DIGITAL FINANCE

This project has received funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101119635

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

State Secretariat for Education, Research and Innovation SERI

Funded by the European Union

MARIE CURIE ACTIONS

DIGITAL

Evaluation frameworks & Applied Work (SAFE AI framework)

Golnoosh Babaei
Researcher
University of Pavia

DIGITAL

Funded by
the European Union

AI: opportunities and risks

-AI can improve operational efficiency, consumer experience and inclusion.

-AI can also bring risks, such as cyber security risks, model risks, governance risks, discrimination risks.

-For this reason, codes of practice and international standards are being developed, by different organisations, for providers and deployers of AI systems.

-The EU AI Act is a risk based law, which distinguishes:

- -Prohibited AI practices (e.g. social scoring)
- -High risk AI practices: allowed, subject to a risk management system (e.g. credit scoring, life insurance pricing)
- -Limited risk practices: allowed, subject to transparency requirements (e.g. chatbots)

DIGITAL

Responsible AI

- There is a need to move from AI to a SAFE AI
- A responsible AI should be able to measure and manage the risk of making harms.
- The aim of our research is the development of statistical metrics for a SAFE AI, in line with the aims of the association iaseai.org

DIGITAL

SAFE Artificial Intelligence

In Babaei et al. (2025) we have mapped AI risk in a S.A.F.E. AI risk management system based on four risk classes:

Security: AI systems should achieve an appropriate level of robustness and cybersecurity, and be resilient to internal anomalies and external attacks (AI Act 15.3,15.4). They should also be environmentally sustainable.

Accuracy: AI systems should achieve an appropriate level of accuracy, and the relevant accuracy metrics should be disclosed (AI Act 15.1, 15.2);

Fairness: Data for AI systems should be relevant, representative and complete, and have appropriate statistical properties, as regards to the persons and groups on which the AI system is used (AI Act 10);

Explainability: AI systems should be designed so that they can be effectively overseen by natural persons (AI Act 14).

DIGITAL

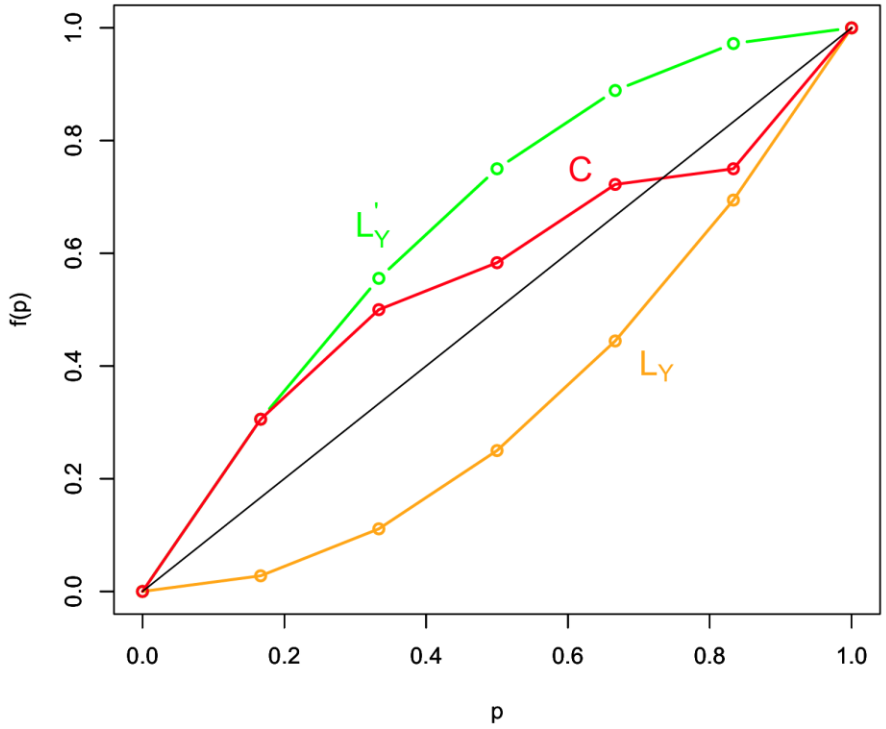
SAFE AI Metrics

- We would like to measure the four principles in a consistent manner, and produce metrics that are model agnostic and transparent.
- To this aim, in Babaei et al. (2025), we have proposed four metrics, that extend the well known Area Under the ROC Curve (AUC) to all principles, and all types of variables, using the framework of the Lorenz curve and of the Gini coefficient.
- The metrics can be mapped to the probability that an AI system cause a harm, due to lack of S., A., F., E., for different categories of stakeholders.

DIGITAL

Ingredients for SAFE AI metrics - I

The Lorenz curve (L_Y), Dual Lorenz curve (L'_Y) and the concordance curve (C) are the main ingredients for the SAFE AI metrics. Here, p (on the x-axis) and f (p) (on the y -axis) are the cumulative values of the x and y coordinates of the L_Y , L'_Y and C curves.



DIGITAL

Ingredients for SAFE AI metrics - II

- the Lorenz curve L_{Y^*} is obtained ordering the Y^* values in a non-decreasing sense: $(i/n, \sum_{j=1}^{i} y^*_{r_j^*}/(n\bar{y}^*))$, where r_j^* indicates the non-decreasing ranks of Y^*.
- the dual Lorenz curve $\bar{L'}_{Y^*}$ is obtained ordering the Y^* values in a non-increasing sense: $(i/n, \sum_{j=1}^{i} y^*_{r^*_{n+1-j}}/(n\bar{y}^*))$, where r^*_{n+1-j} indicates the non-increasing ranks of Y^*.
- the concordance curve C is obtained ordering the Y^* values with respect to the ranks of Y^{**}, r_i^{**}: $(i/n, \sum_{j=1}^{i} y^*_{r_j^{**}}/(n\bar{y}^*))$, where r_i^{**} indicates the non-decreasing ranks of Y^{**}.

DIGITAL

Proposed metric to measure AI risks

Considering the location of the Concordance curve along with the Lorenz and Dual Lorenz curves, RG metric is the ratio of the area between the Concordance curve and the Dual Lorenz curve over the area between the Lorenz and Dual Lorenz curves (Lorenz Zonoid).

The ratio can be expressed mathematically as follows:

$$RGmetric = \frac{\sum_{i=1}^{n}\left\{\frac{1}{n\bar{x}}\left(\sum_{j=1}^{i} x_{r_{n+1-j}} - \sum_{j=1}^{i} x_{\hat{r}_j}\right)\right\}}{\sum_{i=1}^{n}\left\{\frac{1}{n\bar{x}}\left(\sum_{j=1}^{i} x_{r_{n+1-j}} - \sum_{j=1}^{i} x_{r_j}\right)\right\}},$$

DIGITAL

Measuring Accuracy - I

- The measurement of predictive accuracy is well known in the data science community. E.g. RMSE can be used for a continuous response; AUROC for a binary response.
- Using the RG metric (Equation 1) we can have a universal metric: the Rank Graduation Accuracy (RGA) measure, which extends the AUROC to all response variables.
- To this aim, define a concordance curve C by ordering the Y values with respect to the ranks of the predicted values.
- Dividing the area between the dual Lorenz and the concordance curves by its maximum value, we obtain:

DIGITAL

Measuring Accuracy - II

$$RGA = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{\hat{r}_j} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{y}} \left(\sum_{j=1}^{i} y_{r_{n+1-j}} - \sum_{j=1}^{i} y_{r_j} \right) \right\}}.$$

- It can be shown that: $0 \leq RGA \leq 1$, with RGA= 1 for a perfectly concordant model; RGA= 0 for a perfectly discordant model; RGA= 0.5 for random predictions;
- When the response Y is binary, RGA=AUROC;
- RGA can however be calculated, in the same way, for all types of response variables.

DIGITAL

Measuring Security (specifically Rbustness): Rank Graduation Robustness (RGR)

A Rank Graduation Robustness measure can be obtained by: considering the predicted values obtained applying a model with data without perturbations $\hat{y}$; considering the predicted values obtained applying the same model with perturbed data $\hat{y}_p$; re-ordering the $\hat{y}$ values with respect to the non-decreasing ranks of the $\hat{y}_p$ values, denoted with r_p. Normalising the area between the concordance curve and the dual:

$$RGR = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{\hat{y}}} \left(\sum_{j=1}^{i} \hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i} \hat{y}_{r_j^p} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{\hat{y}}} \left(\sum_{j=1}^{i} \hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i} \hat{y}_{r_j} \right) \right\}}.$$

$0 \leq RGR \leq 1$, with RGR= 1 for a perfectly robust mode

DIGITAL

Measuring Fairness: RGA Imparity

- Considering model parity concept and the proposed RGA metric, we can find the difference in the model accuracy between the protected groups to measure the fairness of the model.

- Therefore, RGA imparity score proposed in this framework represents the difference between the RGA values in the privileged and unprivileged groups. If this score is close to zero it means that the considered model has a similar performance towards the protected groups while when it is close to one it shows an unfair decision-making framework.

DIGITAL

Measuring Explainability: Rank Graduation Explainability (RGE)

A Rank Graduation Explainability measure can be obtained by: considering the predicted values provided by a full model (including all the K predictors) → $\hat{y}$; considering the predicted values provided by a reduced model (excluding the k-th predictor under evaluation) → $\hat{y}_{-xk}$; re-ordering the $\hat{y}$ values with respect to the non-decreasing ranks of the $\hat{y}_{-xk}$ values, denoted with r_{-xk}. Normalising the area between the concordance and the dual:

$$RGE_k = \frac{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{\hat{y}}} \left(\sum_{j=1}^{i} \hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i} \hat{y}_{r_j^{-x_k}} \right) \right\}}{\sum_{i=1}^{n} \left\{ \frac{1}{n\bar{\hat{y}}} \left(\sum_{j=1}^{i} \hat{y}_{r_{n+1-j}} - \sum_{j=1}^{i} \hat{y}_{r_j} \right) \right\}}$$

0 ≤ RGEk ≤ 1, with RGEk = 1 if the k-th predictor explains all

DIGITAL

SAFE AI application

All proposed metrics are implemented in a Python package, available at the GitHub repository: https://github.com/GolnooshBabaei/safeaipackage.

The package can be installed using the following command:

```
pip install safeaipackage
```

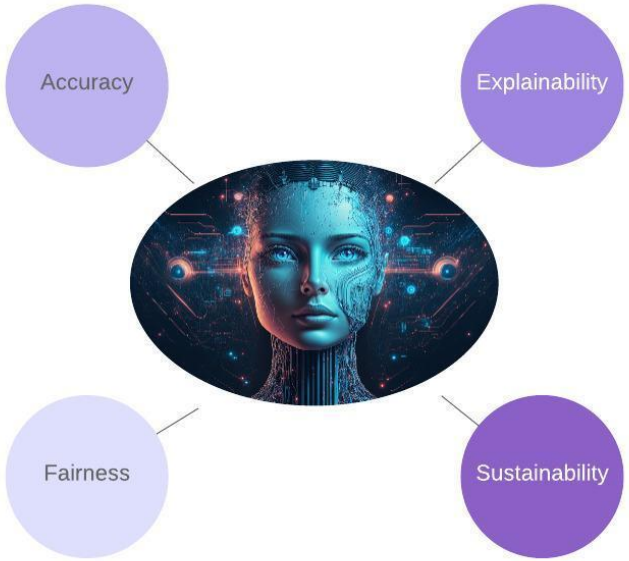
Modules:

- Core
- Check_explainability
- Check_robustness
- Check_fairness

DIGITAL

The safeaipackage

- Each metric measures the "probability" component of the risk: a higher value of the metric indicates a lower risk. The "severity" component of the risk is domain specific, and can be assessed by the user.
- The package can be installed easily. It is also being implemented in the AMELIA platform of the GRINS project: https://grins.it/progetto/piattaforma-amelia.



DIGITAL

core and check_explainability modules:

Core:

```
## Accuracy measured by axa_safeai package

RGA = core.rga(y_val, pred_prob_p1)
RGA
```

```
0.9290676839591755
```

```
## Compare with AUC

AUC = roc_auc_score(y_val, pred_prob_p1)
AUC
```

```
0.9290676839591756
```

check_explainability:

```
check_explainability.compute_single_variable_rge(X_transformed, X_val_transformed, pred_prob_p1, catboost_model,
                                                 ["neta"])
```

	RGE
neta	0.035

DIGITAL

Check_fairness module:

check_fairness:

```
check_fairness.compute_rga_parity(X, X_val, y_val, pred_prob_p1, clf, "cdesc_sesso")
```

As the result of this function, the difference between RGA values is reported:

```
'The RGA-based imparity between the protected gorups is 0.0016847920813110004.'
```

DIGITAL

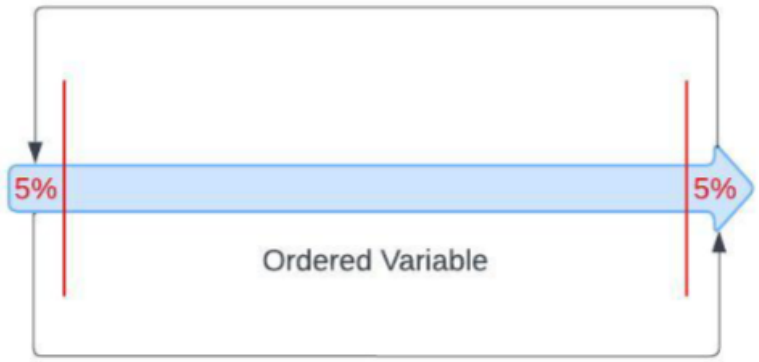
Check_robustness module:

- check_robustness:

```
check_robustness.compute_single_variable_rgr(X_val_transformed, pred_prob_p1, catboost_model, ["neta"])
```

	RGR
neta	0.964

By default, the percentage of perturbation is set to be equal to 5%:



DIGITAL

SAFE Agentic AI

What are agents? What is a multi agent AI system?

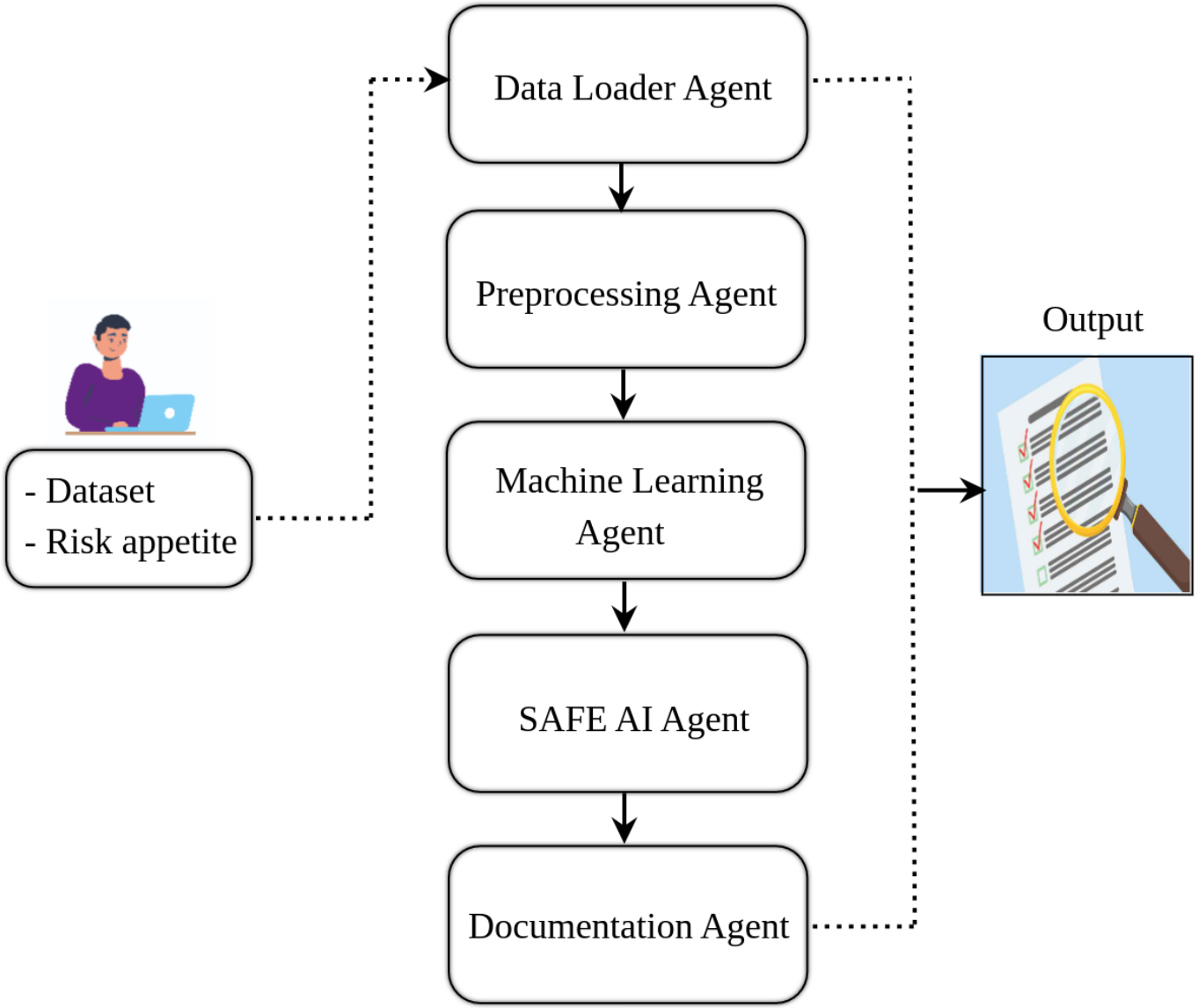
- A *Multi-Agent AI System* is a framework where multiple AI agents collaborate, each performing specialized tasks to solve complex problems more efficiently.

- An autonomous entity that perceives its environment, makes decisions, and acts to achieve specific goals.

- **Key Characteristics of Multi-Agent Systems**
- **Decentralization:** No single point of control
- **Collaboration or Competition:** Agents may work together or independently
- **Specialization:** Each agent can focus on a specific function
- **Communication:** Agents often exchange information to coordinate

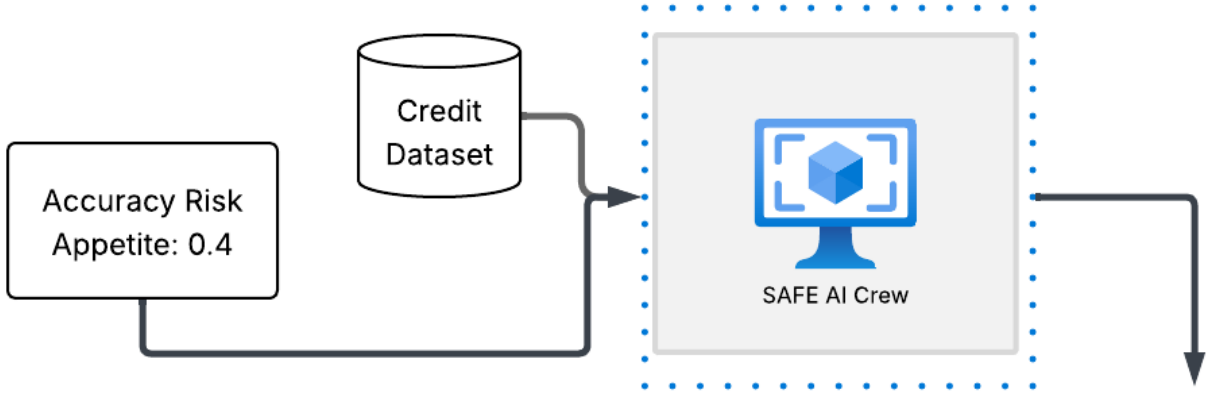
DIGITAL

SAFE AI Crew



DIGITAL

Agentic Credit Risk Management




```
### Summary of RGA Findings:
- **RGA Score**: 0.6875
- **Threshold**: 0.4 (Significantly surpassed)
- **Model Accuracy**: Logistic Regression model accuracy is 0.6667.
- **Conclusion**: The model efficiently classifies data, offering reliable predictions.
```

```
### Summary of Explainability Findings:
- **Highest Explainability**: Gender (0.1673) and Loan Purpose (0.1106).
- **Lowest Explainability**: Race_1 (0.0001), indicating a lack of understanding of the influence of this feature on model predictions.
- Features exhibit varying levels of contribution to model decisions, with some being more interpretable than others.
```

```
### Summary of Robustness Findings:
- The computed RGR score is  **0.91**.
- This indicates the model's strong performance in making predictions, regardless of variations in the data.
```

DIGITAL

Concluding remarks

- We have mapped the requirements of the EU AI Act into four main risk drivers that can be used to assess the probability that a specific AI application causes a harmful event.
- The four risk drivers define our (S.A.F.E.) AI model and are measured by means of consistent metrics, implemented in a Rank Graduation Box that extends the well known AUC by means of concordance measures which can be easily integrated, and are easy to interpret, as all expressed in a [0, 1] probability scale.
- Doing so, AI can measure its own risks and, thus, become more responsible and SAFEr.
- Agents should be SAFE as well.

DIGITAL

References

Babaei, Golnoosh and Giudici, Paolo (2025). A statistical package for safe artificial intelligence: G. Babaei, P. Giudici. Statistical Methods & Applications, 1-19.

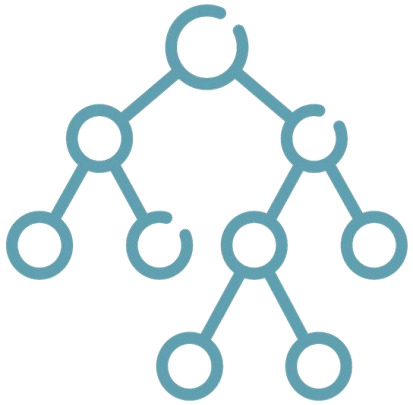
Babaei, Golnoosh, Giudici, Paolo and Raffinetti, Emanuela (2025). A Rank Graduation Box for SAFE AI. Expert Systems with Applications 259, 125239.

Giudici, Paolo and Emanuela Raffinetti (2024). RGA: a unified measure of predictive accuracy. Advances in Data analysis and classification.

Giudici, Paolo and Emanuela Raffinetti (2023).SAFE Artificial Intelligence in Finance. Finance Research Letters , volume 56, 104088.

Giudici, Paolo and Emanuela Raffinetti (2021). Shapley-Lorenz eXplainable Artificial Intelligence Expert Systems With Applications, 167, 114104.

DIGITAL



DIGITAL

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or Horizon Europe: Marie Skłodowska-Curie Actions. Neither the European Union nor the granting authority can be held responsible for them.

DIGITAL

This project has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101119635