AI Tools for Retailers: Assessment Sheet

Disclaimer: The assessment was conducted in the scope of the *Increasing the Uptake of AI in Retail (INAIR)* project, which received funding from the European Union's Horizon Europe Research and Innovation programme - Grant Agreement No. 101133847. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. The assessment of the tools was conducted based solely on vendor documentation accessible at the time of review. The project consortium does not guarantee the accuracy, completeness, or reliability of the information and assumes no responsibility for any decisions or actions taken based on this assessment.

Acknowledgment: The methodology used to evaluate this tool was adapted from the *AI Trustworthiness Framework* developed by the consortium of the STAR project (Horizon-2020-funded project, Grant Agreement No. 956573). The Framework is based on the Assessment List for Trustworthy AI (ALTAI), created by the High-Level Expert Group on AI established by the European Commission.

SOLUTION	**CB4**
PROVIDER	CB4
WEBSITE	https://cb4.com

STRENGTHS	
AREA	**CHARACTERISTICS**
Human Agency & Oversight	• There are clear policies and guidelines for human oversight, outlining when and how human intervention is required during the operation of the AI system. • The system provides explanations for its decisions in a human-understandable manner. • There are triggers or thresholds that prompt human intervention when certain conditions are met or when certain risks are materialised. • The system is designed to include humans in the decision-making process. • The system incorporates redundant systems and safety checks that require human approval before certain actions are taken. • Human oversight is employed to review and correct potential biases. • There are feedback mechanisms for end-users to report concerns or disputes, which can trigger human review and intervention. • There are established review panels or teams consisting of experts and stakeholders that periodically evaluate AI decisions and make necessary adjustments. • Users can customise AI behaviour within certain limits, enabling them to align the system with their values and intentions. • The system adheres to ethical AI frameworks and principles (e.g., IEEE 7000).
Technical Robustness & Safety	• The system employs adversarial training i.e., AI models are trained on data that includes adversarial examples in order to improve their resistance to attacks. • The training dataset is augmented with diverse and challenging examples to expose the model to a wider range of scenarios. • Predictions and decisions from multiple models are combined to reduce the impact of errors and increase robustness. • Features are carefully selected or engineered to make the model more resilient to variations and adversarial input. • The system uses data pre-processing techniques to remove noise and irrelevant information that might make the model more susceptible to adversarial inputs. • The system uses loss functions that are less sensitive to adversarial inputs, such as robust variants of

	cross-entropy loss. • The system employs mechanisms that detect when the input data is out of the model's training distribution, which mitigates the impact of adversarial inputs. • The system employs explainability to gain insights into model decisions and identify potential issues or adversarial attacks. • The system is subject to security audits to identify vulnerabilities and potential attack vectors. • The system monitors AI system behaviour and performance towards responding to any issues or adversarial attacks nearly in real time. • The system is deployed in a secure environment, and access to the model and data is restricted. • The system encrypts data at rest and in transit using strong encryption algorithms. • The system implements robust access controls to restrict who can access the AI system and what actions they can perform. • The system employs the principle of least privilege, ensuring that users and processes have the minimum level of access necessary. • The system employs strong authentication methods, such as multi-factor authentication (MFA). • The system remains up-to-date with respect to security patches and updates. • The system is integrated with intrusion detection and prevention systems that monitor network traffic and detect and block suspicious activities. • There are regular security audits and vulnerability assessments associated with the systems and the infrastructure that supports its operation. • Firewalls and network segmentation are used to isolate the AI system from other parts of the network. • There is a comprehensive incident response plan in place that outlines how to detect, respond to, and recover from cybersecurity incidents against the AI system. • The users of the system are trained on security best practices such as how to identify and report phishing and other social engineering attacks. • The system incorporates security considerations from the early stages of its development in-line with "security by design" approaches. • There are regular security processes in place, including penetration testing, vulnerability scanning, and code reviews.
Privacy & Data Governance	• The system collects data based on the data minimization principle, i.e., it collects only the data necessary for the AI system's intended purpose. No sensitive or personal information that is not directly relevant to the operation of the AI system is collected. • Data collection is based on informed consent, i.e., personal data is collected only after obtaining informed consent from individuals to ensure that they understand how their data will be used and for what purposes. • Data collection anonymizes or pseudonymizes data whenever possible. This includes the removal or encryption of personally identifiable information (PII) to protect individual identities. • During data collection the system uses encryption techniques (e.g., SSL/TLS) when transmitting data over networks to prevent interception and eavesdropping. • The AI system ensures data quality during data collection by validating, cleaning, and sanitizing incoming data to reduce errors and inaccuracies. • Data at rest is encrypted using strong encryption methods to protect it from unauthorised access in storage. • The system implements access control policies to limit who can decrypt and access the data. • The system implements role-based access control (RBAC) and least privilege principles to restrict data access to only those who need it for their specific roles. • Data is regularly backed up, and the backup copies are encrypted and stored securely. • Data retention policies have been developed and enforced to determine how long data is stored, while no longer needed data is deleted. • The system implements robust logging and monitoring systems to track who accesses the data and what changes are made. • Access to data is monitored continuously to identify potentially malicious and/or suspicious activities. • The system classifies the various data assets based on their sensitivity and importance while applying appropriate security measures to each classification level.

	• Data access and usage are regularly audited to ensure compliance with privacy and security policies. • When sharing data, masking techniques are used to replace sensitive information with fictional or obfuscated data. • When sharing data with third parties or between systems, secure methods such as secure APIs and encrypted file transfers are used. • There are established ethical guidelines for data handling and use to ensure the behaviour of the AI system aligns with ethical principles and regulations. • Users are educated about data privacy and security best practices as part of measures to promote a culture of security within the organisation. • There is a comprehensive incident response plan to address data breaches or security incidents promptly. • The system adheres to applicable data protection regulations (i.e., GDPR) and relevant industry-specific standards, while data policies and procedures have been updated to meet compliance requirements. • There are clear and well-defined guidelines for data collection. • There are precise and consistent data annotation standards, including clear instructions for human annotators. • Data cleaning processes are in place to remove duplicate records, correct inaccuracies, and handle missing data. • Data are verified for accuracy and reliability based on proper checks that identify anomalies or errors. • The system keeps track of different versions of datasets to maintain a history of changes and updates. • There are measures for identifying and handling outliers in the data. • There are data quality metrics defined and regularly measure and monitor data quality against these metrics. • Their system employs bias mitigation measures, especially for sensitive attributes. • The system documents metadata of the various datasets, including data sources, collection methods, and any pre-processing steps. • Data retention and data disposal policies are in place to ensure efficient and secure data management. • Data is backed up regularly to prevent data loss due to accidental deletions or technical issues.
Transparency	• The system employs XAI techniques (e.g., LIME, SHAP) to interpret decision-making processes and make them more understandable to humans. • The system uses feature importance analysis to identify which factors or features the AI model relies on the most when making decisions. • The system possesses user-friendly interfaces that provide insights into the AI system's behaviour and allow users to interact with the system while understanding its decision-making process. • The system comes with auditing tools and dashboards allowing real-time AI system performance monitoring, including model accuracy and fairness metrics. • The system complies with applicable and emerging regulations, such as the GDPR, the AI Act and industry-specific standards. • The system leverages specialized XAI techniques and tools to explain complex AI models that operate as black-boxes. • The system employs feature importance analysis, i.e., it can present the importance of individual features or variables in the model's decision-making process. • The system provides explanations on a per-instance basis, which explains why the AI system makes a specific decision for a given input. • The system provides visualizations illustrating how the model processes data and arrives at conclusions. • The system provides natural language explanations. • The system provides sensitivity analysis demonstrating how input data changes affect the model's output. • The system supports counterfactual explanations that demonstrate how a slight change in input data would have led to a different result. • The system has interactive interfaces allowing users to explore and experiment with the AI system's decision-making process. • The purpose, scope, and key objectives of the AI system are properly documented. • There are visual representations of the AI system's architecture, including components, data flows, and interactions.

	• There is adequate documentation about the algorithms, models, and techniques used in the AI system. • There is documentation of all data sources used by the systems, including information about their types, formats, and how they are accessed or collected. • There is documentation about all data pre-processing steps, including data cleaning, normalisation, and feature engineering. • There is documentation about the AI model's training process, including hyperparameters, training data, and validation procedures. • There is documentation about the evaluation metrics used to assess model performance. • There is documentation for the APIs used to interact with the AI system, including input and output formats. • The documentation of the system includes external libraries, frameworks, and services used in the AI system, including information about their versions and licenses. • There is detailed documentation about how the AI system complies with relevant regulations and ethical guidelines, including the GDPR, the AI Act and the guidelines of the HLEG. • There is adequate documentation of the measures taken to protect user data and ensure data privacy (e.g., encryption and access controls). • There is version control for the system's documentation. • The documentation is accessible to all relevant stakeholders, including developers, users, and compliance officers. • The documentation includes references to external resources, research papers, and documents that influenced the AI system's design. • The system includes legal disclaimers, terms of use, and licensing information.
Diversity, Non-discrimination & Fairness	• Collection of diverse and representative training data to reduce bias during AI system training and development. • Careful annotation of data based on structured guidelines to avoid stereo types and biases. • Generation of synthetic data to increase the diversity of the datasets used for the system's training. • Augmentation of possible under represented groups or data regions towards balancing the dataset. • Conduct of subgroup analysis to identify bias against specific demographic groups. • Use of feature selection mechanism to remove potentially biased features and/or creation of new features to counteract biases. • Standardisation and normalisation of data to mitigate the influence of outliers. • Adjusting the importance of data samples or features to give more weight to underrepresented groups. • Implementation of fairness-aware machine learning algorithms (e.g., adversarial training) that consider fairness constraints during training. • Addition of fairness-related regularisation terms to the objective function to penalise biased predictions. • Analysis of the model's sensitivity to different features or groups to detect and correct bias. • Use of metrics like disparate impact, equal opportunity, and calibration to assess the fairness of AI systems. • Application of algorithms that adjust the predictions or decisions post-training to reduce bias. • Specification and use of classification thresholds to achieve fairness (e.g., equal false-positive rates for different groups). • Bias auditing by external organisations or experts. • Support for explanations for decisions to allow for external scrutiny. • Collection of user feedback to identify and address bias in AI systems. • Promotion of diversity in AI development teams to reduce the risk of unintentional bias. • Education and training about bias, fairness, and ethics to AI developers and other stakeholders. • Model development is driven by clear and measurable fairness metrics, such as equal opportunity, demographic parity, and predictive parity. • The system has incorporated fairness constraints during model training to ensure that the model's output adheres to fairness objectives. • The system implements fairness-aware machine learning algorithms that reduce disparate impact and enhance fairness in AI decisions. • The system's model(s) are trained using adversarial networks to make them resistant to adversarial attacks and to improve fairness. • Human reviewers and subject matter experts engage in the model development and evaluation processes.

	• AI system outputs are continually monitored for fairness, and corrective actions are taken if needed. • There is diversity in AI development teams to bring a wide range of perspectives and reduce the risk of unintentional bias. • There are mechanisms for users to report and provide feedback on potential fairness issues.
Environmental & Societal Well-being	• The system has been developed in line with a set of ethical AI development principles that align with retail industry standards. • The system has been developed and deployed in line with a comprehensive code of conduct that outlines the organisation's ethical principles for AI development. • The development and operation of the system are overseen in terms of ethical and responsible AI principles by an AI ethics committee or advisory board. • The system is developed, deployed and operated in ways that are up-to-date with relevant laws and regulations governing AI and retail. • End-users are educated and trained on AI ethics, privacy, and responsible use of the AI system. • Any AI components and technologies used in the system meet ethical standards, including labour practices and environmental responsibility. • The system's behaviour and performance are regularly monitored to detect and rectify ethical issues. • There are mechanisms for individuals to report ethical concerns and violations related to AI systems without fear of retaliation. • The development and operation of the system consider environmental ethics in AI strategies towards reducing the environmental impact of AI.
Accountability	• The system maintains audit trails i.e., detailed records of its activities, including data inputs, model parameters, and decision outputs. • The system supports model versioning i.e., it keeps track of different versions of AI models, along with the changes made to each version. • The system supports algorithm logging, i.e., it logs the specific algorithms and techniques used in the AI system. • The system offers data provenance and traceability functionalities by documenting the origin and history of data assets, including their sources, transformations, and any pre-processing. • The system records the explanations and interpretations generated by the AI system for specific decisions, including information about the reasons behind the system's choices. • The system enables associating every action or decision made by the AI system with a timestamp, allowing for temporal tracking and analysis. • The system maintains user interaction logs, i.e., records of interactions between users or operators and the AI system. • The system maintains error and exception logs that record cases where the AI system diverges from expected behaviour. • The system documents the process of training data annotation, including the actors involved, the annotations provided, and any guidelines provided to human annotators. • The system keeps track of feedback and correction logs, i.e. feedback from users or experts, and documents corrective actions taken to address the received feedback. • The system comes with model validation reports i.e., records of model validation processes such as testing, validation datasets, and evaluation metrics used to assess model performance. • The system offers security incident reports that provide information about security incidents, breaches, or attempts to compromise the AI system's integrity, along with responses and mitigation efforts. • The system supports change management processes, which ensure that any changes made to the AI system's configuration, code, or parameters, along with the rationale for these changes. • There are training and certification records for the personnel involved in AI system development and operation, including information on their roles and responsibilities.

Assessed by: B.S.D. S.r.l.