Applied Human Factors and Ergonomics (AHFE2025), Vol. 199, 2025, 1243–1253
https://doi.org/10.54941/ahfe1006935

Human-Centered Artificial Intelligence for Automotive Systems: Towards Explainable, Intercultural, and Standardized Integration

Rüdiger Heimgärtner

Intercultural User Interface Consulting, Undorf, Bayern, Germany

ABSTRACT

The integration of intelligent systems into socio-technical environments requires not only technical excellence but also systematic consideration of human, cultural, and organizational factors. This paper proposes a Human-Centered Artificial Intelligence (HCAI) framework tailored to the automotive domain, bridging the gap between international standards (e.g., Automotive SPICE, ISO 26262, ISO 9241-221) and the practical deployment of AI-enabled systems. The approach is based on three complementary dimensions: (1) Explainability and Transparency, ensuring that AI-supported decision-making processes are comprehensible to engineers, managers, and auditors; (2) Intercultural Design Integration, incorporating cultural user interface design principles to enhance acceptance in global development teams; and (3) Standardized Assessment, leveraging process models such as Automotive SPICE PAM 4.0 to establish consistent, auditable practices. The research employs context-augmented generation (CAG) techniques with local large language models to assess AI outputs against normative requirements. A multi-agent framework supports evidence extraction, classification, and compliance checking. We introduce Explainable Document Labelling (EDL) to enhance transparency through structured annotations of assessment outputs. Evaluation through industry presentations at the VDA Automotive SYS 2025 Conference, structured demonstrations with automotive suppliers, and systematic reproducibility tests demonstrate that this approach generates standardized outputs with a consistency high enough to work with, addressing one of the major challenges in AI-assisted assessments. Beyond automotive, the findings contribute to understanding how people and intelligent systems can work together effectively in safety-critical industries, illustrating how HCAI principles, intercultural design, and standardized process assessment can jointly advance the reliability, acceptance, and sustainability of intelligent human systems integration.

Keywords: Human-centered AI, Intercultural design, Automotive SPICE, Explainability, Human-autonomy teaming, Standardization, Context-augmented generation, Explainable document labelling (EDL)

INTRODUCTION

The automotive industry faces a fundamental challenge: how to integrate AI into safety-critical systems while maintaining the rigorous standards that protect human life. In our work with automotive development teams across Germany, China, and other countries, we have observed that this challenge extends far beyond technical implementation. It encompasses human factors, cultural differences, and the need to align AI practices with well-established automotive standards. These problems do not stem from technical failures but from a fundamental mismatch between how AI systems operate and how people in different roles and cultures need to interact with them. Current automotive standards like ISO 26262 for functional safety and Automotive SPICE for process capability (VDA, 2023) provide robust frameworks for conventional software. However, these standards were developed before the current wave of generative AI adoption. AI-specific characteristics such as probabilistic outputs, data-driven learning, and the "black box" nature of many machine learning models make it difficult to transparently apply and fulfil these standards with crystal clear reasoning and sound argumentation (Koopman and Wagner, 2017). Furthermore, the globalized nature of automotive development introduces intercultural dimensions that can lead to divergent interpretations of AI system behavior and quality criteria (Heimgärtner, 2019). This paper addresses these challenges through a Human-Centered AI framework specifically designed for automotive systems. Our approach integrates three complementary dimensions that we identified as critical through our work with international automotive teams: explainability and transparency, intercultural design integration, and standardized assessment methodologies. We demonstrate the practical viability of this framework through a demonstrator application evaluated by industry experts and Automotive SPICE assessors.

PROBLEMS IN AI DEPLOYMENT CAUSED BY CULTURAL AND ORGANIZATIONAL DIFFERENCES

Much more than technical configuration must be considered when deploying AI systems in global automotive organizations. Successful AI integration encompasses significantly more than merely implementing algorithms or training models. It requires understanding different mentalities, thought patterns, and decision-making processes that are anchored in organizational and national culture. Moreover, we must know exactly what users need and want from AI systems in different cultural contexts. This knowledge can be determined most precisely through communication-based methods. Therefore, we must systematically address comprehension problems to improve cooperation across cultural boundaries, embracing intercultural HCAI (IHCAI) principles (Heimgärtner, 2025; Xu, 2025).

SOLVING AI DEPLOYMENT PROBLEMS THROUGH HUMAN-CENTERED DESIGN

Our framework addresses AI deployment challenges through three interconnected dimensions. Each dimension targets specific aspects of the socio-technical system while maintaining alignment with established

automotive practices. We developed these dimensions based on our experience implementing AI systems in automotive organizations and analyzing where deployments succeeded or failed. Figure 1 illustrates the three-dimensional framework architecture.

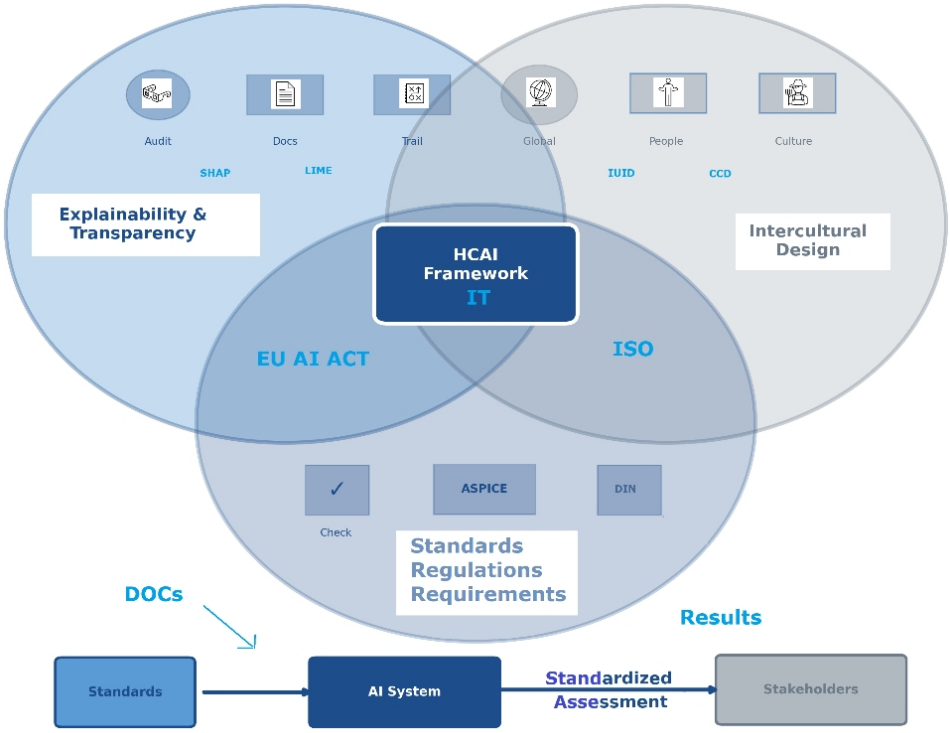

Figure 1: Three-dimensional HCAI framework showing the integration of explainability & transparency, intercultural design, and standardized assessment dimensions.

Dimension 1: Explainability Requires Understanding Different Stakeholder Needs

Successful AI deployment requires stakeholder-specific explainability. System engineers need detailed model architecture and metrics to validate and verify AI components. Managers need higher-level explanations focused on capability assessment, risk identification, and resource implications. Regulatory auditors demand standardized documentation that maps AI characteristics to specific requirements in ISO 26262 and Automotive SPICE. We employ context-augmented generation techniques to achieve this multi-stakeholder explainability. By maintaining traceability between detailed technical artifacts and stakeholder-specific explanations, we ensure consistency while improving communication efficiency. The explainability layer also incorporates provenance tracking, documenting the lineage of training data, model versions, and configuration parameters to support safety argumentation.

Dimension 2: Intercultural Design Goes Beyond Translation

Intercultural design for AI systems addresses cultural factors affecting both user interfaces and development team collaboration. At the interface level, we apply intercultural user interface design principles to adapt information presentation, interaction patterns, and feedback mechanisms to cultural preferences. This includes not just obvious elements like language and visual design, but also subtle factors like communication style, decision-making patterns, and trust-building mechanisms. Globally distributed teams bring diverse cultural perspectives to quality assessment and decision-making. Our framework makes cultural assumptions explicit and provides mechanisms for negotiating shared interpretations across teams.

Dimension 3: Standardized Assessment Builds on Automotive SPICE

Rather than creating new assessment approaches, we extend Automotive SPICE PAM 4.0 to accommodate AI-specific characteristics, leveraging existing organizational knowledge. We define AI-specific process attributes addressing machine learning lifecycle activities (data management, model training, performance monitoring, validation) with adapted capability indicators, evidence requirements, and assessment criteria (cf. VDA, 2023: Automotive SPICE PAM 4.0, MLE.1-4, SUP.11). Our assessment methodology employs a multi-agent framework where specialized agents extract evidence from development artifacts (extract, load and transform), classify evidence according to process requirements (machine learning), and check compliance against standard specifications (generative AI) – which can be stripped down to a static and simple deterministic AI workflow which is often more reliable and effective and will do the work more efficiently than a dynamic complex multi-agent system. In any case, an automated approach improves consistency and reduces subjective interpretation variations – challenges we observed repeatedly in manual assessments.

THE DEMONSTRATOR: PRACTICAL IMPLEMENTATION

To validate our framework, we developed a demonstrator application that implements all three framework dimensions in an integrated assessment environment. The system architecture employs local large language models with extended context windows (up to 128k tokens) to process comprehensive development artifacts including requirements specifications, design documents, test protocols, and safety analyses. We chose local deployment rather than cloud-based AI to address data security concerns that are paramount in automotive development. Figure 2 illustrates the system architecture.

How the Multi-Agent System Works

The demonstrator implements specialized agents that perform distinct assessment functions, working together to analyze process compliance. The Evidence Extraction Agent analyses development artifacts to identify relevant evidence items. We trained this agent to recognize not just explicit statements

but also implicit evidence – for example, inferring process execution from document timestamps and author information. The Classification Agent maps extracted evidence to specific Automotive SPICE PAM 4.0 process requirements. This agent employs domain knowledge we encoded from expert assessors, including rules about evidence sufficiency and quality. For instance, the agent knows that a single test report may provide weak evidence for comprehensive testing, while multiple test reports with traceability to requirements provide strong evidence. The Compliance Checking Agent evaluates whether available evidence satisfies capability level criteria. This agent generates structured assessment reports with supporting rationale, making its reasoning process transparent. Critically, the agent identifies borderline cases where human judgment is needed rather than making autonomous decisions in ambiguous situations. We enhance large language model capabilities through context-augmented generation (CAG). The system maintains a structured knowledge base containing process definitions, capability indicators, and evidence examples extracted from ISO 26262, Automotive SPICE, and ISO 9241-221. When performing assessments, the system retrieves relevant context and incorporates it into agent prompts. This context injection proved crucial for assessment quality. In early trials without CAG, the LLM generated plausible sounding but sometimes incorrect ratings and outcomes. With CAG, the system grounds its outputs in authoritative reference materials, dramatically improving accuracy and consistency. The context also enables the system to explain its reasoning by citing specific standard clauses and evidence patterns.

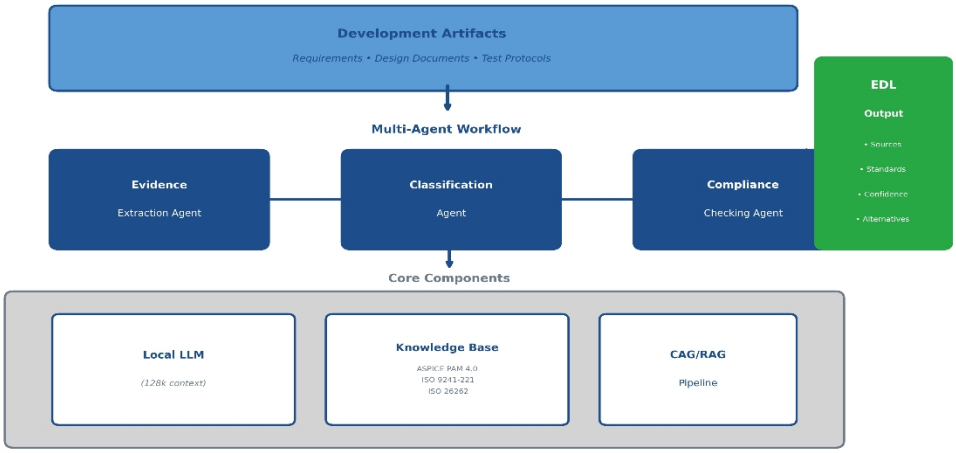

Figure 2: System architecture showing the workflow-based design with local LLM, CAG/RAG pipeline, knowledge base, and multi-agent framework.

Making AI Decisions Transparent Through Explainable Document Labelling (EDL)

We introduce Explainable Document Labelling (EDL), annotating AI-generated assessment outputs with structured metadata documenting reasoning and evidence. While related concepts exist in explainable AI

research, including local explanations for model predictions (Ribeiro et al., 2016; Lundberg and Lee, 2017) and data provenance tracking for machine learning pipelines (Moreau et al., 2008; Samuel et al., 2020), EDL specifically addresses the distinct challenge of making process assessment conclusions auditable and verifiable in safety-critical automotive contexts. Traditional XAI approaches focus primarily on explaining individual model predictions or feature importance, typically at the training or inference stage. In contrast, EDL operates at the document output level, annotating complete assessment statements with contextual evidence that enables human assessors and auditors to verify conclusions. This distinction is crucial in automotive process assessment, where regulatory compliance requires not just model interpretability but complete traceability from normative requirements through evidence evaluation to capability determinations. Specifically, EDL annotates each generated assessment statement with four types of structured metadata:

- Source Evidence References: Direct citations to specific development artifacts, document sections, and evidence items that support the assessment conclusion. For example, "[Project Plan v2.3, Section 4.2; Meeting Minutes 2025-03-15; Review Protocol QA-123]". This provenance information enables auditors to trace assessment rationale back to original documentation.
- Standard Clause Mapping: Explicit linkage to applicable clauses in automotive standards such as ASPICE PAM 4.0 base practices and capability indicators, or ISO 26262 safety requirements. For example, "[ASPICE PAM 4.0, MAN.3.BP1, MAN.3.BP5; ISO 26262-6, Clause 8.4.2]". This mapping ensures traceability to normative requirements and facilitates compliance verification.
- Confidence Indicators: Quantitative confidence scores (0–100%) reflecting the degree of certainty in the assessment, derived from evidence completeness, consistency, and alignment with capability indicators. These scores help assessors prioritize cases requiring deeper review and maintain appropriate epistemic humility about AI-generated conclusions.
- Alternative Interpretations: Documentation of borderline cases, ambiguities, or alternative rating conclusions that warrant human review. For example, "Evidence suggests borderline achievement between Capability Level 1 and Level 2; continuous application of process documented only partially." This explicit acknowledgment of uncertainty maintains human agency in final determinations.

Figure 3 demonstrates an example of EDL annotations applied to a capability assessment statement, showing how source evidence, standard mappings, confidence indicators, and alternative interpretations combine to create a transparent, verifiable assessment record.

The EDL approach draws conceptually on provenance research in scientific workflows (Moreau et al., 2008) and recent work on establishing data provenance for responsible AI systems (Arnold et al., 2019). However, while provenance typically tracks data lineage through computational

pipelines, EDL focuses on documenting the reasoning chain from normative requirements through evidence to assessment conclusions. This shift in focus reflects the distinct requirements of process assessment, where the challenge is not primarily algorithmic transparency but rather establishing auditable connections between standards, evidence, and capability determinations. This annotation approach transforms opaque AI outputs into transparent, auditable documentation, enabling verification and maintaining accountability. EDL principles apply across standards including e.g., ISO 9001, ISO 26262, ISO 27001/21434, ISO 9241-221, ISO 42001, EU AI Act, and GDPR.

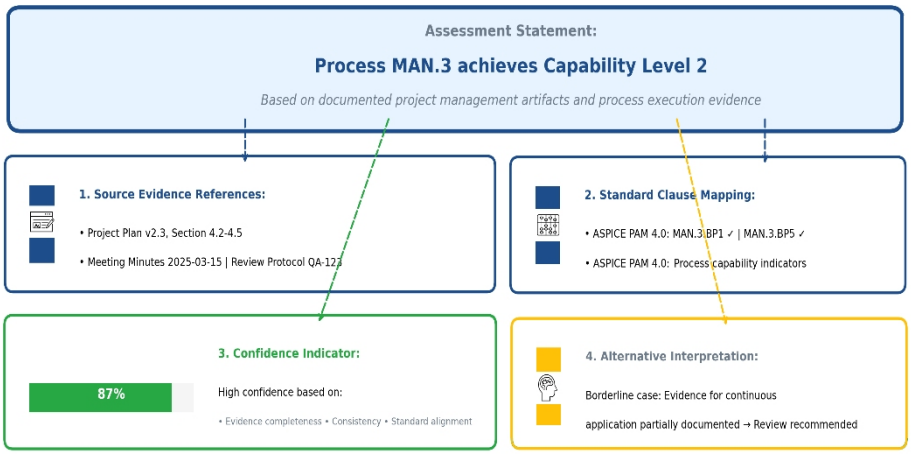

Figure 3: Example of explainable document labelling (EDL) showing structured annotations including source evidence references, standard clause mappings, confidence indicators, and alternative interpretations for an ASPICE process capability assessment.

EVALUATION BY INDUSTRY EXPERTS AND ASSESSORS

To evaluate the framework's practical applicability and validate the approach, we conducted a multi-faceted assessment involving industry experts and ASPICE assessors. The evaluation strategy combined expert evaluations through conference presentations, structured demonstrations with automotive supplier assessment teams, and technical reproducibility tests. This comprehensive approach enabled us to assess both the practical acceptance by domain experts and the technical performance of the system in generating standardized, reproducible outputs – addressing one of the major challenges in current AI applications for process assessments.

Conference Presentation and Expert Feedback

We presented the demonstrator at the VDA Automotive SYS Conference (June 25–27, 2025, Berlin), where ASPICE assessors, quality managers, and engineers from leading OEMs/suppliers experienced live demonstrations.

Reception was notably positive, with participants highlighting EDL's transparency advantages and expressing concrete deployment interest.

Structured Assessment With Automotive Supplier Experts

We conducted an extensive structured demonstration and assessment session with a team of ASPICE assessment experts from a leading Tier-1 automotive supplier. This expert assessment involved live processing of real SPICE assessment artifacts and a structured feedback session, providing valuable insights into practical requirements from experienced assessors.

The assessors confirmed fundamental system capability and valuable SPICE assessment support, viewing AI-assisted support for this complex process as significant progress. They identified requirements for enhanced precision through optimized aggregation strategies and organization-specific evaluation schemes. This structured expert assessment demonstrates both the principal suitability of our approach for real automotive assessment environments and defines concrete development steps required for productive deployment.

Technical Reproducibility Assessment

As an ASPICE Principal Assessor, I conducted systematic technical evaluations of the system's reproducibility. I processed identical development artifacts multiple times to measure output consistency, addressing one of the major challenges in AI-assisted assessment: ensuring that the same inputs produce consistent outputs. The results showed high reproducibility, with assessment conclusions remaining stable across repeated evaluations. The CAG approach contributed significantly to this consistency by providing standardized context that constrained model outputs within acceptable boundaries. Variations in generated text primarily reflected stylistic differences rather than substantive assessment disagreements. For example, the system might describe the same evidence gap as "insufficient test coverage" in one run and "inadequate testing documentation" in another, but the underlying capability determination remained consistent. Table 1 summarizes the evaluation metrics.

Table 1: Evaluation metrics and results.

Metric	Result
Reproducibility (Consistency)	>95% (<5% mostly stylistic variance)
Expert Assessment Alignment	~85%
ASPICE Process Coverage	All PAM 4.0 processes
Test Document Sources	Real assessor-created + generated
Expert Evaluators	5 Automotive SPICE Assessors
Conference Presentations	VDA Automotive SYS 2025
Supplier Demonstrations	1 Tier-1 supplier (structured feedback)

Discussion of Evaluation Results

Expert evaluations, structured assessments, and reproducibility tests provide initial validation in automotive contexts. Positive reception and deployment interest underscore practical relevance. Technical tests demonstrate successful standardized output generation with improved reproducibility through deterministic configuration and CAG, provides the foundation for auditable, reliable AI-assisted assessments meeting safety-critical requirements. This evaluation primarily demonstrates principal feasibility, expert acceptance, and technical reproducibility. Comprehensive empirical evaluation with controlled comparative studies across diverse contexts remains an important next step. Current results provide strong proof-of-concept validation.

LIMITATIONS AND FUTURE WORK

Our work is subject to several limitations that define directions for future research:

- Empirical Validation: While the expert evaluations and conference presentations showed positive reception, a systematic empirical evaluation with controlled comparative studies is still outstanding. Future work should conduct comprehensive studies that compare AI-assisted with manual assessments under controlled conditions, involving larger numbers of assessors and diverse project contexts.
- Intercultural Validation: The framework was primarily conceptualized based on experiences with German and Chinese development teams. Validation with teams from additional cultural contexts is needed to confirm the generalizability of our intercultural design approaches and identify potential adaptations required for other cultural combinations.
- Longitudinal Studies: The evaluations conducted to date occurred within short time frames. Longitudinal studies are necessary to evaluate the acceptance and effectiveness of the system in productive use over multiple assessment cycles, including analysis of how organizations adapt their processes around AI-assisted assessments.
- Technical Enhancement and Integration: Future development should address precision improvement (chunk aggregation, organization-specific assessment schemes, ambiguity handling) and tighter integration with development environments (automated import from other tools, consistency checking, GDPR-compliant processing).
- Human-in-the-Loop Mechanisms: While the current system identifies borderline cases requiring human judgment, more sophisticated mechanisms are needed to maintain accountability while leveraging AI capabilities. This includes better support for human decision-making in ambiguous situations and clear delineation of system versus human responsibility.
- Broader Applicability: The framework can extend beyond process assessment to requirements analysis, design validation, and safety

argumentation in automotive, and to other safety-critical industries (aerospace, medical devices, industrial automation) facing similar challenges of balancing AI capabilities with human values and regulatory requirements.

CONCLUSION

Through developing and evaluating this Human-Centered AI Framework for Automotive Systems, we have demonstrated that successful AI deployment in automotive systems requires careful integration of explainability, intercultural design, and standardized assessment. These three dimensions address the multifaceted challenges of deploying AI in safety-critical, globally distributed development environments. Our demonstration implementation and expert evaluations validate the practical feasibility of this approach. The presentations at VDA Automotive SYS 2025 Conference and structured evaluations with automotive suppliers confirm industry interest and identify concrete requirements for productive deployment. Technical evaluations demonstrate that AI-assisted process assessment can achieve high reproducibility consistency, addressing a critical challenge in AI-based systems. The context-augmented generation approach effectively generates standardized, traceable outputs supporting regulatory compliance, while intercultural design principles enhance global applicability. Most importantly, this work demonstrates that human-centered design principles apply not just to end-user interfaces but fundamentally to how we integrate AI into complex organizational processes. As AI continues to permeate safety-critical systems, frameworks that balance automation capabilities with human centrality, cultural awareness, and regulatory alignment will become increasingly essential. The question is not whether to use AI in automotive systems but how to use it in ways that genuinely serve human needs across diverse cultural contexts while maintaining the rigorous safety standards our industry depends upon.

ACKNOWLEDGMENT

I acknowledge the valuable feedback provided by automotive industry practitioners at the VDA Automotive SYS 2025 Conference and during supplier demonstrations, as well as the ongoing support of research collaborators in developing and validating this framework.

REFERENCES

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilović, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J., Tsay, J., and Varshney, K. R. (2019). FactSheets: Increasing trust in AI services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5), pp. 6:1–6:13.

Heimgärtner, R. (2019). Intercultural User Interface Design. Springer International Publishing.

Heimgärtner, R. (2025). *Intercultural Design for Human-Centered AI Solutions*. In: Xu, Wei (Ed.) (2025). Handbook of Human-Centred Artificial Intelligence. Springer International Publishing. (In Press).

ISO (2018). ISO 26262: Road vehicles–Functional safety. International Organization for Standardization.

ISO (2023). ISO 9241–221: Ergonomics of human-system interaction - Part 221: Human-centred design process assessment model. International Organization for Standardization.

Koopman, P. and Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. IEEE Intelligent Transportation Systems Magazine, 9(1), pp. 90–96.

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, pp. 4765–4774.

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and Van den Bussche, J. (2008). The Open Provenance Model: An overview. In: Proceedings of the International Provenance and Annotation Workshop (IPAW 2008), Lecture Notes in Computer Science, vol. 5272, Springer, pp. 323–326.

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, pp. 1135–1144.

Samuel, S., Löffler, F., and König-Ries, B. (2020). Machine learning pipelines: Provenance, reproducibility and FAIR data principles. In: Proceedings of the 2021 International Provenance and Annotation Workshop, Charlotte, NC. arXiv preprint arXiv:2006.12117.

VDA (2023). Automotive SPICE Process Assessment Model (PAM) 4.0. Verband der Automobilindustrie.

Xu, W. (2025). A User Experience 3.0 (UX 3.0) Paradigm Framework: Designing for Human-Centered AI Experiences. *arXiv preprint arXiv:2506.23116*.

Xu, W. (2025). Handbook of Human-Centred Artificial Intelligence. Springer International Publishing. (In Press).