Summary – Workshop “System of AI Accountability in Financial Services: Quantifying AI Ethics Principles”

The TUM Institute for Ethics in Artificial Intelligence’s team, as part of their collaborative project with Fujitsu Global: "Towards an Accountability Framework for AI systems", conducted a workshop at the TUM Think Tank addressing the **quantification of ethical principles for AI in finance**. The workshop aimed to foster exchange with specialists regarding their perception of scalable characteristics to evaluate the ethicality of AI applications in finance. Participants from the financial industry joined the event, bringing diverse backgrounds and perspectives to the discussion. The workshop enabled participants to collectively formulate and reach a consensus on **critical scalable characteristics** that reflect how AI products align with ethical principles, focusing on the investigated principle of 'explainability & transparency' for the AI Credit Scoring use case. A numeric scale was devised to score the defined characteristics. The essential elements of our discussions and preliminary findings are presented here.

Use Case: Credit Scoring (CS)

A CS AI tool developed and tested in the EU and based on Neural Networks models (making it quite obscure) is put on the market. The company proposing it claims that their tool expands access to capital and financial services for marginalized communities and uses both financial and non-specified alternative data for decision-making when the client gives consent to disclose its data, as required to comply with GDPR.

Procedure

To begin, we conducted an intuitive survey to gain an initial understanding of the participants' perspectives. This survey aimed to gather insights and opinions regarding the characteristics crucial for assessing transparency and explainability in credit-scoring AI systems. We then ran a collaborative discussion where participants shared their thoughts and ideas in groups, identifying and listing various characteristics contributing to the transparency and explainability of credit-scoring AI systems. We then gathered together and let the expert pick the five most relevant characteristics from the fifteen generated during the group discussions. The participants were then asked to individually state scores for each state (critically low to excellent) in fulfilling the characteristics given. This framework enabled us to calculate a quantitative assessment for each characteristic, which ultimately facilitated the creation of a scale for evaluating transparency in credit-scoring AI systems. To further refine our assessment, participants were asked to report the importance of each characteristic concerning the principle of transparency and explainability. The expert opinions were carefully integrated into the final calculation of the general scale for adherence to the principle of transparency and explainability in credit scoring AI systems.

By following this systematic approach, we aim to establish a robust methodology for assessing the quantification of adherence to ethical principles in AI systems.

Results

1. Intuitive Assessment

A first exercise was conducted to gauge the participants' intuitive assessment of the ethicality of typical credit scoring applications regarding the principle of explainability and transparency. Participants were asked to provide their opinions on the **current state of adherence** to this principle, using a scale ranging from critically low (1) to excellent (5). The results revealed that **60% of the participants perceived adherence as low**, 10% considered it satisfactory, 10% rated it as good, and 20% believed it to be critically low. In other words, according to our participants' intuition, transparency and explainability in European AI Credit Scoring systems is, at this time, rather low.

2. Characteristics Definition

Participants were then split into three groups and asked to define five scalable characteristics per group to evaluate the adherence to the principle of transparency and explainability for credit scoring systems.

Group 1	Group 2	Group 3
Share of relevant data points that were used in a decision-making of AI CS that was disclosed and explained to the customer.	Weight of data source and type	Share of documentation of relevant steps in the AI tool lifecycle (defined by standards and including post-hoc adjustments)
Share of AI CS decisions that was reviewed by a credit analysis' domain expert	Share of cases where human intervention was needed	Share of cases for which output is reproducible within acceptable standards (defined by standards)
Share of reviewed decisions by a AI CS, explanations on which were found satisfactory by a domain expert	Share of (sensitive) features used	Share of group of users (reporting) understanding of the tool (UX research)
Share of predictions correctly explained by a local interpretation method	Model metrics (accuracy, confidence level, fairness metrics)	Share of known potential limitations presented to the public
Share of complaints/incidents asked on a AI CS decision after a customer asked for clarification on his/her decision	Number of different data sources/share of trustworthy data sources	Share of information about the system that is publically available (based on internal documentation)

Table 1: Summary of the characteristics defined per group.

3. Scale Definition

After a quorum discussion, five characteristics were defined as the most representative ones. Participants were asked to determine which implementation characteristics would be fulfilled at a critically low to excellent status. In addition, participants were asked the ratio of importance for each characteristic. The averaged results are presented in this table:

Ratio of importance	Assessment of the state – Characteristics	critically low	low	satisfactory	good	excellent
0,27	(1) Share of relevant features that are involved in the AI CS decision that were disclosed and explained to the customers	0,2	0,4	0,5	0,6	0,8
0,25	(2) Share of relevant data that comes from trustworthy data sources	0,3	0,5	0,6	0,7	0,9
0,18	(3) Share of prediction performance metrics and limitations correctly explained to the target group	0,34	0,43	0,52	0,62	0,77
0,13	(4) Ratio of inquiries on AI CS relating to understandability	0,4	0,5	0,7	0,8	0,9
0,17	(5) Share of AI CS decisions that were reviewed by a domain expert (credit analyst)	0,0	0,1	0,1	0,2	0,3
	Generalised Scale	**0,26**	**0,36**	**0,48**	**0,61**	**0,75**

Table 2: Final characteristics, the ratio of importance, associated assessment of state scale, and generalised scale based on experts' opinions.

We can see the possibility of quantifying the threshold between different states of characteristic fulfilment based on experts' opinions. This methodology thus seems relevant for the future of the quantification of AI ethics. However, the concrete values presented in this table are based on a hypothetical scenario and do not build on a practical, specific tool. Thus, the proposed scale needs to be understood in relation to its context when used to assess the AI CS tool at this stage of scale development.

Conclusion and Outlook

The workshop effectively yielded valuable insights and initial findings on measuring ethical principles in AI systems used in financial contexts. The workshop outcomes identified five distinct characteristics that exemplify compliance with the principles of transparency and explainability for AI credit scoring systems. Additionally, a quantifiable scale was developed to assess the extent of implementation of each of these characteristics and, consequently,

to evaluate the overall adherence to the principle of transparency and explainability for AI credit scoring systems.

Building on the initial intuitive assessment from our expert participants regarding the strong lack of transparency and explainability for AI CS systems at this time, we confirmed the need to develop clear scalable characteristics to evaluate at which level of ethicality in a given context a tool is. With our methodology, we propose a first step towards a solution for systematically evaluating the ethicality of AI technologies by developing clear scalable characteristics being context-dependent.

These preliminary results pave the way for further steps in quantifying AI systems' adherence to ethical principles. Continued evaluation and refinement of the defined characteristics will contribute to developing a comprehensive framework for assessing the ethicality of AI applications in chosen sectors and use cases. The workshop's outcomes provide a solid foundation for ongoing research and collaborative efforts to ensure AI's responsible and ethical development, implementation and use.