The FATE System Iterated

Fair, Transparent and Explainable Decision Making in a Juridical Case

de Boer, Maaike H.T.; Vethman, Steven; Bakker, Roos M.; Adhikari, Ajaya; Marcus, Michiel; de Greeff, Joachim; van der Waa, Jasper; van Zoelen, Emma M.; Kamphorst, Bart; More Authors

Publication date
2022

Document Version
Final published version

Published in
CEUR Workshop Proceedings

The FATE System Iterated: Fair, Transparent and Explainable Decision Making in a Juridical Case

Maaike H.T. de Boer[1], Steven Vethman[1], Roos M. Bakker[1], Ajaya Adhikari[1], Michiel Marcus[1], Joachim de Greeff[1], Jasper van der Waa[1,2], Tjeerd A. J. Schoonderwoerd[1], Ioannis Tolios[1], Emma M. van Zoelen[1,2], Fieke Hillerström[1] and Bart Kamphorst[1]

[1]*TNO, Anna van Buerenplein 1, 2595 DA, The Hague, The Netherlands*
[2]*Delft University of Technology, Mekelweg 5, 2628 DE, Delft, The Netherlands*

Abstract

The goal of the FATE system is decision support with use of state-of-the-art human-AI co-learning, explainable AI and fair, secure and privacy-preserving usage of data. This AI-based support system is a general system, in which the modules can be tuned to specific use cases. The FATE system is designed to address different user roles, such as a researcher, domain expert/consultant and subject/patient, each with their own requirements. Having examined a Diabetes Type 2 use case before, in this paper we slightly iterate the FATE system and focus on a juridical use case. For a given new juridical case the relevant older court cases are suggested by the system. The relevant older cases can be explained using the eXplainable AI (XAI) module, and the system can be improved based on feedback about the relevant cases using the Co-learning module through interaction with a user. In the Bias module, the use of the system is investigated for potential bias by inspecting the properties of suggested cases. Secure Learning offers privacy-by-design alternatives for functionality found in the aforementioned modules. These results show how the generic FATE system can be implemented in a number of real-world use cases. In future work we plan to explore more use cases within this system.

Keywords

FAIR AI, Hybrid AI, Explainable AI, Bias, Secure Learning, Knowledge Engineering, Co-Learning

1. Introduction

More and more AI systems are used in real world cases in all types of domains. Those systems are often highly specialized towards one user and one specific application. These personalized systems often work with sensitive data, which makes it essential that they handle data in a privacy-preserving manner. In our previous paper [1], we proposed the FATE system; a system that combines state-of-the-art AI tools. The FATE system aims to provide decision support with AI capabilities in a fair, understandable, trustworthy, controllable and secure manner [2]. The main areas of research include human-AI co-learning, explainable AI and the fair and secure, privacy-preserving usage of data. Especially in co-learning and explainable AI, hybrid AI plays

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), Stanford University, Palo Alto, California, USA, March 21–23, 2022.

maaike.deboer@tno.nl (M. H.T. d. Boer)
0000-0002-2775-8351 (M. H.T. d. Boer)

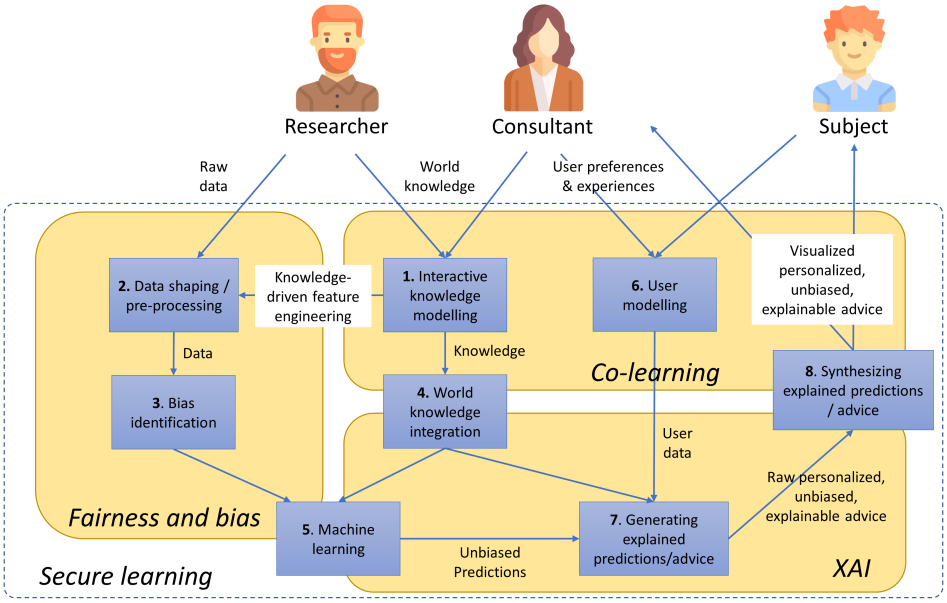


Figure 1: The Iterated FATE system

a big role, as symbolic and sub-symbolic information needs to be combined. The prototype of the FATE system is set-up in a generic, model-agnostic and modular way, so that it can be used in different use cases with different users and different instantiations of the modules.

In this paper, we slightly iterate the FATE system, mainly in the definition of the user roles, and apply this revised version of the system in a juridical use case in which for one (court) case the relevant older cases have to be retrieved. The novelty in this paper is mainly the application of the FATE system on this use case including a storyboard. In the next section, we describe the revised FATE system. Section 3 provides information about the juridical use case, including the task and the different user roles. In section 4, we present the research and results for the research areas. Section 5 contains the conclusions and future work.

2. The Iterated FATE System

Figure 1 shows the FATE system. The largest improvement to its predecessor is the renewed definition of the user roles. AI systems that provide elaborate interaction with users and that include aspects of Explainable AI typically differentiate with respect to the user roles they cater for, such as e.g. [3] and [4]. To provide the various users with the best fitting advice, we follow a similar approach with the FATE system in which we differentiate on the system's functionalities for various user types. In the previous version of the FATE system the users were specified as AI developer, expert user and lay user. The roles are now redefined as researcher, consultant and subject to better reflect our experienced usage of decision support systems in practice.

The domain researcher typically has experience in (data) science and related disciplines. She wants to learn and obtain knowledge from the relation between the (historical) data of subjects

and a phenomenon of interest. In addition, the researcher is interested in how trustworthy the system is in practice. As such, the subject-phenomenon relation in the use case should account for a set of legal, ethical and policy conditions. The researcher balances these context conditions with the utility in predicting the phenomenon. For the AI-system, the researcher is the only user-role with the ability to balance these essential conditions, i.e. the settings of the researcher hold for the consultant and subject as well. In practice, the researcher will do so with the help of respective legal, ethical, and policy experts.

The consultant is a domain expert who holds expert knowledge in a particular field. She wants to advise or intervene on the data subject's activities or behavior. Based on the subject's data and with the help of the system, an individualized outcome of the subject-phenomenon relation is established. The consultant uses the system to obtain contextual information and can question the subject for further information. The system provides the consultant with its confidence and justification regarding advice or interventions targeting the subject. The purpose of this support is on providing actionable information to the consultant. For instance, information on how the subject's behavior should change in order to have another outcome.

The subject is a "naive" user who is neither schooled in AI and data science, nor holds any domain knowledge per se. The subject is either subjected to the system's output or has an intrinsic interest in this output. As such, she uses the system for her own situation. The system functions as an online consultant that monitors the subject's data and generates actionable advice according to the subject-phenomenon relation model. Personal conditions in relation to the context conditions of the use case will be communicated when applicable.

Based on feedback, other changes include that module 2 is now more general data shaping / pre-processing instead of human-driven data shaping, the arrow from module 5 to 6 is now only predictions instead of also advice, the arrow from module 7 is now 'raw' advice instead of advice and the output of module 8 is not only pointed towards the subject but also the consultant. Furthermore, we added in the figure the link to the research areas of XAI, Co-learning, Fairness and bias and secure learning.

3. Juridical Use Case

For the juridical use case, the FATE system should provide decision support in a court case, i.e. providing support for a judge, lawyer or defendant. This does not mean that the FATE system passes a sentence autonomously, but rather it supports a user in the juridical process. In this process there are multiple roles: the researcher can be working for the public prosecution scientific office, a law faculty at a university, or an NGO interested in fair judgements, the consultant can be a lawyer and the subject can be the defendant. These users have different goals and different ways to use the FATE system. The goals of the researcher could be to explore innovation in case law (broad goal), to recognize trends, to increase quality and efficiently of case law (through automation), or to control or audit the juridical system. The goals for the consultant could be to determine whether a case should be started against a defendant, getting fair ruling or to inform the subject. The goals for the subject could be getting the lowest possible sentence as outcome by for example finding out if you need a lawyer. The FATE system aims to help the different users with their respective goals.

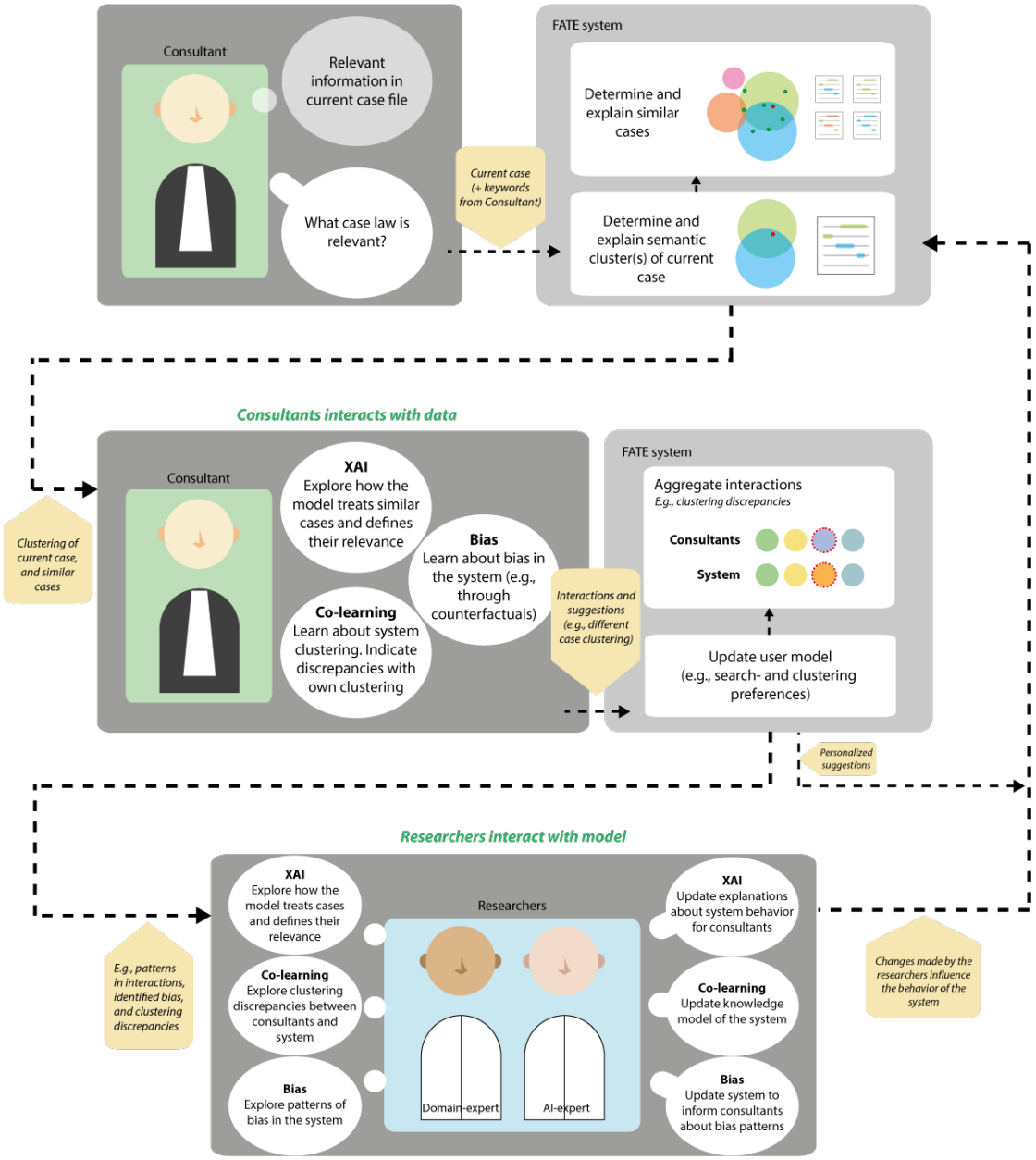


Figure 2: Storyboard of the juridical use case and FATE system

4. Methodology and Results

Figure 2 shows a storyboard illustrating our approach to the juridical use case, and describes a high-level workflow for the consultant and researcher roles. This storyboard was constructed and verified based on input from a domain expert (a solicitor general in the Dutch high council), collected through an interview. The consultant, being a public prosecutor in this use case, has

the goal of constructing an argument and advise on the verdict given a certain juridical case. The prosecutor searches for relevant information in the current case, and wants to find relevant case law. The FATE system supports this process by categorizing the current case, using a topic tree - a hierarchical knowledge structure created by domain experts -, and suggesting similar cases based on the same categories and the vector distance (doc2vec embeddings with a cosine distance metric). The predicted category of the current case is returned to the public prosecutor, together with a top 5 of similar cases. In the storyboard, an example is shown for visualizing the semantic clusters by means of a scatter plot that shows individual cases within their assigned clusters (colored circles). The explainable AI module explains the cluster assignment of the current case by highlighting the paragraphs in the case that are found to be indicative of the cluster. It provides a similar explanation for the cases that are found to be semantically similar to the current case, while also presenting a counterfactual explanation that indicates which textual changes in the current case would lead to a different cluster assignment. Next, the consultant interacts with the data visualisation and explanation of the FATE system, in order to achieve different goals relating to XAI, bias, or co-learning. Through interaction, the consultant provides feedback to the system. For example, the consultant might disagree with the system about a case clustering, and suggest another category for this case. Such feedback is saved into a user model containing individual search- and clustering preferences that changes the clustering behavior of the system. Moreover, the feedback from consultants is aggregated and shown to the researcher(s). The researchers explores the system's model and the consultant feedback in order to identify (bias) patterns in the clustering by system and consultants, and understand clustering discrepancies between consultant and system. Researchers might decide to alter the model by providing assignment feedback (i.e., assigning individual cases to a particular cluster) and description feedback (i.e., change the cluster label that is used by the system). The model updates that are made by the researchers alter the clustering behavior of the FATE-system, triggering new feedback from consultants.

In this use case, a publicly available dataset of semi-structured Dutch textual data[1] is used.

The following sections explain the research done in the different modules.

4.1. Explainable AI

The XAI module focuses on explaining how the system determines case similarity, used when determining case suggestions. It aims to explain which words, sentences or segments play a role in determining why the current document is similar to others. Counterfactual explanations are used for this [5], to convey which text needs to be omitted for the case text to be treated as significantly different to result in different case suggestions. As such it differs from regular counterfactual explanations that deal with a classification task and what input changes are required to alter the classification [6].

To construct these counterfactual explanations, a topic clustering approach is used. A hierarchical topic tree is provided, and cases are assigned to such topics based on the system's used similarity. An example of such topics is the parent topic *crimes against life* with sub-topics of *murder* and *manslaughter*. This allows the XAI module to explain the system's behaviour in terms of more generic behaviour, as opposed to one document versus another.

[1] https://www.rechtspraak.nl/Uitspraken/Paginas/Open-Data.aspx

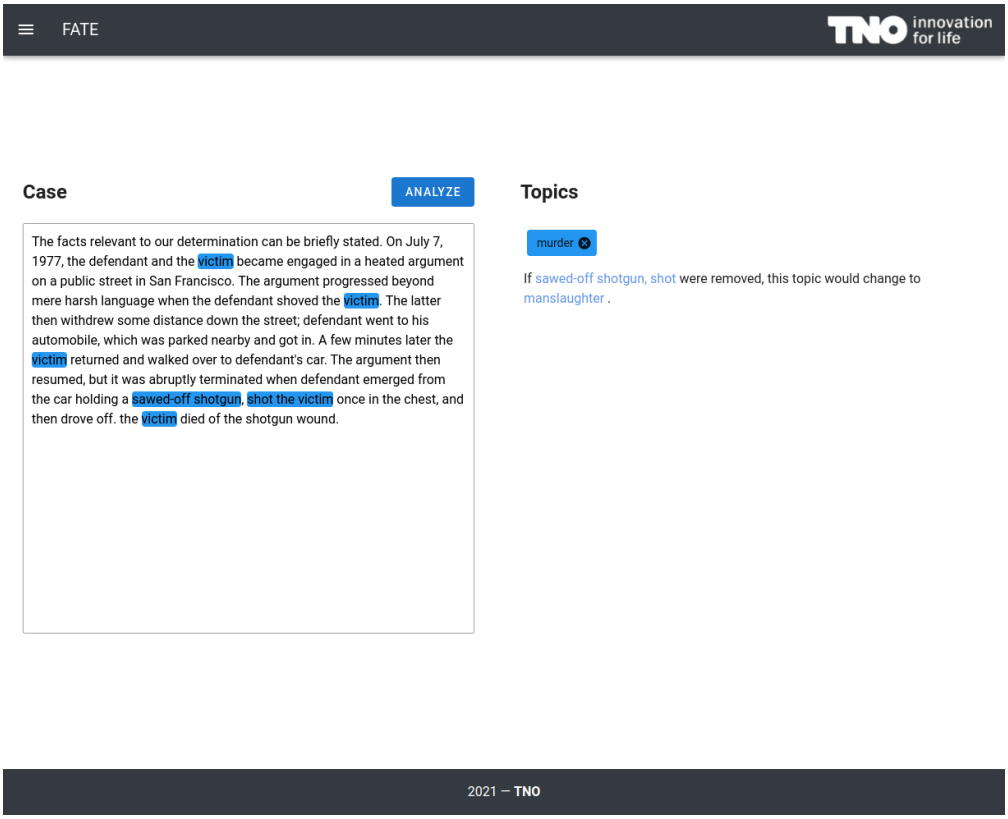


Figure 3: User interface to show how the system determines case similarity using counterfactual explanation. The example snippet[7] comes from a real case-law text without personal details.

With the help of this topic clustering, counterfactual explanations explain what textual omissions are required for the case to change from one topic-cluster to another. These explanations aim to convey what aspects of the case text the system finds important in its case suggestions with respect to meaningful topics. This allows users to determine whether a case should indeed be treated as belonging to a certain topic, and thus whether the case suggestions are in any way meaningful. In addition, the use of topic clusters combined with counterfactual explanations allows for a highly interactive interaction between user and the FATE system to facilitate exploration and a deeper understanding.

The module itself supports the use of existing methods to identify counterfactuals based on searching the embedding space. These are methods such as C-LIME or C-SHAP [8], FACE [9] and SEDC [10]. In practice, these methods do not scale well. The case texts in our juridical use-case vary from a few hundred words to tens of thousands words. The current counterfactual methods all perform a heuristic search on the word-level, which proved to be intractable on our real-world text documents. Instead, we incorporate domain knowledge from the researcher to identify text segments relevant for the use case. In the juridical use case this is the identification of various legal arguments and decisions. This allows for a more efficient approach in the search of an appropriate counterfactual.

Figure 3 shows a view of the demonstrator currently under development. The user provides text of a court case as input to the text area element on the left. The system automatically shows different topics, such as *murder* found in the text. It provides information per topic, such as "If the words *sawed-off gunshot* and *shot* were removed, the topic of *manslaughter* would apply instead of *murder*". Moreover, words which triggered the assignment of this topic are highlighted in the court case. Our goal is to enrich the interface to review this type of data both spatially, using a dimensional reduced scatter plot and textually using the topic hierarchy and relevant case text snippets. In this way, the user will be able to understand, through exploratory interaction, what case texts the system deems similar or dissimilar and what in those texts causes this.

4.2. Co-learning

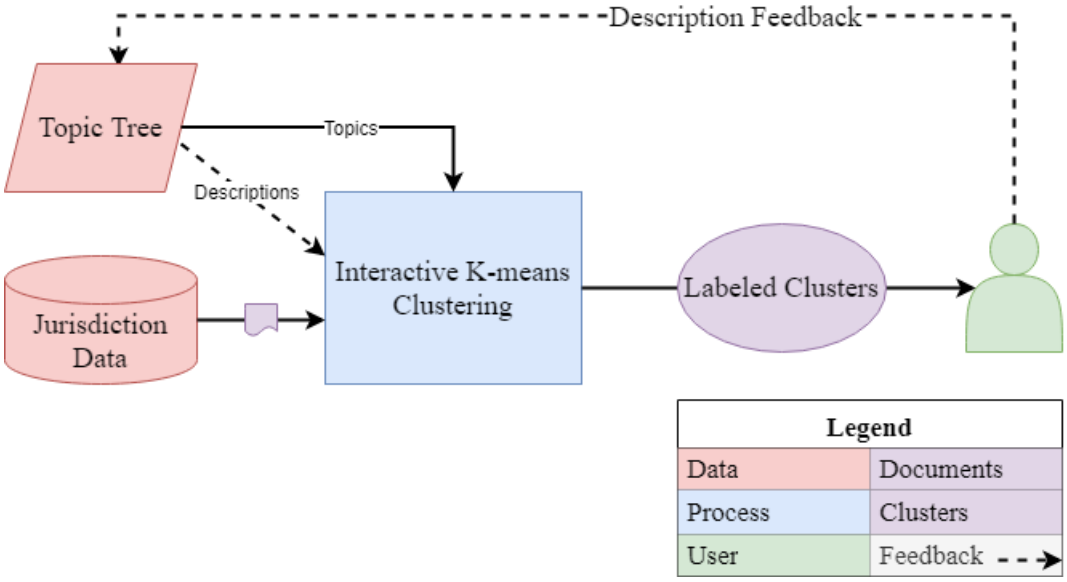


Figure 4: Co-learning technical architecture

Human-AI co-learning stands for both collaborative and continuous learning. This is a process in which a human and an AI system both learn (i.e. acquire new knowledge, change behavior and/or build meaning) through their collaborative interactions, in an ongoing manner. The term 'co-learning' has relatively recently been introduced as a vital process in human-AI collaboration and interaction [11], and can be positioned next to similar terms such as co-adaptation and co-evolution [12]. A key aspect of co-learning is the creation of intuitive and fluent interaction between humans and machines such that the humans can learn from the machines, and the other way around.

In this module, we enable co-learning by allowing users of the FATE system to extensively explore the clustering, an automatic grouping of documents. The clustering is made by the system and presented through a dashboard-like GUI, thereby allowing the human to learn how

data is clustered. The users can then provide feedback on different aspects of the clustering, thereby allowing the system to learn about human expert knowledge. Users can provide feedback by changing the textual description of clusters and by reassigning case laws to a new cluster. Integrating this feedback from the user is done with an automatic structuring approach of the textual data in a continuous process. Case recommendations are improved by adding feedback from the user to the topic tree. The topic tree is then combined with an interactive clustering method, allowing users to also directly provide feedback on how specific cases are clustered. The clustering provides insights from the data, while the topic tree contains the domain knowledge. With this method, the FATE system will benefit from both human and machine knowledge and make better recommendations to the users.

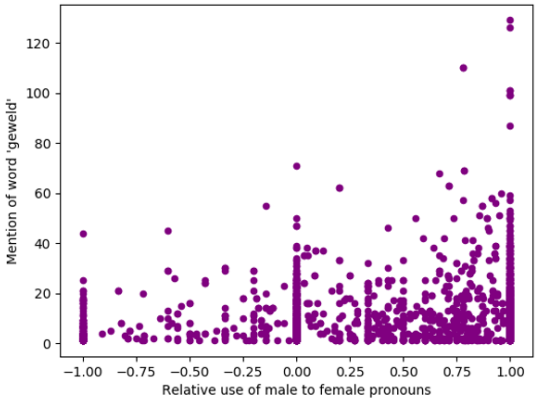
For the juridical use case, we have a large amount of unstructured textual data: past case laws. We can automatically create clusters to organize and provide insight to this data. However, the topics, or labels, of these automatically created clusters are not always intuitive. This can be improved by involving a user. The user can influence the cluster topics by adding feedback about the cluster descriptions. The user can also assign a case law to a cluster. Such semi-supervised clustering models, or interactive clustering models, become increasingly more popular due to their ability to produce more meaningful categories for the user. Dubey et al. [13] show how a k-means clustering framework can be improved by using assignment and description feedback. With assignment feedback, the user can reassign a data point from one cluster to another. With description feedback, the user can modify the vector of a cluster by providing a description for the cluster. We elaborate on this work by linking the topic tree to the clusters, such that the clusters not only contain insights from the data, but are labeled using human knowledge. In figure 4 an architecture containing these technical components is shown. The jurisdiction data is clustered in two ways, by using the topics from the topic tree, and by using the interactive K-means clustering. Afterwards they are aligned by using a contingency matrix, such that the clusters now are labelled with the topics. This alignment is presented to the user, who can provide feedback. The new descriptions are incorporated in the topic tree, and then used to update the interactive clustering.

The approach described above to enable co-learning in the FATE system has strengths and weaknesses. This method uses assignment feedback and description feedback as proposed by Dubey et al. [13]. In practice, the assignment feedback does not work as expected from their work. Another weakness is that evaluating our system is a tedious task without labeled data and user interactions to use as a ground truth. For future evaluation, either labeled data needs to be created or domain experts should be interviewed to determine the performance. The strength of this system is that meaningful and intuitive clusters are created for the user, which helps them to be able to find and compare unlabeled documents. These clusters are created by both human and machine; the user created a topic tree which provides context to the clusters, and they provide feedback on these clusters, thereby continuously improving the results.

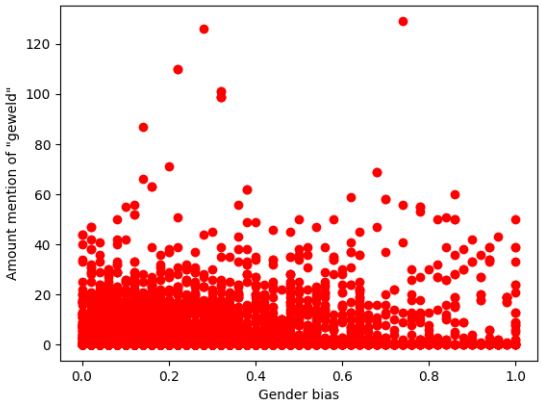
4.3. Fairness and Bias

Different from our previous paper on this topic [1], we focus this year on the practical perspective of bias. Based on the use case, three hypotheses were observed:

1. the association between men and violence may be embedded in the data, e.g. criminals



(a) Experiment 1: Representation of gender pronouns and the number of times violence is described in case law.



(b) Experiment 2: Language bias measured based on the method of Bolukbasi et al. (2016) and the number of times violence is described in case law.

are more often male [14], , *i.e. bias in data*

2. and these associations can be captured in the language model trained on the data, e.g. the link between violence and masculinity in language can be captured in correlations or in their semantic similarity [15, 16] , *i.e. bias in the model*
3. in turn, the bias in the language model may affect the suggestions, e.g. AI predictions for men may have more false positives than predictions for women , *i.e. bias in outcomes*

From the dataset, the cases from 2020 concerning criminal law in the Dutch courts were collected and annotated for gender and violence [17]. Annotation of gender was first explored based on whether the majority of pronouns (such as 'he','she', 'him', 'her') were female or male, however, this was found not accurate enough. Therefore gender was annotated per case based on reading the text and finding pronouns directly indicating the gender of the defendant. Annotation of the case pertaining physical violence from the defendant to another human was approximated by counting the occurrences of the word 'geweld' (violence in Dutch). A case containing more than two mentions of the word 'geweld', is annotated as violent, while cases with zero mentions of 'geweld' are annotated as non-violent. Cases with 1 or 2 mentions of 'geweld' were excluded. A closer inspection by a random sample of 50 non-violent and 50 violent cases based on this annotation approach showed that 96% of the 'non-violent' cases did not contain indication of a violent defendant and 98% of the 'violent' cases contained written indication of violent behavior of the defendant. From the set of 8240 case law of 2020, cases were annotated until a 100 cases were found for each combination of a male or female defendant concerning non-violent or violent criminal behavior. Based on this data, we set up three experiments to investigate bias in the case law.

Experiment 1 concerned bias in the data (*hypothesis 1*), in which all cases from 2020 were investigated to see whether cases with dominantly male pronouns contained more mentions of violence ('geweld'). In Figure 5a, it is shown that cases with relatively more male pronouns contain more mentions of violence. This skewness in the data might be or might not be

representative of the current justice system. Nonetheless, the skewness creates a risk for bias in the use of the system, as the association between male and violence might be captured and relied upon by the system. This may not always be desirable in the application. Thereafter, the annotated violent cases for male and female defendants were compared but both sets had an average of 5.02 and 5.21 mentions of violence, respectively. A two sample t-test showed with a p-value of 0.849 that the hypothesis of equal means could not be rejected. Hence, by inspecting the number of times violence is described and the gender of the defendant, the hypothesized risk of bias in data between male and violence is not found in this experiment.

Experiment 2 concerned bias in the model (*hypothesis 2*), a doc2vec model was trained on the case law of 2020 and as a first iteration the Direct Bias measurement of the well-known paper by Bolukbasi [18] was used to measure gender bias. In brief, the gender direction in the vector representation of the language model was measured by the Dutch equivalents of the following word-pairs 'Man-Woman', 'Male-Female', 'Father-Mother', 'Brother-Sister' given that these were present in the text. The similarity of the vectors of other words in a case with respect to this gender direction in the vector space is then used to measure whether a case has gendered wording within the lens of the language model. Figure 5b does not indicate a strong relation between the language bias measured and mentions of violence. The averages of the bias scores for the violent and non-violent cases are very close, 0.278 and 0.268, respectively. The results of experiment 2 suggest that the measure by Bolukbasi does not indicate that a model trained on case law has captured a strong association between men and violence.

Experiment 3 concerned the bias in use of the model and data (*hypothesis 3*), i.e. the bias in the outcomes. To measure the bias in the outcomes, non-violent cases with male and female defendants, one hundred each, were put into the downstream task of finding 10 suggestions of similar case law. Unwanted bias may prevail when non-violent cases with a male defendant more often receive violent suggestions than female non-violent cases. Results showed that the suggestions based on both female and male non-violent cases received very little violent suggestions. The difference for female defendants 4.1% and male defendants 3.0% was statistically insignificant, with a p-value of 0.21.

Hence, these set of experiments: (1) identified a potential risk for bias in the data between the association between male and violence, (2) did not identify a strong association between male language bias and violence in the model trained on the data and (3) did not identify unwanted bias when using the system to suggest similar cases for non-violent cases.

In the future, more iterations on these experiments are necessary to investigate the hypotheses stemming from the use case and make strong claims on bias and fairness present in the use case. To start, next to gender bias, risks for bias related to nationality, migration status and social background as well as their intersection are also likely to be present in the use case [19, 20]. Next to that, we also advice the investigation concerning gender bias to be extended. One opportunity is to add different bias measurement techniques for Experiment 2, such as [21]. Another is to add meta-data of the verdict of the case to the data, such that associations between gender and verdicts can be demonstrated. Additionally, bias measured in the use of data-driven AI on the text for case law may be compared to more expert-driven modelling approaches of case law, e.g. by extracting the key features to represent a case. On top of that, to investigate bias mitigation in this use case, we want to inspect the effect of changing the gender representation of the data used for training the model on bias measured in the use of

the AI. This fits closely with our aim of future work in FATE: to further our understanding on how to mitigate the impact of unwanted bias of AI in many different scenarios, where bias may be mitigated in the data, the model or the use of the AI. More use cases are necessary to show the multiple facets of bias and give direction when and which approaches for mitigation of unwanted bias are suitable.

4.4. Secure Learning

The problem that the area of Secure Learning aims to solve, is that of privacy violation in AI applications. The solutions created in the areas of explainable AI, co-learning and fair AI explicitly work on centralized data that is freely available to be processed, but this assumption is often unrealistic in today's world due to two main issues, namely:

1. Data containing privacy-sensitive information cannot simply be used for AI purposes, due to privacy concerns and laws such as the GDPR. This is the case in, for example, the health care and financial domain and any other domain where personal data is present.
2. Data is often distributed over different entities. This data cannot always simply be sent to a central entity first, because either the data volume is too large - resulting in issues concerning bandwidth, battery life etc. - or because it is not desired due to the reason provided above.

In order to tackle both issues, the area of Secure Learning offers techniques called Privacy-Enhancing Technologies (PETs) to create privacy-by-design solutions to AI algorithms. Privacy-by-design is a design principle that ensures that negligible information about sensitive data is disclosed throughout a certain process. In our case, we are interested in algorithms that enable certain AI functionalities on sensitive data, without leaking information about this sensitive data. Even though some research has been done on such algorithms for general machine learning applications [22][23][24], the combination with explainable AI, fair AI and co-learning has yet remained completely unexplored.

Nowadays, there is a variety of PETs that can be used to create privacy-by-design solutions. However, the FATE system provides personalized decision support, which brings about new privacy-related challenges that are not trivially captured by privacy-by-design solutions. For example, some explainable AI algorithms, such as foil trees [25], output a data point from the training set of the model to be explained. Such an algorithm can be converted into a privacy-by-design solution using PETs, but the output would still violate privacy.

In the juridical use case, all information is public and preservation of privacy plays a minor role, so our work continued to focus on the Diabetes type 2 use case from 2020. Nevertheless, the developed algorithms and insights gained are applicable to many other contexts.

We set out to design a privacy-by-design algorithm that provides the same functionality as the foil tree algorithm [25] but provides the strongest privacy guarantees possible, by keeping the sensitive training data *and the model* encrypted throughout the entire algorithm, since attacks are known that can reconstruct training data from a model [26].

Our first contribution is a synthetic data generation subroutine, that generates synthetic data based on the sensitive training data that contains negligible information on the sensitive data. This synthetic data then completely replaces the sensitive data during the rest of the algorithm.

Our second contribution is a cryptographic procedure for training a decision tree on data with encrypted target variable values. After the synthetic data is generated, the encrypted model is securely applied in a black-box manner to obtain encrypted target values. The synthetic data and corresponding encrypted labels are then used to train a decision tree called the foil tree. There are algorithms that can train decision trees on encrypted data [27][28]. However, these algorithms generally assume that all the data is encrypted, but in our situation only the target variable is encrypted. Our scenario therefore lends itself to a more efficient solution.

Our third contribution is an algorithm that can extract a contrastive explanation from the encrypted foil tree trained before. As the input to the foil tree training algorithm was purely synthetic data, we can provide data points from that set as part of the explanation without revealing any information about the sensitive training data or the model.

Our experiments show that the algorithm can run with only seconds of delay, which is critical to interactive personalized decision support. The solution works for numerical and categorical data, but it is still an open problem to extend this to textual data for use cases such as the juridical one. Additionally, improvements could be made to the synthetic data algorithm to make the data more realistic using domain knowledge. This would ensure that the user experience is affected minimally by adopting this algorithm.

Our work only touched upon the surface of an intriguing pool of problems where personalized AI and privacy-preservation coexist. We believe that this seemingly contradictory combination has much potential and hopefully our work motivates others to investigate privacy-preserving explainable AI.

5. Conclusion and Future Work

In this paper we have described the next iteration of the the FATE system, which is being developed as a generic decision support system, taking into account aspects of bias and fair AI, Explainable AI, co-learning and secure learning. In particular, we described how the system can aid different types of users in a juridical use case, in which the system is monitored for certain aspects of fairness by making possible biases explicit, presenting the results in a user-understandable manner (XAI) and exploring related case law by identifying relevant clusters through interaction with a user. Additionally, we described how a privacy-by-design approach – deemed a necessity in many cases due to restrictions on data use from multiple sources – impacts on the algorithms used, and how possible mitigation strategies, e.g. through the use of synthetic data, might remedy potential violations of privacy.

By taking up a new (juridical) use case we explored on a per research area basis (bias, XAI, co-learning and secure learning) whether modules previously developed for the Diabetes use case would generalize to this new context. This required (amongst others) an adaptation of low level data and ML components, as in the Diabetes use case we dealt with (tabular) patient data and in the juridical use case with textual data. Additionally, various improvements were made within the different modules to be able to support the new user roles in their decisions, such as the topic tree and clustering in XAI and co-learning.

Future work will focus on the refinement of the FATE system around the research areas of XAI, Co-learning, Fairness and bias and secure learning, through the adoption of the next

use case in yet another domain. By adopting a series of use cases from different domains, the general applicability of the system is exemplified.

Acknowledgments

The FATE project is funded by the TNO Appl.AI program (internal AI program). We would like to thank the AI4J project for their use case and data and the other members of the FATE project for their valuable feedback, especially Klamer Schutte and Thijs Veugen.

References

[1] J. de Greeff, M. H. de Boer, F. H. Hillerström, F. Bomhof, W. Jorritsma, M. A. Neerincx, The fate system: Fair, transparent and explainable decision making., in: AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering, 2021.

[2] K. H. Tae, Y. Roh, Y. H. Oh, H. Kim, S. E. Whang, Data cleaning for accurate, fair, and robust models: A big data-ai integration approach, in: Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning, 2019, pp. 1–4.

[3] G. Ras, M. van Gerven, P. Haselager, Explanation methods in deep learning: Users, values, concerns and challenges, in: Explainable and interpretable models in computer vision and machine learning, Springer, 2018, pp. 19–36.

[4] S. Hepenstal, D. McNeish, Explainable artificial intelligence: What do you need to know?, in: International Conference on Human-Computer Interaction, Springer, 2020, pp. 266–275.

[5] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[6] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[7] People v. satchell, https://law.justia.com/cases/california/supreme-court/3d/6/28.html, 2021.

[8] Y. Ramon, D. Martens, F. Provost, T. Evgeniou, A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c, Advances in Data Analysis and Classification 14 (2020) 801–819.

[9] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, P. Flach, Face: Feasible and actionable counterfactual explanations, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 344–350.

[10] Y. Ramon, D. Martens, F. Provost, T. Evgeniou, Counterfactual explanation algorithms for behavioral and textual data, arXiv preprint arXiv:1912.01819 (2019).

[11] K. van den Bosch, T. Schoonderwoerd, R. Blankendaal, M. Neerincx, Six challenges for human-ai co-learning, in: International Conference on Human-Computer Interaction, Springer, 2019, pp. 572–589.

[12] E. M. Van Zoelen, K. Van Den Bosch, M. Neerincx, Becoming team members: Identifying interaction patterns of mutual adaptation for human-robot co-learning, Frontiers in Robotics and AI 8 (2021).

[13] A. Dubey, I. Bhattacharya, S. Godbole, A cluster-level semi-supervision model for interactive clustering, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2010, pp. 409–424.

[14] CBS, Jaarrapport integratie 2020 - criminaliteit, https://longreads.cbs.nl/integratie-2020/criminaliteit/, 2020.

[15] R. Sarre, et al., Men are more likely to commit violent crimes. why is this so and how do we change it?, https://theconversation.com/men-are-more-likely-to-commit-violent-crimes-why-is-this-so-and-how-do-we-change-it/-157331, 2021.

[16] APA, Harmful masculinity and violence, https://www.apa.org/pi/about/newsletter/2018/09/harmful-masculinity, 2018.

[17] R. S. Centrum, Dutch case law search engine: "uitspraken, een deel van alle rechterlijke uitspraken wordt gepubliceerd op rechtspraak.nl. dit gebeurt geanonimiseerd.", 2020. URL: https://uitspraken.rechtspraak.nl/.

[18] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, A. T. Kalai, Man is to computer programmer as woman is to homemaker? debiasing word embeddings, Advances in neural information processing systems 29 (2016) 4349–4357.

[19] M. Tonry, The social, psychological, and political causes of racial disparities in the american criminal justice system, Crime and justice 39 (2010) 273–312.

[20] A. S. Hartry, Gendering crimmigration: The intersection of gender, immigration, and the criminal justice system, Berkeley J. Gender L. & Just. 27 (2012) 1.

[21] H. Gonen, Y. Goldberg, Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them, arXiv preprint arXiv:1903.03862 (2019).

[22] P. Mohassel, Y. Zhang, Secureml: A system for scalable privacy-preserving machine learning, in: 2017 IEEE Symposium on Security and Privacy (SP), 2017, pp. 19–38. doi:10.1109/SP.2017.12.

[23] S. de Hoogh, B. Schoenmakers, P. Chen, H. op den Akker, Practical secure decision tree learning in a teletreatment application, in: Proceedings of the 18th International Conference on Financial Cryptography, Lecture Notes in Computer Science, Springer, Netherlands, 2014, pp. 179–194. URL: https://ifca.ai/fc14/. doi:10.1007/978-3-662-45472-5_12.

[24] R. Shokri, V. Shmatikov, Privacy-preserving deep learning, in: 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2015, pp. 909–910. doi:10.1109/ALLERTON.2015.7447103.

[25] J. van der Waa, M. Robeer, J. van Diggelen, M. Brinkhuis, M. A. Neerincx, Contrastive explanations with local foil trees, CoRR abs/1806.07470 (2018).

[26] Q. Wang, D. Kurz, Reconstructing training data from diverse ML models by ensemble inversion, CoRR abs/2111.03702 (2021). URL: https://arxiv.org/abs/2111.03702. arXiv:2111.03702.

[27] S. de Hoogh, B. Schoenmakers, P. Chen, H. op den Akker, Practical secure decision tree learning in a teletreatment application, in: N. Christin, R. Safavi-Naini (Eds.), Financial Cryptography and Data Security - 18th International Conference, FC 2014, Christ Church, Barbados, March 3-7, 2014, Revised Selected Papers, volume 8437 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 179–194. URL: https://doi.org/10.1007/978-3-662-45472-5_

12. doi:10.1007/978-3-662-45472-5_12.
[28] M. Abspoel, D. Escudero, N. Volgushev, Secure training of decision trees with continuous attributes, Proc. Priv. Enhancing Technol. 2021 (2021) 167–187. URL: https://doi.org/10.2478/popets-2021-0010. doi:10.2478/popets-2021-0010.