Opportunities, challenges (and threats) of the use of forgetting in Symbolic XAI

First Author[a,*,1], Second Author[b,1] and Third Author[b,c]

[a]Short Affiliation of First Author
[b]Short Affiliation of Second Author and Third Author
[c]Short Alternate Affiliation of Third Author
ORCID (First Author): https://orcid.org/....-....-....-...., ORCID (Second Author): https://orcid.org/....-....-....-....,
ORCID (Third Author): https://orcid.org/....-....-....-....

Abstract. The increasing use of artificial intelligence systems in making decisions that affect humans has led to legislation requiring that these decisions are explained (XAI) and aligned with human values (value awareness). In this context, systems based on symbolic reasoning, such as Answer Set Programming (ASP), have re-emerged as an alternative because they generate explanations and their models are auditable. Since the justifications can leak sensitive information, forgetting is not only a legal right, but has become an important aspect of XAI. Specifically, ASP forgetting techniques have been used to preserve the privacy of victims of gender violence in the automated allocation of school places. However, forgetting sensitive information can be seen as a way to obfuscate the model (and its justifications) so that it can become a threat. In this work, we focus on this threat, so we have designed a framework in which we learn a symbolic model in ASP, from biased data, so that by forgetting the biased feature, the model maintains the bias but without it being detectable when analyzing the model or its justifications. From the evaluation with sex- and race-biased data we can conclude that automated systems that intentionally discriminate without detection are a realistic threat.

1 Introduction

The automation of all sorts of processes through Artificial Intelligence (AI) systems have made significant progress over the last few years. More recently, it has become apparent that this development needs to be accompanied by means that guarantee, as much as possible, the protection of the people who are affected by the decisions generated by such systems. Whether through self-regulation or soft law (guidelines, codes of conduct, declarations, ethical charters, etc.) or through legal regulation (e.g. the General Data Protection Regulation or the Regul concern has increased for safeguarding the fundamental rights and safety of people affected by AI systems. Promoting a reliable AI, focused on humans, is of foremost importance because autonomous AI systems may cause significant harm. Note that this assumption is valid even it the designers' intentions are good, but it is a real threat when that is not the case.

To this respect, the novel field of value-awareness engineering [5] is emerging, which claims that it is possible to formally represent values, and to reason with and about them, paving the way for future *machine morality*. AI systems, even if they are value-aware, can only be trustworthy if they can explain the *reasons* for their decisions, so they can be validated and/or audited. The DARPA Explainable Artificial Intelligence (XAI) [4], for instance, aims to create AI systems whose learned models and decisions can be understood by end users. This includes seeking methods to increase the interpretability of models, designing effective explanation interfaces, and understanding the psychological requirements of effective explanations. In particular, value-aware systems must be capable of explaining the models and justifying decisions taken in a human-understandable manner, in terms of the values and norms that influenced the reasoning process, among others.

To this respect, we can draw upon work on explanation generation in computational legal reasoning, create machine learning techniques, and develop human-computer interaction principles, strategies and techniques to generate effective explanations. In addition, and related to robustness, we must address the problem of explainability of the learned models and of the automatically generated decisions based on these models. As we have already commented above, this last concern is of special interest for systems with automated decision making that directly influence people's lives, to justify in human-understandable terms the decision taken [1].

Since the justifications can leak sensitive information, forgetting is not only a legal right, but has become an important aspect of XAI. Specifically, ASP forgetting techniques have been proposed to preserve the privacy of victims of gender violence in the automated allocation of school places [3]. However, forgetting sensitive information can be seen as a way of obfuscating the model (and its justifications) that if we use it to hide evidence in deliberately discriminatory decision-making, can become a threat.

In this paper, we describe an attack scheme in which a machine learning framework can learn a discriminative decision-making model such that it does not need to receive the discriminative feature to make the decision. We validated the existence of this vulnerability with several use cases (using different learning algorithms and different datasets). With these experiments, we have found that the proposed framework outperforms (in terms of discrimination) models where the discriminative feature is not used during the training phase. Note that no model uses the discriminative feature to classify the instances, during the testing phase. Given that in most use cases

[*] Corresponding Author. Email: somename@university.edu.
[1] Equal contribution.

we used sex- and race-biased data sets, we argue that this threat is not only realistic, but will also increase distrust of automated systems as long as the authorities do not require verification methods (beyond requiring traceability and a readable explanation) to demonstrate that a high-risk AI system ensures reliability (robustness, security, and accuracy) and that it does not entail risks or generate discriminatory results. Importantly, this threat is already available for commercial exploitation.

2 Background

2.1 (Symbolic) Machine Learning Algorithms

2.2 XAI based on Answer Set Programming

ANNEX III High-risk AI systems referred to in article 6(2) (a) the AI systems are intended to be used in any of the areas listed in points 1 to 8 of Annex III; (b) the AI systems pose a risk of harm to health and safety, or an adverse impact on fundamental rights, and that risk is equivalent to or greater than the risk of harm or of adverse impact posed by the high-risk AI systems already referred to in Annex III.

2.3 Special categories of data

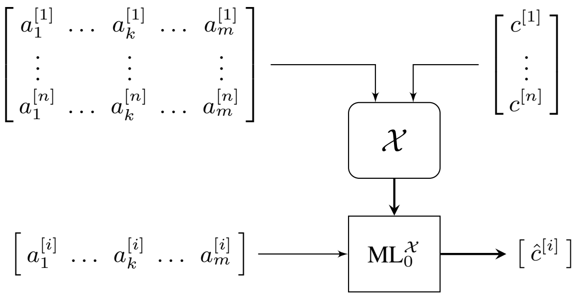
Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

3 Cyber Threat Landscape for XAI

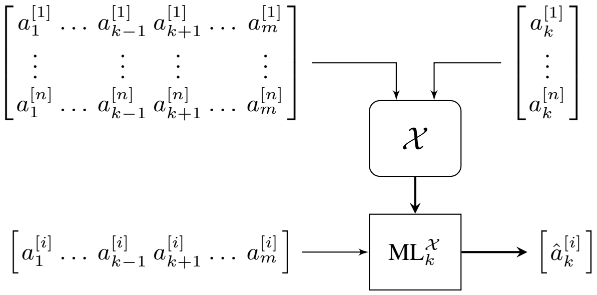
While the AI Act, the European regulation that assigns AI systems to three risk categories: (i) unacceptable risk, banned; (ii) high-risk, subject to requirements; and (iii) low-risk, unregulated, we claim that current requirements for high-risk AI systems can be hacked with non-sophisticated machine learning frameworks.

First, let us remind that high-risk systems are those that include AI technology used in the areas described in Table. 1. Note that, Biased-ML, the framework described in Section 3.1, will be evaluated with examples included in application areas 3 and 4 (e.g., the COMPAS Recidivism Racial Bias data set) while its combination with f_{CASP}, proposed in Section 3.2, targets applications in area 6, e.g., to automate administrative legal reasoning with discretion to act [2].

- Biased-ML$^{\mathcal{X}}$ is a parameterizable framework, where $\mathcal{X}$ represents the machine learning classifier used to learn the models used in the framework. The framework simulates a single classifier model but internally, as we describe in detail in Section 3.1, it contains two models:
  ML$_0^{\mathcal{X}}$, this model has been trained taking into account all attributes available in the data set. Since we assume that the dataset is biased, i.e., it discriminates depending on the attributes a_k.


$\begin{bmatrix} a_1^{[1]} & \dots & a_k^{[1]} & \dots & a_m^{[1]} \\ \vdots & & \vdots & & \vdots \\ a_1^{[n]} & \dots & a_k^{[n]} & \dots & a_m^{[n]} \end{bmatrix}$
$\begin{bmatrix} c^{[1]} \\ \vdots \\ c^{[n]} \end{bmatrix}$
$\mathcal{X}$
$\begin{bmatrix} a_1^{[i]} & \dots & a_k^{[i]} & \dots & a_m^{[i]} \end{bmatrix}$
$\text{ML}_0^{\mathcal{X}}$
$\begin{bmatrix} \hat{c}^{[i]} \end{bmatrix}$

ML$_k^{\mathcal{X}}$: This model is trained to learn the attribute we want to hide during the inference phase because it is known to be a discriminative attribute, i.e., it learns a_k:


$\begin{bmatrix} a_1^{[1]} & \dots & a_{k-1}^{[1]} & a_{k+1}^{[1]} & \dots & a_m^{[1]} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_1^{[n]} & \dots & a_{k-1}^{[n]} & a_{k+1}^{[n]} & \dots & a_m^{[n]} \end{bmatrix}$
$\begin{bmatrix} a_k^{[1]} \\ \vdots \\ a_k^{[n]} \end{bmatrix}$
$\mathcal{X}$
$\begin{bmatrix} a_1^{[i]} & \dots & a_{k-1}^{[i]} & a_{k+1}^{[i]} & \dots & a_m^{[i]} \end{bmatrix}$
$\text{ML}_k^{\mathcal{X}}$
$\begin{bmatrix} \hat{a}_k^{[i]} \end{bmatrix}$

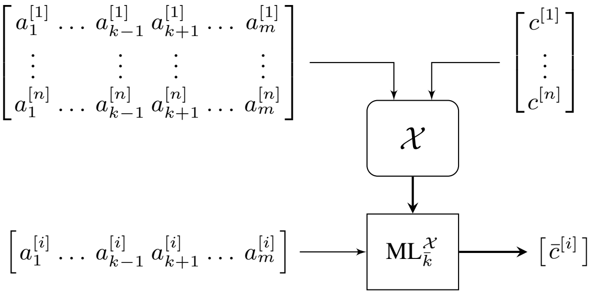
- Biased-ML$^{\text{ILP}}$ is a specific framework in which we use Inductive Logic Programming (ILP) which, as classification models, generates logic programs. Recent ILP proposals, including the FOLD family of algorithms [7, 6], can run the learned programs under s(CASP) to generate explanations (in natural language) of the inferred classification. In addition, we could provide the learned logic program(s) to be audited –this possibility is one of the main advantages of ILP over black-box based learning models in building trust in the system. However, the architecture of the framework would reveal the hacking scheme. In Section 3.2 we explain how we can use f_{CASP} [3] to obfuscate this hacking scheme and hide the use of discriminatory features to auditors.

3.1 Biased-ML$^{\mathcal{X}}$: A classifier that discriminate

3.2 Biased-ML^{ILP} + f_{CASP}: A real threat for XAI

4 Evaluation

To validate that the threat of hiding discriminative attributes is not only real but also worthy, let us present **ML**$_{\bar{k}}$, a classifier that is trained without considering the discriminative attribute a_k from the data set:


$\begin{bmatrix} a_1^{[1]} & \dots & a_{k-1}^{[1]} & a_{k+1}^{[1]} & \dots & a_m^{[1]} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_1^{[n]} & \dots & a_{k-1}^{[n]} & a_{k+1}^{[n]} & \dots & a_m^{[n]} \end{bmatrix}$
$\begin{bmatrix} c^{[1]} \\ \vdots \\ c^{[n]} \end{bmatrix}$
$\mathcal{X}$
$\begin{bmatrix} a_1^{[i]} & \dots & a_{k-1}^{[i]} & a_{k+1}^{[i]} & \dots & a_m^{[i]} \end{bmatrix}$
$\text{ML}_{\bar{k}}^{\mathcal{X}}$
$\begin{bmatrix} \bar{c}^{[i]} \end{bmatrix}$

We will use different datasets to validate the accuracy of our proposal by comparing it with **ML**$_{\bar{k}}^{\mathcal{X}}$ (baseline classifier), and **ML**$_0^{\mathcal{X}}$ (optimum classifier):

$$\text{Acc}(\text{ML}_0^{\mathcal{X}}, \hat{c}^{[i]}) > \text{Acc}(\text{Biased-ML}^{\mathcal{X}}, \hat{\hat{c}}^{[i]}) >> \text{Acc}(\text{ML}_{\bar{k}}^{\mathcal{X}}, \bar{c}^{[i]})$$

Table 1. AI systems identified as high-risk include AI technology used in these areas

	AI Technology	Example
1	Critical infrastructures that put the life/health of citizens at risk	Transport
2	Safety components in products	AI application in robot-assisted surgery
3	Employment and workers management	CV-sorting software for recruitment procedures
4	Essential private and public services	Credit scoring to access a loan
5	Migration, asylum and border control management	Verification of authenticity of travel documents
6	Administration of justice and democratic processes	Applying the law to a concrete set of facts

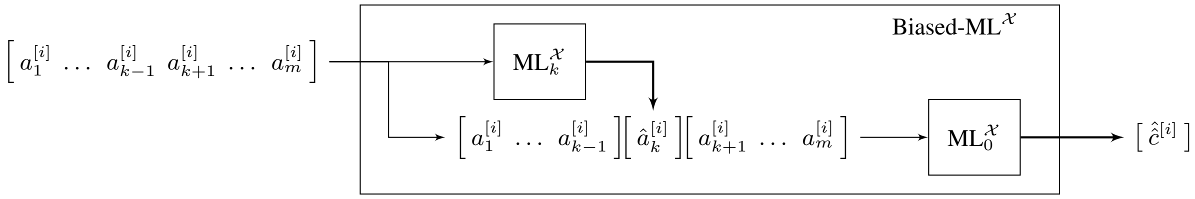

$\left[a_1^{[i]} \; \dots \; a_{k-1}^{[i]} \; a_{k+1}^{[i]} \; \dots \; a_m^{[i]} \right]$
Biased-ML$^{\mathcal{X}}$
ML$_k^{\mathcal{X}}$
$\left[a_1^{[i]} \; \dots \; a_{k-1}^{[i]} \right]\left[\hat{a}_k^{[i]} \right]\left[a_{k+1}^{[i]} \; \dots \; a_m^{[i]} \right]$
ML$_0^{\mathcal{X}}$
$\left[\hat{c}^{[i]} \right]$

Figure 1. Architecture of Biased-ML$^{\mathcal{X}}$

5 Conclusions

Note that high-risk AI systems will be subject to strict obligations before they can be put on the market:

Ob1 Adequate risk assessment and mitigation systems
Ob2 High quality of the datasets feeding the system to minimize risks and discriminatory outcomes
Ob3 Activity logging to ensure traceability of results
Ob4 Detailed documentation providing all information on the system and its purpose for authorities to assess its compliance
Ob5 Clear and adequate information to users
Ob6 Appropriate human oversight measures to minimize risk
Ob7 High level of robustness, security, and accuracy.

However, obligations Ob1 and Ob2 cannot be audited once the model is built, and obligations Ob4 and Ob7 are independent of the system used and any proposal (whether it has good or bad intentions) would want to be able to satisfy them. Therefore, malicious systems would only have to cover obligations Ob3, Ob5 and Ob6.

As we have shown in this paper, the combination of Biased-ML^{ILP} with forgetting allows (Ob3) tracing the results –the justification is a trace of the inference based on the model, (Ob5) providing the user with clear and adequate information – one can query why a certain classification is not obtained and the corresponding justification is generated, and (Ob6) the justifications can be generated in natural language to facilitate interaction with humans.

References

[1] J. Arias, M. Carro, Z. Chen, and G. Gupta. Justifications for Goal-Directed Constraint Answer Set Programming. In *Proceedings 36th ICLP (TC)*, volume 325, pages 59–72. EPTCS, 2020. doi: 10.4204/EPTCS.325.12.

[2] J. Arias, M. Moreno-Rebato, J. A. Rodriguez-García, and S. Ossowski. Automated legal reasoning with discretion to act using s(LAW). *Artificial Intelligence and Law*, 23:1–24, 2023. doi: 10.1007/s10506-023-09376-5.

[3] L. Fidilio-Allende and J. Arias. f_{CASP}: A forgetting technique for XAI based on goal-directed constraint ASP models. In *Proceedings of XXIII Jornadas sobre Programación y Lenguajes (PROLE 2024)*. EPTCS, 2024.

[4] D. Gunning and D. Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40(2):44–58, Jun. 2019. doi: 10.1609/aimag.v40i2.2850.

[5] N. Montes, N. Osman, C. Sierra, and M. Slavkovik. Value engineering for autonomous agents. *CoRR*, abs/2302.08759, 2023. doi: 10.48550/arXiv.2302.08759.

[6] H. Wang and G. Gupta. FOLD-R++: A Scalable Toolset for Automated Inductive Learning of Default Theories from Mixed Data. In *International Symposium on Functional and Logic Programming*, pages 224–242. Springer, 2022. doi: 10.1007/978-3-030-99461-7_13.

[7] H. Wang, F. Shakerin, and G. Gupta. FOLD-RM: A scalable, efficient, and explainable inductive learning algorithm for multi-category classification of mixed data. *Theory and Practice of Logic Programming*, 22(5): 658–677, 2022. doi: 10.1017/S1471068422000205.