2023 IEEE Conference on Artificial Intelligence (CAI) | 979-8-3503-3984-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAI54212.2023.00085

The *Right to Be Forgotten* in Artificial Intelligence: Issues, Approaches, Limitations and Challenges

Jesus L. Lobo*, Sergio Gil-Lopez*, and Javier Del Ser*,†
*: TECNALIA, Basque Research and Technology Alliance (BRTA), 48160 Derio, Bizkaia, Spain
†: University of the Basque Country (UPV/EHU), 48013 Bilbao, Bizkaia, Spain
Email: jesus.lopez@tecnalia.com

***Abstract*—The *Right To Be Forgotten* is widely conceived as a fundamental principle of the human being. It has become a subject of capital importance in domains where sensitive information is collected from individuals, requiring the provision of monitoring, governance and audit tools to control where such information is used. Artificial Intelligence models are not an exception to this statement: since they are learned from data, this fundamental right should allow individuals to have their personal information erased from AI-based systems. However, the application of this right is not straightforward: what does *erasing* mean in the context of a model learned from data? Is it just a matter of removing the concerned data and retraining the models? This manuscript provides a brief overview of these and more issues, proposing a desiderata for technical advances noted in this direction, and outlining research directions for prospective studies.**

***Index Terms*—AI ethics, trustworthy AI, data governance, data privacy, right to be forgotten.**

I. Introduction

Nowadays, information systems process enormous volumes of personal data (e.g., medical, social or economical) in many domains of activity. Even though this upsurge of personal data is enabling disruptive advances in Machine Learning (ML) and Artificial Intelligence (AI), applications leveraging such advances in real scenarios can threaten the users' data privacy. Current regulations across the world require that upon request, private data must be removed without undue delay from databases. This is known as the *Right To Be Forgotten* (RTBF) or the *Right To Erasure*, i.e. the right of erasure provision. This right in particular not only poses some challenges in the AI field, but also have implications in international relations due to the regulation differences in the protection of data privacy between countries. The RTBF is usually a topic under hot debate between the European Union (EU) and the United States of America (USA) when applied to cross-border data flows. Despite it is assumed that the reach of the jurisdiction of each country is constrained by its territory, interactions on Internet are carried out across multiple territories, and then sowing doubt in the traditional concept of territorial sovereignty. This issue forces EU and USA to work together on their regulatory differences and negotiate the implications of data flow[1].

This manuscript aims to delve into the impact of the RTBF on AI-based models. To this end, we first pause at the definitions and regulations established by authorities in the EU and USA as their scope of application and implementation has served as a reference and attracted the attention of many other countries. Next, we identify several issues that arise when guaranteeing the RTBF in current AI-based models, together with an enumeration of the properties (*desiderata*) that should be met to realize this right. Finally, several ongoing research areas are highlighted due to their connection and potential to cover this niche, concluding that even if steps are being taken lately towards RTBF-compliant models, a solution tackling the many facets of this problem is still lacking and much in need.

II. General Data Protection Regulation (GDPR)

The GDPR is a regulation in EU law on data protection and privacy, and also addresses the transfer of personal data outside EU. One of the most controversial articles in the GDPR corresponds to *Article 17*[2], which establishes the bases for the RTBF application, and where is shown that is far from being a simple process in which an individual simply requests an organization to erase his/her personal data from its databases. This erase satisfies the RTBF, but the capability of ML to learn a representation of its training data may require further legal guarantees that such modeled knowledge has not memorized personal data from the requester. We will later show that this is not straightforward to ensure from a legal perspective.

Before proceeding further, it is worth discussing on the case of USA and the GDPR. A significant disparity in the concept of privacy prevails between EU and the USA. In the EU privacy is a fundamental right and a core of the RTBF established in the GDPR. However, in USA privacy is not considered a fundamental right. The most remarkable difference between EU and USA legislation is probably the lack of a comprehensive data privacy law that applies to all types of data and all USA. Law in USA is fragmented with various regulations governing different sectors and types of data, e.g. the Health Insurance Portability and Accountability Act (HIPAA), the Gramm-Leach-Bliley Act (GLBA), or the Federal Information Security Management Act (FISMA), among others. Perhaps the USA law more similar to GDPR is the California Consumer Privacy Act (CCPA).

[1] https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/eu-us-data-transfers_en

[2] https://gdpr.eu/article-17-right-to-be-forgotten/

979-8-3503-3984-0/23/$31.00 ©2023 IEEE
DOI 10.1109/CAI54212.2023.00085

III. Issues created by RTBF on AI-based Models

Even in the case when personal data have been effectively erased from a given database, the process might have not finished yet if AI-based models have been learned therefrom before a user requests the application of the RTBF. Should the trained model be deleted? Would it be enough to retrain the model with the rest of the data? Several issues may arise at this point such as those addressed in [1], which basically claims that:

- the trained model is not affected (and should not be deleted) by the RTBF, since the data representation learned by the model is no longer personal (section 3.5.2 of [1]); and
- the GDPR does not specifically address AI-based processing (section 4.2.5 of [1]), but there are still uncertainties about the unlawfulness of this particular data processing.

Assuming that the model has not to be deleted, three further aspects motivated by the RTBF must be considered:

1) Erasing the individual part of the requester from the training data may affect the coherence between the learned model and the data from which it is trained. This can be of relevance, for instance, when attributing wrong decisions issued by the model over the data from which it was learned, or when eliciting explanations of the knowledge captured by the model.
2) If the model has to be retrained without such individual part, and assuming that the rest of training data is suitable to produce a new model with good properties (in terms of e.g., performance, complexity and/or interpretability), rebuilding a new model from scratch would compromise the overall sustainability of the AI-based application at hand, especially for models requiring long training latencies.
3) One cannot a priori ensure that a newly produced model performs resiliently without that part of its training data. The removal of data may result in e.g. a worse performance, overfitting, bias, less explainable decisions, less interpretable models or less relevant features, among others. Erasing individual data could imply no chance to fit a new model on the remaining data performing at the same levels of actionability than before the RTBF was claimed.

The conclusion is that the model is always affected in some way by the application of the RTBF. In the first and third aspects discussed above, there is still no clear responses in the literature. However, the second aspect has been central for a recent and promising paradigm called *unlearning* [2], where AI-based models can be obliged to forget about particular data without retraining again over the ablated training data [3]. Even so, the application of unlearning techniques is not straightforward and undergoes challenges on its own.

IV. RTBF-aware AI Approaches and Limitations

We finish this manuscript by traversing several AI research areas motivated by different aspects of the RTBF:

a) Machine Unlearning: Ideally, the *unlearning* paradigm [2] would allow users to remove their data from the model without needing to retrain it from scratch. However, the application of this paradigm is full of difficulties: 1) the relevance of each data point in the training of a model may be unknown due to the stochastic nature of the training procedure, e.g. in neural networks; 2) in a streaming fashion (e.g. continuous learning or real-time machine learning) the training process is incremental, and then the update of the model with a concrete data point at a time step ts_i affects the performance of the model at $ts_{i+1}, ts_{i+2}, ...ts_{i+n}$; and 3) the performance of an unlearned model may reach a point where the degradation is undesirable if the data to be unlearned is too large (*catastrophic unlearning*).

b) Generative Modeling: By reducing the data volume required to train AI-based models, we are decreasing the possibility that a model is affected by the RTBF application. This can be achieved by generative models, including Generative Adversarial Networks or the modern Stable Diffusion Models. Although they do not eliminate the necessity of training, they can promote the use of less data but more efficiently. The generation of synthetic [4] and anonymous data in allowed scenarios opens up new possibilities for replacing data affected by the application of the RTBF. However, the problem arising here is the trade-off between utility and privacy, since synthetic data cannot be generated without control. Furthermore, most generative models are learned from data and thereby, are subject to the application of RTBF.

c) Federated Learning and Transfer Learning: Although Federated Learning can avoid many of the GDPR challenges, private data are still used for training purposes. When a RTBF application arrives, it can be more complicated to remove such data from multiple decentralized servers. A similar problem arises when reusing pretrained models: is the knowledge of a pretrained model subject to consideration due to a RTBF claim affecting the data from which it was learned?

V. Conclusions and Future Work

This paper has addressed the issue of guaranteeing the application of the Right To Be Forgotten (RTBF) in AI-based systems. AI-based models learn from data that are subject to partial removal by the application of this right. We have herein shown several promising RTBF-friendly research areas, but are not exempt of limitations. We definitely advocate for more attention placed on this issue in the short term, particularly in mature regulatory contexts and risk-aware applications dealing with personal sensitive digital footprints.

References

[1] E. Parliament, D.-G. for Parliamentary Research Services, F. Lagioia, and G. Sartor, *The impact of the general data protection regulation on artificial intelligence*, G. Sartor, Ed. Publications Office, 2021.

[2] T. T. Nguyen, T. T. Huynh, P. L. Nguyen, A. W.-C. Liew, H. Yin, and Q. V. H. Nguyen, "A survey of machine unlearning," *arXiv preprint arXiv:2209.02299*, 2022.

[3] E. F. Villaronga, P. Kieseberg, and T. Li, "Humans forget, machines remember: Artificial intelligence and the right to be forgotten," *Computer Law & Security Review*, vol. 34, no. 2, pp. 304–313, 2018.

[4] J. Hradec, M. Craglia, M. Di Leo, S. De Nigris, N. Ostlaender, and N. Nicholson, "Multipurpose synthetic population for policy applications," *No. JRC128595*, 2022.