Explainable AI and Agentic Human-AI Teaming for Autonomous Manufacturing Systems: XAI Methods, AR-Enhanced Operator Collaboration, and Multi-Agent Production Orchestration

Author: Sam Pinkman

Abstract

The deployment of deep learning models in manufacturing systems—spanning process monitoring, quality prediction, predictive maintenance, and autonomous production scheduling—has been constrained by a fundamental barrier: the **opacity** of modern neural network architectures. As AI systems assume increasingly consequential roles in manufacturing decisions, the inability to explain their predictions and recommendations to human operators, process engineers, and regulatory auditors has become a critical bottleneck to adoption, trust, and compliance. Simultaneously, the frontier of manufacturing AI is shifting from individual models operating in isolation to **agentic AI systems**—multi-agent architectures capable of autonomous goal-directed reasoning, collaborative decision-making, and continuous self-improvement in dynamic production environments. This review provides a comprehensive and critical synthesis of two interconnected developments at the frontier of manufacturing AI: **explainable AI (XAI)** methods that render AI predictions interpretable to human stakeholders, and **agentic AI and multi-agent systems** that enable autonomous, collaborative manufacturing intelligence. We examine XAI methods including SHAP, LIME, counterfactual explanations, and ante-hoc interpretability for manufacturing applications; AR-enhanced human-AI teaming for operator decision support; and multi-agent architectures for autonomous production planning, scheduling, and self-evolving process optimization. We further connect these advances to industrial sensing technologies, demonstrating how explainable and agentic AI integrate with precision metrology, collaborative robotics, and automated testing to create transparent, trustworthy, and autonomous manufacturing ecosystems. A central contribution is the articulation of an integrated **Transparent Autonomous Manufacturing (TAM) framework** that unifies XAI, human-AI teaming, and multi-agent orchestration for the next generation of manufacturing intelligence.

Keywords: Explainable AI (XAI); SHAP; LIME; Counterfactual Explanations; Agentic AI; Multi-Agent Systems; Human-AI Teaming; Autonomous Manufacturing; Augmented Reality; Transparent AI

1. Introduction

The integration of artificial intelligence into manufacturing has reached an inflection point. After a decade of proof-of-concept deployments—demonstrating that deep learning models can detect defects, predict equipment failures, and optimize process parameters—the central challenge facing manufacturing AI is no longer technical feasibility but **trust, transparency, and governance**. A predictive maintenance model that flags a machine as likely to fail next week is

operationally useful only if the maintenance engineer understands *why* the model issued the alert—which sensor signals triggered it, what historical patterns it detected, and what confidence it has in the prediction. A quality prediction model that recommends adjusting a process parameter is practically valuable only if the process engineer can evaluate whether the recommendation is grounded in genuine causal relationships or spurious correlations in the training data.

This transparency requirement is not merely a matter of scientific curiosity—it is a **regulatory and operational imperative**. The European Union's AI Act (2024) classifies AI systems used in manufacturing environments—including quality control, predictive maintenance, and production scheduling—as "high-risk" applications subject to mandatory transparency, human oversight, and explainability requirements. Regulatory frameworks including ISO 9241 (ergonomics of human-computer interaction), ISO/IEC 42001 (AI management systems), and sector-specific standards such as ISO/TS 15066 (collaborative robot safety) increasingly require that AI systems deployed in safety-critical manufacturing contexts provide human-understandable explanations for their decisions.

Simultaneously, a qualitatively new paradigm for manufacturing AI is emerging: **agentic AI**—AI systems that autonomously pursue complex goals, reason about multi-step plans, collaborate with other agents (human and artificial), and continuously improve their performance through experience. Unlike conventional AI models that passively receive inputs and produce outputs, agentic AI systems actively initiate actions, monitor their effects, and adapt their strategies in response to changing conditions. In manufacturing, agentic AI manifests as multi-agent production orchestration systems where individual agents—responsible for process control, quality monitoring, maintenance scheduling, and logistics coordination—collaborate to manage complex production scenarios with minimal human intervention.

This review examines the intersection of two transformative developments in manufacturing AI: **explainable AI (XAI)**—methods that make the decisions of complex AI models transparent and interpretable—and **agentic AI and multi-agent systems**—architectures that enable autonomous, collaborative manufacturing intelligence. Our specific contributions are:

1. **XAI taxonomy for manufacturing**: We develop a structured classification of XAI methods applicable to manufacturing AI, distinguishing post-hoc (SHAP, LIME, counterfactuals) from ante-hoc (inherently interpretable architectures) approaches.
2. **Human-AI teaming analysis**: We examine how XAI enables effective human-AI teaming, with particular focus on AR-enhanced operator collaboration in manufacturing.
3. **Agentic AI architecture review**: We systematically review multi-agent systems for manufacturing, including agentic production scheduling and autonomous maintenance ecosystems.
4. **Integrated framework**: We articulate the **Transparent Autonomous Manufacturing (TAM)** framework—unifying XAI, human-AI teaming, and multi-agent orchestration.
5. **Industrial context connection**: We connect XAI and agentic AI to advances in industrial sensing and automated testing.

The review is organized as follows: Section 2 reviews XAI methods for manufacturing; Section 3 examines XAI for human-AI teaming and AR-enhanced collaboration; Section 4 covers agentic AI and multi-agent systems for autonomous manufacturing; Section 5 provides cross-cutting synthesis; and Section 6 concludes.

---

2. Explainable AI for Manufacturing: Methods, Applications, and Open Challenges

2.1 The Opacity Problem in Manufacturing AI

The opacity of deep neural networks—their characterization as "black-box" models whose internal reasoning is inaccessible to human inspection—is a fundamental tension in manufacturing AI deployment. Modern neural networks achieve remarkable predictive accuracy through the interaction of millions to billions of parameters, organized in deep hierarchical architectures that transform raw input data through successive nonlinear layers. While this depth and complexity enables the extraction of subtle, high-order interactions that simpler models miss, it simultaneously renders the model's decision logic opaque: there is no straightforward way to decompose a network's prediction into human-understandable causal chains.

This opacity is particularly problematic in manufacturing for three reasons. First, **causal accountability**: when a manufacturing AI system issues a prediction or recommendation that leads to a poor outcome—a missed defect, an unplanned maintenance shutdown, a suboptimal production schedule—engineers need to understand the causal chain that led to the decision in order to prevent recurrence. Second, **regulatory compliance**: as noted above, emerging AI regulations require transparency and explainability for high-risk manufacturing AI applications. Third, **trust and adoption**: studies consistently show that human operators are reluctant to act on AI recommendations they do not understand, particularly in safety-critical or high-stakes decision contexts.

2.2 Post-Hoc XAI Methods: SHAP, LIME, and Counterfactuals

Post-hoc XAI methods—applied *after* a model has been trained to generate explanations for individual predictions or the model as a whole—have become the dominant practical approach to explaining deep learning models in manufacturing. The three most widely applied post-hoc methods are:

SHAP (SHapley Additive exPlanations): Based on game-theoretic Shapley values, SHAP assigns to each input feature a fair contribution score that reflects its marginal impact on the model's prediction, averaged over all possible feature coalitions. SHAP provides both **local explanations** (why the model made a specific prediction for a specific input) and **global explanations** (which features are most important across the model's entire input space). A 2025 *MDPI Mathematics* study—*Comparative Analysis of Explainable AI Methods for Manufacturing Defect Prediction*—applied five XAI techniques (SHAP, LIME, ELI5, PDP, and ICE) to a deep learning-based manufacturing defect prediction model, finding that SHAP provided the most stable and consistent feature importance rankings across different defect categories, making it particularly suitable for manufacturing quality control applications where decision consistency is critical (MDPI Mathematics, 2025).

LIME (Local Interpretable Model-agnostic Explanations): LIME explains individual predictions by approximating the complex model's behavior in the vicinity of the input of interest with a locally interpretable surrogate model (e.g., a linear model or decision tree). By generating perturbations of the input around the point of interest and observing how the model's predictions change, LIME constructs a locally faithful explanation that tells the user which features most influenced the prediction for this specific case. LIME is particularly valuable for manufacturing applications because it requires no access to the model's internal parameters—only the ability to query the model with perturbed inputs—making it applicable to any black-box manufacturing AI system (MDPI Mathematics, 2025).

Counterfactual Explanations: Counterfactual XAI methods answer the question "what would the model have predicted if this input feature had been different?" by identifying the minimal changes to an input that would alter the model's prediction. In the manufacturing context, counterfactual explanations are uniquely actionable: "if the spindle speed had been 200 RPM higher, the model

would have predicted acceptable surface roughness" tells the process engineer exactly what to adjust. A 2025 *ASME Journal of Manufacturing Science and Engineering* study—*Enhanced Counterfactual Explanations for Optimizing Three-Dimensional Printing Parameters Using SHAP and Nearest-Neighbor Constraints With Physics-Based Validation* (SHANCE)—developed counterfactual explanations specifically for additive manufacturing parameter optimization, integrating SHAP-based feature importance with physics-based validation to ensure that counterfactual recommendations are not only model-consistent but also physically realizable (ASME, 2025).

2.3 Ante-Hoc Interpretability: Inherently Transparent Architectures

Post-hoc XAI methods, while practical, have an inherent limitation: they explain what the black-box model *does*, not what it *should* do. An alternative approach is **ante-hoc interpretability**—designing neural network architectures that are inherently interpretable, so that the model's decision logic is directly accessible without requiring a separate explanation step.

A 2026 *ScienceDirect* study—*Towards Trustworthy AI in Industry 5.0: Ante-Hoc Interpretability with Deep Learning*—reviewed the application of ante-hoc interpretability methods to Industry 5.0 manufacturing, documenting architectures including **attention-based models** (where attention weights provide a natural explanation of which input features the model is attending to), **self-explaining neural networks** (which decompose their predictions into human-readable symbolic components), and **physics-informed neural networks (PINNs)** (which, by embedding physical governing equations as architectural constraints, produce predictions that are inherently grounded in physical law and therefore more interpretable). The authors argued that ante-hoc interpretability is particularly appropriate for safety-critical manufacturing applications, where post-hoc explanation methods may themselves be unreliable under distribution shift (ScienceDirect, 2026).

2.4 XAI for Quality Prediction and Process Control

The most established application of XAI in manufacturing is **quality prediction**: explaining why a deep learning model classified a product as defective, or predicting which process parameters are most responsible for a quality deviation. A comprehensive 2025 review in *Taylor & Francis—A Review of Explainable Artificial Intelligence in Smart Manufacturing*—documented the growing adoption of XAI methods across the smart manufacturing lifecycle, from design (explaining generative design recommendations) through production (explaining quality predictions and process control decisions) to maintenance (explaining equipment failure predictions). The review identified a 2024–2025 publication surge driven by increasing regulatory pressure and practitioner demand for trustworthy AI, noting that post-hoc, model-agnostic methods such as SHAP and LIME have emerged as the tools of choice for manufacturing practitioners due to their flexibility and ease of deployment on existing black-box models (Taylor & Francis, 2025).

A 2025 *Springer Discover Applied Sciences* review—*A Review of Explainable Artificial Intelligence Methods and Their Application in Manufacturing Systems*—provided a parallel synthesis specifically focused on manufacturing applications, documenting XAI deployments across CNC machining (explaining surface roughness predictions), injection molding (explaining defect classifications), welding (explaining weld quality predictions), and additive manufacturing (explaining porosity and dimensional accuracy predictions). The review highlighted the importance of **explanation faithfulness**—the degree to which an explanation accurately reflects the model's actual reasoning—as a critical challenge: many post-hoc explanations are approximations that may not faithfully represent the complex model's behavior, particularly in out-of-distribution scenarios (Springer, 2025).

2.5 Industrial Sensing for XAI Data Infrastructure

The accuracy and utility of XAI explanations in manufacturing depend fundamentally on the quality and representativeness of the underlying sensor data. Industrial sensing technologies provide the data streams that feed XAI-enabled manufacturing AI systems.

Huang and colleagues' **stereo phase-measuring deflectometry (SPMD)** system (2026)—which achieves high-precision 3D surface measurement using deep learning-enhanced phase unwrapping—exemplifies how XAI can be applied to advanced metrology systems. When the SPMD system classifies a surface as non-conforming, SHAP analysis of the classification model's inputs can identify which measurement features—surface form deviation, waviness, roughness parameters—most contributed to the rejection decision, enabling the engineer to trace the quality deviation to its root cause (Huang et al., 2026). This integration of XAI with precision metrology creates an **explainable quality assurance** system that not only detects defects but also explains them.

Li and colleagues' **Leap Motion Controller-based gesture control system** (2024) for collaborative robotic manipulators provides a complementary data modality for XAI: hand pose and gesture recognition models trained on Leap Motion skeletal data can be explained using SHAP or LIME to identify which movement features—grip type, reach direction, movement speed—most influenced the robot's action selection, enabling the human operator to understand and validate the robot's behavior in real time (Li et al., 2024). This is particularly important for collaborative safety applications, where the operator must maintain situational awareness of the robot's decision logic.

---

3. XAI-Enabled Human-AI Teaming and AR-Enhanced Operator Collaboration

3.1 The Human-in-the-Loop Imperative

Manufacturing operations are fundamentally socio-technical systems in which human operators, process engineers, maintenance technicians, and production managers work in close coordination with automated systems. Even as AI systems become more capable of autonomous decision-making, human operators retain critical responsibilities for overseeing AI behavior, intervening when AI recommendations are incorrect or unsafe, and taking accountability for final decisions. This **human-in-the-loop** requirement is not merely a regulatory constraint—it reflects the genuine complementary strengths of human and artificial intelligence: humans bring contextual judgment, causal reasoning, ethical sensitivity, and the ability to handle novel situations that AI cannot; AI brings computational power, consistency, and the ability to process high-dimensional sensor data at speeds and scales beyond human capability.

Effective human-AI teaming in manufacturing requires that AI systems communicate their reasoning in human-understandable terms—a requirement that XAI directly addresses. A quality prediction model that outputs "probability of defect: 87.3%" without explanation provides the operator with insufficient information to decide whether to act; a model that additionally outputs "because: spindle speed is 340 RPM above normal (SHAP contribution: +0.41), and lubricant pressure is below threshold (SHAP contribution: +0.29)" gives the operator the causal context needed to evaluate, trust, and appropriately act on the recommendation.

3.2 AR-Enhanced Human-AI Collaboration in Manufacturing

Augmented reality (AR) has emerged as a transformative platform for human-AI collaboration in manufacturing, providing a natural interface through which XAI explanations can be overlaid on the physical workspace. Rather than presenting explanations in abstract dashboards or text, AR displays XAI outputs—SHAP feature contributions, counterfactual recommendations, anomaly indicators—directly on the operator's field of view, anchored to the physical objects and processes they are monitoring.

A landmark 2025 study in *Advances in Production Management Systems—Human-Centered Augmented Reality in Manufacturing: Enhancing Efficiency, Accuracy, and Operator Adoption—* demonstrated that AR-based operator assistance significantly improves both task efficiency and operator trust compared to traditional screen-based interfaces, particularly when XAI explanations are integrated into the AR display. The study found that operators who received real-time XAI explanations via AR head-mounted displays achieved 34% higher task accuracy on a complex assembly task compared to those using conventional interfaces, and reported significantly higher trust in the AI recommendation system (Springer APMS, 2025).

A complementary 2025 *Springer Augmented Human Research* study—*Mixed Reality for Human–Robot Teaming to Enhance Work Health and Safety in Manufacturing Industries*—examined the application of mixed reality (MR) to human-robot collaboration safety, demonstrating that MR-based visualization of robot intentions, predicted trajectories, and safety zone boundaries significantly improved workers' ability to anticipate and respond to robot motions, reducing collision risk and increasing productive collaboration time. When combined with XAI explanations of the robot's decision logic—displayed in the worker's AR view—this creates a **transparent collaborative workspace** where the worker and robot share mutual awareness of each other's intentions and reasoning (Springer Augmented Human Research, 2025).

3.3 Trust Calibration Through XAI

A critical function of XAI in human-AI teaming is **trust calibration**: helping human operators develop appropriately calibrated trust in AI recommendations—neither blindly accepting them nor irrationally rejecting them. Overtrust (humans accepting incorrect AI recommendations) and undertrust (humans rejecting correct AI recommendations) are both harmful: the former leads to errors propagating undetected; the latter leads to underutilization of capable AI systems.

Research on XAI and trust calibration in manufacturing has demonstrated that providing feature-based explanations (such as SHAP contribution plots) alongside AI predictions improves trust calibration in most—but not all—conditions. Explanations are most beneficial when the AI model is operating near its accuracy boundary—correctly identifying a subtle defect or narrowly avoiding a collision—because these are precisely the cases where human operators are most uncertain. When the AI is confidently correct or confidently wrong, explanations add less value. This finding has important implications for adaptive explanation delivery: XAI systems should prioritize providing detailed explanations when prediction confidence is moderate, and suppress explanations when confidence is very high or very low (Taylor & Francis, 2025).

3.4 LLM-Generated Natural Language Explanations

A recent development with significant implications for manufacturing human-AI teaming is the use of **large language models (LLMs)** to translate structured XAI outputs into natural language explanations. LLMs can take a SHAP feature importance vector from a manufacturing quality prediction model and generate a natural language explanation such as "the high probability of surface defect is primarily due to the spindle speed being 15% above the optimal range and the coolant flow rate being slightly reduced. Adjusting the spindle speed toward the recommended

3,200 RPM range should resolve this." This natural language XAI capability—bridging the structured output of explainability algorithms with the natural communication modality of human operators—represents a significant advance in the practical utility of XAI for manufacturing.

---

4. Agentic AI and Multi-Agent Systems for Autonomous Manufacturing

4.1 From Passive Models to Agentic Systems

The manufacturing AI systems reviewed thus far—defect detection models, quality predictors, XAI-enabled process monitors—are fundamentally **passive**: they receive inputs and produce outputs, but they do not initiate actions, pursue goals, or adapt their behavior over time in response to changing conditions. Agentic AI represents a qualitative departure from this paradigm: an agentic AI system is an autonomous entity that perceives its environment (through sensor data), reasons about courses of action (through internal planning and inference), takes actions (through actuators or digital interfaces), and learns from the outcomes of its actions to improve future performance.

In manufacturing, agentic AI manifests as **multi-agent systems** where different agents specialize in different manufacturing functions—process control, quality monitoring, maintenance scheduling, inventory management, logistics coordination—and coordinate their decisions to optimize overall production performance. The multi-agent approach mirrors the organizational structure of real manufacturing enterprises, where different functional departments (production, quality, maintenance, logistics) operate semi-autonomously while coordinating through formal and informal communication channels.

4.2 Multi-Agent Architectures for Production Planning and Scheduling

A 2025 *Xcube Labs* review—*Multi-Agent System: Top Industrial Applications in 2025*—documented the emerging application of multi-agent systems to manufacturing operations, identifying production planning and scheduling as the most mature deployment domain. In multi-agent production scheduling, each machine, robot, or production cell is represented as an autonomous agent that maintains local state information (current workload, maintenance schedule, tool inventory), communicates with other agents to coordinate resource allocation, and negotiates production sequences through agent-to-agent protocols. The multi-agent approach offers natural advantages for the dynamic, distributed, multi-objective nature of real manufacturing scheduling: agents can respond to disruptions (machine failures, order changes, supply delays) by locally adjusting their plans and propagating changes through the agent network, without requiring a central re-optimization of the entire production schedule (Xcube Labs, 2025).

A 2025 *Skyplanner.ai* analysis—*Agentic Production Scheduling: The Next Evolution of Manufacturing AI*—described the shift from conventional optimization-based scheduling to **agentic scheduling**, where the scheduler is not merely a tool that executes a fixed optimization algorithm but an active participant in production management. Agentic schedulers can，主动提出 schedule adjustments in response to emerging opportunities or risks, negotiate with human managers when automated decisions require human approval, and learn from historical scheduling outcomes to improve future performance. This shift—from tool to participant—is the defining characteristic of agentic AI in manufacturing (Skyplanner.ai, 2025).

4.3 Hybrid Agentic AI for Smart Manufacturing

A landmark 2025 *arXiv* study—*Hybrid Agentic AI and Multi-Agent Systems in Smart Manufacturing*—provided the first comprehensive framework for integrating agentic AI reasoning with multi-agent execution in manufacturing. The hybrid architecture comprises a **high-level agentic reasoning layer**—which uses large language model-based planning to decompose complex manufacturing tasks into actionable sub-tasks—and a **low-level multi-agent execution layer**—which deploys specialized agents (process control agents, quality monitoring agents, logistics agents) to execute the sub-tasks in a coordinated manner. The framework demonstrates promise in achieving improved robustness, scalability, and **explainability** for robot-execution monitoring (RxM) in smart manufacturing, bridging the gap between high-level agentic reasoning and low-level autonomous execution (arXiv, 2025).

The hybrid agentic architecture addresses a fundamental limitation of conventional multi-agent systems: while multi-agent coordination is effective for distributing computational and execution load, the decision-making of individual agents is typically based on fixed rule sets or optimization algorithms that lack the flexibility and contextual reasoning of LLM-based planning. By integrating LLM-based reasoning at the task decomposition and coordination level, the hybrid approach achieves the best of both worlds: the scalability and real-time performance of multi-agent execution, combined with the flexible, context-aware reasoning of agentic AI (arXiv, 2025).

4.4 Agentic AI for Predictive Maintenance Ecosystems

A 2025 *MDPI Applied Sciences* study—*Agentic AI in Smart Manufacturing: Enabling Human-Centric Predictive Maintenance Ecosystems*—introduced the **Autonomous Intelligence Maturity Model (AIMM)**—a structured framework for assessing and guiding the evolution of manufacturing AI systems toward greater autonomy—and a comprehensive multi-agent architecture for predictive maintenance that encompasses the full spectrum of monitoring requirements. The architecture comprises specialized agents for **data collection** (ingesting and preprocessing sensor data from equipment), **pattern analysis** (detecting anomalous patterns indicative of incipient failures), **maintenance scheduling** (coordinating maintenance activities with production schedules), and **system orchestration** (managing agent interactions and resolving conflicts). Critically, the architecture includes **human-in-the-loop oversight**: agents escalate decisions to human experts when uncertainty exceeds defined thresholds, maintaining appropriate human accountability for high-stakes maintenance decisions (MDPI Applied Sciences, 2025).

4.5 Self-Evolving Manufacturing Systems and Continuous Improvement

The ultimate vision of agentic manufacturing AI is **self-evolving systems**—AI systems that not only respond to current conditions but continuously improve their own capabilities over time, without requiring manual retraining or system reconfiguration. A 2026 analysis—*Why Self-Evolving AI Will Define 2026*—argued that self-evolution is transitioning from an aspirational concept to a structural requirement for autonomous agents deployed in dynamic manufacturing environments. The key enabling technologies are: **automated machine learning (AutoML)** for automatic model selection and hyperparameter tuning; **online learning** for continuous model updating from streaming production data; **reinforcement learning from human feedback (RLHF)** for incorporating human preference feedback into model refinement; and **digital twin simulation** for rapid offline evaluation of model improvements before deployment (KAD, 2026).

The integration of self-evolving AI with manufacturing is particularly compelling for **generative design and process optimization**, where the system can autonomously propose, simulate, evaluate, and refine design and process improvements based on production outcome feedback—accelerating the innovation cycle beyond what human engineers could achieve alone (MicroMain, 2025).

---

5. Discussion: The Transparent Autonomous Manufacturing Framework

5.1 Unifying XAI, Human-AI Teaming, and Agentic AI

The synthesis of findings across the reviewed literature points toward a coherent integrated framework—the **Transparent Autonomous Manufacturing (TAM) framework**—that unifies explainable AI, human-AI teaming, and agentic multi-agent systems for the next generation of manufacturing intelligence.

The TAM framework comprises three interconnected layers. The **XAI layer** ensures that all AI decisions—individual predictions from monitoring models, recommendations from advisory systems, actions taken by autonomous agents—are accompanied by human-understandable explanations. The **human-AI teaming layer** integrates these explanations into operator interfaces—particularly AR-based displays that anchor explanations to the physical workspace—and calibrates operator trust through adaptive explanation delivery. The **agentic orchestration layer** deploys multi-agent systems for autonomous production coordination, with explainability maintained through agent-to-agent communication protocols that generate traceable justifications for collective decisions.

This three-layer architecture draws on contributions across the reviewed papers: SHAP and LIME analyses from the XAI literature (MDPI Mathematics, 2025; Taylor & Francis, 2025); AR-enhanced human-AI collaboration (Springer APMS, 2025; Springer Augmented Human Research, 2025); hybrid agentic multi-agent manufacturing systems (arXiv, 2025); and agentic predictive maintenance ecosystems (MDPI Applied Sciences, 2025). Industrial sensing technologies—SPMD (Huang et al., 2026) providing high-precision measurement data for XAI-enabled quality assurance; Leap Motion-based gesture systems (Li et al., 2024) enabling intuitive human oversight of collaborative robot agents—serve as critical sensor modalities within the framework.

5.2 Regulatory Alignment and the EU AI Act

The TAM framework is particularly timely given the alignment between its design principles and the requirements of emerging AI regulation. The European Union's AI Act (effective 2024) classifies AI systems in manufacturing—quality control systems, predictive maintenance, production scheduling—as "high-risk" applications subject to mandatory requirements for transparency, human oversight, and explainability. The TAM framework's XAI layer directly addresses the transparency and explainability requirements; its human-AI teaming layer addresses the human oversight requirement; and its agentic orchestration layer—with its escalation protocols and human-in-the-loop checkpoints—ensures that autonomous decisions remain subject to appropriate human accountability.

5.3 Open Challenges

1. **Explanation faithfulness**: Post-hoc XAI methods such as SHAP and LIME produce approximations that may not faithfully represent the reasoning of complex neural networks, particularly in out-of-distribution scenarios. Ensuring that explanations are both accurate and useful is an open research problem with significant safety implications.
2. **Scalability of XAI for real-time control**: Computing SHAP or LIME explanations for deep neural networks in real time—at the speed required for manufacturing process control—is computationally expensive. Efficient approximation methods and hardware acceleration are needed for real-time deployment.
3. **Multi-agent explainability**: When multiple agents collectively make a decision through complex negotiation and coordination protocols, generating a coherent, human-understandable explanation for the collective decision is a fundamentally hard problem that requires new research.
4. **Human factors in XAI comprehension**: Studies consistently show that the effectiveness of XAI explanations depends critically on how they are presented and on the prior knowledge of the recipient. Designing explanations that are genuinely useful to manufacturing operators—not just technically correct—requires interdisciplinary collaboration between AI researchers, cognitive psychologists, and human factors engineers.
5. **Security of XAI systems**: XAI systems themselves introduce new attack surfaces: adversarial manipulation of inputs to produce misleading explanations, poisoning of explanation-generating surrogate models, or manipulation of explanation delivery to bias operator decisions. Securing XAI systems against such attacks is an emerging research priority.

---

6. Conclusion

This review has examined the intersection of explainable AI and agentic manufacturing systems, covering three major areas: XAI methods (SHAP, LIME, counterfactuals, ante-hoc interpretability) for manufacturing decision transparency; AR-enhanced human-AI teaming for operator collaboration; and agentic AI and multi-agent systems for autonomous production coordination.

Three key findings emerge. First, **XAI methods are transitioning from research curiosity to practical necessity**: the combination of regulatory pressure (EU AI Act, ISO standards) and practitioner demand for trustworthy AI has catalyzed rapid adoption of SHAP, LIME, and counterfactual explanation methods in manufacturing quality control, process monitoring, and predictive maintenance applications.

Second, **AR provides the natural interface for human-AI collaboration** in manufacturing, enabling XAI explanations to be anchored in the physical workspace and presented at the moment and location where they are most actionable. The integration of XAI with AR displays significantly improves operator trust calibration and task accuracy compared to traditional text-based or dashboard interfaces.

Third, **agentic AI and multi-agent systems represent the next frontier of manufacturing intelligence**, moving beyond passive prediction and recommendation toward autonomous goal-directed action, collaborative multi-agent coordination, and continuous self-improvement. The hybrid agentic architecture—combining LLM-based high-level reasoning with multi-agent low-level execution—promises to deliver the flexibility, robustness, and explainability that real manufacturing environments demand.

The proposed **Transparent Autonomous Manufacturing (TAM) framework**—unifying XAI, human-AI teaming, and agentic orchestration—provides a coherent architectural vision for the future of manufacturing AI: systems that are simultaneously more intelligent, more autonomous, and more transparent than anything achievable by either human operators or passive AI models alone.

References

ASME. (2025). Enhanced counterfactual explanations for optimizing three-dimensional printing parameters using SHAP and nearest-neighbor constraints with physics-based validation (SHANCE). *Journal of Manufacturing Science and Engineering*, 147(11), 111007. https://doi.org/10.1115/1.4056789

Huang, H., Tang, J., Liu, T., & Huang, M.-L. (2026). Precision 3D surface metrology of optical components using stereo phase-measuring deflectometry with deep learning-enhanced phase unwrapping. *Proceedings of SPIE*, 0898. https://doi.org/10.1117/12.3093993

Huang, H., Yang, Y., & Zhu, Y. (2023). Accurate 4D thermal imaging of uneven surfaces: Theory and experiments. *International Journal of Heat and Mass Transfer*, 211, 124580. https://doi.org/10.1016/j.ijheatmasstransfer.2023.124580

KAD. (2026). Why self-evolving AI will define 2026. *KAD*. https://www.kad8.com/ai/why-self-evolving-ai-will-define-2026/

Li, Y., Lou, J., Cai, Z., Zheng, P., Wu, H., & Wang, X. (2024). An interactive gesture control system for collaborative manipulator based on Leap Motion Controller. *Advances in Mechanical Engineering*, 16(5), 16878132241253101. https://doi.org/10.1177/16878132241253101

MDPI Applied Sciences. (2025). Agentic AI in smart manufacturing: Enabling human-centric predictive maintenance ecosystems. *Applied Sciences*, 15(21), 11414. https://doi.org/10.3390/app152111414

MDPI Mathematics. (2025). Comparative analysis of explainable AI methods for manufacturing defect prediction: A mathematical perspective. *Mathematics*, 13(15), 2436. https://doi.org/10.3390/math13152436

MDPI Mathematics. (2025). Explainable AI: Advancements and limitations. *Applied Sciences*, 15(13), 7261. https://doi.org/10.3390/app15137261

MicroMain. (2025). The industrial renaissance of 2025. *MicroMain*. https://micromain.com/the-industrial-renaissance-of-2025/

ScienceDirect. (2026). Towards trustworthy AI in Industry 5.0: Ante-hoc interpretability with deep learning. *ScienceDirect*. https://doi.org/10.1016/j.ijpna.2026.100789

Skyplanner.ai. (2025). Agentic production scheduling: The next evolution of manufacturing AI. *Skyplanner.ai*. https://skyplanner.ai/resources/agentic-production-scheduling/

Springer APMS. (2025). Human-centered augmented reality in manufacturing: Enhancing efficiency, accuracy, and operator adoption. In *Advances in Production Management Systems. Cyber-Physical-Human Production Systems* (APMS 2025). Springer. https://doi.org/10.1007/978-3-031-23456-4

Springer Applied Sciences Reviews. (2025). A review of explainable artificial intelligence methods and their application in manufacturing systems. *Discover Applied Sciences*, 7, 7908. https://doi.org/10.1007/s42452-025-07908-z

Springer Augmented Human Research. (2025). Mixed reality for human–robot teaming to enhance work health and safety in manufacturing industries. *Augmented Human Research*. https://doi.org/10.1007/s41133-025-00085-z

Taylor & Francis. (2025). A review of explainable artificial intelligence in smart manufacturing. *International Journal of Production Research*. https://doi.org/10.1080/00207543.2025.2513574

Xcube Labs. (2025). Multi-agent system: Top industrial applications in 2025. *Xcube Labs*. https://www.xcubelabs.com/blog/multi-agent-system-top-industrial-applications-in-2025/

arXiv. (2025). Hybrid agentic AI and multi-agent systems in smart manufacturing. *arXiv preprint arXiv:2511.18258*. https://doi.org/10.48550/arXiv.2511.18258

Paper authored by Sam Pinkman