SHOULD (A)I STAY OR SHOULD (A)I GO?
BLACK BOX AI CHALLENGES DATA PROTECTION LAW

Francesca Mazzi

INTRODUCTION

Black box Artificial Intelligence (AI) produces outcomes that might be optimal but are not explainable. They can affect individuals' rights if biased and engaged in decision-making processes, since it is not possible to identify and correct the biases in a black box. From a legal perspective, such scenario concerns *inter alia* data protection law. In Europe, for example, the General Data Protection Regulation provides specific warranties to data subjects not only in case of solely automated decision-making processes, but also in cases where a human intervention can still be found.[1] Such system, made of guarantees for individuals and compliance requirements for companies, has generated conflicting opinions.

Some authors that focus on the potential negative effects of AI on individuals argue that it would be desirable to have further regulation of the black box phenomenon. Other authors that focus on the potential positive effects of AI state that the requirements set out by GDPR might obstacle innovation in the field.

I suggest that the GDPR should not be seen as stifling innovation, but rather as directing innovation towards explainable AI, which is the desirable scenario to balance the interests at play. Moreover, I argue that further general regulation of the Black Box phenomenon might be an obstacle for AI innovation for two reasons: firstly, the speed of the technological advancement allows only a partial understanding of the phenomenon, and secondly, a sectorial regulation would better suit the diversity of scopes of AI technologies. I consider data protection law in Europe as a sufficient legal instrument to protect data subjects from automated decision-making processes as per the state of the art. Finally, I argue that similar legislation would be desirable in the United States, in order to avoid a decrease of data flow from the European Union to the United States that would affect Research and Development in AI in both countries, and cooperation between them in the field.

[1] Regulation (EU) 2016/679 of the European Parliament and Council of the European Union 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) (Apr. 27, 2016), http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679&qid=1490179745294&from=en [hereinafter GDPR].

This article is structured as follow: part 1 describes the black box issue and part 2 explains the one of the challenges deriving from the application of black box AI from a data protection perspective, i.e. biased decisions. Part 3 focuses on the legal framework that currently regulate the described challenges, with a specific focus on the GDPR as an example of a detailed and recent regulation. Part 4 and part 5 analyse different opinions of academics regarding the suitability of the GDPR to regulate the black box phenomenon, with part 4 considering the potential needs for further regulation and part 5 concerning the potential negative effects of a detailed legislation such as the GDPR on AI innovation. Finally, part 6 offers some considerations highlighting the importance of innovation in the field of explainable AI and the risks of overregulating the AI phenomenon.

PART 1

"Paint it black": Certain AI logic cannot be traced back

This part is aimed at describing why commonly used AI systems may be a black box to humans.[2] AI can be defined as a category of computer systems that are able to learn from their experiences in order to reach goals and solve complex problems.

The term "AI" includes, *inter alia*, the concept of machine learning, which consists in computers that are able to create mathematical algorithms based on training data and therefore "create" autonomously, without the need of human input.[3] Two aspects of AI are relevant for the present discussion: the first one is that the software is able to make decisions independently, the second one is that the system learns how to make decisions from its experience.[4] Such experience is gained through the information that is provided to the software, by which the machine learns and create a model finding patterns or similarities in the selected information. Once the machine has a model, it is able to process data that is similar to the training data and to identify the similar pattern that the new data resembles. Therefore, the machine can make a prediction producing an estimated result autonomously after having learnt from its experience.

One of the consequences of such ability is that the logic behind it is not always explainable. Indeed, for example, the model used might have weighted certain combination of features differently from others. Since the relevance of features in the process is part of the self-learning process, and when the result is produced it is often

[2] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, HARV. J. L. & TECH. 897 (2018).

[3] DATATILYSNET, ARTIFICIAL INTELLIGENCE AND PRIVACY, NORWEGIAN DATA PROTECTION AUTHORITY (Jan. 2018), https://www.datatilsynet.no/globalassets/global/english/ai-and-privacy.pdf.

[4] *Id.*

produced without any explanation of the self-learning process, a human user might not be able to explain how the machine weighted different features.

Such explainability issue depend on the model engaged. While for certain machine learning, for example those engaging decision trees, is simple to have a high degree of transparency in decision making, other types of machine learning, such as neural networks, are hard to examine. In fact, deep neural networks consist of three parts, an input layer, two or more hidden layers, and an output layer. Input data is processed though hidden layers and emerge as a result. Such hidden layers are difficult to understand for various reason, such as the number of artificial neurons involved and the complexity of their interconnections.[5] Hence, it is often difficult, if not impossible, to determine precisely the logic that led to a certain decision or prediction.[6]

PART 2

"Black or White": Automated-decisions vs biases fight.

Black box AI is often engaged in automated decision-making processes. As mentioned, such decision-making processes are based on training data, and not only such data are often personal (if not sensitive) according to GDPR and other data protection laws, but the outcome of such processes might also influence individuals' rights. Therefore, one of the most relevant issue concerns the fact that such automated decisions might represent a distort, incomplete or misleading reality because of biases.[7] Indeed, biases can come not only from training data but also from the design of the algorithm or the outcome itself.[8] Specifically, AI outputs risks contamination from two types of biases: cognitive biases and statistical biases.[9]

Cognitive biases concern erroneous collection of data that generates an inaccurate representation of the reality, whereas statistical biases relate to structural discrimination and leads to

[5] *See* Bathaee, *supra* note 2, at 891 n.9 (citing Davide Castelvecchi, *Can We Open the Black Box of AI?*, NATURE (Oct. 5, 2016) (describing the black box explainability as the "equivalent of neuroscience to understand the networks inside" the brain)).

[6] *Id.*

[7] Gianclaudio Malgieri & Giovanni Comandé, *Why a Right To Legibility of Automated Decision-Making Exists in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 243, 248 (2017).

[8] *See* Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 680 (2016); James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CAL. L. REV. ONLINE 164 (2016).

[9] KATE CRAWFORD & MEREDITH WHITTAKER, THE SOCIAL AND ECONOMIC IMPLICATIONS OF ARTIFICIAL INTELLIGENCE TECHNOLOGIES IN THE NEAR TERM 6, AI NOW INST. (2016), https://ainowinstitute.org/AI_Now_2016_Report.pdf.

perpetuation of existing inequalities.[10] While human decisions do have a person behind that is accountable for explaining how he or she reached such decision, which allows to evaluate the fairness of the underlying principles of the latter, automated decisions coming from black boxes do not. Hence, if black box AI is involved in automated decision-making processes, data controllers do not have the opportunity to identify possible biases in their automated processing, and they do not have instruments to correct them. Therefore, two competing interests come into play: on one side, black box AI, such as deep neural networks, has the potential to bring revolutionary innovation in a wide range of sectors, including for example healthcare. On the other side, an uncontrolled use of black box AI is likely to create societal risks *inter alia* in relation to fundamental rights, for example perpetuating discrimination towards women. The next part describes the safeguards provided by data protection law, in view of discussing its role in balancing such competing interests.

PART 3

"Law can't get no satisfaction": Black Box AI challenges data protection.

The described issue of black box AI in decision-making processes concerns *inter alia* data protection law. Indeed, as mentioned, not only training data are often personal data, but also AI decisions might affect individuals' lives and privacy.

In general, data protection laws mostly reflect long-established principles, and include requirements based on them. [11] The OECD Guidelines on the Protection of Privacy and Transborder Flows of Personal Data, adopted in 1980, list eight basic principles of data protection: collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation and accountability.[12]

The GDPR is no exception. Nonetheless, it goes further: according to the GDPR, data subjects have the right not to be subject to decisions based solely on automated processing,[13] together with the right to human intervention and explanation.[14] Such right of Article

[10] *See* Malgieri & Comandé *supra* note 7, at 249.

[11] ARTIFICIAL INTELLIGENCE AND DATA PROTECTION: DELIVERING SUSTAINABLE AI ACCOUNTABILITY IN PRACTICE. FIRST REPORT: ARTIFICIAL INTELLIGENCE AND DATA PROTECTION IN TENSION, CTR. INFO. POL'Y LEADERSHIP (Oct. 10, 2018), https://www.informationpolicycentre.com/uploads/5/7/1/0/57104281/cipl_ai_first_report_-_artificial_intelligence_and_data_protection_in_te....pdf.

[12] *See* OECD REVISED GUIDELINES ON THE PROTECTION OF PRIVACY AND TRANSBORDER FLOWS OF PERSONAL DATA, OECD (2013), http://oecd.org/sti/ieconomy/oecd_privacy_framework.

[13] GDPR, art. 22(1).

[14] *Id.* art. 22(3).

22 (1) does not apply in exceptional cases listed in Article 22(2), nonetheless data controllers must still "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests", even when exempted.[15] Moreover, under Article 15, individuals have the right of access to their controlled personal data, and if subject to solely automated decision-making they have the right to be informed about the existence and they can request meaningful information about the logic involved, the significance and the envisaged consequences of such processing.[16] Additionally, the preface states that the data subject is entitled to receive an explanation of how the decision was made, even if recitals of the Regulation are not binding.[17]

The described legal framework seems to suggest a right to explainability according to certain authors, while other authors argue that it simply requires controllers to provide information about the technology involved.[18] Article 29 Working Party released guidelines on interpretation of such provisions, acknowledging the difficulties in explaining AI processes from a controllers' perspective.[19] Nonetheless, regardless of the whether a right to explainability exists or not in the GDPR, it is clear that the data subject should be able to understand the decision sufficiently in order to exercise his or her rights.[20] Moreover, it should be underlined that even though part of the mentioned provisions regard solely automated decision-making processes, the data controller should provide an explanation of the decision according to the transparency principle even in the case of human decision based on recommendation of an algorithmic model.[21]

In the United States, there is not a federal legislation on data protection. Privacy regulations are mostly sectorial, one example of it is the HIPAA in relation to health data.[22] Moreover, privacy laws in the United States lack some of the requirements set by the GDPR.[23] The GDPR might constitute a precedent for the United States to enact

[15] *Id.*
[16] *Id.* art. 15.
[17] *Id.*, Recital 71.
[18] *See e.g.*, Sandra Wachter, Brent Mittelstadt, & Luciano Floridi, *Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation*, 7 INT'L DATA PRIVACY L. 76, 78 (2017).
[19] Article 29 Working Party, Guidelines on Automated Individual decision-making and Profiling for the purpose of regulation 2016/679, EUR. COMM'N (Oct. 3, 2017).
[20] ARTIFICIAL INTELLIGENCE AND PRIVACY, *supra* note 3, at 21.
[21] *Id.* at 22.
[22] The Health Insurance Portability and Accountability Act was enacted by the 104th United States Congress in 1996. Health Insurance Portability and Accountability Act of 1996, 104 P.L. 191, 110 Stat. 1936 (1996).
[23] Mélanie Bourassa Forcier, Hortense Gallois, Siobhan Mullan, & Yann Joly, *Integrating artificial intelligence into health care through data access: Can the GDPR act as a beacon for policymakers?*, 6 J. L. & BIOSCIENCES 317 (2019).

more solid legislation in the sector. In fact, a coherent legal framework would incentivise data flow between the European Union and United States and cooperation in the field of AI.

PART 4

“(Show me) A little respect”—Is the GDPR protecting enough the data subject?

Some academics and experts focus their attention on the societal risks that are likely to arise from the use of black box AI in different fields.[24] Consequentially, part of those belonging to such current of thought calls for further regulation of the phenomenon. Such trend is visible to various extents depending on the industry and the potential risks.

Some legal scholars specifically argue in favour of further strengthening the rights of data subjects already provided by the GDPR, for example by structuring a legibility test of machine learning involved in decision-making processes.[25] Nonetheless, the GDPR is a sufficiently detailed instrument to protect European citizens. Further safeguards should eventually come from a sectorial regulation, rather than by increasingly burdening data protection law. In medicine, for example, many authors recognise the potential positive effects of AI, but nonetheless stress the need for further safeguards to prevent potential distributed biases.[26] Such safeguards could be conceived and structured in a certain way that is peculiarly connected to specific uses of AI in healthcare. Similar calls for regulation and guarantees for individuals have arisen in relation to auto driven cars and liability issues.[27] Similarly, insurance schemes and further requirements of transparency could be designed specifically for the automotive sector.

PART 5

“Money for nothing” —Is GDPR obstructing?

Contrary to what was discussed in the previous section, part of legal scholars claim that GDPR provides such a rigid framework to protect individuals that it threatens to stifle innovation in the AI

[24] *See* James Vincent, *Elon Musk Says We Need to Regulate AI Before It Becomes a Danger to Humanity*, VERGE (Jul. 17, 2017), https://www.theverge.com/2017/7/17/15980954/elon-musk-ai-regulation-existential-threat [https://perma.cc/EY2Q-2R2P].

[25] *See* Malgieri & Comandé, *supra* note 7, at 4–5 (suggesting the legibility test).

[26] *See, e.g.*, Robert Challen et al., *Artificial intelligence, bias and clinical safety*, 28 BMJ QUALITY & SAFETY 231–37 (2019).

[27] Keri Grieman, *Hard Drive Crash: An Examination of Liability for Self-Driving Vehicles*, 9 J. INTELL. PROP., INFO. TECH. & E-COM. L. 294 (2019).

field.[28] According to Humerick, for example, several aspects of the GDPR, including the provisions regarding solely automated decision-making processes, pose concerns regarding the impact of such Regulation on AI research and development in Europe.[29] He states that due to the territorial scope of the GDPR, certain AI technologies might be developed independently from Europe and without the use of personal data of European citizens, potentially stifling the ability of such AI to function in Europe.[30] Humerick also compares the European approach towards data protection in relation to AI with the approach of United States, China and India. He highlights the fact that, oppositely to Europe, such countries tend to maintain a neutral approach in terms of regulating data protection in order not to obstacle research and development innovation in AI.[31]

Although this approach appears sensible, it has certain limitations. It is based on the erroneous consideration that individuals' rights and AI innovation can be compared in terms of weight from a legal perspective. Indeed, while undoubtedly AI innovation is highly desirable for progress, data protection law concerns fundamental rights, such as the right to a private life, protected by Article 8 of the Human Rights Convention.[32] Hence, the current GDPR framework in Europe should not be seen as an obstacle to AI innovation, but rather as a guidance to innovate in the field of interpretable AI, as it will be described in the following section. The GDPR should also be considered as a precedent for meaningful reform in data protection legislation in the United States.[33]

PART 6

"A little less legislation": The answer might not be the regulation

Potential challenges to data protection deriving from the use of black box AI should be addressed with solutions coming from the field of machine learning rather than with further regulation of the

[28] *See, e.g.*, Matthew Humerick, *Taking AI Personally: How the E.U. Must Learn to Balance the Interests of Personal Data Privacy & Artificial Intelligence*, 34 SANTA CLARA HIGH TECH. L. J. 393 (2018).

[29] *Id.* at 414.

[30] Indeed, the regulation applies to foreign companies as well if they process personal data of EU citizens. *See* NICK WALLACE & DANIEL CASTRO, THE IMPACT OF THE EU'S NEW DATA PROTECTION REGULATION ON AI, CTR. DATA INNOVATION 1–4, 25–27 (Mar. 27, 2018), http://bit.do/Wallace_Impact.

[31] *See* Rishi Iyengar, *These three countries are winning the global robot race*, CNN (Aug. 21, 2017), https://money.cnn.com/2017/08/21/technology/future/artificial-intelligence-robots-india-china-us/index.html.

[32] Council of Europe, *European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 14*, 4 November 1950, ETS 5, https://www.refworld.org/docid/3ae6b3b04.html

[33] Forcier et al., *supra* note 23.

phenomenon. Such solution appears to be desirable for two main reasons.

Firstly, overregulating the AI phenomenon could actually stifle innovation. Indeed, as supported by Reed, current laws and regulation should be adapted to deal with risks created by AI.[34] This solution would work as per the current state of the art in terms of technology, and it would allow not to obstacle AI innovation by regulating it at a too early stage.[35]

Secondly, innovation in the field of explainable AI is already fertile and it is desirable for multiple reasons.

Explainable AI is an emerging field of machine learning aimed at making AI results understandable by humans. Explainability of AI solve the black box problem by rendering AI processes and results of such processes interpretable to humans. By "reading" the logic behind the machine, AI users are ensured of a high level of transparency, and therefore accountability and trust. Private and public entities are enthusiastically investing in the field of explainable AI. One example is the XAI program that has been launched by Defense Advanced Research Projects Agency ("DARPA") in the United States.[36] The XAI program focuses on the production of "glass box" models that are aimed to be both efficient and explainable: efficient in avoiding a painful trade-off with AI performances (i.e. maintaining the level of accuracy and potential of the algorithms), and explainable to a human user during the processing and *ex post*.

CONCLUSIONS

Certain sub-categories of AI, such as deep neural networks, have a revolutionary potential in decision-making processes but come with a transparency problem, the so-called black box phenomenon. Black box AI, for example, might perpetuate biases and discriminations without the users being able to identify them, because the logic behind the decision is not explainable to them. Such potential risk is regulated by data protection law in the European Union through the GDPR. The paper proposed an analysis of the criticisms levelled at the GDPR on two levels.

[34] Chris Reed, *How should we regulate artificial intelligence?*, 376 PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A: MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES 9 (Aug. 6, 2018), https://royalsocietypublishing.org/doi/pdf/10.1098/rsta.2017.0360.

[35] Reed states that "[m]asterly inactivity in regulation is likely to achieve a better long-term solution than a rush to regulate in ignorance." *Id.*

[36] For additional information, see Matt Turek, *Explainable Artificial Intelligence (XAI)*, DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence (last updated Jan. 7, 2019).

First, the criticisms that calls for further safeguards in relation to the black box phenomenon were analysed. The GDPR is a sufficiently detailed data protection instrument, and that further guarantees should eventually come from sectorial regulation, in relation to risks associated to specific AI applications. Secondly, the criticisms towards the GDPR concerning its potential negative effect on AI innovation were examined. The GDPR should be perceived as an incentive towards explainable AI, which is desirable to solve the transparency problem *inter alia* from a data protection perspective, in line with the traditional OECD principle of transparency. Hence, the solution to the black box problem should not come from further regulation of the phenomenon, and specifically it should not be searched by burdening the GDPR. The solution should rather come from the technology field, by stimulating investments in explainable AI. It is critical to stress the importance of a significant data protection reform in the United States, in order to align it with the GDPR. Such reform would enhance data flow and cooperation in the AI field between the European Union and United States.