 'decisional input', ii) the 'normativities' contained by the system both at the inference and decision (rule-making) level, iii) the context and further implications of the decision, iv) the accountable actors.

Keywords: algorithmic transparency, data-driven decision-making, GDPR Article 22, techno-regulation

Introduction

Theorising transparency to see automated decision-making systems "at work" is a territory ever expanding as we attempt to map it (Leese 2014; Burrell 2016). The opacities and informational asymmetries inherent in machine learning (ML) result in a "mental invisibility" on the side of individuals that may only be counteracted through a visibility of different type. For the purposes of normative contestation, e.g. the one provided under Article 22 of the GDPR, this visibility should be an 'actionable transparency', an instrument to an effective and practical enforcement of rights (Hildebrandt 2017). Based on this, the provocation in hand proposes a follow-up on Ruben Binns's premise that 'algorithmic decision-making necessarily embodies contestable epistemic and normative assumptions' (2017, 4). The aim is to provide a systemisation of transparency requirements that enables the contestation of automated decisions, based on a 'reconstruction' of the system as a regulatory process containing different types of 'normativity'.

Normativity as a key to understand automated decisions

Regulatory systems are goal-oriented. Their behaviour may eventually be attributed to the values and assumptions that are implied in the rules and standards which guide the systems' response to a given input. This allows us to expect a related 'normativity' in the system's output. Since, by themselves, facts (input data) cannot provide "reasons for action" (Raz 1979), looking through the lens of normativity informs us about the decisional criteria (norms) underlying the system, and thus opens the way to a rule-based (normative) evaluation of the observed behaviour/action.

Accordingly, challenging the truth claim or the accuracy of a decision, thus contesting 'what ought to be' in a given situation, will initially require a conceptualisation of the outcome as the result of a 'rule-based' process where certain input is rightfully matched with certain results— akin to a legal system where rules (norms) are applied to facts (input data) to make decisions (output data). In the context of automated decisions based on personal data processing, this would refer to how and why a person is classified in a certain way, and what consequences follow from that. As Leenes noted in *Profiling the European Citizen*: '[…]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 notice and possibly remedy unjust judgements' (Leenes 2008, 299).

Rule-based modelling (RbM): reverse engineering the 'normativity' in machine learning

A 'rule-based explanation' of a decision means that given certain decisional ("factual") input data, the decision (output) should be verifiable, interpretable, and thereby contestable with reference to the rules (normative framework) that are operational in the system. Following from above, the concrete transparency requirements of such a model entail an "explanation" about the following aspects of the system, to redefine it as a regulatory process:

Features as decisional cues: Any normative contestation will start with the knowledge of what the system relies upon about the world in order to make decisions. This requires a perspective which treats the concept of “data” not as a tool of insight, but simply as certain representational or constructed input for decisional purposes.

In a ML process, data instances exist as variables of descriptive features where each feature such as age, height and weight is a dimension of the problem to be modelled (Sorelle A. Friedler et. al 2016). Depending on the nature of the analysis and the type of data available, features may also contain more constructed and computed representations such as one’s habit of eating deep-fried food, educational level, speaking a dialect, or the level of intimacy between parties of a phone conversation. Features as decisional cues refers to the totality of the relevant data representations extracted from a set of variables. In case of personal data processing, a feature space maps how people will be represented as inputs to the algorithm. The objective of a ML model is the identification of statistically reliable relationships between the feature variables and some target variable (e.g. healthy or not, or at least 70% healthy). The features that a system infers to be significant and their relevant weightings help us understand which inputs (inferences) factored into a decision to get to the final result.

Normativity: Normative contestation of automated decisions can be based on two grounds, scrutinising two different types of ‘normativity’. First, decisions may be contested on the basis of the selection and construction of the relevant features that the decision relies upon. What is questioned here is whether inferences made by way of selected features are sufficiently informative and causally reliable for the given purpose, e.g. whether one’s search for deep-fryers suffices for the inference of one’s eating deep fried food, and consequently being classified as risky. The normativity of decisional cues (features) lies in their being formal constructions by way of if-then rules. Both the accuracy and suitability of the features together with the methodology used for their selection and construction could be subject to normative scrutiny.

Second, normativity operates as a set of rules (decisional norms) for the determination of the ensuing effects. Decisional norms describe how a certain ML outcome (target value) is translated into concrete results in a wider decision-making framework, e.g. a certain health risk resulting in an increased insurance premium in an automated health insurance system. The question is: what is the meaning of the target variable(s) obtained? For instance, what score (in numeric or other quantified form) would suffice for a successful loan, and most importantly why? This type of scrutiny eventually reaches back to the goals and values encoded in the system, together with the underlying assumptions and justifications (ratiocinations).

The ‘context’ and further consequences: To fully evaluate the automated decisions for the purposes of contestation, the context of the decision—the particular situation, environment or domain in which the decision is to be made—is a key piece of information. This primarily involves informing of the data subject about where the decision starts and ends, and whether the system interoperates with other data processing operations. Accordingly, which other entities and authorities are informed of the decision; and for what other purposes or in which other contexts the results could be used, are all crucial for a normative assessment**.** More importantly, the implementation of a transparency model, with contestation in mind, requires not only the knowledge of why a decision was made but also why a different decision was not made (Miller 2017; Lipton 2004).

Responsible actors: This is an essential component of an actionable transparency model, meaning that the implications of automated decisions must be situated and analysed in an institutional framework, revealing the parties and the interests behind the decisions. The ‘agency’ behind automated decisions is not necessarily monolithic but often related to a plethora of conflicting, competing and partially overlapping interests and objectives which are linked to multifarious commercial frameworks and stately functions. This highly fragmented and obscure landscape requires a purposeful mapping of the institutional structures and the intricate web of relations among those who may be responsible for different parts or aspects of a decision, i.e. the data brokers, public and private clients, service providers, regulators, operators, code writers and system designers. Lacking this particular dimension, the transparency model remains incomplete.

Impediments and pitfalls

Both the determination of the decisional cues and the ensuing results are normative undertakings which, in theory, may be reconstructed in the if-then form (if $\text{condition}_1 \wedge \text{condition}_2 \wedge \text{condition}_3$, then outcome). Thus, theoretically every decision that is claimed to be “rational” can be decomposed to infer which rules have been followed in what order. However, in case of automated decisions, neither the input inferred nor the rules that produce the outcome reveal themselves easily. Problems are not always as straightforward or easily verifiable as is the relation between eating habits and increased health risk—a plausible assumption based on common sense or past data.

In many cases, decisional cues do not exist as readily available features as they need to be constructed from a multi-dimensional data set. This increased dimensionality of the feature space (meaning that a great many variables are repeatedly correlated), entails that features are further selected and extracted to reduce the complexity of the data and consequently the model. In this process, physical meanings of features may not be retained, and thus it may not be possible to clarify how the final output of the system relates to any specific feature (Li, 2017). The result is a set of overly constructed and computed features where correlations between feature variables and the target variable do not depend on the conventional understanding of 'cause and effect'—introducing seemingly irrelevant input. Think of e.g. using spelling mistakes for predicting overweight in a health insurance scheme, or the length of the screen name of a social media account for credit scoring. This implies that the assumed link between the input and the actual behaviour may not only turn out to be intrusive, incorrect, or invisible, but may even be non-existent due to spurious correlations. Especially in case of deep learning models, normative scrutiny of these overly constructed features may not be possible primarily because these systems have not been designed with such an assessment in mind.

A viable scheme

Based on the transparency model developed above, we propose the following set of questions as the basics of a viable contestation scheme, that may contribute to the contestability of automated data-driven decisions.

- Is the training data that was used to develop the decisional cues (input) representative of the data subject? If not, to what extent do the discrepancies matter, considering the purposes and the further impact of the decision as well as the regulatory context?
- Based on the decisional cues (selected and weighted features), are the consequences 'explainable' by providing legally, ethically and socially acceptable reasons?
- Are the results interpreted and implemented in line with the declared purposes of the system (purpose limitation principle)?
- Are data subjects made aware of how they can contest the decisions and who is liable for insufficient transparency?

Where those responsible fail to respond to these contestability requirements, their automated decisions may be regarded as *per se* unlawful (Hildebrandt 2016, 58), or as ethically questionable, depending on whether or not they violate legal norms.

*Emre Bayamlıoğlu is researcher at the Tilburg Institute for Law, Technology, and Society (TILT) at Tilburg University, the Netherlands. He is also an external fellow of the Research Group on Law Science Technology & Society (LSTS) at Vrije Universiteit Brussels.

References

Binns, Ruben. 2017. "Algorithmic Accountability and Public Reason" Philosophy & Technology :1-14 DOI 10.1007/s13347-017-0263-5.

Burrell, Jenna. 2016. "How the machine 'thinks': Understanding opacity in machine learning algorithms" Big Data & Society (1): 1-12.

Friedler, A. Sorelle, Carlos Scheiddegger, Suresh Venkatasubramanian. 2016. "On the (im)possibility of fairness" arXiv:1609.07236v1

Hildebrandt, Mireille. 2019 (forthcoming). "Privacy as Protection of the Incomputable Self: From Agnostic to Agonistic Machine Learning". Theoretical Inquiries in Law 19(1). doi: 10.2139/ssrn.3081776.

Hildebrandt, Mireille. 2016. "The New Imbroglio. Living with Machine Algorithms" In The Art of Ethics in the Information Society: Mind You, edited by Liisa Janssens, 55–60. Amsterdam: Amsterdam University Press.

Koops, Bert-Jaap. 2013. “On Decision Transparency, or How to Enhance Data Protection after the Computational Turn.” In Privacy, Due Process and the Computational Turn, edited by M. Hildebrandt & K. De Vries, 196-220. Abingdon: Routledge.

Leenes, Ronald. 2008. “Reply: Addressing the Obscurity of Data Clouds.” In Profiling the European Citizen: Cross-disciplinary Perspectives edited by Mireille Hildebrandt and Serge Gutwirth, 293-300. Dordrecht: Springer Science.

Leese, Matthias. 2014. “The new profiling: Algorithms, black boxes, and the failure of anti-discriminatory safeguards in the European Union” Security Dialogue 45 (5): 494-511.

Li, Jundong, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang,, Huan Liu. 2017. “Feature Selection: A Data Perspective” ACM Computer Surveys 50, 6, Article 94. https://doi.org/10.1145/3136625

Lipton, Peter. 2004. Inference to the Best Explanation, London: Routledge.

Miller, Tim. 2017. “Explanation in Artificial Intelligence: Insights from the Social Sciences” arXiv:1706.07269v2https://ssrn.com/abstract=2928293

Raz, Joseph. 1979. The Authority of Law. Oxford: Clarendon Press.