Regulatory Mechanisms and Algorithms towards Trust in AI/ML

Eva Thelisson
Massachusetts Institute of Technology

Regulatory Mechanisms and Algorithms towards Trust in AI/ML

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

Kirtan Padh
EPFL, Switzerland
kirtan.padh@epfl.ch

L. Elisa Celis
EPFL, Switzerland
elisa.celis@epfl.ch

Abstract

Recent studies suggest that automated processes that are prevalent in machine learning (ML) and artificial intelligence (AI) can propagate and exacerbate systemic biases in society. This has led to calls for regulatory mechanisms and algorithms that are transparent, trustworthy, and fair. However, it remains unclear what form such mechanisms and algorithms can take. In this paper we survey recent formal advances put forth by the EU, and consider what other mechanisms can be put in place in order to avoid discrimination and enhance fairness when it comes to algorithm design and use. We consider this to be an important first step – enacting this vision will require a concerted effort by policy makers, lawyers and computer scientist alike.

1 Introduction

Computer science has developed a wealth of algorithms for increasingly difficult problems, creating efficiency in the world around us, and making the unimaginable possible. Machine learning (ML) and Artificial Intelligence (AI) in particular are projected to yield the highest economic benefits for the United-States, on a worldwide comparison, culminating in a 4.6% growth rate by 2035 [Purdy and Daugherty, 2016]. Using ML/AI, Japan could triple its gross value added growth during the same period, raising it from 0.8% to 2.7%, and Germany, Austria, Sweden and the Netherlands could see their annual economic growth rates double. This is all due to AI/ML's unique ability to drastically improve efficiency by making use of the vast amounts of data currently being generated, collected, and stored in a myriad of business applications. Besides its immense contribution to economic growth, AI/ML has found its place in the daily fabric of our lives, pervading everything from our social interactions (e.g., Facebook) to our news consumption (e.g., Google News and Twitter) to our entertainment (e.g., YouTube and Netflix). Furthermore, decision-making based on algorithms has disseminated to fundamental aspects of everyday life from the finance industry (e.g., credit scoring), to transportation, housing, education, policing, insurance, health, and political systems.

Despite the incredible boon that computational techniques have been to society, certain red flags have recently appeared which demonstrate that algorithms, in particular AI/ML techniques that rely on data, can be biased. A growing number of global leaders and experts including Bill Gates, Elon Musk, Georges Church and Stephen Hawking have publicly voiced their concern regarding the speed and pervasiveness of the developments of AI/ML. In the US, President Obama's administration produced a report which states that "*big data technologies can cause societal harms beyond damages to privacy*" [Executive office of the President *et al.*, 2014]. In particular, it expressed concerns about the possibility that decisions informed by big data could have discriminatory effects, even in the absence of discriminatory intent. The 2017 edition of the World Economics Forum Global Risks Report, which surveyed 745 leaders in business, government, academia and members of the Institute of Risk Management, listed AI as "the emerging technology with the greatest potential for negative consequences over the coming decade".

Many negative instances have now been demonstrated [O'Neil, 2016; Kirkpatrick, 2016; Barocas and Selbst, 2015]. For instance, Google's online advertising system displayed ads for high-income jobs to men much more often than it did to women [Datta *et al.*, 2015], and ads for arrest records were significantly more likely to show up on searches for distinctively black names or a historically black fraternity [Sweeney, 2013]. Recent events have shown that such algorithmic bias is affecting society in a multitude of ways, e.g., exacerbating systemic bias in the racial composition of the American prison population [Angwin *et al.*, 2016], inadvertently promoting extremist ideology [Costello *et al.*, 2016] and affecting the results of elections [Baer, 2016; Bakshy *et al.*, 2015]. Despite these serious concerns, algorithms, at a fundamental level, pervade everything we do. Simply eliminating them is not an option. Hence it is essential to design algorithmic tools and regulatory mechanisms to empower society at large to mitigate any resulting discrimination, inequality and bias.

For AI/ML to remain beneficial, we must build trust in the systems that are transforming our social, political and business environments and are making decisions on our behalf. We consider at the technical aspect of how bias and discrimination can creep into decisions made by AI, often despite the best intentions of the developers of the algorithm, and how can we prevent such negative outcomes. We then outline the necessary regulatory mechanisms and techniques that must be developed in order to prevent such biases in the future.

2 Algorithmic Bias

One must first understand how such biases occur. Indeed, computers are inherently impartial, and computer scientists and programmers are not malicious. The problem lies at all points in the cycle of collecting, encoding, modeling and optimizing the data.

2.1 Sources of Algorithmic Biases

Input Data

The problem begins with the data that the algorithms build upon, or even the realities of the world itself. Unconscious and systemic biases, rather than intentional choices, account for a large part of the disparate treatment observed in employment, housing, credit, and consumer markets [Pager and Shepherd, 2008]. Such biases can lead to misrepresentation of particular groups in the training data. If the set of examples in the training data do not fairly represent the data on which the algorithm is supposed to run then misrepresented groups could be disadvantaged [Barocas and Selbst, 2014].

Data Vectorization and Cleaning

The raw data must be converted into a digital form (i.e., represented by some kind of *vector*) that an algorithm can use. This process can also introduce biases. This effect is most striking when the training data is labeled manually; the inherent subjectivity in labeling the data can naturally lead to a bias in the dataset. Consider the real life example of St. Georges Hospital in the United Kingdom in where an algorithm for admission decision was developed based on the previous decisions by the admissions committee [Lowry and Macpherson, 1988]. This algorithm simply learned existing biases in the admissions process and resulted in being systematically unfavorable towards minorities.

Model Building

AI/ML algorithms then take as input a subset of vectorized and/or labeled data, and output a model that can take decisions or make predictions. In making these predictions, algorithms can not only propagate biases as discussed above, but in fact amplify them. One potential solution would be to strip away any identifying information that could lead to discrimination, intended or otherwise. However, this could unnecessarily (or undesirably) hamstring the algorithm itself, rendering it useless.

Behavioral Impact

This in turn affects users' actions, feeding back into the real world. For example, it has been hypothesized that increasingly polarized content in search results and online feeds such as Facebook and Twitter can lead to increasingly polarized opinions and behavior [Epstein and Robertson, 2015].

Hence, the steps in the AI/ML life cycle become a destructive feedback loop that can not only propagate, but also exacerbate, societal biases. Thus, if approached without care, algorithms can end up duplicating or even aggravate existing patterns of discrimination that persist in society.

2.2 A Rising Level of Awareness in the EU

On 25 May 2018, the General Data Protection Regulation (GDPR) will be directly applicable in all Member States of the European Union. It brings some substantial changes on data protection and decision making based on algorithms. The GDPR aims at creating a free data flow market in the EU, while making the rules on data protection in the EU consistent, reinforcing data subject's fundamental rights and increasing the liability of companies that control and process such data. Its scope is global (Art. 3, §1). In particular, it reaffirms the data subject's right to explanation and places restrictions on automated decision-making. The GDPR will be applicable in all EU countries and will introduce EU-wide maximum penalties of €20 million or 4% of Global revenue, whichever is greater (Art. 83, Paragraph 5).

Data processors (i.e., entities who process personal data) will now be obliged to comply with data protection requirements which previously only applied to data controllers (i.e., entities who determine why and how personal data are processed). The GDPR will apply regardless whether the processing takes place in the EU or not, and applies processing activities that are related to the offering of goods or services and monitoring their behavior. This regulation gives data subjects the right to access information collected about them, and also requires data processors to ensure data subjects are notified about the data collected (Articles 13 – 15).

It further recognizes that transparency is a key principle. Data must be treated in a transparent manner (Art. 5, §1a)), transparency may occur in the treatment itself (Art. 13, §2 and Art. 14, §2), and the information communicated by the data controller to the data subject must be transparent (Art. 12, §1). The codes of conduct and certification mechanisms must also respect this transparency principle (Art. 40, §2a) and (Art. 42, §3), and transparency also applies to decision-making (Art. 22). Furthermore, this article gives individuals the right to object to decisions made about them purely on the basis of automated processing when those decisions have significant/legal effects. Other provisions in the Regulation 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. In addition, the article 22 of the regulation provides the obligation for the data processor to add additional "safeguards for the rights and freedoms of the data subject", when profiling takes place. Although the article does not elaborate what these safeguards are beyond "the right to obtain human intervention", Articles 13 and 14 state that, when profiling takes place, a data subject has the right to "meaningful information about the logic involved".

Towards satisfying various points of this regulation, and more generally ensuring that the worst fears about AI and ML do not come into effect, we propose various types of solutions which must be developed in collaboration between lawyers, policy makers, and computer scientists in order to ensure a fair and balanced society in the presence of algorithms.

3 Proposed Solutions

To begin, we draw a comparison between the regulation of algorithms and regulations ensuring food safety. Consumers must trust the food that producers and distributors provide on the market. The EU General Food Law Regulation establishes basic criteria for whether a food item is safe. If we instead think of data and algorithms instead of food, one could similarly build a system that is meant to guarantee safety to the functioning of algorithms, following the same reasoning as the EU General Food Law Regulation. Figure 1 draws this parallel between the food law regulation and our proposed regulation of algorithms.

Regulation (EC) No. 178/2002 of the European Parliament and of the Council of 28 January 2002 lays down the general principles and requirements of food law, establishing the European Food Safety Authority and laying down procedures in matters of food safety. On a similar basis, we propose that an EU Regulation dedicated to algorithms, accompanied with European Algorithms Safety Authority laying down procedures in matters of algorithms. This could involve establishing codes of conduct (such as the Food Law Practice guidance), developing third party quality control labels (such as organic certification) and establishing transparency by careful regulation and monitoring of data use as it propagates through various algorithms and tools (as is done when tracing food through the food chain).

Lastly, we call on algorithm designers to further push towards developing the technical tools required to detect, prevent, and correct algorithmic and data biases.

3.1 Codes of Conduct

On 27 June 2017, the European Commission fined Google a record-breaking €2.42 billion for antitrust violations pertaining to its shopping search comparison service. It ordered Google to comply with the simple principle of giving equal treatment to rival comparison shopping services and its own service. Competition commissioner Margrethe Vestager said that "Google has given its own comparison shopping service an illegal advantage by abusing its dominance in general internet search. It has promoted its own service, and demoted rival services. It has harmed competition and consumers. That's illegal under EU antitrust rules." In effect, Google systematically gave disproportionately prominent placement to its own shopping service in its search results. As a result, Google's comparison shopping service is much more visible to consumers in Google's search results, whilst rival comparison shopping services are much less visible. This appeared to be the result of an *explicit* code in Google's algorithm whose *intent* was to discriminate against other services.

Burrell identifies between three *barriers* to transparency [Burrell, 2016]: *1)* intentional concealment on the part of corporations or other institutions, *2)* gaps in technical literacy which, for most people, mean that having access to underlying code is insufficient, and *3)* a lack of interpretability of the decisions made by the algorithm even to experts. For barrier 1, clear codes of conduct that are enforceable, as demonstrated in the example with Google above, is a crucial first step.

3.2 Quality Labels and Audits

To increase transparency, one possibility could be to open the code to public scrutiny. The main drawback to this approach would be the harm it could cause to the valuable intellectual property exposed, and barriers 2 and 3, which state that, even if made public, the results would not be interpretable. As [Lisboa, 2013] notes, "machine learning approaches are alone in the spectrum in their lack of interpretability". Hence, we instead propose that quality labels – similar, e.g., to organic certification, Minergie label, quality management systems and insurance certification (9001 ISO norms), IT security certification (ISO 27 001 norms or Information Technology Infrastructure Library) be made available on a voluntary basis.

The GDPR allows the data controller or processor to draft approved codes of conduct or get a certification on data protection to demonstrate the fulfillment of its duties. The codes of conduct will be approved by the competent authority. The monitoring of compliance with a code of conduct pursuant to Article 40 of GDPR may be carried out by a body which has an appropriate level of expertise in relation to the subject-matter of the code, and is accredited for that purpose by the competent supervisory authority.

The certification can be done by a limited number of certification bodies (Art. 43 GDPR) or by the competent supervisory authority, on the basis of criteria approved by that competent supervisory authority pursuant to Art. 58, §3 GDPR or by the Board (Art. 63 GDPR). Where the criteria are approved by the Board, this may result in a common certification - the European Data Protection Seal. Certification may be issued for a maximal period of three years (renewable). The Board shall collate all certification mechanisms and data protection seals and marks in a register and shall make them publicly available by any appropriate means (Art. 42 GDPR).

The GDPR empowers the regulator to conduct audits and inspections of companies on demand. Strict new compliance requirements are imposed. For example, entities have to perform "Privacy Impact Assessments" and privacy audits as a matter of course. They have to implement "Privacy by Design" methodologies into their business, so that compliance is baked-in to everything they do. They also have to deliver on a new "Accountability" obligation, which means creating written compliance plans, which they will have to deliver to regulators on demand.

3.3 Transparency in the Data Chain

Algorithms must be designed so that a human can interpret the outcome [Goodman and Flaxman, 2016]. However, there is a trade-off between the representation and interpretation of algorithms. Simpler models are easier to explain, but also fail to capture complex interactions among many variables. This also happens to be one of the biggest issues with neural networks, because while they give excellent results in practice, we have very sparse theoretical understanding for them and therefore they are almost completely uninterpretable.

Making reference to the GDPR, [Goodman and Flaxman, 2016] highlighted that "while this law will pose large challenges for industry, it highlights opportunities for computer scientists to take the lead in designing algorithms and evaluation frameworks which avoid discrimination and enable explanation".

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 Articles 15.1.h and 14.2.g). Appropriate safeguards should include the ability of data subjects "to obtain an explanation of the decision reached after such assessment" (recital 71).

Data controllers will have to provide satisfactory explanations for specific automated decisions, i.e., they will have to give the reason why the AI/ML model gives the outputs it does. This will be especially difficult for AI/ML systems, whose outcome may vary from one test to another even if the attributes remain the same. Providing transparency to machine learning systems and black boxes will be a significant technical challenge. 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. Log files may help bringing those guarantees.

We propose to create a data chain traceability, based on the same pattern as the food chain cycle (see Figure 1).

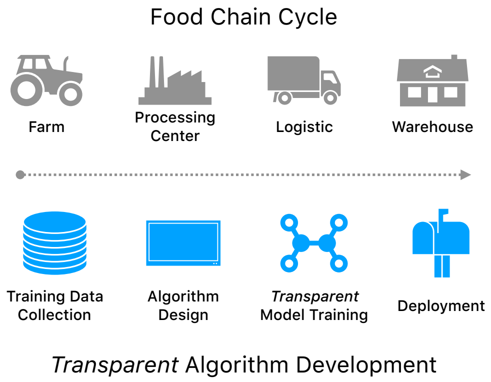

Figure 1: This figure illustrates the symmetries between the food chain cycle and the *transparent* algorithm development. Different regulations and codes of conducts can be devised for each of the steps in algorithm development to ensure overall transparency.

3.4 De-biasing Datasets and Algorithms

According to [Žliobaitė, 2017], "Discrimination-aware data mining studies how to make predictive models free from discrimination, when historical data, on which they are built, may be biased, incomplete, or even contain past discriminatory decisions". There are two main parts to discrimination-aware machine learning, namely discrimination detection and discrimination prevention. Discrimination detection involves finding discriminatory patterns in the training data. Discrimination prevention, on the other hand, entails the development of algorithms which are free from discrimination even on datasets on which standard AI models may discriminate.

The traditional approach to discrimination detection is to fit a regression model to the training data and look at the regression coefficients of the potentially discriminating features such as race, gender etc. The magnitude and the statistical significance of these coefficients can tell us about the possibility of discrimination in the dataset. Discrimination prevention on the other hand can be applied in one of the following three stages of the data processing pipeline according to [Žliobaitė, 2017]: *a)* data preprocessing, *b)* model post-processing, and *c)* model regularization. Data preprocessing is when the training data is preprocessed to remove the discrimination from it and then standard AI models are used for prediction on the cleaned data. Model post-processing starts with standard model and modifies it to incorporate the non-discrimination condition in it. And model regularization adds some constraints to the optimization problem to ensure non-discrimination.

Discrimination-aware machine learning is still in its nascent stage of research and much more needs to be done before it can be incorporated as part of the law.

4 Conclusion

As the new economic business models worldwide are based on data mining and algorithms, a balance has to be found between encouraging innovation with a flexible regulation while protecting the fundamental rights and freedom of people. In the EU, the Charter of Fundamental Rights became legally binding on the European Union in December of 2009, with the entry into force of the Treaty of Lisbon. The Charter contains rights and freedoms under six titles: Dignity, Freedoms, Equality, Solidarity, Citizens' Rights, and Justice.

Building AI Safeguards in order to ensure the respect of those fundamental rights as well as a proper, safe, and reliable functioning of algorithms must be a priority. These safeguards should consider designing accountable algorithms in a way that ensures that ethical principles are encoded in the algorithms. Transparency and trust of algorithms is of key importance to ensure the equal treatment among people and the adequate functioning of a true democratic system.

In this paper we surveyed recent formal advances, and consider what other mechanisms should be put in place. We consider this to be an important first step – enacting this vision will require a concerted effort by policy makers, lawyers and computer scientist alike.

References

[Angwin *et al.*, 2016] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23, 2016.

[Baer, 2016] Drake Baer. The 'Filter Bubble' Explains Why Trump Won and You Didn't See It Coming, November 2016. NY Mag.

[Bakshy *et al.*, 2015] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[Barocas and Selbst, 2014] Solon Barocas and Andrew D. Selbst. Big Data's Disparate Impact. *SSRN eLibrary*, 2014.

[Barocas and Selbst, 2015] S. Barocas and A.D. Selbst. *Big Data's Disparate Impact*. SSRN eLibrary, 2015.

[Burrell, 2016] Jenna Burrell. How the machine thinks: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

[Costello *et al.*, 2016] Matthew Costello, James Hawdon, Thomas Ratliff, and Tyler Grantham. Who views online extremism? individual attributes leading to exposure. *Computers in Human Behavior*, 63:311–320, 2016.

[Datta *et al.*, 2015] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[Epstein and Robertson, 2015] Robert Epstein and Ronald E Robertson. The search engine manipulation effect (seme) and its possible impact on the outcomes of elections. *Proceedings of the National Academy of Sciences*, 112(33):E4512–E4521, 2015.

[Executive office of the President *et al.*, 2014] United States Executive office of the President, John Podesta, Penny Pritzker, Ernest J. Moniz, John Holdren, and Zients Jeffrey. *Big data: Seizing opportunities, preserving values*. White House, 2014.

[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.

[Kirkpatrick, 2016] Keith Kirkpatrick. Battling algorithmic bias: how do we ensure algorithms treat us fairly? *Communications of the ACM*, 59(10):16–17, 2016.

[Lisboa, 2013] Paulo JG Lisboa. Interpretability in machine learning–principles and practice. In *International Workshop on Fuzzy Logic and Applications*, pages 15–21. Springer, 2013.

[Lowry and Macpherson, 1988] Stella Lowry and Gordon Macpherson. A blot on the profession. *British medical journal (Clinical research ed.)*, 296(6623):657, 1988.

[O'Neil, 2016] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown/Archetype, 2016.

[Pager and Shepherd, 2008] Devah Pager and Hana Shepherd. The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. *Annual review of sociology*, 34:181, 2008.

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

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

[Sweeney, 2013] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

[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.

[Žliobaitė, 2017] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, Jul 2017.