December 2018

Trust by Discrimination: Technology Specific Regulation & Explainable AI

Jakub HARAŠTA [a,1],

[a] *Institute of Law and Technology, Faculty of Law, Masaryk University, Czech Republic*

Abstract. Regulation of emerging technologies such as AI is partially controversial, because of the strive towards 'technology neutral' regulation. This paper summarizes the different approaches hidden behind the grand term of 'technology neutrality' to unravel its competing meanings. One of those meanings is then used for proposal of discriminatory approach towards deployment of AI in different services where society requires more trust and hence explanation. Regulatory barriers should be put forth to prohibit deployment of non-explainable AI into crucial services, such as medical diagnostics and triage. Paper argues that trust can be build by discrimination non-explainable machine learning models.

Keywords. explainable AI, technology neutrality, discrimination of technology, technology specific regulation, critical services

1. Introduction

The term *technology neutrality* became widely used at the end of the 20th and the beginning of the 21st century. It was meant to describe the desirable mode of regulation of emerging technologies, mainly in ICT. However, as stated by Reed [15, p. 265] the term was used without deep discussion about its meaning and about specific regulatory tools related to its use. Only in 2006, Koops deconstructed this policy one-liner to describe its hidden complexity of often mutually exclusive regulatory approaches and competing values [9].

This paper presents certain issues related to technology neutrality in different contexts. I argue for a technology specific approach to technology neutrality ([9, p. 85] and [7, p. 247]) as a way to stimulate the wide use of explainable AI in society. I approach this issue from the perspective of regulating AI as social innovation, and not purely market innovation [20]. Therefore, I consider it axiomatic that in some cases innovation does not present a value *per se* and requires regulation to be valuable for society. However, to prevent hampering innovation throughout the whole field of AI, I suggest distinguishing different AI uses. The use in crucial services, such as diagnostic or medical triage tools, or the use on vulnerable groups, such as minors or ethnic minorities, should be regulated towards explainability.

[1] E-mail: jakub.harasta@law.muni.cz.

2. Technology Neutrality

2.1. Different Faces of Technology Neutrality

Koops dissected technology neutrality into four following distinct legislative goals [9]. The first goal aims to regulate actions of users and consequences of these actions. This happens regardless of the technology used to mediate these actions. This form of technology neutrality is rather extreme, as it focuses on lowest common denominator and does not distinguish between different technologies. As an example, the purpose of both a hand-written and an electronic signature is the expression of will or identification of an individual. Following this aim, technology neutrality does not focus on technology, but on the act of the expression of will.

The second goal aims to achieve functional equivalence between different technologies. This type of technology neutral regulation contains plethora of technology-specific norms. These norms are used to compensate for differences between technologies in terms of their use, the effect of such use or related costs. To follow the example set above, this type of technology neutrality would set up specific norms for hand-written and electronic signatures to compensate for reasonable differences between those two means of expressing one's will.

The third possible aim of technology neutrality is the non-discrimination of technology. The framework following this mode is forbidden from selecting technological winners at any given point in time and also over longer periods by subsidizing specific technologies. The regulation following this maxim has strong competition and innovation aspects. It serves the purpose of opening any given field for further innovations and aims to lower the barriers to entry. To follow the example set above, this aim of technology neutrality prevents us from creating legal framework requiring hand-written signatures and not recognizing legal validity of electronic signatures.

The fourth possible aim of technology neutrality aims to create a flexible framework accounting for future changes and accommodating future innovations. This regulation is formulated in general terms and is often accompanied by open textured formulations and flexible tests.

Very broadly put, technology neutrality serves as an aspiration to enact laws that could be sustainable over time and would not require frequent reviews. Similarly, it can be interpreted as a duty not to make technological choices by creating a restrictive legal framework, but contrarily to leave these choices to market actors. Framework legislation or general soft-law guidelines should stay clear of technological concepts, relying on functional or economic ones instead. On the other hand, a technology-specific reasoning should be implemented at the lowest possible level, e.g. the level of individual decisions of regulatory authorities [1, p. 75–76].

2.2. Regulating Innovation

The general approach towards technology seems to be to avoid regulation out of fear of failing behind countries with less strict regulation or no regulation at all. Zarsky, among others, recalled an anecdote claiming that data protection framework hampers EU innovative potential, as the connection between the strength of privacy laws and the level of ICT innovation is often evident [23, p. 154-155]. However, Zarsky goes beyond this

anecdote and brings forth Stewart's distinction between market innovation and social innovation [20, p. 1277-1279]. The market innovation allows firms to offer new and/or improved products to customers [20, p. 1279], while the social innovation leads to social gains beyond a pure market-oriented approach. Stewart invokes technology leading to cleaner air [20, p. 1279], while Zarsky applies this to actors offering stronger protection of privacy [23, p. 127]. This is, in my opinion, the line between different approaches to technology neutrality as captured by Koops [9]. The first, the third and the fourth aim as explained in subsection 2.1 seem to aim primarily towards market innovation. The second, functional equivalence, seems to be concerned with social innovation. The claim that any regulation will hamper the innovative potential of AI is, without any doubts, correct. However, in the domain of regulating AI as a social innovation, it carries little value for further discussion.

As is evident, technology neutrality contains different and often contradictory approaches. A specific regulatory approach towards technology neutrality is based on a specific context. Arguing for technology neutrality without providing such context could lead to confusion in the creation, application and interpretation of legal framework. It is possible to stimulate the use of explainable AI by implementing the technology neutral legislation, but only if such legislation follows the second aim - functional equivalence. Human and algorithmic decision-making carries certain differences and these differences should make it into legal framework of the future use of AI, through the requirement of explainable AI.

3. Regulation of Artificial Intelligence

I will stay away from trying to define what Artificial Intelligence is. First of all I do not feel competent enough to even try, and second of all, there is an apt pragmatic definition available, which will serve my purpose: *"[AI is] the science of making machines capable of performing tasks that would require intelligence if done by [humans]."* [11, p. V]. Essentially, AI aims towards replacing humans in certain activities. Or at least towards mimicking them and perform certain tasks either well or well enough for it to be economically viable under specific circumstances.

If we are to put forth a legislation that would push AI towards explainability, there are two main concerns. First, how to maximize the social innovation potential of AI through legislation. Second, how to minimize the impact this will inevitably have on AI as a market innovation.

3.1. Explainable AI for Achieving and Maintaining Trust

When it comes to decision-making and reasoning, the blackest box of all is human mind. It is theoretically possible to apply technology neutrality *stricto sensu* (the first aim as described in Subsection 2.1) and require no more from AI than we already require from humans. We have settled for decision-making based partly on intuition and poor reasoning skills. However, in order to build trust in the AI decision-making, we require different standards. We demand an explanation.

Trust in technology is determined by multiple factors - human characteristics, such as the user's personality and ability to understand the technology and deal with related

risks [19]; environment characteristics, such as culture (including socioeconomic status), the task for which the technology is used, and institutional factors including existing regulation and its enforcement [13]; and technology characteristics, such as performance of the technology, transparency of its process and purpose of its use [17].

A transparent decision-making by technology is one of the precursors for developing and maintaining trust ([14],[4],[21],[16]). It addresses human characteristics, namely the ability to understand technology. For general population, AI is currently too opaque not only in its decision-making, but also in its general functioning. Requiring transparency also directly promotes certain technology characteristics by pushing specific (non-transparent) designs off the market. Also, a legal framework for AI or its use would affect institutional factors of trust building, because it would send a message that legislators understand the technology well enough to require certain changes in its design.

At this point in time, different algorithms are used for machine learning purposes. Machine learning classifiers include the use of decision trees or decision lists, neural networks, *k*-nearest neighbour algorithms, support vector machines, Bayesian networks etc. Usually, we tend to think about artificial intelligence in terms of a trade-off between accurate and black-box models, and inaccurate and white-box models. Lipton argues against this belief by stating that linear models are *per se* no more interpretable than deep neural networks [10, p. 7]. According to Lipton, the issue of explainable AI is not predominantly concerned with the use of different machine learning models, but rather with the issue of implementation of these models. Additionally, the discussion about interpretability of specific models is made difficult by missing a clear definition of *interpretability* [10, p. 7-8].

Without aiming for proper definition of interpretability, it is important to note one of the basic motivations. The explainable AI allows us to verify the results of decision-making. As reported in [2], due to bias in training data, AI can achieve incorrect results. In this specific case, the model learned that patients with asthma and heart problems have lower risk of dying of pneumonia. This result is directly opposite to what current medicine knows about pneumonia. This particular bias leading to wrong classification was easy to spot once the model was evaluated by medically trained personnel. However, in a lot of areas, decision-making is currently not as linear as the relation between heart disease, asthma, and pneumonia. Taking a lot of variables in account does not necessarily lead to better results. We tend to be rather surprised when complex models get outperformed by simple linear predictors [3]. However, in my opinion, a lot of variables should mean stronger pressure towards explainability regardless of model performance.

The right to explanation (as a legal requirement for explainability) was heavily debated over the course of drafting the new EU Data Protection Framework, mainly of the General Data Protection Regulation. This attracted significant attention by legal scholars debating the extent of obligation to provide the logic of automated decision-making ([18],[22],[8]). Formulation of specific requirements for the explainable AI in legislation specifically targeted to use of AI in society would without any doubts lead to a significant opposition. However, at the same time, requiring specific design – in this case the explainable design – of AI would arguably lead to greater trust in automated decision-making.

3.2. Minimizing Negative Effect on Innovation

Directly resulting from the conclusion to require explainable AI by law in Subsection 3.1 is the issue of potential AI winter.

We surely desire explainable AI. However, if all AI is to be explainable before it could be implemented, we might be facing another AI winter, at least in countries with strict (or any) regulation. Unregulated development and deployment means more innovative potential, however at cost of social benefits, such as dissolution of trust. On the other hand, strictly regulated development and deployment means less success in AI research compared to other countries, e.g. China. But we cannot possibly have a pie and eat it too.

However, I believe there is a partial solution to this conundrum. Harwich and Laycock [6] laid out certain issues in health industry where AI is beneficial or could be beneficial in the future. These issues are wide-ranging. AI could be useful in boosting general well-being of population through health promotion or targeted prevention. It could be used to augment cognitive capacities of medical professionals and provide us with improved and fast diagnostics and triage. Additionally, AI could make our health system more efficient overall by taking away some of the administrative burden through automation, or by allowing for better management and self-care for chronically ill patients. These implementations require different levels of duty of care as they carry with them different consequences in case of a mistake. Some areas, such as diagnostics, should be made more transparent, even if only for the issue reported in [2]. Diagnostics is inherently more complex than automation of repetitive administrative tasks or reminding to patients with cognitive impairment to take their medicine. The crucial services supported by AI decision-making require more trust and this can be achieved by requiring transparency. Other areas can be deemed as less important or less crucial.

This sort of distinction among various services for purpose of legal regulation has a precedent. Critical infrastructure protection currently works along the same line. Some sectors, such as energy, water management, food industry and agriculture, health services, financial market etc., are deemed as critical and as a result are regulated differently. Different countries approach these issues in multiple ways, labelling different sectors as critical. However, labelling specific sectors as critical attracts an increased attention in terms of specific applicable policy and legal framework [5].

Different areas of AI deployment described in [6] require different levels of duty of care. Mistakes lead to different consequences, from administrative nuisance to life-threatening situations. Any regulation should, in this regard, prescribe the use of explainable AI in some areas that we deem crucial for society – either within the scope of the current critical infrastructure protection framework or outside of it. This approach allows us to maximize the benefit of AI as social innovation (because we opt for regulation) without stopping AI as market innovation dead in its tracks (because we opt for targeted regulation following risk analysis).

4. Conclusion

Lipton [10] formulated different levels of explainable AI. He noted the transparency at the level of entire model (*simulatability*), transparency at the level of individual components (*decomposability*), and at the level of training algorithm (*algorithmic transparency*).

I argue that different levels of explainability should be required from AI deployed into different areas, different services or different individual decision-making tasks. A legal requirement for explainable AI design could help us build trust in automated decision-making. However, requiring explainable AI everywhere could seriously hamper the innovative potential of AI. There has to be a regulation requiring explainable AI models to ensure that we develop and maintain trust. However, the two-tier classification of services ensures proportionality of such regulation and prevents unreasonable barriers to block further development of AI. Similarly to the existing framework of critical infrastructure protection, we require a different level of protection once the asset is considered a critical infrastructure. Discriminating between explainable and non-explainable AI in access to those areas makes perfect sense.

Approaching AI in a technology neutral way – or more precisely in its first, third and fourth aim as described in Subsection 2.1 – means regulating it as market innovation. This approach is related to the perspective of law and economics, which aims to regulate market failures. Innovation and welfare from constantly innovating services is of utmost importance and considered critical for modern society.

However, AI mimics human intelligence and is increasingly deployed into areas typical for human decision-making. Regulating AI from perspective of technology neutrality seeking functional equivalence adds reasonable regulatory burden to new technology. This aims to protect the values and tame the market-driven innovation. This approach targets AI as social innovation.

What I described in this paper should serve as a compromise between these two forces and two regulatory approaches. We should strive to tame the flame of innovation, so it would serve us. But extinguishing it completely would leave us in the dark.

Acknowledgment

Author gratefully acknowledges support by the Czech Science Foundation under grant no. GA17-20645S.

References

[1] A. Butenko, P. Latouche. Regulation for Inovativeness or Regulation of Innovations. *Law, Innovation and Technology*, 2015, vol. 7, no. 1, pp. 52–82.

[2] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730.

[3] J. Dressel, H. Farid. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*, 2018, vol. 4, no. 1, eaao5580.

[4] A. Glass, D.L. McGuiness, M. Wolverton. Toward Establishing Trust in Adaptive Agents. *Proceedings of the 13th International Conference on Intelligent User Interfaces IUI*, 2008, pp. 227–236.

[5] J. Harasta. Legally Critical: Defining Critical Infrastructure in an Interconnected World. *International Journal of Critical Infrastructure Protection*, 2018, vol. 21, s. 47–56.

[6] E. Harwich, K. Laycock. *Thinking on its own: AI in the NHS*. Reform, 2018.

[7] M. Hildebrandt. Legal Protection by Design: Objections and Refutations. *Legisprudence*, 2011, vol. 5, no. 2, pp. 223–248.

[8] M. Kaminski. The Right to Explanation, Explained. 2018. SSRNID 3196985.

[9] B.-J. Koops. Should ICT Regulation be Technology Neutral? In: B.-J. Koops, M. Lips, C. Prins, M. Schellekens (eds.). *Starting Points for ICT Regulation. Deconstructing Prevalent Policy One-Liners*. The Hague: T.M.C. Asser Press, 2006. Pp. 77–108.

[10] Z.C. Lipton. The Mythos of Model Intepretability. 2017. arXiv:1606.03490.

[11] M. Minsky. *Semantic Information Processing*, Cambridge: MIT, 1968.

[12] P. Ohm. The Argument against Technology-Neutral Surveillance Laws. *Texas Law Review*, 2010, vol. 88, no. 7, pp. 1685–1714.

[13] K. Oleson, D. Bilings, V. Kocsis, J.Y. Chen, P.A. Hancock. Antecedents of Trust in Human-Robot Collaborations. *Proceedings of IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support CogSIMA*, 2011, pp. 175–178.

[14] P. Pu, L. Chen. Trust building with explanation interfaces. *Proceedings of the 11th International Conference on Intelligent User Interfaces IUI*, 2006, 93–100.

[15] C. Reed. Taking Sides on Technology Neutrality. *SCRIPTed: A Journal of Law, Technology and Society*, 2007, vol. 4, no. 3, pp. 263–284.

[16] M.T. Ribeiro, S. Singh, C. Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135-1144.

[17] K.E. Schaefer, J.Y. Chen, J.L. Szalma, P.A. Hancock. A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors*, 2016, vol. 58, no. 4, pp. 377–400.

[18] A. Selbst, J. Powles. Meaningful Information and the Right to Explanation. *International Data Privacy Law*, 2017, vol. 7, no. 4, pp. 233–242.

[19] K. Siau, Z. Shen. Building Customer Trust in Mobile Commerce. *Communications of the ACM*, 2003, vol. 46, no. 4, pp. 91–94.

[20] R. Stewart. Regulation, Innovation and Administrative Law: A Conceptual Framework. *California Law Review*, 1981, vol. 69, no. 5, pp. 1256–1377.

[21] N. Tintarev, J. Masthoff. Designing and Evaluating Explanations for Recommender Systems. *Recommender Systems Handbook*, 2010, pp. 479–510.

[22] S. Wachter, B. Mittelstadt, L. Floridi. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 2017, vol. 7, no. 2, pp. 76–99.

[23] T. Zarsky. The Privacy-Innovation Conundrum. *Lewis & Clark Law Review*, 2015, vol. 19, no. 1, pp. 115- -168.