# The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing

Andrea Agiollo* Luciano C. Siebert†
Pradeep K. Murukannaiah† Andrea Omicini*

*Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum – Università di Bologna
{andrea.agiollo, andrea.omicini}@unibo.it

†Delft University of Technology, Delft, The Netherlands
{P.K.Murukannaiah, L.CavalcanteSiebert}@tudelft.nl

$5^{th}$ International Workshop on EXplainable and TRAnsparent AI and Multi-Agent Systems (EXTRAAMAS)

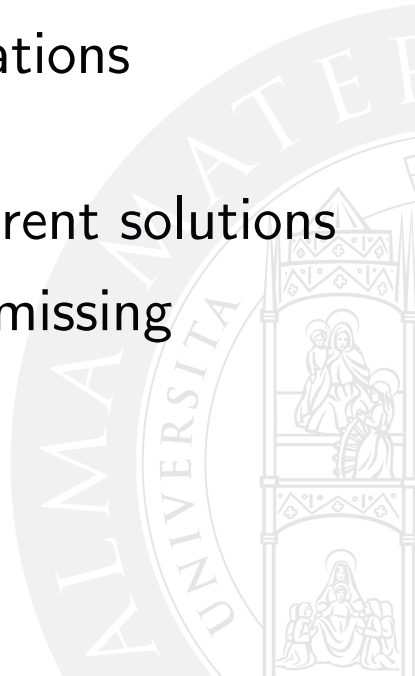
$29^{th}$ May 2023, London, UK

# Next in Line...

1. Context & Motivation
2. Similarity between LPEs
3. Experiments & Results

# Context & Motivation

- *Large language models* (LLM) de-facto solution for Natural Language Processing (NLP) tasks
- LLMs characterised by performance vs. transparency trade-off that hinders their applicability to sensitive scenarios
- XAI research approaches focus on local post-hoc explanations (explanations for single inputs after model is trained)
- different local post-hoc explanainers (LPEs) rely on different solutions
- comparison among available explainability techniques is missing



# Contribution of the paper

## In this paper...

- we propose novel framework for comparing state-of-the-art LPEs
- we apply such framework to the Moral Foundation Twitter Corpus (MFTC) dataset
- we identify the extent of explanations correlation when applied over large set of inputs

# Next in Line...

1 Context & Motivation

2 Similarity between LPEs

3 Experiments & Results

# Background on LPEs

- analysis of the similarity between state-of-the-art *Local Post-Hoc Explainers* (LPEs)

## Local explanations

*Local* $\rightarrow$ explainability approaches that given a single input, produce an explanation of the reasoning process followed by the black-box model to output its prediction for the given input[Luo et al., 2021]

## Post-Hoc explanations

*Post-hoc* $\rightarrow$ explainability approaches which apply to an already optimized black-box model for which it is required to obtain some sort of insight[Madsen et al., 2022]

# Defining LPEs comparison I

## Assumption

comparing LPEs over single local explanations is complex
(lack of well-defined metrics)



we consider measuring how much LPEs correlate with each other over a set of fixed samples

- LPEs are applied to the same black-box model
- explanations may differ over local samples, reasonable to assume for Local Post-hoc explainer (LPE)s to converge to similar results when large number of samples

# Defining LPEs comparison II

## Definitions

- $\epsilon_{NN}$ defines LPE applied to a NN model
- $\epsilon_{NN}$ applied to input $x_i$ produce output one impact score for each input component
- $j\left(l_k, \epsilon_{NN}(x_i)\right)$ defines output impact score for single component
- $\epsilon_{NN}(\mathcal{S})$ defines aggregated LPE impact scores over NN model and set of samples $\mathcal{S}$
- $\mathcal{A}$ defines an aggregation operation

## LPEs correlation metric

$$\mathcal{C}\left(\epsilon_{NN}\left(\mathcal{S}\right), \epsilon'_{NN}\left(\mathcal{S}\right)\right) = \mathcal{C}\big(\mathcal{A}\left(\{\epsilon_{NN}(x_i) \text{ for each } x_i \in \mathcal{S}\}\right), \mathcal{A}\left(\{\epsilon'_{NN}(x_i) \text{ for each } x_i \in \mathcal{S}\}\right)\big) \quad (1)$$

# State-of-the-art LPEs I

- Gradient Sensitivity Analysis (GS), gradient over input components
- Gradient $\times$ Input (GI), GS multiplied by input components[Kindermans et al., 2019]
- Layer-wise Relevance Propagation (LRP), conservation of relevance scores (gradient) across the layers[Ali et al., 2022, Bach et al., 2015]
- Layer-wise Attention Tracing (LAT), analyse relevance scores of attentions heads[Abnar and Zuidema, 2020, Wu et al., 2020]
- Integrated Gradient (HESS), integrate gradient along trajectory of input space[Sundararajan et al., 2017]
- SHapley Additive exPlanations (SHAP), Shapley values to identify contribution of input components[Kokalj et al., 2021]
- Local Interpretable Model-agnostic Explanations (LIME), identify relevant components learning explainable surrogate[Ribeiro et al., 2016]

# Impact scores aggregation I

$\mathcal{A}$ defines an aggregation operation

- *Sum* – summation to obtain aggregate score

$$\mathcal{A}\left(\{\epsilon_{NN}(\mathrm{x}_i) \text{ for each } \mathrm{x}_i \in \mathcal{S}\}\right) = \left\{ \sum_{i=1}^{N} j\left(l_k, \epsilon_{NN}\left(\mathrm{x}_i\right)\right) \text{ for each } l_k \in \mathcal{S} \right\} \quad (2)$$

- *Absolute sum* – summing absolute values to increase awareness of high impact scores

$$\mathcal{A}\left(\{\epsilon_{NN}(\mathrm{x}_i) \text{ for each } \mathrm{x}_i \in \mathcal{S}\}\right) = \left\{ \sum_{i=1}^{N} \left| j\left(l_k, \epsilon_{NN}\left(\mathrm{x}_i\right)\right) \right| \text{ for each } l_k \in \mathcal{S} \right\} \quad (3)$$

# Impact scores aggregation II

- *Average* – averaging scores to account for lemmas frequency

$$\mathcal{A}\left(\{\epsilon_{NN}(\mathsf{x}_i) \text{ for each } \mathsf{x}_i \in \mathcal{S}\}\right) = \left\{ \frac{1}{N} \cdot \sum_{i=1}^{N} j\left(l_k, \epsilon_{NN}\left(\mathsf{x}_i\right)\right) \text{ for each } l_k \in \mathcal{S} \right\} \tag{4}$$

- *Absolute average* – similar to absolute sum, for better-managing lemmas with skewed impact

$$\mathcal{A}\left(\{\epsilon_{NN}(\mathsf{x}_i) \text{ for each } \mathsf{x}_i \in \mathcal{S}\}\right) = \left\{ \frac{1}{N} \cdot \sum_{i=1}^{N} \left| j\left(l_k, \epsilon_{NN}\left(\mathsf{x}_i\right)\right) \right| \text{ for each } l_k \in \mathcal{S} \right\} \tag{5}$$

# Comparing aggregated LPEs scores I

- aggregated global explanation $J$ depends on corresponding label $\tau_m$
- define set of aggregated global scores depending on label:

$$\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S}) = \{J(I_k, \epsilon_{NN}) \,|\tau_m \text{ for each } I_k \in \mathcal{S}\} \tag{6}$$

- use Pearson correlation to compare aggregated global scores
- overall LPEs correlation is average Pearson correlation over labels

$$\mathcal{C}\left(\epsilon_{NN}(\mathcal{S}), \epsilon'_{NN}(\mathcal{S})\right) = \frac{1}{M} \cdot \sum_{m=1}^{M} \rho\left(\|\mathcal{J}_{\tau_m}(\epsilon_{NN}, \mathcal{S})\|, \|\mathcal{J}_{\tau_m}\left(\epsilon'_{NN}, \mathcal{S}\right)\|\right) \tag{7}$$

# Next in Line...

1. Context & Motivation
2. Similarity between LPEs
3. Experiments & Results

# The Moral Foundation Twitter Corpus Dataset (MFTC)

- MFTC dataset composed of 35,108 tweets (sentences)[Hoover et al., 2020]
- considered a collection of seven different datasets, collected following a certain event or context target:
- each tweet labelled (by multiple annotators) following moral theory
- one or more of 11 moral values
  - care/harm
  - fairness/cheating
  - loyalty/betrayal
  - authority/subversion
  - purity/degradation
  - non-moral

# Model Training

- morality classification as multi-class multi-label classification
- BERT as the LLM to be optimised[Devlin et al., 2018]
- 70% training and 30% testing
- pre-trained *bert-base-uncased* model – from Hugging Face python library* – as starting point
- each model trained for 3 epochs, binary cross entropy loss,[Zhang and Sabuncu, 2018] learning rate $5 \times 10^{-5}$, batch size 16, maximum sequence length 64

| | ALM | BLM | BLT | DAV | ELE | MT | SND | TOT |
|---|---|---|---|---|---|---|---|---|
| $F_1$ score | 63.04% | 82.59% | 64.51% | 88.12% | 63.14% | 52.16% | 56.85% | 69.10% |

Table: BERT performance over MFTC datasets

*https://github.com/huggingface

# Are Local Post-hoc Explainers Aligned? I

- compute pairwise correlation $\mathcal{C}\left(\epsilon_{NN}\left(\mathcal{S}\right), \epsilon'_{NN}\left(\mathcal{S}\right)\right)$ for each pair of LPE

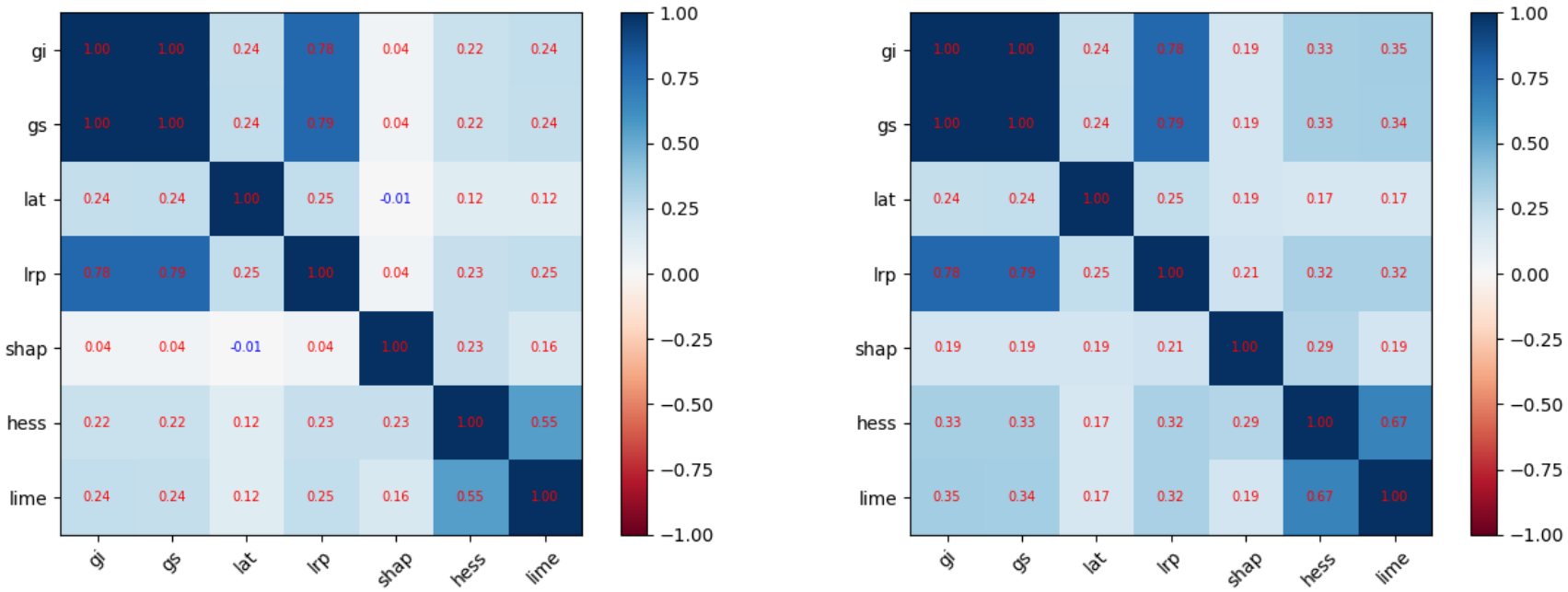



Figure: $\mathcal{C}\left(\epsilon_{NN}\left(\mathcal{S}\right), \epsilon'_{NN}\left(\mathcal{S}\right)\right)$ using average aggregation (left) and absolute average aggregation (right) as $\mathcal{A}$ over the Black Lives Matter (BLM) dataset

# Are Local Post-hoc Explainers Aligned? II

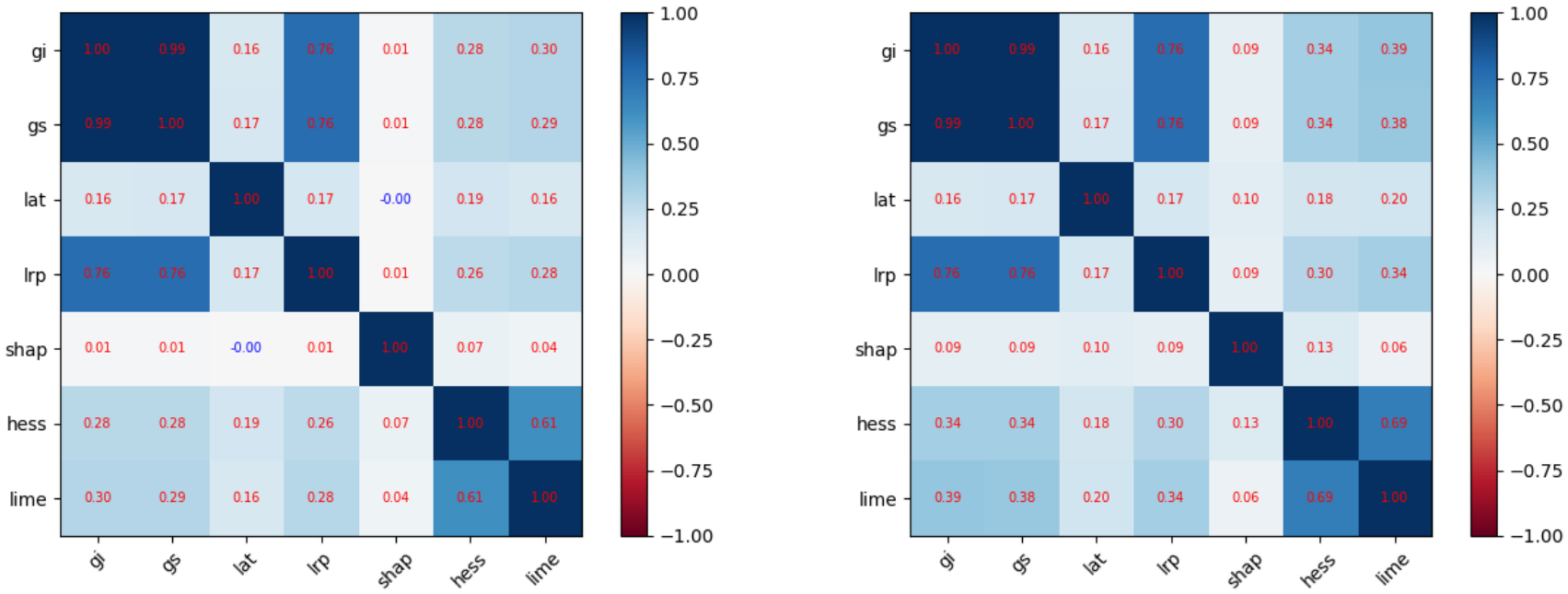



Figure: $\mathcal{C}\left(\epsilon_{NN}\left(\mathcal{S}\right), \epsilon'_{NN}\left(\mathcal{S}\right)\right)$ using average aggregation (left) and absolute average aggregation (right) as $\mathcal{A}$ over the 2016 presidential election (ELE) dataset

# Are Local Post-hoc Explainers Aligned? III

## Interesting results

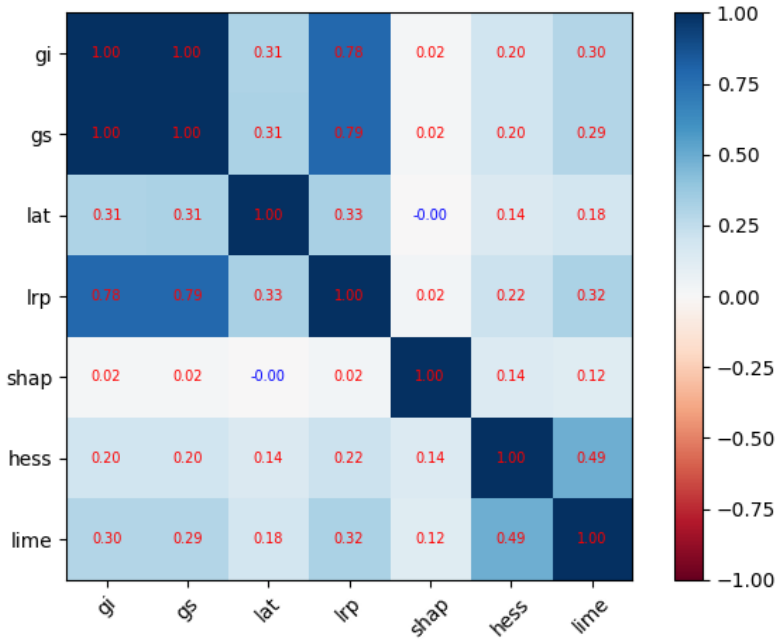
- very weak correlation score between most LPEs over different datasets (*LPEs quarrel*)
- few specific couples or clusters of LPEs highly correlated (rely on similar approach)
- GI and HESS little to no correlation, although both gradient-based
- SHAP and LIME no correlation even both rely on input perturbation

## Why quarrel?

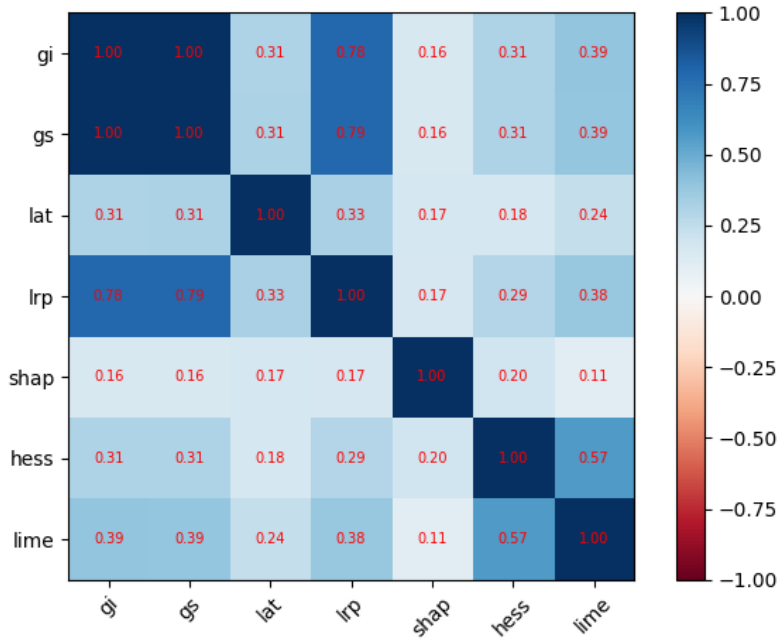
- unreliability of LPEs for identifying LLMs reasoning principles
- uncorrelated LPEs highlight different subsets of LLMs reasoning principles

# How Does Impact Scores Aggregation Affect Correlation? I

- LPEs correlation metric is dependent on $\mathcal{A}$

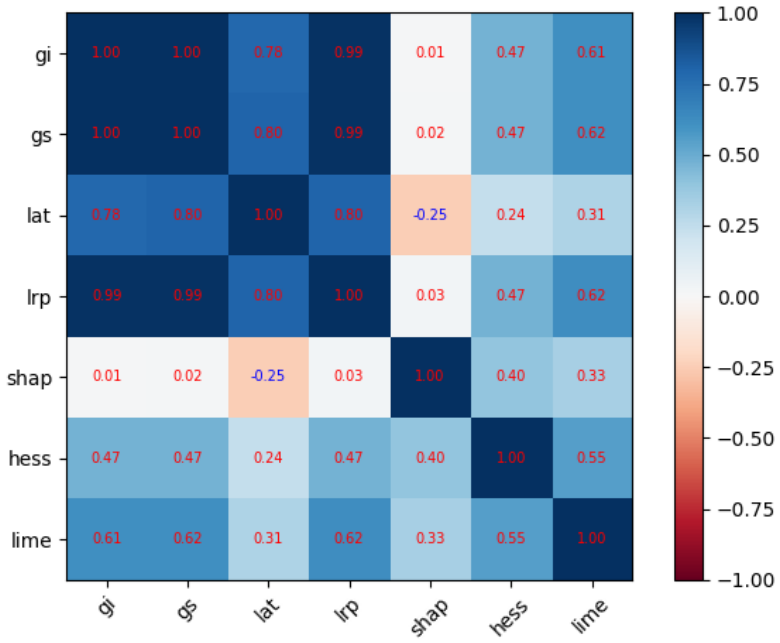



(a) average aggregation

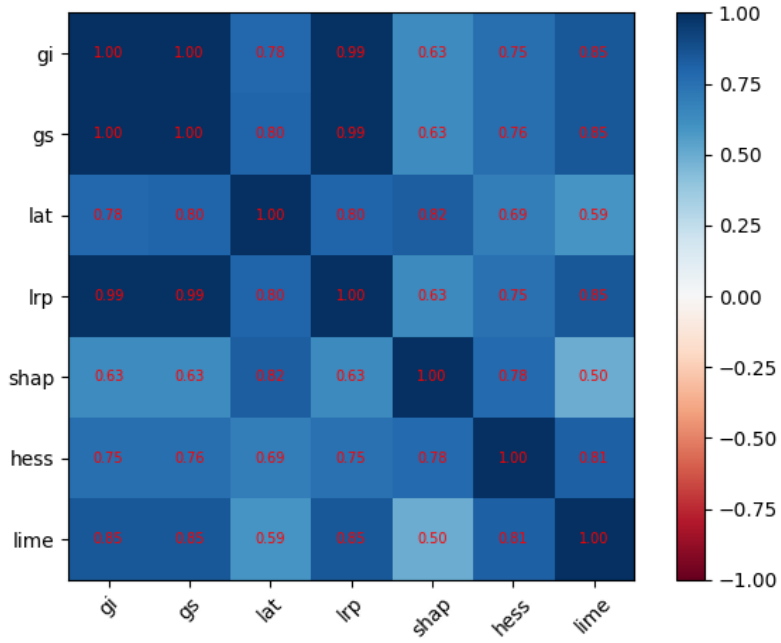



(b) absolute average aggregation

# How Does Impact Scores Aggregation Affect Correlation? II




(c) sum aggregation




(d) Absolute sum aggregation

Figure: $\mathcal{C}\left(\epsilon_{NN}\left(\mathcal{S}\right), \epsilon'_{NN}\left(\mathcal{S}\right)\right)$ using different aggregations over the All Lives Matter (ALM) dataset.

# How Does Impact Scores Aggregation Affect Correlation? III

## Interesting results

- summation leads to strong correlation
  - ? contrast with previous findings?
- susceptibility of summation to tokens frequency
  - → overestimate the relevance of popular concepts
- absolute value leads to higher correlation scores
- absolute value shrinks differences between positive and negative scores
- LPEs rely heavily on scoring positive contributions and give less focus to negative impact scores

# Conclusions

## Summing up

- we propose new approach for comparison among LPE mechanisms, identifying the extent of explanations correlation
- novel framework for LPEs comparison applied to MFTC dataset
- experiments show LPEs explanations are far from correlated when applied over large set of inputs

## Future Work

- analysis of the alignment between LPEs explanations and human morality
- in-depth investigation of LPEs robustness
  - $\rightarrow$ can we rely on them to get the model from a global perspective?

# The Quarrel of Local Post-hoc Explainers for Moral Values Classification in Natural Language Processing

Andrea Agiollo* Luciano C. Siebert†
Pradeep K. Murukannaiah† Andrea Omicini*

*Dipartimento di Informatica – Scienza e Ingegneria (DISI)
Alma Mater Studiorum – Università di Bologna
{andrea.agiollo, andrea.omicini}@unibo.it

†Delft University of Technology, Delft, The Netherlands
{P.K.Murukannaiah, L.CavalcanteSiebert}@tudelft.nl

$5^{th}$ International Workshop on EXplainable and TRAnsparent AI and Multi-Agent Systems (EXTRAAMAS)

$29^{th}$ May 2023, London, UK

# References I

[Abnar and Zuidema, 2020] Abnar, S. and Zuidema, W. (2020).
Quantifying attention flow in transformers.
*arXiv preprint arXiv:2005.00928*.

[Ali et al., 2022] Ali, A., Schnake, T., Eberle, O., Montavon, G., Müller, K.-R., and Wolf, L. (2022).
Xai for transformers: Better explanations through conservative propagation.
In *International Conference on Machine Learning*, pages 435–451. PMLR.

[Bach et al., 2015] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015).
On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.
*PloS one*, 10(7):e0130140.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018).
Bert: Pre-training of deep bidirectional transformers for language understanding.
*arXiv preprint arXiv:1810.04805*.

[Hoover et al., 2020] Hoover, J., Portillo-Wightman, G., Yeh, L., Havaldar, S., Davani, A. M., Lin, Y., Kennedy, B., Atari, M., Kamel, Z., Mendlen, M., et al. (2020).
Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment.
*Social Psychological and Personality Science*, 11(8):1057–1071.

[Kindermans et al., 2019] Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. (2019).
The (un) reliability of saliency methods.
*Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 267–280.

# References II

[Kokalj et al., 2021] Kokalj, E., Škrlj, B., Lavrač, N., Pollak, S., and Robnik-Šikonja, M. (2021).
Bert meets shapley: Extending shap explanations to transformer-based classifiers.
In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 16–21.

[Luo et al., 2021] Luo, S., Ivison, H., Han, C., and Poon, J. (2021).
Local interpretations for explainable natural language processing: A survey.
*arXiv preprint arXiv:2103.11072*.

[Madsen et al., 2022] Madsen, A., Reddy, S., and Chandar, S. (2022).
Post-hoc interpretability for neural nlp: A survey.
*ACM Computing Surveys*, 55(8):1–42.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
"why should i trust you?" explaining the predictions of any classifier.
In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017).
Axiomatic attribution for deep networks.
In *International conference on machine learning*, pages 3319–3328. PMLR.

[Wu et al., 2020] Wu, Z., Nguyen, T.-S., and Ong, D. C. (2020).
Structured self-attention weights encode semantics in sentiment analysis.
*arXiv preprint arXiv:2010.04922*.



# References III

[Zhang and Sabuncu, 2018] Zhang, Z. and Sabuncu, M. (2018).
Generalized cross entropy loss for training deep neural networks with noisy labels.
*Advances in neural information processing systems*, 31.