ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Fairness and Accountability of machine learning Models in Railway Market: are Applicable Railway Laws Up to Regulate Them?*

Charlotte Ducuing[1], Luca Oneto[2], Renzo Canepa[3]

[1]Centre for IT & IP Law - Katholieke Universiteit Leuven - Belgium
[2]DIBRIS - University of Genova - Italy
[3]Rete Ferroviaria Italiana - Italy

Abstract. This paper discusses whether the law is up to regulate machine learning ("ML") model-based decision-making in the context of the railways. We especially deal with the fairness and accountability of these models when exploited in the context of train traffic management ("TTM"). Railway sector-specific regulation, in their quality as network industry, hereby serves as a pilot. We show that, even where technological solutions are available, the law needs to keep up to support and accurately regulate the use of the technological solutions and we identify stumble points in this regard.

1 Introduction

ML models are now pervading every aspect of life and industry especially the transportation industry and the railway sector. In particular, ML models can be designed to optimize TTM. TTM is performed by the railway infrastructure manager (IM) to decide in real time upon the priorities and directions of the trains run by its customers - the railway undertakings (RU) -, especially in case of disruptions.

Regulatory context: like other network industries, the IM and its activities are regulated as part of sector-specific law deriving from the liberalization. Infrastructure management was legally unbundled to various extents from carriage (of goods and passengers) in order to open carriage activities to competition. The IM is therefore subject to specific regulation as a natural monopoly in the absence of the spur of competition, to the benefit of RUs on the one hand, but also because railways are considered as a public utility which should be used to the benefit of society at large, therein subject to public service obligations, on the other hand. At EU level, TTM as performed by the IM is specifically regulated by Directive 2012/34 as modified by the 4th Railway Package[1]. TTM shall be "exercised in a transparent and non-discriminatory manner [...]". In case of "disruption concerning them, [the RUs shall be entitled to] full and timely access to relevant information"[2]. Besides, TTM decisions of the IM fall within the scope for which the RUs may appeal to the railway regulatory body[3] where they consider having been "unfairly treated, discriminated against or [...] aggrieved"[4]. The regulatory body has extensive competences to investigate and remedy the

*This research has been supported by the European Union through the projects IN2DREAMS (European Union's Horizon 2020 research and innovation programme under grant agreement 777596).

[1]Directive 2012/34/EU of the European Parliament and of the Council of 21 November 2012 establishing a single European railway area ("Recast Directive") OJ L 343 14.12.2012, p. 32 as amended by Directive (EU) 2016/2370 of the European Parliament and of the Council of 14 December 2016 amending Directive 2012/34/EU as regards the opening of the market for domestic passenger transport services by rail and the governance of the railway infrastructure OJ L 352, 23.12.2016, p. 1-17.

[2]Article 7 (b) of the Recast Directive.

[3]See article 56 of the Recast Directive.

[4]Article 56 (1) (h) of the Recast Directive.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

alleged problem. Then, is railway law fit for the purpose of regulating decisions made or based on ML models? There is no railway-specific case law, precedent or legal doctrine so that broader picture of algorithmic decision-making shall be taken into account.

Relevance - railway law as pilot: this questioning reaches beyond the railways as the scholarship reflects more broadly upon regulation of algorithmic decision-making. In that sense, it can serve as a pilot to feed the scholarly debate, especially on two aspects. Firstly, should algorithms be regulated *as such* (technology as regulatory target [1]), or rather *indirectly* when deployed to perform a regulated activity? Railway law is an illustration of the latter, defined as functional regulation [2]: TTM is regulated as an activity, indifferently from the (technological) means used. Secondly, how should the law be best designed to accommodate the fact that technologies - especially algorithmic models - are evolving fast? The observation of this "pacing problem" of the law led some to advocate in favor of flexible legal norms, namely "principle-based regulation", rather than rigid "rule-based regulation" [3]. Unlike rule-based regulation which prescribes or prohibits specific behaviors, principle-based regulation "emphasizes general and abstract guiding principles for desired regulatory outcomes" [3]. The regulation of TTM - which prescribes principles of fairness, "non-discrimination" and "transparency" subject to interpretation by the regulatory body and the judiciary authority - undoubtedly qualifies as principle-based regulation.

The legal challenges arising from the use of ML models to make or support TTM decisions can be generally classified in two categories. With regard to the merits of the decision (1), the operation of ML models, based on correlations between the trains' profiles rather than on individual assessment of the respective trains, could result in inaccurate decisions and particularly to uneven or discriminatory decisions. (2) The operation of algorithms may be more or less obscure depending upon their design, especially in the case of ML. These procedural issues in turn challenge the ability of third parties to contest the decisions especially in the case where third parties are non-experts (RUs and the railway regulatory body in our case).

2 Fair Models for Automated Decision Making

There are many cases where the above mentioned problems may arise in the railways, in particular with regard to (lack of) fairness of decisions. Take for instance the case of a model designed to optimize TTM, and particularly delays and deviations from the planed timetable as well as penalties due by the IM in case of faulty delay. Sometimes two or more trains are in the wrong relative position on the railway network because of maintenance, delays or other causes. When an event like this occurs it is required to predict the best place where to make an overtake and enforce it as soon as possible in order to the correct relative position of the trains for the purpose of minimizing delays and deviations from the timetable. This decision must be made with the main goal to provide the highest possible level of service to the final user. Unfortunately, due to higher penalty costs of High Speed trains with respect to Regional or Freight trains, it may happen that High Speed trains are favourite and receive more priority for the overtake. This fact results in biased historical data that, if exploited to make automated decisions, may lead to even more unfair behaviour of the automated decision system [4]. For this reason, in recent years, researcher (see [5] and reference therein) have tried to reduce the unfairness of these data-driven models with various techniques. The question that arises is whether these approaches are really able to be as effective also in the railways. To give a preliminary answer, we mapped the train overtaking prediction problem into a binary classification one, namely, when two trains are in the wrong relative position we try to predict, exploiting the same feature mapping described in [6], if in the next station is convenient or not to make the overtake. In order to built this model

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

we exploited the approach proposed in [5] exploiting one year of data provided by Rete Ferroviaria Italiana (RFI), the Italian IM, about the Italian Railway Network. We use the first 8 months of data for building the model and the remaining 4 month of data for testing it. In Table 1, similarly using the same experimental protocol described in [5], we reported the result when Linear Support Vector Machines or the Non Linear one (using the Gaussian Kernel) are exploited, when the sensitive features (in our case the type of train: Regional, High Speed, and Freight) is known or not to the model during the prediction phase, and when the fairness constraint is present or not in the model. In particular, Table 1 reports the classification accuracy in percentage (ACC) and the fairness measured with the Difference of Equal Opportunity (DEO) [5] on the test set.

LIN	Yes				No			
SF	No		Yes		No		Yes	
FC	No	Yes	No	Yes	No	Yes	No	Yes
ACC	89.3	87.8	91.5	88.3	95.3	93.3	97.3	94.3
DEO	0.41	0.03	0.51	0.05	0.31	0.01	0.41	0.03

Table 1: Fairness in Train Overtaking Prediction Problem.

From Table 1 it is possible to note that (i) non linear models are more effective but less fair, (ii) using the sensitive feature increases the accuracy but diminish the fairness of the model, and (iii) the fairness constraint helps in increasing the fairness but they also reduce the accuracy.

These numbers tells us the that it is feasible, for example, to make fairer models also in the railways - in this case in TTM. Given that the applicable law imposes *inter alia* principles of fairness and non-discrimination onto the decision-maker (the IM), how do both notions of fairness fit? While fairer models can technically be built, how far does the law impose "fairness" and "non-discrimination" with regard to the merits of the decisions made? Next section will attempt to assess whether the law is fit for answering this question.

3 Legal challenges related to the merits of the decisions

The specificity of ML models lies in their data-driven character: decisions are not based on an individual assessment of a train, but on a data-based profiling and subsequent likelihood of future action (in our case of train delays and subsequent penalties). To enhance accuracy, the models are provided with large amounts of input data, sometimes without consideration for causality between the input and the "target decision". (Under which conditions) is such decision-making process compliant with fairness and non-discrimination principles?

Is ML a legitimate means to make TTM decisions? The absence of individual assessment by a ML model-based decision-making process was found discriminatory by the National Non Discrimination and Equality Tribunal of Finland in the situation of a credit institution company refusing to grant credit to an applicant based on non-deterministic algorithmic profiling[5]. In this case, ML was found to be inherently discriminatory and unfair because of the misalignment between the inductive reasoning of the model (based on statistical accuracy) and the individual situation of the applicant, where the latter would have resulted in more beneficial decision for the applicant. In our case, the principles of non-discrimination (to be interpreted as *impartiality*)[6] and fairness incumbent onto the IM to the benefit of the RUs do not exist in isolation but are balanced by the principle of management independence that the IM enjoys

[5]Decision of 21 March 2018, n^o 16/2017, which can be found here: `https://www.yvtltk.fi/en/index/opinionsanddecisions/decisions.html`, last visited 14th November 2018.

[6]see article 7b of the Recast Directive, and especially its title ("Impartiality of the infrastructure manager in respect of TTM [...] where "impartiality" echos the "non-discrimination" obligation enshrined in the core of the article

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

in infrastructure capacity management[7], which is in turn instrumental to the objective of optimization of the use of the infrastructure capacity[8]. The IM shall thus independently choose *how* traffic management decisions are made, so as to optimize the use of infrastructure capacity, with the legal limits set by the principles of (1) fairness and (2) non-discrimination (or impartiality) in the sense that decisions shall in principle be made indifferently from the customer(s) at stake. Given their purpose to optimize the use of infrastructure capacity, the use of ML models appears not to be illegitimate as such.

Is disparate impact discriminatory? Deliberate bias may be maliciously introduced at different stages of the operation of the model [7], which would obviously qualify as discriminatory and/or unfair, such as the manipulation of the learning loop by the historical human-made decisions provided to the models [8]. Direct discrimination or proxy (by means of an apparently neutral factor) discrimination - by favoring an RU or a market segment - would also be illegitimate. However, railway law does not clarify whether the *mere finding* that the operation of the model results in “disparate impact” [8] on the RUs or respectively on market segments, where the input data would not appear to be related to these criteria, would suffice to qualify discrimination. This finding more generally meets the difficulty of the law to deal with data-driven decision-making processes which, in the big data context, leverage large and diversified datasets as input [8, 9, 10]. When disparate impact reflects the fact that “capacities or risks are unevenly distributed between [in casu, RUs or market segments]” [8], as revealed by ML models, it could arguably be justified by the objective of optimum use of infrastructure capacity, given the interpretation of the non-discrimination principle as impartiality obligation. It however remains unclear whether the observation of disparate impact could trigger procedural obligations (e.g. obligation to state reasons?), which will be analyzed below.

ML as enhancement and fairness: the optimization of the use of infrastructure capacity is the *raison d'être* of the use of ML models in TTM and may serve to some extent as justification for potential detrimental effects that it could occasionally occur. Out of fairness, the model should therefore *genuinely* aim to reach that objective. The model could technically reveal patterns and therein enable the decision-maker (the IM) to trump the parameters [11]: e.g. the IM could use the model to optimize the level of penalties due to the RUs *to the detriment* of the overall diminution of delays, while both would be inserted as parameters in the model. As shown in section 2 above, this could result in disparate impact on the market segments (namely High Speed trains on the one hand and Regional or Freight trains on the other hand) as well as in sub-optimum overall use of the infrastructure capacity. The principle of fairness - in light of the objective to optimize the use of infrastructure capacity - would arguably be found to result in the obligation falling onto the IM as decision-maker to conduct regular audits and where appropriate adaptations of the model given its dynamic character, to make sure that its operation is and remains in line with this legal objective. Besides, penalties due in case of train delay are based on the "performance scheme" that the IM has to establish with the agreement of the RUs in order to"[...] improve the performance of the railway network" and which is more generally *instrumental* to the objective of optimum use of infrastructure capacity [9]. Therefore, the creation of *additional information* by the model, where revealing misalignment between the parameters of the perfor-

[7]Article 4 (2), 7 and 7a of the Recast directive and recital 43 of Directive 2012/34.

[8]Article 26 of the Recast Directive, as interpreted by the CJEU in judgment of 28 February 2013, Commission v Spain, C-483/10, EU:C:2013:114, paragraph 44 and judgement of 9 November 2017, CTL Logistics GmbH v DB Netz AG, C-489/15, ECLI:EU:C:2017:834, para 40 and 80.

[9]See article 26, 35 and Annex VI (2) of the Recast Directive.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

mance scheme and the overall diminution of delays, should be found to result in the obligation for the IM to take steps to *revise* the performance scheme. In this context, the ML model would not only optimize TTM, but also one of its parameters, namely the performance scheme based on which penalties are due. It should be noted that such misalignment could have occurred *in the absence of* ML models, but they would have probably gone unnoticed for lack of available information. In this specific case, the fact that ML models *constitute an enhancement*, in that they produce information that was until then not available, appears not to raise substantive gaps in railway law as opposed to other branches of law, e.g. competition law with regard to algorithmic tacit collusion [12, 13]. It may however pose enforcement issues which are now being analyzed.

4 Procedural Challenges

The legal issue of opacity: setting aside legal challenges posed by ML on the merits of decisions, algorithmic decision-making is also generally blamed for its more or less opaque character [14]. This results in third parties - and particularly in this case the RUs and the regulatory body - being *de facto* prevented to contest the decisions [15]. The challenge is all the more striking when the outcome depends not only on the original input data but also on the various interactions that the model operates dynamically with its environment (ML), for example by progressively providing the model with the historical decisions made by the model and/or human operators. While human decisions are essentially individual, the decisions made by the model would in such case be *fundamentally intertwined*: they would both constitute the outcomes of the processing as well as input data for further decisions, turning the model into a “moving target” [16]. The legal scholarship has discussed the desirability and existence of various regulatory tools in order to break the algorithmic opacity, such as algorithmic transparency [17], algorithmic accountability [7], which especially includes a right to a more or less specific explanation of the decision made [18], especially where personal data are processed.

Unclear legal regime on procedural challenges posed by ML: railway law - described as principle-based - provides for principles of transparency and fairness: while they undoubtedly imply procedural obligations, it remains highly unclear *what concrete measures* should be taken in order to comply with them in the case of algorithmic - and particularly ML model-based - decision-making. Railway law does also not prescribe the (specific) means by which procedural transparency and fairness should be complied with by the IM - contrarily to e.g. the General Data Protection Regulation for instance which notably imposes ”compliance by design”[10]). However, the very nature of ML models is so that, if not technically designed so *as from the design*, most of procedural compliance measures would be simply impossible to set up *ex post*. As a result, compliance “by design” appears to be practically needed [7]. This is all the more so because railway actors - such as the RUs and the regulatory body - are not ML experts.

The timeline for compliance: this constraint related to the technological means used to make decisions has major consequences in this case. The challenge relates to the timeline for compliance: in order to be legitimately *used*, the model needs to be *designed* so. However, while the IM is the regulated entity, the knowledge and means to organize compliance “by design” would lie with its contractor (the model developer) so that, practically, organization of compliance is *moved upstream*. In such a case where (1) the user of the model is not the developer and (2) the law (described as ”principle-based”) does not provide

[10] Article 25 of Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (“General Data Protection Regulation” or “GDPR”) OJ L 119, 4.5.2016, p. 1-88.

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

clear *ex ante* guidelines, this results in an inconsistency about who organizes compliance and how. It places the regulated entity (the IM) at risk of non-compliance, which could also have a chilling effect on the acceptance of such technologies.

5 Conclusion

The legal objective of optimum use of infrastructure capacity was found to be such as to justify the deployment of ML models to optimize TTM decisions. This is so even when this entails accidental disparate impacts on the customers to some extent, although it remains unclear how far and under which legal conditions. The legal interpretation of fairness and non-discrimination in ML demonstrably *differs upon the regulatory context* (e.g. its interpretation here undeniably differs from cases where fundamental human rights are at stake). The major problem identified lies in the practical need to comply "by design" with (procedural) obligations consisting of high-level principles (such as transparency and fairness). Where the user of the model - as regulated entity (in casu, the IM) - is not the designer (in casu, its contractor) such as in Industry 4.0 context, this results in an inconsistency, which detrimental to the regulated entity and to the acceptance of the technology.

References

[1] L. B. Moses. How to think about law, regulation and technology: Problems with "technology" as a regulatory target. *Law, Innovation and Technology*, 5(1):1–20, 2013.

[2] A. Jabłonowska, M. Kuziemski, A. M. Nowak, H. W. Micklitz, P. Palka, and G. Sartor. Consumer law and artificial intelligence: challenges to the eu consumer law and policy stemming from the business' use of artificial intelligence: final report of the artsy project. In *European University Institute - LAW Working Papers*, 2018.

[3] M. Fenwick, W. A. Kaal, and E. P. M. Vermeulen. Regulation tomorrow: Strategies for regulating new technologies. In *Transnational Commercial and Consumer Law*, 2018.

[4] A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[5] M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, 2018.

[6] A. Lulli, L. Oneto, R. Canepa, S. Petralli, and D. Anguita. Large-scale railway networks train movements: a dynamic, interpretable, and robust hybrid data analytics system. In *IEEE International Conference on Data Science and Advanced Analytics*, 2018.

[7] J. A. Kroll, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *U. Pa. L. Rev.*, 165:633, 2016.

[8] P. Hacker. Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under eu law. *Common Market Law Review, Forthcoming*, 2018.

[9] J. Grimmelmann and D. Westreich. Incomprehensible discrimination. *California Law Review Online*, 2017.

[10] P. G. Picht and B. Freund. Competition (law) in the era of algorithms. *Max Planck Institute for Innovation & Competition Research Paper*, 2018.

[11] A. Gal. It's a feature, not a bug: On learning algorithms and what they teach us. In *OECD Background Paper, Roundtable on Algorithms and Collusion, DAF/COMP/WD*, 2017.

[12] A. Ezrachi and M. E. Stucke. Artificial intelligence & collusion: When computers inhibit competition. In *U. Ill. L. Rev.*, 2017.

[13] F. Marty. Algorithmes de prix, intelligence artificielle et équilibres collusifs. *Revue internationale de droit économique*, 31(2):83–116, 2017.

[14] D. R. Desai and J. A. Kroll. Trust but verify: A guide to algorithms and the law. In *Harvard Journal of Law & Technology*, 2017.

[15] A. Vedder and L. Naudts. Accountability for the use of algorithms in a big data environment. *International Review of Law, Computers & Technology*, 31(2):206–224, 2017.

[16] N. W. Price. Regulating black-box medicine. *Mich. L. Rev.*, 116:421, 2017.

[17] M. Ananny and K. Crawford. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, 20(3):973–989, 2018.

[18] G. Malgieri and G. Comandé. Why a right to legibility of automated decision-making exists in the general data protection regulation. 2017.