Full AI Autonomous Governance Factory (Full-AGF): Implementation Pathway of the Universal Production Line

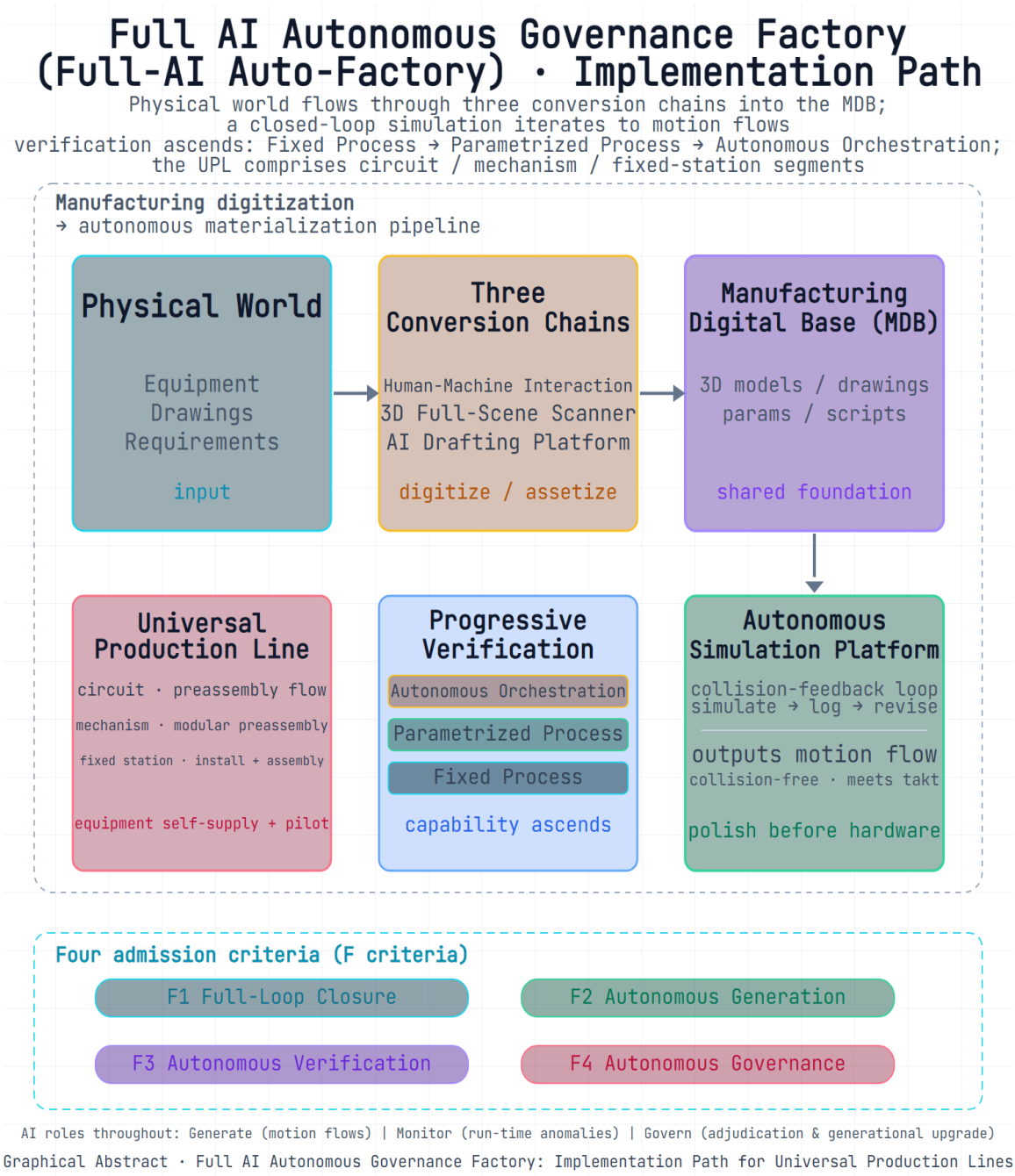
【Wei Jili】

【ORCID: 0009-0006-5696-1873】

【Corresponding Author Email: jiligzs@qq.com】

Abstract

We propose a normatively verifiable construction pathway for the Universal Production Line (UPL), targeting equipment self-supply and prototyping during the factory-construction phase. The implementation line is as follows: the Manufacturing Digital Base (MDB) transduces the physical world into AI-operable digital assets; four categories of integration actions give rise to an autonomous simulation closed loop fed by collision logs, enabling action flows to be autonomously refined before ever running on the physical machine; capability is then verified progressively along the graduated ladder of "fixed flow → parameterized flow → autonomously orchestrated flow," culminating in a universal production line that integrates three segments—circuitry, mechanisms, and fixed workstations. To render "full" autonomy decidable, we introduce four measurable admission criteria—full-chain closed loop, autonomous generation, autonomous verification, and autonomous governance—and instantiate them as collectable metrics with recommended acceptance thresholds (representative recommended values: first-pass rate (FPR) ≥ 95%, generation success rate (GSR) ≥ 90%, abnormal self-closure rate (ASCR) ≥ 80%, steady-state human intervention rate (HIR) ≤ 5%; all are suggested values pending calibration). The decision criterion is whether AI forms a complete closed loop covering the six links of sensing, drafting, interaction, simulation, decision-making, and execution, and autonomously governs its own operation—not whether AI technology is used. Throughout the paper, an evidence-boundary table explicitly separates constructed engineering prototypes from research agendas yet to be validated, and an end-to-end concept validation (PoC) protocol is provided. This paper is positioned as a definitional contribution and a research agenda, not as a claim of achieved full autonomy.



Graphical Abstract: Pathway of the Full AI Autonomous Governance Factory (Full-AGF) and its Universal Production Line (UPL)

Keywords: Full AI Autonomous Governance Factory (Full-AGF); Universal Production Line (UPL); Manufacturing Digital Base (MDB); Autonomous Simulation Platform; sim-to-real transfer; progressive verification; autonomous governance

Contents

1 Introduction: From Concept to a Testable Implementation Pathway

Manufacturing is undergoing a paradigm shift from automation to autonomy. Existing notions such as “lights-out factories” and “unmanned factories” emphasize unmanned execution but do not systematically answer the fundamental question of “where a factory’s autonomous capability comes from and who governs it” [1]. In prior work, we proposed the concept of the Artificial Intelligence Autonomous Governance Factory (AI-AGF), emphasizing that AI systems autonomously complete the full process of sensing, decision-making, and execution without human intervention, while systematically governing their own goals, rules, boundaries, and risks; we also provided a four-dimensional governance framework and a four-stage launch pathway [25].

Against this background, this paper advances a contestable claim: the prevailing tendency to conflate “the factory uses AI” with “the factory is autonomous” constitutes a fundamental conceptual error. The history of automation has shown that deploying machine-learning or large-model components in a production line may yield only localized intelligence enhancement, leaving entirely unaltered the overarching structure in which “goals are set by humans, flows are orchestrated by humans, and risks are borne by humans.” If “whether AI technology is used” serves as the criterion for autonomy, any factory that inserts an AI module could claim to be on the path to full autonomy, and such a claim is neither falsifiable nor comparable. We argue that the correct unit for measuring “full” autonomy is the **completeness of the closed loop**: whether sensing, drafting, interaction, simulation, decision-making, and execution form a coherent closed loop, whether each link is autonomously performed by AI, and whether governance is autonomously implemented by AI. This criterion turns “full autonomy” from a rhetorical goal into an engineering boundary that can be adjudicated, measured, and audited at both the equipment level and the production-line level. The four admission criteria, measurement framework, and verification pathway presented in the following sections all serve to bring this position into a testable form.

1.1 Criteria for “Full” Admission

Accordingly, we propose four “full” admission criteria as operational definitions that connect with the aforementioned concepts and can be measured in engineering practice. Criterion F1 (full-chain closed loop) requires that sensing, drafting, interaction, simulation, decision-making, and execution are all carried by AI; the absence of any link degenerates the system into localized intelligence. Criterion F2 (autonomous generation) requires that AI autonomously generates action flows and executable programs under a given goal, rather than merely matching pre-set templates. Criterion F3 (autonomous verification) requires that AI autonomously completes the verification, iteration, and optimization of action

flows in a simulation environment, and takes “passed simulation verification” as a necessary condition for going onto the physical machine. Criterion F4 (autonomous governance) requires that, during production operation, AI monitors execution status in real time, performs root-cause analysis, alternative-plan substitution, and experience accumulation for anomalies, thereby forming a self-improving closed loop. Together, the four criteria constitute the necessary-and-sufficient boundary of a Full-AGF; a facility that fails to meet them should be objectively described as a “smart manufacturing factory deploying AI technology.” To make the criteria testable, Section 3 maps F1–F4, one by one, onto collectable metrics, decision rules, and recommended acceptance thresholds, converting them from qualitative commitments into decidable conditions.

To ensure that the relationship between the judgment of “full autonomy” and empirical results is not misread, we explicitly split the criteria into two layers. The “criterion structure” refers to the combination of six-link closed-loop completeness, metric definition formulas, and decision rules, which contains no numerical parameters; it defines “what is full autonomy” and can therefore be evaluated independently of any empirical result. The “parameter instance” refers to the specific recommended thresholds given in Table 2, which are objects of engineering heuristics and statistical calibration; it only adjudicates “whether a given factory currently meets the standard,” rather than determining “what the boundary is.” Threshold calibration after the PoC only updates the parameter instance and does not revise the criterion structure; threshold drift only changes a line’s current compliance status and does not move the conceptual boundary of “full autonomy.” The whole paper follows its graphical abstract: the physical world is transduced by the digital base, the autonomous simulation platform iterates in a closed loop, capability is verified step by step along the graduated verification ladder, and finally the F criteria give a decidable boundary for the completeness of the loop and its autonomous governance.

1.2 Research Questions, Positioning, and Evidence Boundary

The prior research focused on “why to do it, what form it takes, and how to launch it,” belonging to the conceptual and macro-pathway level. This paper takes up its four-stage launch pathway and focuses on the most engineering-challenging part—the Universal Production Line (UPL)—to answer a more specific question: within an established conceptual framework, how should the universal production line of a Full-AGF be actually built, and how can this pathway be objectively verified? “Universal” here is bounded: the UPL discussed in this paper refers to the production-line form capable of undertaking equipment self-supply and prototyping tasks during the factory-construction phase—covering the spectrum of action flows required for equipment manufacturing and trial production with finite kinematic combinations, rather than unlimited adaptation to arbitrary manufacturing tasks. This qualification is consistent with the UPL definition in the glossary (Appendix C).

Here we must clarify the genre and evidence boundary of this paper. This paper is a **position paper with a concrete research agenda**, rather than an experimental study with large-scale empirical validation. The "implemented" parts of the paper refer to foundational work such as the engineering prototype of the Manufacturing Digital Base and the construction of the simulation closed-loop environment, whose status and evidence are explicitly given in the evidence-boundary table in Section 2.3; the "expected / to-be-validated" parts are presented in the form of a research agenda and a validation protocol, together with a primary validation milestone and an end-to-end concept validation (PoC) protocol (Appendix A). We deliberately avoid describing not-yet-completed empirical work with the wording of "proven conclusions"—anything not actually measured is not asserted as completed. Notably, the definitional contribution of this paper—namely the F1–F4 criteria and their decidability—can be independently evaluated without any empirical result; the paper's commitment to empirical evidence is explicitly scoped by the protocol in Appendix A, and the two do not obscure each other.

To make the boundary of "full autonomy" locatable in cross-domain discussion, we further clarify the relationship between our criteria and existing autonomy-level frameworks. Existing characterizations of autonomy in industry and robotics largely follow two lines: one is ordinal grading for vehicles (e.g., SAE J3016 divides six levels according to the distribution of driving responsibility between human and system [26]); the other is autonomy capability/maturity assessment oriented to tasks and missions, which grades capability tiers by the functions a system can autonomously perform at a given task complexity. The four criteria here do not constitute a new ordinal scale but rather a **capability-completeness boundary**: they do not ask "which level is the system at," but "whether the closed loop of sensing–drafting–interaction–simulation–decision-making–execution is complete, whether its links are carried by AI, and whether governance is autonomously implemented by AI." On both spectra, our definition can be understood as replacing the meaning of "highest level" with an engineering statement of "closed loop complete and every link measurable"; for systems that do not reach the boundary, we provide an intermediate label of "progress state toward full autonomy" (Section 3.2), compatible with the transitional semantics of "partial automation" under ordinal spectra. This mapping shows that "full autonomy" is, in this paper, a decidable engineering boundary rather than a borrowing of the top tier of any external classification.

Methodologically, this paper combines conceptual deduction (deriving boundaries and base-design requirements from the F criteria), engineering reasoning (deducing, from a systems-integration perspective, the chain from data collection to action-flow generation), and case analysis (using a custom non-standard handling device as a complete case to present a reproducible implementation flow). It also draws broadly on existing results in digital twins [3-4], industrial AI [5], LLM-driven robot task planning and program generation [6-7,17-19], simulation-feedback repair [8],

and sim-to-real transfer [20-21], and systematically compares with them (Chapter 6).

It is also necessary to delimit four existing paradigms. First, regarding “lights-out/unmanned factories”: the latter emphasize unmannedness, whereas this paper emphasizes autonomy plus governance; a lights-out factory can be fully automated while still having all flows defined by humans, whereas a Full-AGF requires AI to possess the capability of autonomously generating and governing flows (criteria F2–F4). Second, regarding “factories deploying agents”: deploying agents is merely localized intelligence, lacking a full-chain closed loop and an autonomous verification mechanism, and therefore fails the full criteria. Third, regarding digital twins: a digital twin is an enabling technology of our system (modeling and virtual verification), while the Manufacturing Digital Base is the data source of digital-twin models—the two have a “supply–consumption” relationship [3-4]. Fourth, regarding Reconfigurable Manufacturing Systems (RMS): RMS achieves process flexibility through modularity and reconfiguration [14]; this paper further delegates the act of “reconfiguration” to AI, moving from “structurally reconfigurable” to “capability self-organizing.”

2 Manufacturing Digital Base and Capability Emergence

The premise of building a Full-AGF is to let AI “see, draft, and communicate”—that is, to understand the physical world, generate engineering objects, and collaborate efficiently with humans. We collectively call the software set supporting these three capabilities the **Manufacturing Digital Base (MDB)**, namely the software collection that transforms objects, behaviors, and knowledge in the physical manufacturing world into AI-understandable, AI-operable digital assets. It must be distinguished from the commonly used “data foundation/base platform”: generic data platforms are responsible for storing, flowing, and aggregating data, whereas the function of MDB is **transduction**—converting equipment geometry, drawing semantics, and displacement behaviors into models, parameters, and executable scripts that AI can directly call; its output is manufacturing digital assets rather than datasets themselves. The system boundary of MDB is: at the input end it receives physical-world information (scan data, drawings, parameters, requirement texts); at the output end it provides digital assets (3D models, drawing data, displacement parameters, executable scripts); internally it serves the autonomous simulation platform, and externally it supports human–machine interactive review. It should be emphasized that MDB is not a specific application but a common base that carries the upper-layer simulation, planning, and governance capabilities; its “digital” attribute highlights the essence of digitizing and assetizing the physical manufacturing world, and it connects with digital twins [3-4]: a digital twin is a “model,” while the digital base is a “production line for manufacturing digital assets.” The three-layer architecture of the Manufacturing Digital Base is shown in Figure 1.

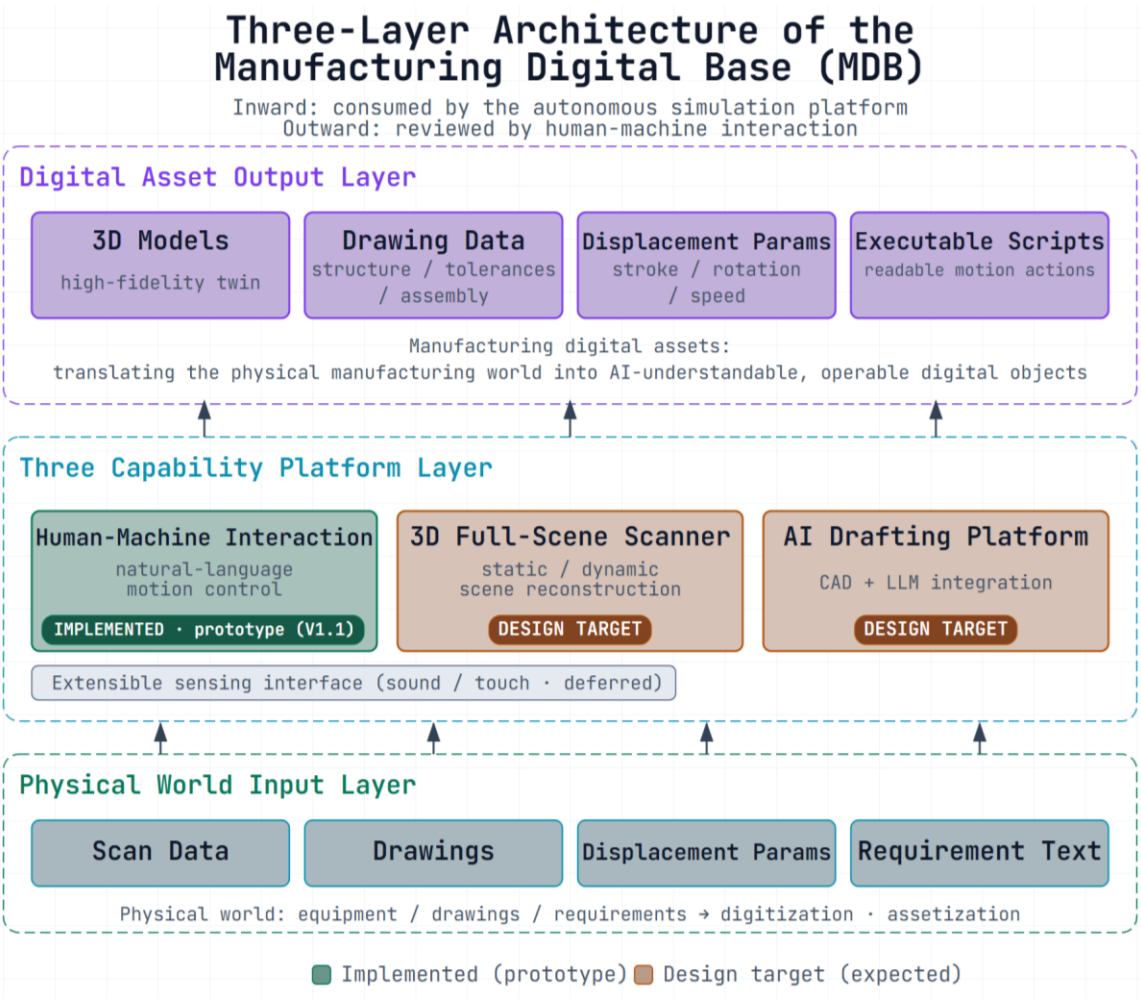


Figure 1 | Three-layer architecture of the Manufacturing Digital Base (MDB)

2.1 Design of the Three Categories of Foundational Software

The Manufacturing Digital Base consists of three categories of software, which respectively address three foundational problems—"humans can understand the code, machines can be seen, and drawings can be generated"—and reserve a unified sensing-extension interface.

Human–Machine Interaction Platform (natural-language script motion control). The products of AI-driven action-flow generation must ultimately be realized as executable code, and human engineers need to review, understand, and correct that code; if AI directly generates obscure low-level instructions, the human-in-the-loop review stage cannot hold. Accordingly, the core requirement of the interaction platform is that the action code produced by AI be readable, reviewable, and editable by humans. This design is carried by the Chinese-script motion control platform prototype we have already built (a prior engineering result, registered under software copyright V1.1): it encapsulates low-level motion control (point-to-point motion, linear interpolation, circular interpolation, I/O control, multi-axis

coordination, etc.) into script instructions close to natural language, forming a two-layer structure of "motion-control function library + readable scripts." This design is consistent with the mainstream paradigm of using LLMs for robot control: Vemprala et al. argued that a "high-level function library + natural-language interaction" architecture allows LLMs to adapt to different robot morphologies and simulation environments without retraining, translating user intent into executable code [9]. In the industrial adoption of this paradigm for custom automation, three properties are particularly critical. The first is auditability: every line of action script generated by AI can be read directly by engineers without relying on proprietary vendor protocols, enabling them to serve as reviewers. The second is portability: scripts, as an intermediate representation, can be mapped to the low-level instructions of different motion-control cards, avoiding tight coupling between AI output and hardware. The third is stackability: process parameters, timing-avoidance rules, and safety constraints can be stacked on top of scripts, reserving interfaces for subsequent autonomous simulation and governance. In addition, this platform is an important vehicle for anomaly tracing and audit forensics—each executed script and its parameters constitute a traceable "action archive."

3D Full-Scene Scanner (static and dynamic scene reconstruction). The premise of an AI-governed factory is to "see" the real equipment. The design goal of the 3D full-scene scanner is to achieve 3D reconstruction of equipment and scenes with handheld-device- or lightweight-camera-level hardware, covering two capabilities. Static 3D scene scanning builds high-fidelity 3D geometric models of complete machines, modules, fixtures, and products, used to construct simulation environments and generate point clouds to support pose recognition; dynamic 3D scene scanning captures scene changes in real time during equipment motion, used to validate displacement parameters of components, such as axis stroke ranges, cylinder upper/lower limits, and the actual trajectories of moving parts, thereby calibrating "parameters written on the drawing" into "parameters actually running." Dynamic scanning is a signature stage of our methodology: it serves not only simulation modeling but also reverse data collection (Section 2.5), enabling the factory to continuously generate action data usable for AI learning during operation. On the technical-support side, pose-estimation methods for industrial grasping and assembly can already achieve robust six-degree-of-freedom (6-DoF) pose regression on point clouds [10], providing a mature base for scan data to directly drive manipulator operations; digital-twin research also shows that constructing twin models with real-time sensing data is key to supporting runtime monitoring and decision-making [3]. The design and performance goals of this component are yet to be validated through prototyping and are treated as "design goals" per the Section 2.3 convention.

AI Drafting Platform (CAD and LLM integration). Traditional drawing design relies heavily on engineer experience and is a bottleneck restricting a factory's rapid response to new orders. The design goal of the AI drafting platform is to integrate Computer-Aided Design (CAD) with AI large language models so that, upon

receiving a customer requirement, AI autonomously completes the entire process from requirement understanding and solution design to outputting all drawings. On the implementation path, the platform consolidates design knowledge (standard-part libraries, tolerance systems, structural constraints, process specifications) into a retrievable engineering knowledge base—research on manufacturing-domain knowledge graphs provides a systematic methodological foundation for building such knowledge bases [12]—and uses an LLM to perform semantic parsing of requirement texts and to call parametric-modeling interfaces for progressively generating part drawings, assembly drawings, bills of materials (BOM), and mechanism kinematic models. The platform's output does not exist in isolation but serves as one of the input data sources for subsequent simulation (Section 2.4), forming a "design–simulation–verification–optimization" loop. Like the 3D scanner, the current status of this platform is a design goal, and all performance statements are hypotheses to be validated.

Extensible Sensing (deferred capability and unified interfaces). Sound (vibration noise, pneumatic action sounds, abnormal acoustics) and touch (force, pressure, torque) are valuable in assembly force control and fault diagnosis, but given engineering investment and current validation priorities, we classify them as deferred capabilities to be "developed when needed," and reserve a unified sensing-access interface in MDB to avoid duplicate construction.

2.2 Four Categories of Integration and Capability Emergence

Once the foundational software is built, capability does not emerge automatically; platform integration is required to organize digital assets into decision-making capability. The core of this component is four categories of integration actions, whose emergent product is the "autonomous simulation platform closed loop fed by collision logs."

Data Integration: High-Fidelity Digital-Twin Models. The scene data obtained from 3D scanning, all drawings of the corresponding equipment, and the displacement parameters of components (standard equipment is accompanied by user manuals) are uniformly fed into the simulation platform; after registration, assembly, and semantic annotation, the goal is to obtain a high-fidelity simulation model whose three properties form the basis of subsequent collision detection. First, structural completeness: every component belongs to a well-defined module, and assembly and hierarchical relationships are queryable. Second, semantic annotation: each module can be labeled with name, function, specification, and movable range (stroke/rotation/angle). Third, behavioral imitation: combined with displacement parameters, the model's kinematics and motion range are consistent with the physical machine. This process is consistent with the "virtual–real mapping" philosophy of industrial digital twins, one of whose core challenges is precisely model accuracy and data validity [3]; this paper mitigates the challenge in engineering through a "scanned geometry + drawing structure + displacement

parameters” three-source integration, giving the simulation model a “verification-grade” accuracy orientation. Data- and knowledge-hybrid-driven thinking has been widely practiced in intelligent control of discrete manufacturing systems [16] and can serve as a reference for model construction in this section.

AI Integration: Autonomous Simulation Platform and Closed-Loop Iteration. If data integration is compared to “function definition” before writing a program, this item is the “AI develops action flows from requirements” stage. Its core mechanism is a five-step closed loop (Figure 2): AI receives a high-level goal to “complete a certain action” (e.g., take a product out of the bin and transfer it to a workstation); based on module definitions and kinematic constraints in the simulation model, it autonomously plans an action sequence and generates a simulation-executable flow; the simulation platform runs the flow and outputs collision alerts and interference logs; AI reads the collision logs as abnormal-debugging information, corrects action timing, speed, and paths, and re-runs the simulation; iteration continues until a collision-free flow satisfying cycle time and process requirements is obtained. This “generate–feedback–correct” loop is consistent with the direction of “correcting plans with execution feedback” in LLM-based industrial task planning [7-8]. For example, existing work has proposed a trajectory-guided domain repair framework that lifts LLM-generated plans from “syntactically correct” to “physically executable” through precise attribution of execution failures and structured prompting [8]. The contribution of the autonomous simulation platform is to mechanize this idea into a generic closed loop: each simulation collision feedback is equivalent to a code-debugging log, and AI continuously optimizes with logs as input until convergence, giving the factory the ability to autonomously polish action flows “without going onto the physical machine.” The convergence, termination, and fidelity-error propagation of this loop are developed as analytical propositions in Chapter 5 and are designed as empirically testable metrics in the PoC protocol of Appendix A.

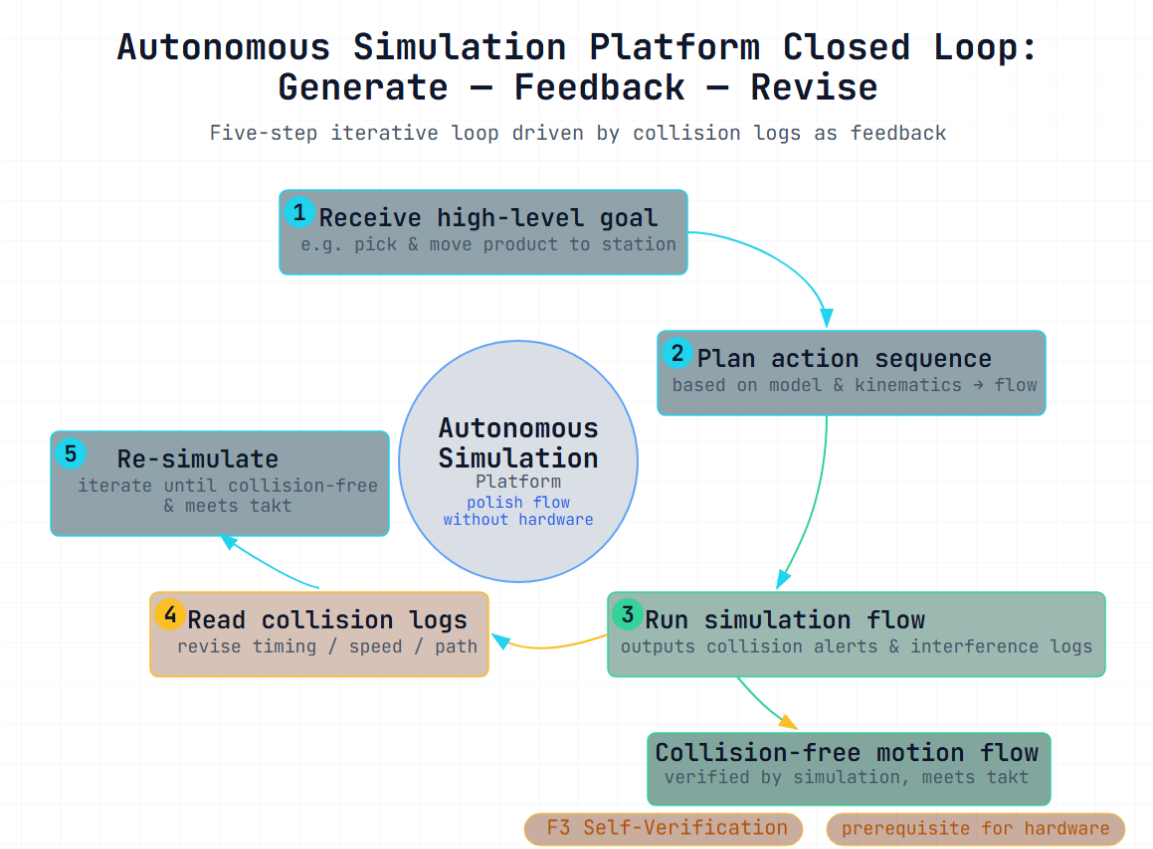


Figure 2 | Autonomous simulation platform closed loop

Human–Machine Integration: Review Closed Loop before Physical Operation. A flow produced by the autonomous simulation platform still requires a "human–machine co-governance" review stage before actual operation. Integrating the human–machine interaction platform enables: readable presentation—translating AI-generated flows into Chinese scripts and visualized steps for engineers to quickly understand; preemptive review—engineers confirm high-risk actions and abnormal-condition contingency plans before the first physical operation, avoiding anomalies in the equipment's first operating cycle (first-ever run); dynamic convergence—once the same category of flow becomes stable after multiple physical validations, item-by-item human review can be waived and handed over to autonomous AI execution; and archive retention—interaction records are continuously preserved as the basis for tracing and audit forensics when anomalies occur. This stage is consistent with the "AI acts, humans review" human–machine co-governance idea in the previously proposed "four-stage launch pathway" [15,25], and echoes the "human in the loop" notion emphasized by augmented-reality-assisted mutual-cognitive human–robot safety interaction [11]; human-centric fusion technologies for Industry 5.0 provide methodological support for this co-governance pattern in production scenarios [15].

Drafting Integration: Through the Design–Simulation–Manufacturing Chain. The AI drafting platform is not an isolated end-stage but a source capability running throughout. Its mechanism is: obtain customer requirements and generate equipment drawing data; combined with displacement parameters, drawing data enter the simulation platform for kinematic simulation and mechanism interference checking; simulation results feed back points where the mechanism is unreasonable (interference, insufficient stroke, insufficient rigidity, etc.), and AI optimizes the drawings accordingly; the optimized drawings re-enter simulation verification until the mechanism design is sound. In this way, the AI drafting platform and the autonomous simulation platform form a "design–simulation–optimization–redesign" collaborative loop, exposing and correcting equipment design errors in virtual space and significantly reducing the rework risk at the physical stage.

2.3 Evidence Boundary and Current Status

To ensure the above narrative is not misread as completed empirical work, we explicitly declare the current status and verifiability of each major component. Table 1 is the evidence-boundary table of the whole paper: components marked "prototype" have their existence verifiable through software copyright registration or a runnable local environment; components marked "design" only provide design goals and implementation paths, and any statement about their performance is a hypothesis to be validated. The whole paper accordingly uses a unified wording convention—"prototype" components use already-implemented/already-built phrasing, while "design" components always use design goal/expected/planned and equivalent phrasing.

Component	Current status	Verifiable evidence	Performance wording convention
Chinese-script motion control platform	Built prototype	Software copyright registration V1.1	Runnable; related performance metrics to be quantified by PoC
3D full-scene scanner	Design goal	No standalone deliverable	Design goal and integration vision, to be prototyped
AI drafting platform	Design goal	No standalone deliverable	Design goal and integration vision, to be prototyped
Autonomous simulation platform closed	Mechanism design + environment vision	No standalone deliverable	Mechanism design and convergence propositions, to be

Component	Current status	Verifiable evidence	Performance wording convention
loop			evidenced by PoC
F1–F4 criteria and measurement framework	Definitional layer (formalized)	Chapter 3 of this paper	Definitions and rules directly decidable in engineering

Table 1. Evidence boundary and current status of components. No component that has not reached the "built prototype" status is asserted as completed anywhere in this paper.

2.4 Forward and Reverse: Bidirectional Complete Action Flows

The final output of platform integration is two complementary pathways for obtaining complete action flows. The forward path (human → AI) takes equipment definitions and action goals from engineers or customers; AI computes the complete action flow of the real equipment from drawings, displacement parameters, and simulation verification, serving as an automated replacement for current manual action-flow development. The reverse path (physical machine → AI) uses 3D scanning of real-time equipment operation to extract the equipment's actual action flow in reverse, forming training data; it serves two purposes: first, when equipment fails, comparing the "expected flow" with the "actual flow" traces the failure cause; second, real action data are accumulated into datasets for AI continuous learning and model fine-tuning. The forward and reverse paths constitute a bootstrapping progression of "generate forward first, verify in reverse, and reinforce forward capability with reverse data"—the mechanistic basis for the factory's self-strengthening autonomous governance capability.

2.5 Extension: Generalized Control of Standard Equipment

The above capabilities apply to both custom and standard equipment (manipulators, robots), with standard-equipment control being more direct to implement. After scanning a manipulator or robot and obtaining its operating methods and related drawings, it can be normally controlled in the simulation environment; the operation of standard equipment is fully defined in its manual, and AI can directly parse the manual and confirm the control method with real-time poses, without the complex reasoning of "mechanism-drawing matching → module identification → displacement calibration" needed for custom equipment. Because the work goal of custom equipment is programmable, it essentially downgrades "action-flow development" into a job-shop capability; standard equipment, with its complete manuals, is naturally easier for AI to take over. Our capability system thus exhibits a "reverse-convergence" pattern: the more difficult custom equipment is,

the earlier it needs full-chain AI support; once the full-chain capability of custom equipment is running, generalized control of standard equipment follows naturally.

3 Operationalization of the Criteria and Measurement Framework

The four criteria are qualitative descriptions that are hard to measure. This chapter converts them into the form of "definition + measurement function + decision rule," enabling data collection and adjudication at both equipment-level and production-line-level granularities. It should be noted that the acceptance thresholds given in this chapter are currently **recommended values** set on the basis of engineering heuristics; their reasonableness is itself part of the research agenda and will be calibrated on real data through the PoC protocol in Appendix A.

3.1 Formal Definition of the Metrics

We give a reviewable formal definition for each of the core metrics, including the definition formula, observation unit, statistical window, and collection/audit party. Table 2 provides the complete mapping of criterion–metric–definition–data source–threshold–threshold basis. The statistical window of every metric is "all relevant events within the most recent natural quarter," automatically exported from operation logs and subject to random-sampling audit.

To eliminate ambiguity in observation scope, we first define this paper's basic observation unit—the **production operating cycle**: a continuous operating window on a line from the start of a task/order to its acceptance completion, an observable interval bounded by the task/order identifier; all sensing, planning, execution, and governance events within the same window are attributed to that cycle. All metrics counted per "cycle" follow this definition.

The metrics are defined as follows (expressed in terms of "observation unit" and "decision formula"; any non-formula text can be verified in equivalent wording):

- **AI Link Coverage (CoC)** (mapped to F1): The observation unit is all links of one production operating cycle (sensing, drafting, interaction, simulation, decision-making, execution). It is defined as "the number of links carried by AI / the total number of links," with the requirement that all six links be non-vacuously carried, i.e., CoC = 1.
- **Human Intervention Rate (HIR)** (mapped to F1): The observation unit is one production operating cycle. It is defined as "the number of cycles with human intervention / the total number of operating cycles." To distinguish two qualitatively different kinds of intervention—"human in the loop for gating" versus "insufficient AI capability"—this metric is split into two sub-metrics: the review-and-release intervention rate (HIR_r), referring to interventions where humans only perform pre-set release/confirmation actions (e.g., approving a flow submission) and the intervention then ends; and the correction-and-takeover intervention rate (HIR_c), referring to

interventions where humans modify flow content or take over/intervene in execution. The recommended steady-state threshold is HIR (total) ≤ 5%; HIR_c is a trend metric, required to be reported separately each quarter and to show a downward trend, with two consecutive quarters of increase triggering governance-record-level review (under the Section 3.3 duration convention); its convergence target (approaching 0) will be given in magnitude after PoC calibration.

- **Generation Success Rate (GSR)** (mapped to F2): The observation unit is one action-flow generation task. It is defined as "the number of flows that pass review and simulation verification without substantive human modification after one generation / the total number of generated flows" (the proportion of template-matched non-independent generation is recorded as an auxiliary metric). "Without substantive human modification" is bounded by an operable binary criterion: if human review of a flow produces only the pre-set release action without changing the action sequence or parameters, the flow is counted as "unmodified"; any change to the action sequence or parameters (even a single one) reclassifies the flow as "modified." The recommended threshold is GSR ≥ 90%.
- **First-Pass Rate (FPR)** (mapped to F3): The observation unit is the first on-machine execution of one action flow. The "passing acceptance" criterion is bound to the final-level acceptance checklist of the six-level commissioning ladder in Section 4.2: a loaded trial run completed within cycle time and accuracy tolerances, with no collision/interference and no safety interlock trigger, is deemed passing. It is defined as "the number of action flows that satisfy the acceptance criterion on their first physical execution / the total number of flows put on the physical machine." The recommended threshold is FPR ≥ 95%.
- **Simulation Convergence Rounds** (mapped to F3, auxiliary): defined as "the number of iteration rounds consumed from the first simulation to satisfying the termination criterion," recorded and reported without a hard threshold.
- **Zero-Reference First-Success Task-Planning Rate (ZFS-FPR)** (mapped to F2/F3, advanced extension): The observation unit is an assembly or collaborative task whose input has "never been seen"—i.e., no existing flow template and no manual flow example. It is defined as "the number of tasks independently generated by AI under zero-reference conditions that pass simulation verification on the first round and satisfy the final-level acceptance conditions of the six-level commissioning ladder / the total number of such tasks." Data sources are flow-generation logs, simulation platform logs, and physical-machine records. The recommended threshold is ZFS-FPR ≥ 60% (equipment level); this threshold is likewise a recommended value to be calibrated in pilot runs at the autonomous-orchestration level; the statistical window and audit convention are the same as Table 2.

- **Abnormal Self-Closure Rate (ASCR)** (mapped to F4): The observation unit is one governance event. It is defined as "the number of anomalies for which AI autonomously completes root-cause analysis and closes the loop (including alternative plans verified by re-run) / the total number of anomalies during the audit period." The recommended threshold is ASCR ≥ 80%.
- **Generational Escalation Time to Human (ToE)** (mapped to F4, auxiliary): defined as "the median time interval between two generational escalations of same-category anomalies escalated to humans because of insufficient governance capability," required to decrease month over month and be recorded.

Criterion	Metric	Definition formula and statistical convention	Data source	Recommended threshold (equipment level)	Threshold basis
F1 full-chain closed loop	CoC; HIR (HIR_r / HIR_c)	six-link coverage = 1; HIR = intervention cycles / total cycles, with review-and-release and correction-and-takeover reported separately	operation logs	CoC = 1; steady-state HIR ≤ 5% (HIR_c reported additionally)	engineering heuristics, recommended
F2 autonomous generation	GSR	proportion passing review and simulation without substantive human modification (binary criterion)	flow-generation logs	GSR ≥ 90% (template-matched proportion reported additionally)	engineering heuristics, recommended
F3 autonomous	FPR; convergence rounds	proportion passing final-level acceptance	simulation platform logs + physical-	FPR ≥ 95%; convergence rounds recorded	engineering heuristics, recommen

Criterion	Metric	Definition formula and statistical convention	Data source	Recommended threshold (equipment level)	Threshold basis
verification		on the first run; number of convergence iterations	machine records	and reported	ded
F4 autonomous governance	ASCR; ToE	autonomous-closure anomaly ratio; median escalation time	governance event repository	ASCR ≥ 80%; ToE median decreasing monthly	engineering heuristics, recommended
F2/F3 (advanced extension)	ZFS-FPR	proportion of zero-reference first-sight tasks passing simulation on the first round and physical first run	flow generation + simulation + physical-machine logs	ZFS-FPR ≥ 60% (recommended)	engineering heuristics, recommended

Table 2. Formal definition of the full admission criteria F1–F4: metrics, definition formulas, data sources, and recommended acceptance thresholds (recommended values, pending PoC calibration).

3.2 Decision Rule: Graded Admission

For the decision rule, we recommend “graded admission” rather than “all-or-nothing”: a factory may partially satisfy the criteria on a given line (e.g., F3 already met while F4 is still accumulating), in which case it should be described as in a “progress state toward full autonomy” with the gaps explicitly noted in governance records. This design is intended to preserve the conceptual boundary while avoiding denial of overall progress because a single metric has not been met.

The rules are as follows. For each production line, compute the metric criterion by criterion and compare with the threshold, yielding a three-state result of “met /

partially met / not met"; the line status is classified by the "weakest criterion" principle—only when F1–F4 are all met can the line be labeled "fully autonomous"; if any criterion is not met, it is labeled "progress state toward full autonomy (with the unmet criteria listed)"; partially met criteria must have their gaps and improvement plans stated in the governance record. Each status decision produces a governance record whose fields include: line identifier, metric values for each criterion, decision result, gap list, and whether incorporated into next-quarter improvement goals. The complete decision flow and governance-record requirements of graded admission will be referenced by the PoC protocol in Appendix A.

3.3 Aggregation from Equipment Level to Production-Line Level

Criterion metrics are collected at the equipment level, while "full autonomy" is an assertion about the production-line or factory-level property; an explicit aggregation rule is therefore required. Line-level metrics are defined as follows: for all equipment constituting the line, weight by each equipment's share of working time in line tasks, and aggregate to obtain line-level CoC, GSR, FPR, ASCR, etc.; HIR and ToE are computed on the aggregate basis of line operating cycles. The evaluation window is the natural quarter, and both upgrade and downgrade behaviors of the autonomy level (upgrade/downgrade, as distinct from fault-report escalation) are triggered by meeting or falling below the corresponding threshold for two consecutive quarters and are logged. This aggregation mapping resolves the granularity mismatch of "a single equipment meeting the standard ≠ the line meeting the standard": only when all constituent equipment of a line reach the baseline threshold can its weighted metrics be legitimately interpreted as line-level "fully autonomous" progress.

4 Capability Validation and Empirical Roadmap

The autonomous-generation capability of equipment action flows directly determines whether an AI-governed factory can be trusted to operate reliably. To achieve progressive validation from easy to hard and from fixed to free, trading deterministic gain for controllable risk, this chapter provides a graduated verification ladder, a reproducible complete case, and a roadmap that converges all empirical commitments into a single highest-priority end-to-end concept validation (PoC).

4.1 The Graduated Verification Ladder

In the low stage (fixed flows), equipment action paths are relatively fixed and flows are composed of finite kinematic combinations; the validation focus is whether correct flows can be stably reproduced. In the middle stage (parameterized flows), action paths vary with product and process parameters; the validation focus is whether flows can be correctly matched and adjusted according to parameters. In

the high stage (autonomously orchestrated flows), AI must understand module capabilities and assembly goals from drawings, physical objects, and real-time states, autonomously select parts, orchestrate action flows, and coordinate multiple execution units to complete assembly; the validation focus is whether completely new tasks can be autonomously planned and completed. Only after the previous stage passes sufficient validation does the next stage proceed, thereby progressively proving the core proposition that "well-trained simulation results also perform well on the physical machine." Equipment-level validation covers four representative custom non-standard devices with progressively increasing difficulty (Table 3).

Case	Device	Validation content	Key difficulty
1	Handling device	Bin lifting, in/out bin, and simple pick-and-place flows	Fixed flow, few action combinations
2	Laminating device	Pick-and-place flow with higher positioning accuracy	Higher accuracy requirement, stricter pose control
3	Assembly device	Adding assembly actions on top of pick-and-place	More action types, force-position coordination
4	Dispensing device	Adding dispensing process on top of assembly	Further accuracy increase, high trajectory continuity

Table 3. Equipment-level graduated verification ladder (fixed- and parameterized-stage cases).

Note: Case 1 corresponds to the fixed-flow tier, where equipment action paths are relatively fixed; Cases 2–4 correspond to the parameterized-flow tier, where action paths vary with product and process parameters.

The validation tasks of the autonomous-orchestration stage (high tier) are circuit-board assembly, component assembly, and multi-execution-unit (e.g., manipulator/robot) collaborative assembly, acting on a line composed of three segments—circuitry, mechanisms, and fixed workstations; this stage validates the line's overall flow-orchestration capability, rather than binding circuitry, mechanisms, or fixed workstations to the "autonomous-orchestration stage"; its achievement is jointly determined by the zero-reference first-sight task-planning success rate (ZFS-FPR) of Section 3.1 and the completion of milestone M3, thereby providing a measurable acceptance convention for "autonomous orchestration."

At the component-level validation, circuit-board assembly tasks require selecting wires and electrical components according to circuit drawings and completing assembly by itself, and component assembly tasks require selecting parts

according to equipment drawings and completing equipment assembly; such tasks require AI not only to reproduce flows but also to select parts per drawings and organize flows on demand—a critical leap toward full autonomy. Combining equipment-level and component-level validation, the target capability is: to achieve autonomous control of the manipulated object and complete the specified work based on physical 3D models, drawings, displacement parameters, or manuals. Accordingly, we set four measurable milestones: M1 completes autonomous generation and physical validation of the two standard flows of handling and laminating; M2 completes autonomous generation and physical validation of the two complex flows of assembly and dispensing; M3 completes autonomous planning and physical validation of the two autonomous-orchestration tasks of circuit-board assembly and component assembly; M4 compresses, for any new device, the time from delivered data to completion of the physical first-article trial production to an agreed time threshold. Among these, M1–M3 successively accept the three tiers of flow capability (fixed, parameterized, autonomously orchestrated), with the fixed tier incorporated into M1 and accepted together with the first parameterized case; M4 is an end-to-end efficiency criterion independent of flow tiers. The schedule and criteria for achieving all of M1–M4 constitute the backbone of the research agenda. It must be clarified here: F1–F4 are the criteria that define the boundary of "full autonomy"; M1–M3 are agenda milestones that successively accept the three tiers of flow capability and carry the empirical load of the F2/F3 boundary (the fixed tier is merged into M1 and batch-accepted; in Table 3, Case 1 covers the fixed-tier acceptance of M1, and Cases 2–4 cover the parameterized-tier acceptance range of M1–M2); M4 is a performance target for end-to-end delivery efficiency and does not constitute a component of the "full autonomy" boundary. The two sequences respectively answer "what is full autonomy" and "research-agenda completion," and must not be conflated into a single criterion set.

4.2 Complete Case: Full-Process Implementation of a Custom Handling Device

Taking a custom non-standard handling device as an example, we can fully demonstrate the engineering flow from data to production line as a reproducible template of the methodology, whose steps are also data-collection points of the PoC protocol (Appendix A).

The data-preparation stage collects three categories of data: 3D full-scene scan data (static 3D models of the device and environment), all drawing data (general assembly drawings, module drawings, part drawings, and models of the handled product), and component displacement parameters (axis strokes and positive/negative limits, cylinder upper/lower limits, gripper strokes). After format normalization, the data enter the Manufacturing Digital Base and become a unified data source for simulation and verification. The module-check stage performs consistency checks on the digital-twin model before entering action-flow development: the module-affiliation check confirms that the axes/actuators on

each module belong correctly, identifying structural errors such as “Module A’s Y-axis appears on Module C” and reporting them; the coordinate and right-hand-rule check verifies the definitions of positive/negative axis directions to avoid inconsistent directions between simulation and the physical machine; the data-consistency check verifies dimensional and structural inconsistencies between scan data and drawings. Failed checks are output as an anomaly list to serve as the basis for subsequent correction, preventing erroneous data from polluting flow generation.

In the autonomous action-flow generation stage, AI completes this on the autonomous simulation platform: progressive speed configuration from low to full speed, hardware interference checks between parts, and timing-avoidance planning between module actions, ultimately outputting a complete action flow adapted to the device. The physical commissioning ladder adopts a “simulation first, stage-by-stage acceleration” six-level process, each level based on real-time 3D scanning results: hardware verification (scan the current device and confirm the assembly matches the simulation model); device routine check (verify each component’s displacement direction and action open/close states); module commissioning (obtain the current device’s actual position coordinates and adjust open/close limits under loaded conditions); simulation rehearsal (based on current scan and commissioning data, run the full action flow in the simulation platform from low to full speed to identify interference and timing collisions); device dry run (run all flows without product at low speed, then progressively accelerate to full speed and observe interference and collisions); loaded trial run (execute all flows with product and likewise progressively accelerate, validating interference, timing collisions, and positioning stability). Only after all six levels pass can the device enter production. For manipulators and robots, the flow can be simplified: the “module displacement identification” stage is removed, replaced by confirming the control method directly from the manual and real-time poses; consequently, the cost of integrating standard equipment into the Full-AGF system is significantly lower than for custom equipment.

4.3 Research Agenda: PoC and Empirical Route

This section reframes future work as an executable empirical roadmap and clarifies its “expected / to-be-validated” identity.

Primary validation milestone: end-to-end concept validation (PoC) on a single custom device. As the highest priority, we plan to complete the full closed loop of “data preparation–simulation closed-loop convergence–physical commissioning ladder” on a handling device, and collect four categories of metrics: the simulation convergence curve (iteration rounds vs. remaining collision count), the simulation-to-physical performance gap (FPR vs. interference probability), commissioning-hours comparison (AI-assisted vs. traditional manual-programming baseline), and the human review pass rate. On statistical design, the three hypotheses to be

tested (H1: simulation converges within a finite number of rounds; H2: the physical FPR of the experimental group is significantly higher than that of the control group; H3: a 1× scan-accuracy error envelope preserves safety margins without significantly sacrificing cycle time) are all handled as two-proportion tests: with a significance level $\alpha = 0.05$, power $1 - \beta = 0.8$, and a control-group first-pass baseline of 0.5, the effect-size prior set by H2 (FPR difference of the experimental group relative to the control group ≥ 20 percentage points) requires about 25–30 action cycles per group; we accordingly set no fewer than 30 action cycles per class of experiment in Appendix A, and make explicit that this effect size is a study hypothesis rather than a value from existing literature. The complete variable design, control group, sample, and reporting format are given in Appendix A. We explicitly declare that H1–H3 are hypotheses to be validated, not existing conclusions.

Subsequent agenda. It must be emphasized that the single-device PoC of Appendix A is used only to calibrate the recommended thresholds of Tables 2 and 4, verify the operability of the F criteria on a single device, and test H1–H3; it does not constitute any proof of line-level autonomy or autonomous-orchestration-stage capability. After the PoC milestone, the validation agenda aligns level by level with the graduated verification ladder and proceeds in sequence: single-device PoC (this protocol) → multi-segment, multi-device line-level pilot (at least one device of each of the three segments—circuitry, mechanisms, and fixed workstations—connected, with line-level metrics collected per the aggregation rules of Section 3.3) → autonomous-orchestration-level validation (with the zero-reference first-sight task-planning success rate ZFS-FPR and milestone M3 as acceptance conventions, covering circuit-board assembly, component assembly, and multi-execution-unit collaboration) → cross-product knowledge transfer and multi-line linkage trials. The horizontal agenda includes: development and integration of extensible sensing capabilities such as sound and touch in the Manufacturing Digital Base; implementation of reverse-data auto-annotation and continuous AI model fine-tuning; piloting the graduated verification ladder on a real production line and collecting criterion metrics item by item per the measurement framework of Chapter 3; and calibrating the recommended thresholds of Table 2 on real data and publishing the calibration results.

5 Counterarguments, Safety, and Responsibility

This chapter first provides an analytical discussion of the convergence, termination, and fidelity-error propagation of the simulation loop, then presents a framework for safety-standard alignment, failure grading, and responsibility allocation. All propositions in this chapter are open questions that need to be answered experimentally.

5.1 Convergence and Termination

The iteration fed by collision logs can be viewed as a local search over the space of action sequences. Collision feedback is local: a single collision only indicates a conflict between a specific pair of components at a specific time; correcting timing or path accordingly can locally reduce the conflict count but may introduce new conflicts. We therefore do not claim a global convergence guarantee for this iteration on arbitrary problems, but rather define it as "engineering convergence guaranteed by a deterministic rollback strategy": after each correction, the candidate solution with the minimum evaluated collision count is retained, and when there is no improvement for K consecutive rounds (recommended $K \geq 5$), iteration terminates and rolls back to the best historical solution. Judging the "optimal action flow" requires an explicit objective function; we recommend multi-objective weighted acceptance: no collision (hard constraint), cycle time met, safety-distance margin, and readable scripts passing human secondary review, with the multi-objective weights pre-set by engineers; the termination condition is the conjunction of "no improvement for K consecutive rounds" and "human secondary review passed," thereby decoupling "algorithm termination" from "engineering release." These recommended termination criteria and convergence boundary will be empirically tested through the simulation convergence curve (iteration rounds vs. collision count) in the PoC protocol of Appendix A.

5.2 Fidelity-Error Propagation and Hedging

The performance gap from simulation to the physical machine stems from three categories of fidelity error: dynamic error (inconsistency between the physics engine and real friction, inertia, and elasticity), perception error (geometric/normal deviation introduced by 3D scan reconstruction, which in turn affects pose calibration and collision detection), and control/latency error (actuator response and communication latency). We provide three hedging measures: physics-engine calibration (back-calibrating dynamic parameters from physical dry-run data), error-envelope budgeting (reserving a clearance margin on the order of scan accuracy in simulation collision detection), and the stage-by-stage absorption of residual errors by the physical commissioning ladder of Section 4.2.

On the rationale for the error-envelope magnitude: public benchmarks of industrial point-cloud registration show that registration accuracy is typically on the order of millimeters (the root mean square error (RMSE) of common Iterative Closest Point (ICP) variants on real scan data is mostly in the 1–3 mm range), and the error of 6-DoF pose estimation under occluded/sparse scenarios is likewise measured on the order of millimeters and milli-radians [10]; therefore, reserving a "1× scan-accuracy order of magnitude" clearance margin in simulation collision detection absorbs most of the false-negative collision risk introduced by perception error, without introducing excessively large process clearances for each interference and sacrificing cycle time. We design the specific value of this magnitude as an

independent variable of the PoC (error envelope 0 / 1× / 2× scan accuracy, see Appendix A), quantitatively evaluating its impact on safety margins and cycle time through the "simulation-to-physical performance gap" metric.

5.3 Safety-Standard Alignment and Graded Boundary

Robot operations in the universal production line should explicitly align with the existing safety-standard system: human–robot coexistence operations of collaborative robots need to comply with the ISO 10218 series and ISO/TS 15066 restrictions on force, speed, and safety distance [22]; the functional-safety framework follows the Safety Integrity Level (SIL) grading of IEC 61508 to classify the risk of software failures in the autonomous decision chain [23]; the AI-oriented management system can interface with ISO/IEC 42001, incorporating AI risk into organization-level management systems [24]. On this basis, we recommend a graded safety boundary: high-risk actions (high kinetic energy, near-human, critical mechanisms) are by default released by human review with hard constraints, while low/medium-risk actions are autonomously executed by AI but with on-site emergency stops and safety PLC as a fallback.

5.4 Failure Grading, Degradation, and Human Takeover

An autonomous system must define explicit failure modes, degradation strategies, and auditable takeover conditions; otherwise "autonomy" is unverifiable in safety audits. We divide failures into three categories—perception failure, execution failure, and governance failure—and designate for each an auditable trigger criterion, degradation path, human-takeover time limit, and log-trail fields (Table 4). The key principle is that human takeover is the ultimate fallback rather than an optional alternative; any degradation and takeover event must be traceable and replayable.

Failure category	Trigger criterion (monitored signals)	Degradation path	Human takeover time limit	Log-trail fields
Perception failure	abnormal scan/pose updates (point-cloud missing rate, registration RMSE exceeding threshold)	model refuses to generate; the workstation degrades to human-guided scanning or human input	immediate prompt after first trigger; forced escalation if not resolved within 5 minutes	abnormal signal, trigger time, degradation action, takeover person, recovery time
Execution failure	action misses target, over-stroke limit triggered, force/torque	safe stop; generate governance event and	take over at operator console immediately	deviation amount, limit trigger point, stop cause,

Failure category	Trigger criterion (monitored signals)	Degradation path	Human takeover time limit	Log-trail fields
	exceeded, safety PLC stop	invoke contingency plan	after stop; handle per contingency-plan time limits	plan number, handling conclusion
Governance failure	N consecutive abnormal self-closure failures (recommended N ≥ 3) or ToE exceeded	escalate to human arbitration; AI enters restricted read-only mode	escalate as soon as ToE times out; duty engineer takes over within 30 minutes after escalation	iteration records, failure-cause chain, escalation trigger value, takeover person, experience item

Table 4. Failure grading and human-takeover trigger table. The trigger thresholds in the table (perception metric thresholds, N, time limits) are recommended values and will be calibrated by equipment category in the PoC.

5.5 Responsibility Attribution and the AI Management System

When an accident results from autonomous AI decision-making, responsibility should not be left suspended. We advocate making the "human–AI" responsibility allocation explicit: the owner of the equipment and production line is the organizational responsibility subject; the AI system provider bears responsibility for traceability and technical defects; the operating/reviewing engineer bears the responsibility of human confirmation within the authorized scope; and event logs, decision trails, and audit interfaces support post-hoc responsibility determination. This framework is consistent with the AI risk-management loop of ISO/IEC 42001 [24].

5.6 Impact on the Workforce and Skill Structure

The universal production line will significantly change the frontline job structure: repetitive program-choreography jobs decrease, while jobs of "flow review, anomaly arbitration, data governance, and AI operation" increase. As a position paper, we have an explicit responsibility to state: a portion of the automation gains should be invested in retraining systems so that line workers can transition to flow reviewers and AI operators; although this topic is beyond the technical scope of this paper, it can be incorporated into the agenda of subsequent policy–technology coordinated research and is consolidated in the limitations of Chapter 7. It should

be added that the operational quality of these new jobs (flow review, anomaly arbitration, data governance, and AI operation) is itself the audit object of the F4 criteria of Chapter 4 (abnormal self-closure rate ASCR, generational escalation time ToE); changes in the labor structure and the boundary of autonomous governance are two sides of the same coin.

6 Discussion: Comparison with Existing Work

6.1 Link-by-Link Comparison with a Representative Existing System

We take as the comparison object a representative recent public work with a similar goal—IMR-LLM [7], which targets industrial multi-robot task planning and program generation. The selection rationale is: its publication window is the closest to this manuscript (ICRA 2026), its task setting (equipment-definition input → action-flow generation → verification → physical operation) is isomorphic to ours, and it is publicly reproducible; among this range, no public work with a closer goal was found. Under a unified task setting (equipment-definition input → action-flow generation → verification → physical operation), we compare our differences with it link by link. The comparison results are given in Table 5, where “enhancement” means productizing/mechanizing an existing idea, and “addition” means a claim not present in the existing framework and proposed by this paper.

Comparison link	IMR-LLM [7] (representative existing system)	This paper’s claim	Increment type
Input data source	text-instruction-driven multi-robot task definition	three-source integration of scanned geometry + drawing structure + displacement parameters, including custom equipment	enhancement
Action-flow generation	LLM generates action sequences/programs	LLM generation + simulation collision-feedback closed-loop iteration	enhancement
Pre-physical verification	primarily syntax/task-level checks	simulation verification with physical collision and interference detection as a necessary condition	enhancement
Line-level governance	governance not incorporated into the task loop	F1–F4 criteria + graded admission + failure-graded takeover (Sections 3 and 5.4 of this paper)	addition

Table 5. Link-by-link comparison with a representative existing system (using IMR-LLM as the comparison object).

This comparison lands on two points. First, most existing methods target single-task, single-device action policies, whereas the object of this paper is line-level capability emergence and flow-level autonomous governance, with the increments mainly at the facility layer (base integration) and the governance layer (criterion-driven generational upgrade), rather than point algorithms. Second, the increments of this paper are two-layered: on the definitional chain, it provides a decidable boundary for "full autonomy"—a contribution that can be evaluated without the PoC; on the implementation chain, the concrete capabilities and performance metrics are delivered through empirical work in the research agenda. The two-layer boundary was clarified in Section 1.2 and is reconfirmed here from the comparison angle.

The comparison must also give its symmetric side. Relative to IMR-LLM-type methods, this paper remains subject to the following shortcomings. First, the online-validation capability of multi-robot scheduling and heterogeneous-actuator orchestration—the comparison method already has it, whereas this paper currently carries it only through the single-device PoC of a custom handling device, with multi-execution-unit collaboration (e.g., manipulator assembly) as a planning item at the autonomous-orchestration stage. Second, the comparison method can be reproduced offline on existing industrial datasets or standard-machine task settings, whereas our validation ladder relies on a self-built simulation environment whose fidelity calibration (Section 5.2) is itself part of the research agenda. The increments of this paper mainly fall on the facility layer (base integration) and the governance layer (criterion-driven generational upgrade), rather than benchmarking the online multi-robot orchestration capability that existing methods already possess.

6.2 Technology Increment Matrix

To clarify the complete increments of this paper relative to existing work, Table 6 labels the relationship between our components and existing directions with a three-tier notation of "consistent / enhancement / addition."

Existing direction	Representative work and its component	Relationship to this paper	Section
sim-to-real transfer	domain randomization/domain adaptation [20-21]	enhancement	Section 5.2 fidelity analysis and references
closed-loop imitation learning / plan repair	trajectory-guided domain repair [8]	enhancement	Section 2.2
LLM task planning / program generation	IMR-LLM [7], SayCan [17], Code-as-Policies [18],	enhancement	Sections 2.1, 2.2, 6.1

Existing direction	Representative work and its component	Relationship to this paper	Section
	ProgPrompt [19]		
industrial multi-agent / human–robot collaborative planning	LLM-based multi-agent human–robot collaborative assembly [13]	enhancement	Section 2.2
digital-twin factory	industrial digital twins [3-4]	enhancement	Sections 2.2, 1.2
reconfigurable manufacturing systems	RMS [14]	enhancement	Section 1.2
factory-level AI governance / AI management system	ISO/IEC 42001 [24]	consistent	Chapter 5
line-level capability emergence and flow-level autonomous governance	this paper's MDB + criterion-driven generational upgrade	addition	Chapters 2–5

Table 6. Technology increment matrix relative to existing directions ("consistent" means adopted as a positional reference, "enhancement" means productizing/mechanizing an existing method, and "addition" means the line-level increment proposed by this paper). "Addition" refers only to the line-level increment proposed by this paper and does not imply denial of the existing capabilities of the comparison methods; the relative advantages of the comparison methods and this paper's constraints are stated symmetrically in Section 6.1.

7 Conclusion

The core thesis of this paper is: a Full-AGF is not a factory that "uses AI" but a factory in which "AI constitutes a complete autonomous closed loop." To define this boundary, this paper establishes four measurable admission criteria—"full-chain closed loop, autonomous generation, autonomous verification, and autonomous governance"—and instantiates their formal definitions, graded-admission rules, and equipment-to-line aggregation mapping into a directly decidable engineering framework; around the universal production line for equipment self-supply and prototyping, this paper elaborates an implementation pathway that starts from the Manufacturing Digital Base, centers on the autonomous simulation platform, uses progressive verification as the ladder, and constitutes the universal production line from the three segments of circuitry, mechanisms, and fixed workstations, presenting a reproducible implementation flow with a custom handling device. Throughout, an evidence-boundary table explicitly separates built engineering prototypes from to-be-validated design claims,

and an auditable framework of failure grading and human takeover responds to safety concerns. The definitional-layer contribution—the F criteria and the decidable boundary of “full autonomy”—is fully presented and can be evaluated independently of any empirical work beyond this chapter; the implementation-layer contribution converges into a single highest-priority end-to-end PoC and the subsequent empirical route, whose success or failure will be determined in a verifiable manner by the metrics defined in Appendix A; that determination is limited to the operability of the criteria on a single device and threshold calibration, with line-level and autonomous-orchestration-level capability carried by the subsequent two stages (line-level pilot and orchestration-level validation). The criterion structure defines what “full autonomy” is (evaluable without empirical work), while the parameter instance defines whether a given line currently meets the standard (dependent on PoC calibration); the two are decoupled, and threshold drift does not move the conceptual boundary. Next, we will launch the concept validation of the first device under this protocol and make the calibrated thresholds and results public.

7.1 Limitations and Risks

This paper has several limitations that must be candidly stated. First, data- and model-quality dependence: simulation-model accuracy is constrained by scan quality and parameter completeness, and model errors may lead to “simulation passed, physical anomaly,” which requires remediation by the physical commissioning ladder and quantitative evaluation of error impact (Section 5.2). Second, reliability of large language models: generative models carry risks of logical hallucination and incomplete flows, and industrial scenarios must uphold the three-line defense of “simulation validation up front, physical review as fallback, and abnormal closed-loop governance.” Third, initial investment and talent thresholds: building the Manufacturing Digital Base requires composite engineering capability across vision, mechanics, software, and AI, with no trivial short-term cost, and its cost–benefit ratio itself requires empirical data support (the M4 efficiency target is a performance acceptance rather than an autonomy-boundary item; its achievement does not affect the judgment of the “fully autonomous” label). Fourth, safety and responsibility boundary: the responsibility attribution, safety redundancy, and human-takeover mechanisms when AI autonomously governs major anomalies still need deep integration with enterprise safety management systems; Chapter 5 gives a framework but it has not yet been tested by real accidents. Fifth, as a position paper, the current criterion thresholds, failure-trigger parameters, and PoC protocol have not yet been calibrated with data on a real production line—the part most needing careful scrutiny by readers and the first priority of our follow-up work. Sixth, the labor transition described in Section 5.6 involves policy and social dimensions and requires interdisciplinary coordinated research to realize its impact assessment. In addition, this paper’s scope ends at the line-level autonomous closed loop: factory-level management autonomy (e.g., enterprise resource

planning, cross-factory scheduling) belongs to a broader digitalization and organization-level research agenda outside the scope of this paper, and this paper does not constitute a judgment of its feasibility.

Appendix A Concept Validation Design (PoC Protocol)

This protocol is the expanded design of the primary validation milestone of Section 4.3 and is ready to be applied directly during execution.

Objective. Validate on a single custom handling device the end-to-end feasibility of "data preparation–simulation closed-loop convergence–physical commissioning ladder," collect four categories of metrics—simulation convergence curve, simulation-to-physical performance gap, commissioning-hours comparison, and human review pass rate—and convert them into the F1–F4 criterion metrics of Chapter 3.

Object and sample. One medium-sized custom handling device with pick-and-place functionality, with a single-specification workpiece as the product; each class of experiment (simulation, dry run, loaded trial run) is repeated for no fewer than 30 action cycles to provide reportable confidence intervals for success rates. Sample-size basis: under a two-proportion test with $\alpha = 0.05$, $1 - \beta = 0.8$, and a control-group first-pass baseline of 0.5, the H2 effect-size prior (FPR difference ≥ 20 percentage points) requires about 25–30 cases per group; 30 cases are taken to leave margin; this effect size is a study hypothesis rather than a literature value.

Experimental variables. The independent variables are (a) whether the simulation rehearsal closed loop is enabled (experimental group vs. control group using manual programming only) and (b) the simulation fidelity configuration (physics-engine calibration on/off; error envelope 0 / 1× / 2× scan accuracy); the dependent variables are remaining collision count, FPR, interference probability, first-pass-qualified cycle, commissioning hours, and human review pass rate.

Procedure. ① Collect scan/drawing/displacement parameters and build the model; ② run the simulation closed loop and record the convergence curve (iteration rounds vs. remaining collision count), stopping per the termination criterion of Section 5.1; ③ record each level of the six-level physical commissioning ladder; ④ the control group completes the equivalent task by traditional manual programming with hours recorded; ⑤ convert data into F1–F4 metrics per the measurement framework of Chapter 3; ⑥ record perception/execution/governance failure events and log them into the Table 4 trail fields.

Reporting format. The submitted results will include: the simulation convergence curve, the simulation-to-physical metric gap table, the experimental/control commissioning-hours comparison, criterion achievement converted per Table 2, error-envelope sensitivity analysis, and takeover and trail statistics of failure events. Results are reported truthfully regardless of sign.

Expected hypotheses (to be validated). H1: simulation converges within a finite number of rounds; H2: the physical FPR of the experimental group is significantly higher than that of the control group (effect-size prior ≥ 20 percentage points); H3: a 1× scan-accuracy error envelope preserves safety margins without significantly sacrificing cycle time.

Threshold calibration. Both the recommended thresholds of Table 2 and the trigger thresholds of Table 4 will be given calibrated values based on actual data upon completion of this protocol, with the calibration report published simultaneously.

Milestone schedule. To facilitate assessing the executability of this protocol, we give a sequential arrangement of stage–duration–output (durations are estimates, updated on a rolling basis per the device's actual status): data preparation and modeling (about 1 month; output: device modeling and check checklist); simulation closed-loop convergence trials (about 2 months; output: convergence curve and termination-criterion validation results); six-level physical commissioning and control trials (about 2 months; output: physical metric gap, commissioning-hours comparison, and F1–F4 conversion table); criterion threshold calibration and public report (about 1 month; output: calibrated thresholds and full recalculation report).

Appendix B Implementation Parameters and Data-Production Notes

Scan-accuracy metrics. It is recommended to record point-cloud resolution, registration error (RMSE), normal deviation, and sphericity error as reconstruction-quality baselines; pose calibration adopts a coarse-to-fine process of "drawing datum + on-site feature point pairs" followed by ICP fine registration. The 1× basis of the error envelope is taken as the scan-accuracy magnitude estimated from this reconstruction-quality baseline.

Coordinate-system convention. A unified right-hand rule applies; the positive axis directions are consistent with the drawing definitions, and module local coordinate systems belong to the global coordinate tree; modules violating the right-hand rule must be reported and rejected at the module-check stage before entering simulation.

Script-platform essentials. Chinese natural-language scripts consist of a two-layer structure of "motion-control function library + readable scripts," with scripts serving as the intermediate representation mapped to low-level control-card instructions; it is recommended to reserve extension fields for process parameters, timing avoidance, and safety constraints at the script layer. The current platform version is software copyright registration V1.1.

Simulator selection and model strategy. The simulator should support collision-detection-log export and speed/timing parameterization; the model is

recommended to adopt a two-stage prompting strategy of reasoning and reflection, with collision logs structured and injected into the prompt context.

Reverse data production. When dynamic 3D scanning obtains "physical action → action flow" annotation data, it is necessary to define annotation-accuracy requirements (key-pose sampling rate, timestamp-alignment error upper limit) and to handle the sparse abnormal-sample problem—it is recommended to augment rare abnormal samples by fault injection or manual replay, and annotations must undergo engineer spot-check review. The reverse dataset is planned to be selectively released after desensitization and protocol review.

Appendix C Glossary

Abbreviation / term	Full name / definition	Notes
AI-AGF	Artificial Intelligence Autonomous Governance Factory	concept proposed in prior research, see ref. [25]
Full-AGF	Full AI Autonomous Governance Factory	meets the F1–F4 boundary; the research object of this paper
UPL	Universal Production Line	constitution layer: integration of three segments—circuitry, mechanisms, fixed workstations; note the same Chinese word for "segment" has two senses: the three segments and the six links of the full-chain closed loop belong to different coordinate systems and must be distinguished in translation; deployment layer: for equipment self-supply and prototyping during the factory-construction phase
MDB	Manufacturing Digital Base	the collection of three categories of foundational/supporting software; distinguished from generic data platforms by "transduction"
F1–F4	full-chain closed loop / autonomous generation / autonomous verification / autonomous governance	full admission criteria, see Table 2
HIR / GSR / FPR / ASCR / ToE / CoC	human intervention rate / generation success rate / first-pass rate / abnormal self-closure rate /	criterion metrics, see Table 2

Abbreviation / term	Full name / definition	Notes
	generational escalation time / AI link coverage	
HIR_r / HIR_c	review-and-release intervention rate / correction-and-takeover intervention rate	the two sub-metrics of HIR, see Section 3.1
graded admission	—	grading the autonomy status of a line by the degree to which criteria are met, see Section 3.2
ZFS-FPR	Zero-Reference First-Success Rate	advanced extension metric of F2/F3 at the autonomous-orchestration stage, see Section 3.1
evidence-boundary declaration	—	the declaration mechanism that explicitly separates “implemented/expected” via Table 1, see Section 2.3
RMS / DT	Reconfigurable Manufacturing Systems / Digital Twin	relationship to the concepts of this paper, see Section 1.2

References

[1] Zhou J., Li P., et al. Toward a New Generation of Intelligent Manufacturing. Engineering, 2018, 4(1): 11-20. DOI: 10.1016/j.eng.2017.12.006.

[2] ISO/IEC 22989:2022. Information technology — Artificial intelligence — Artificial intelligence concepts and terminology. Geneva: ISO/IEC, 2022.

[3] Tao F, Zhang H, Zhang C. “Advancements and challenges of digital twins in industry”. Nature Computational Science, 2024, 4(3): 169-177.

[4] Tao F, Zhang H, Liu A, et al. “Make more digital twins”. Nature, 2019, 573(7775): 490-491.

[5] Leng J, et al. “Unlocking the power of industrial AI towards Industry 5.0”. Journal of Manufacturing Systems, 2024, 73: 349-363.

[6] Fan H, Liu X, Fuh J Y H, et al. “Embodied intelligence in manufacturing: leveraging large language models for autonomous industrial robotics”. Journal of Intelligent Manufacturing, 2025, 36(2): 1141-1157.

[7] Su X, Xu J, van Kaick O, et al. “IMR-LLM: Industrial multi-robot task planning and program generation using large language models”. In: Proceedings of the IEEE

International Conference on Robotics and Automation (ICRA 2026), Vienna, 2026. [in press]

[8] Liu R, et al. “Bridging the semantic gap: trajectory-guided domain repair for reliable planning”. Robotics and Computer-Integrated Manufacturing, 2026. DOI: 10.1016/j.rcim.2026.103290. [in press]

[9] Vemprala S, Bonatti R, Bucker A, et al. “ChatGPT for robotics: design principles and model abilities”. IEEE Access, 2024, 12: 55682-55696.

[10] Zhuang C, Wang H, Ding H. “AttentionVote: a coarse-to-fine voting network of anchor-free 6D pose estimation on point cloud for robotic bin-picking application”. Robotics and Computer-Integrated Manufacturing, 2024, 86: 102671.

[11] Zheng P., Li C., Yin Y., et al. Augmented-Reality-Assisted Mutual-Cognitive Human-Robot Safety Interaction System. Journal of Mechanical Engineering, 2023, 59(6): 173-184. DOI: 10.3901/JME.2023.06.173.

[12] Tao J., et al. Application Research Status and Frontiers of Manufacturing-Domain Knowledge Graphs. Computer Integrated Manufacturing Systems, 2022, 28(12): 3720-3736.

[13] Wang B, Zheng L, Wang Y, et al. “LLM-based multi-agent task planning for human-robot collaborative assembly balancing operator experience and efficiency”. Journal of Manufacturing Systems, 2025, 82: 1020-1045.

[14] Koren Y, Jovane F, Heisel U, et al. “Reconfigurable manufacturing systems”. CIRP Annals — Manufacturing Technology, 1999, 48(2).

[15] Qiao F., et al. Human-Centric Fusion Technologies for Production Scheduling in Industry 5.0. Journal of Mechanical Engineering, 2025, 61(15): 40-56.

[16] Zhang D., et al. A Data-Knowledge Hybrid-Driven Intelligent Control Architecture for Discrete Manufacturing Systems. Journal of Mechanical Engineering, 2024, 60(6): 1-10.

[17] Ahn M, Brohan A, Brown N, et al. “Do As I Can, Not As I Say: Grounding language in robotic affordances”. In: Proceedings of the 5th Conference on Robot Learning (CoRL 2021). arXiv:2204.01691, 2022.

[18] Liang J, Huang W, Xia F, et al. “Code as Policies: Language model programs for embodied control”. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2023), London, 2023.

[19] Huang W, Abbeel P, Pathak D, Mordatch I. “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents”. In: Proceedings of the 39th International Conference on Machine Learning (ICML 2022). PMLR 162: 9118-9147, 2022.

[20] Tobin J, Fong R, Ray A, et al. “Domain randomization for transferring deep neural networks from simulation to the real world”. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2017), Vancouver, 2017: 23-30.

[21] Zhao W, Queralta J P, Westerlund T. “Sim-to-real transfer in deep reinforcement learning for robotics: a survey”. In: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI 2020), Canberra, 2020: 737-744.

[22] ISO 10218:2011, ISO/TS 15066:2016. Robots and robotic devices — Safety requirements for industrial robots / Collaborative robots. Geneva: ISO.

[23] IEC 61508:2010. Functional safety of electrical/electronic/programmable electronic safety-related systems. Geneva: IEC.

[24] ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. Geneva: ISO/IEC.

[25] Wei J. AI Autonomous Governance Factory: Concept, Architecture, and Launch Pathway. engrXiv Preprint, 2026. DOI: 10.31224/7440.

[26] SAE International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (SAE J3016). SAE Standard, 2021.