Towards Trust, Transparency, and Liability in AI/AS Systems

Eva Thelisson
University of Fribourg, Switzerland
eva.thelisson@unifr.ch

Abstract

The research problem being investigated in this article is how to develop governance mechanisms and collective decision-making processes at a global level for Artificial Intelligence (AI) systems and Autonomous systems (AS), which would enhance confidence in AI and AS.

1 Introduction

[Purdy and Daugherty, 2016], in their article, have estimated that AI should yield the highest economic benefits for the United-States, culminating in a 4.6% growth rate by 2035, Japan could triple its gross value added growth during the same period, raising it from 0.8% to 2.7%. Germany, Austria, Sweden and the Netherlands could see their annual economic growth rates double. Besides their immense contribution to the economic growth of the nations, AI has conveniently found its place in the daily fabric of our lives by assisting us in different tasks through tools such as the spam filters, recommender systems, and geographical navigation.

However, as a disruptive innovation, it will also change the role of humankind in the creation of value, potentially destroying many jobs, while creating new opportunities [Brynjolfsson and McAfee, 2014]. This calls for (urgent) efforts to educate the public about the different aspects of such processes, and systems, as well as ensure that these technologies serve the humanity in beneficial and responsible ways. Therefore in this article, we argue to open a dialogue in the AI community to ground these systems and their consequences within a proper legal framework, to design them in a way that issues related to trust, transparency, and liability are properly addressed, and most importantly the interests of all stakeholders are taken into account.

2 Trust, Transparency, and Liability

This research is based on the new EU General Data Protection Regulation (GDPR) [Regulation, 2016]. Article 22 of this Regulation sets out the rights and obligations around the use of automated decision making. It gives individuals the right to object to decisions made about them purely on the basis of automated processing (where those decisions have significant/legal effects). Other provisions in the GDPR gives data subjects the right to obtain information about the existence of an automated decision making system, the "*logic involved*", and its significance and envisaged consequences.

The right to obtain information was already presented by Poullet as an obligation of transparency for the technology in 2010 [Poullet, 2010]. A regulation of its functioning and a technical standardization would bring some safeguards for the users that may enhance the respect of their fundamental rights and civil liberties.

How to ensure that all decision made on basis an AI or AS can be explainable? The notion of a "right to explanation" [Goodman and Flaxman, 2016] for an automated decision is correlated to the right to obtain an "explanation of system's functionality". Meaningful information must be provided about the logic involved as well as the significance and the envisaged consequences of such a processing to the data subject (under Article 15.1.h and 14.2.g). Recital (71) illustrates that appropriate safeguards should include the ability of data subjects "to obtain an explanation of the decision reached after such assessment". Is it a recognition of a right to explanation? How binding are those provisions? The answer to these questions will be given by the Court. The EU member States and the new European Data Protection in its harmonization role at EU level shall also define to which extent the recital 71 must be interpreted as a right of explanation.

Furthermore, the right of explanation is widely recognized, Data controllers will have to provide satisfactory explanations for specific automated decisions i.e. they will have to give the reason why the (AI or AS) model gives the outputs it does. This will be especially difficult for Machine Learning systems, whose outcome may vary from one test to another even if the attributes remain the same.

Transparency about the personal attributes used by the organizations may allow the data subject to use the decision tree [Rivest, 1987] to follow its logic and gain meaningful information about its significance and the envisaged consequences of such a processing [Wachter *et al.*, 2017]. The data subject could work out what decisions the model would recommend based on a variety of different values for the attributes it considers. Transparency about the logic and likely effects of the automated decision-making system given the person's personal circumstances, transparency about the values used by the algorithm and how it was trained should be guaranteed.

If data controllers manage to provide dynamic, exploratory systems which allow data subject to explore the relationship between inputs and outputs, this would be equivalent to an explanation for a particular decision.

To reinforce the trust of all stakeholders, those dynamic systems may receive a quality label as well as an associated rating to assess its transparency level.

Providing transparency to machine learning systems and black boxes will be a challenge. Errors and biases in the programming as well as underlying values and criteria taken into consideration to program the algorithms shall be identifiable. An example of bias in facial recognition systems was recently presented by Rod McCullom. He confirmed that this area was understudied.[1] But how to create transparency in a dynamic environment, still remains an open question which has to be well thought trough.

3 Conclusion

AI can indeed be qualified as a new factor of production and reveal unprecedented opportunities for value creation. AI has the potential to double economic growth rates across the 12 countries that together generate more than 50% of the worlds economic output. This PhD research examines the question of data governance based on the EU Regulation as well as the question of algorithm governance. Fostering a shared dialogue among stakeholders, developing an ethical framework for AI systems will reinforce the trust in products and services developed by the industry and will increase its social acceptance, which is economically rational. In addition, AI systems' design should incorporate the attributes of transparency and the consequences of their actions should be properly grounded within an ecosystem that is beneficial for all the stakeholders [Chatila *et al.*, 2017].

References

[Brynjolfsson and McAfee, 2014] Erik Brynjolfsson and Andrew McAfee. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company, 2014.

[Chatila *et al.*, 2017] Raja Chatila, Kay Firth-Butterflied, John C Havens, and Konstantinos Karachalios. The IEEE global initiative for ethical considerations in artificial intelligence and autonomous systems [standards]. *IEEE Robotics & Automation Magazine*, 24(1):110–110, 2017.

[Goodman and Flaxman, 2016] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*, 2016.

[Poullet, 2010] Yves Poullet. Conférence internationale des commissaires à la protection des données: Madrid–novembre 2009 internet–quo vadis? *Cahiers Droit, Sciences & Technologies*, (3):281–289, 2010.

[Purdy and Daugherty, 2016] Mike Purdy and Paul Daugherty. Why artificial intelligence is the future of growth. *Accenture, September*, 28, 2016.

[1] https://undark.org/article/facial-recognition-technology-biased-understudied/

[Regulation, 2016] EU Regulation. 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). *Official Journal of the European Union (OJ)*, 59:1–88, 2016.

[Rivest, 1987] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.

[Wachter *et al.*, 2017] 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. *International Data Privacy Law*, 2017.