



Predicting EU Country of Origin Based on Moral Choices Concerning AI

And the possibility of EU wide legislation concerning AI

Rashied Doualibi

Thesis Proposal

Data Science & Society

STUDENT NUMBER

2060116

COMMITTEE

dr. Klincewicz

LOCATION

Tilburg University

School of Humanities and Digital Sciences

Department of Cognitive Science & Artificial Intelligence

Tilburg, The Netherlands

DATE

March 4th, 2024

WORD COUNT

1813

Predicting EU Country of Origin Based on Moral Choices Concerning AI

And the possibility of EU wide legislation concerning AI

Rashied Doualibi

1. **Project Definition, Motivation & Relevance**

* **Project Definition**

This research will focus on countries within the European Union countries specifically. A new development, the AI Act which was recently greenlighted by The European Commission is likely to come into effect in the near future. The EU itself refers to it as the first ever legal framework on AI. This precisely demonstrates the relevance and importance of this subject.

<https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>

This piece of legislation is designed to construct regulatory framework relating to the field of AI and is part of the overall approach to AI by the European Union. <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>

The aim of this paper is to use the moral machine dataset in order to predict country of origin based on other variables present in the data. This will be done in order to find out whether it is possible to predict attitudes towards AI and Moral decision making based on the country of origin. This can have implications for the possibility of EU wide legislation regarding this topic.

* **Motivation**
* **Societal**

With this arguably new field of AI and the governmental sphere that is playing catch up with regards to regulations, it could be an intriguing idea to focus on whether these policies can be generalized on an EU scale or if there are distinct differences between EU countries when it comes to these topics. This is something which could have significant repercussions for the future of European Union legislation concerning artificial intelligence.

* **Scientific**

While there have been studies concerning the topic at hand, given the recent nature of the AI act and associated legislation, the field is still relatively small.

Earlier papers have implemented distinct algorithms when working with the Moral Machine Dataset that this paper will also work with.

Moreover, most of the research concerning this topic is in survey form which can fail to find associations in data. Therefore, this paper will aim to investigate the relationships that can be found between the variables present in the dataset.

* **Relevance**
* **Societal**

Because of the relatively ‘new’ nature of artificial intelligence and the lack of established regulatory frameworks surrounding it, now is the time for research regarding this topic. This is because it can provide an aid for policy makers and to establish whether there are differences between EU countries which has to be taken into account when forming EU wide policies.

Furthermore, this could also be interesting when keeping in mind the recent tensions within the EU between member states (e.g. Hungary, Poland) and the Brexit phenomenon which could be seen as a sign that there are vast cultural, political and economic differences within the union (Szent‐Iványi & Kugiel, 2020).

* **Scientific**

With regard to the scientific relevance of this paper, while there is some research surrounding moral attitudes on AI based on country of origin or personality traits, applying it to the framework of European law and the betterment of governmental policy is something which will set this research apart from most of the academic literature in this field.

1. **Literature Review**

In order to further enhance the relevance of this paper the current academic literature that is available on the topic will be outlined as to show similarities and differences between the current academic literature and what could be added by this research paper.

Awad et al. (2018): Describes moral judgements made by respondents from allover the world regarding autonomous vehicles. The research subcategorizes the respondents into cultural clusters and investigates whether these clusters differ in their moral decision making. This paper may serve as a springboard due to the similar nature of the datasets that are used.

Dong et al. (2023): This study specifically focused on psychological fears regarding artificial intelligence in 20 countries and attempted to predict it. Moreover the study investigated the effects in 6 different fields of application. Countries were from allover the world and culturally highly distinct.

O’Shaughnessy et al. (2022): In this paper the focus was on the effect of artificial intelligence on governance and policy making. It also investigated some factors which could affect attitudes towards AI. Cultural values were perceived to be of importance by this study. This can be tested with regards to the moral machine dataset based on a European framework.

Schepman & Rodway (2020): Came to the conclusion that acceptance of and attitudes towards artificial intelligence differs from traditional technology acceptance that has been observed previously. This leaves a gap in knowledge as to what has caused this difference which may become more clear when investigating the moral machine dataset.

This literature can be used to establish a well defined research scope and adequately focus on the research question. Moreover, it can guide the research in a direction as to fill knowledge gaps that as of yet still exist in the academic literature regarding prediction of AI attitudes among EU citizens. While some research has been done concerning prediction of attitudes towards AI as outlined above, much is still unclear about the specific relationships at play.

Moreover, a significant proportion of knowledge on the topic of attitudes towards AI is merely based on simple surveys such as the research by Gillespie et al. (2023) which investigated in 17 countries. Because of this, although there is some research regarding attitudes towards AI, what this means for the scientific and societal field is yet to be studies and understood. Example studies such as the ones by Dong et al. (2023 ) or O’Shaughnessy et al. (2022) arguably were not specific enough with regards to a well defined scope which is something which will be taken into account in this paper by focusing specifically on the EU and relevant materials.

1. **Research Strategy & Research Questions**

* **Research Strategy**

The aim of this research would be to construct an algorithm which predicts the country of origin of the respondent based on their responses to other questions in the moral machine survey. Based on the success of the algorithm there could be several key takeaways.

If the algorithm is able to accurately predict the country of origin of the respondent then it likely means that there is significant distinctions between EU countries’ citizens responses which could lead to the conclusion that there are differences in moral attitudes between these EU member states. That being said, if the algorithm is unable to accurately predict country of origin based on the input, it is likely that there is insufficient distinction between EU citizens’ responses to the moral machine dataset questions. This might lead to the conclusion that there are not any significant differences in attitudes on AI between EU member states.

Both of these would be key findings since the research aims to find out if recent European Union policy regarding AI can be generalized on a EU scale or whether there are excessive differences in attitudes between EU member states for there to be a single cohesive legislative framework. The implications of this research can then be used for input in governmental decision making and regulation when it concerns AI and these sorts of developments.

Moreover, this paper might be able to probe into the possibility of predicting European attitudes towards AI based on their moral decision making and future research could improve upon the findings of these observations.

* **Research Questions:**

**RQ:** *To what extent can EU country of origin be predicted based on the data variables?*

**S-RQ1:** *What model performs best at predicting the country of origin?*

**S-RQ2:** *What variables are important when it comes to making predictive models?*

**S-RQ3:** *Which relationships can be established between certain variables?*

These research questions will provide a guiding direction during the course of the paper and will delineate the scope and subject of the paper.

1. **Methodology and Evaluation**

* **Methodology**

The research will be conducted using the ‘moral machine’ dataset which can be found at the following URL: <https://osf.io/3hvt2/>

Awad et al. (2018) constructed the dataset for ‘The Moral Machine experiment’. It contains responses to moral scenarios relating to self-driving cars. Respondents are from allover the world and submit the choices they would make in certain scenarios based on their moral judgment.

Two models will be used in order to evaluate the predictive power of the dataset.

According to research by Bulavas et al. (2021) random forests and gradient boosting classification have been shown to perform based on this type of dataset.

Random forests will serve as a baseline since it has been shown to perform well and is arguably an interpretable machine learning model for multi class classification (Bulavas et al., 2021).

Moreover, this will aid in comparing model performance and investigating variable relationships. Hyperparameter tuning will be done in order to find the optimum values for each parameter.

Given the size of the dataset and overall efficiency, training times will be recorded as to improve efficiency and avoid unnecessary complexity.

With regards to class imbalances, differences in origin counts have already been taken into account in the choice of models. All instances of the origin variable occur adequately in order for the findings to be representable, also in proportion to their respective population sizes. That being said, larger countries obviously feature more in the dataset as their population is larger. Nevertheless, this has been taken into account in the set up of the research.

Variables with a NULL value or missing values will be dealt with accordingly, deleted if this does not exacerbate class differences and imputated if this increases the representability of the findings.

* **Evaluation**

In order to effectively asses performance of the models employed several evaluation metrics could be used in order to quantify the results. Some of these would be:

*Accuracy*

*Precision*

*Recall*

*F-Score*

*Confusion matrices*

Due to the categorical nature of the country origin variable these evaluation metrics would likely be most useful in order to asses performance given that these metrics are well suited to this. Moreover, feature selection and feature extraction will be performed in order to shape the data in an optimal way for model training and testing while also aiming to minimize the loss of data. K-fold cross validation will be preferred for assessing parameter performance.

The dataset is too rich in volume to be used fully, therefore a subset of the dataset will be used for the training and testing of the models. This has the potential effect of making the results less representable, however, it is a tradeoff which has to be made due to processing and computational efficiency. Each row symbolizes an individual and their answers to the survey. Experimentation with sub setting will be done in order to find the optimum balance between size and efficiency. This should lead to a sufficient size for the dataset which will be used, while also making its processing significantly smoother and less computationally expensive. This should theoretically benefit the ability to create models for this research.

1. **Milestones and Plan**

|  |  |
| --- | --- |
| Thesis Proposal | 18-03-2024 |
| Finalize Outline of Research | 24-03-2024 |
| Start Model Experimentation | 25-03-2024 |
| Proposal Presentation | 08-04-2024 |
| Evaluation of Models Completed | 21-04-2024 |
| Finish Draft Thesis Version | 06-05-2024 |
| First Submission | 20-05-2024 |
| Second Submission | 24-06-2024 |

**Bibliography:**

Awad, E., Dsouza, S., Kim, R. *et al.* The Moral Machine experiment. *Nature* **563**, 59–64 (2018). <https://doi.org/10.1038/s41586-018-0637-6>

Bulavas, V., Marcinkevičius, V., & Rumiński, J. (2021). Study of Multi-Class Classification Algorithms’ performance on Highly imbalanced network intrusion datasets. *Informatica (Lithuanian Academy of Sciences)*, 441–475. <https://doi.org/10.15388/21-infor457>

Dong, M., Conway, J., Bonnefon, J., Shariff, A., & Rahwan, I. (2023). A Psychological Model Predicts Fears about Artificial Intelligence across 20 Countries and 6 Domains of Application. <https://doi.org/10.31234/osf.io/pjvqt>

Gillespie, N., Lockey, S., Curtis, C., Pool, J., & Akbari, A. (2023). Trust in Artificial Intelligence: A Global Study. The University of Queensland and KPMG Australia. doi:10.14264/00d3c94

O’Shaughnessy, M., Schiff, D., Varshney, L. R., Rozell, C. J., & Davenport, M. A. (2022). What governs attitudes toward artificial intelligence adoption and governance? *Science and Public Policy*, *50*(2), 161–176. <https://doi.org/10.1093/scipol/scac056>

Schepman, A., & Rodway, P. (2020). Initial validation of the general attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports*, *1*, 100014. <https://doi.org/10.1016/j.chbr.2020.100014>

Szent‐Iványi, B., & Kugiel, P. (2020). The Challenge from within: EU Development Cooperation and the Rise of Illiberalism in Hungary and Poland. *Journal of Contemporary European Research*, *16*(2), 120–138. <https://doi.org/10.30950/jcer.v16i2.1078>