

Standards-Aware Deep Reinforcement Learning for Adaptive Quality Control in Smart Manufacturing: A Controlled Simulation Study

Sunday Obokenuenu Emmanuel Inonoje, Tsetimi Jonathan², and Henrietta Ify Ojarikre²

Department of Mathematics, Faculty of Science, Southern Delta University,
Delta State, Nigeria

¹Department of Mathematics, Southern Delta University, Ozoro, Nigeria

²Department of Mathematics, Delta State University, Abraka, Nigeria

*Corresponding author: inononjeso@sdu.edu.ng; emmainonoje29@gmail.com

DOI: 10.29322/IJSRP.16.09.2026.p17728

<https://dx.doi.org/10.29322/IJSRP.16.09.2026.p17728>

Paper Received Date: 20h August 2026

Paper Acceptance Date: 25th September 2026

Paper Publication Date: 30th September 2026

Abstract

We present a standards-aware deep reinforcement learning (S-DRL) approach for AQC in smart manufacturing. The framework represents four heterogeneous process variables on a common standardized scale, uses a $\pm 3\sigma$ quality boundary, and incorporates PPO-based reward function deviations from that boundary. We conducted the comparison in a controlled synthetic manufacturing environment tracking temperature, pressure, production speed, and vibration under three distinct operating conditions: normal, moderate disturbance, and severe disturbance. To ensure statistical robustness, 30 independent runs were conducted per controlled scenario combination, resulting in a dataset of 270 run-level observations. Across the 90 evaluations per controller, S-DRL demonstrated superiority with the lowest mean defect rate (1.94%) and the highest mean standard compliance (95.77%), compared with PPO (3.10%/92.22%) and SPC (4.05%/88.40%). Factorial analysis of variance revealed significant main and interaction effects of both controller type and scenario on all outcomes. Welch-adjusted, bootstrap, trimmed mean, and Kruskal–Wallis analyses yielded the same controller ordering, confirming the robustness of the results. The findings demonstrate the value of integrating SARP in a simulated quality control framework. This stimulated result does not imply immediate industrial efficiency; bridging the gaps requires prospective evaluation or safety-constrained testing with DT environment.

Keywords: standardization; deep reinforcement learning; proximal policy optimization; statistical process monitoring; quality control; smart manufacturing; quality 4.0; adaptive control

1. Introduction

Continuous data generated across sensors, industrial networks, and automation software are increasingly dictating quality control in smart manufacturing. Managing these continuous data streams requires an advice control framework. This real-time influx enables a shift from passive monitoring to proactive, data-driven optimization. This shift redefines the control problem; detecting anomalies is insufficient because the controller must also determine the corrective action and intervention magnitude. Statistical process control (SPC) remains a foundational framework for detecting process deviations and supporting quality improvement (Montgomery, 2019). However, contemporary manufacturing environments present significant complexities. These include heterogeneous variables, nonlinear relationships, multiregime operations, high-dimensional measurements, and time-varying disturbances. Consequently, there is a critical need to integrate statistical monitoring with ML and SM rather than treating them as isolated paradigms.

1.1 Statistical Process and Smart Quality Control

Statistical process control (SPC) remains an important approach for identifying abnormal process variations through control charts. However, conventional SPC methods can become less effective when manufacturing data involve strong correlations among variables, temporal dependence, non-normal distributions, or several quality characteristics. These limitations have motivated the development of multivariate and data-oriented monitoring methods. For example, Ramos et al. (2021) reviewed multivariate SPC techniques for batch processes, including dimensionality-reduction and latent-variable approaches for handling complex process measurements. Related research on big-data manufacturing has also emphasized the need for statistical monitoring methods capable of processing large-scale, correlated, and continuously generated production data.

Recent work has examined the integration of AI with statistical process monitoring for extracting meaningful information from complex and high-dimensional data processes. Chang et al. (2026) provided a comprehensive review of AI-based statistical process monitoring, covering multivariate monitoring, pattern recognition, high-dimensional data analysis, and anomaly detection. They further identify the transition from conventional statistical process monitoring to smart process control when information can support corrective actions. However, an important unresolved question remains: how can statistical control signals be incorporated directly into the sequential decision-making process of a learning agent rather than being used solely for anomaly detection or post hoc diagnosis?

1.2 Quality 4.0 and Data-Driven manufacturing

Quality 4.0 integrates established quality management principles with industrial big data, connected manufacturing technologies, artificial intelligence (AI), and real-time process monitoring. Quality 4.0 promotes the continuous generation, collection, and analysis of manufacturing data rather than relying primarily on inspection after production to support faster quality-related decisions and process improvement. Escobar et al. (2021) associate this transition with the increasing use of industrial big data and artificial intelligence (AI) for process monitors and quality improvement, while also emphasizing the challenges created by complex, high-dimensional, and dynamic manufacturing environments.

This transition has an important methodological implication for adaptive quality control: quality information must be incorporated into the control process while production is occurring. Measurements obtained during runtime can provide information about deviations from desired operating conditions and enable corrective actions before process deterioration develops into nonconforming output. This perspective is consistent with recent developments in intelligent manufacturing, where learning-based control methods are increasingly being investigated for production systems characterized by uncertainty, variability, and changing operating conditions. Recent work, for example, demonstrates the use of DR for adaptive manufacturing control under stochastic production conditions.

Deep learning provides a means of extracting nonlinear patterns and representations from complex manufacturing data, particularly where analytically specifying relationships among process variables is difficult (Goodfellow et al., 2016). Reinforcement learning complements this capability by learning decision policies through interactions between an agent and its operating environment (Sutton & Barto, 2018). Therefore, the combination is relevant to manufacturing control problems in which process observations must be translated into sequential corrective decisions rather than merely classified as normal or abnormal. However, predictive capability alone does not guarantee quality-conforming control. An adaptive controller must also distinguish between acceptable and unacceptable process states and ensure that its actions are consistent with predefined quality requirements. This creates a specific opportunity for directly integrating quality boundaries into the learning and decision-making mechanism. Accordingly, this study builds on the Quality 4.0 paradigm by incorporating quality-standard information into a deep reinforcement learning framework, thereby linking real-time process observations, adaptive control decisions, and statistical quality constraints within unified control architecture.

1.3 Reinforcement learning in industrial and smart manufacturing processes

Reinforcement learning is relevant to manufacturing problems in which quality, throughput, maintenance, and process control decisions influence one another over time. Nian et al. (2020) reviewed reinforcement learning applications in industrial process control and discussed its adaptive potential and deployment difficulties. Li et al. (2023) examined deep RL in smart manufacturing, whereas Dogru et al. (2024) surveyed RL applications across process-industry tasks, including monitoring, fault detection, control, optimization, planning, and scheduling. Taken together, this literature supports the use of RL for AI decisions. This leaves a more specific design question open for this study: how should statistical quality limits affect the learning objective rather than serving only as monitoring thresholds?

1.4 Reinforcement learning for Quality and Integrated Manufacturing Decisions

There is also evidence that RL can be used when quality is coupled with other manufacturing decisions. For example, Paraschos et al. (2020) studied reinforcement-learning policies for production, maintenance, and quality control in a deteriorating manufacturing system. Their work illustrates the value of sequential decision-making when several operational objectives interact. However, for this problem, the representation of quality remains important. Therefore, the proposed S-DRL formulation places standardized process deviations and the explicit quality boundary inside the state and reward structure, allowing compliance information to influence policy learning.

1.5 Safety, Interpretability, and Barriers to Deployment

Moving an RL controller from simulation into an industrial setting introduces constraints that an unconstrained benchmark cannot fully capture. Gu et al. (2024) identified safety as a central issue in safe RL, while Nievas et al. (2024) discussed practical barriers to industrial RL, including model complexity, access to suitable environments, safety constraints, and interpretability. These concerns affect the design of a quality-control controller because a statistically desirable action may still be infeasible or unsafe. Therefore, this study treats the synthetic experiment as a controlled methodological test and does not interpret its results as evidence of readiness for deployment. Safety-constrained learning, human oversight, and plant-level testing remain validation stages.

1.6 Digital twins and external validation

Digital twins offer a possible route from historical manufacturing information to sequential experimentation because they can represent process dynamics and allow controlled what-if analysis. Aniba et al. (2024) demonstrated a digital-twin approach combining deep learning with manufacturing quality control, while Fu et al. (2025) reviewed digital-twin applications for

machining-quality characterization and optimization. A static observational dataset such as SECOM is different: it records process observations and quality labels but does not define how an intervention changes the next process state. Consequently, external validation of an RL controller requires leakage-safe preprocessing and an explicit transition mechanism, such as a calibrated simulator or digital twin, rather than treating the static dataset as direct evidence of control performance.

1.7 Research gaps, objectives, and contributions

The literature relevant to this study can be viewed as three connected but distinct strands. Statistical process monitoring provides established ways of defining acceptable process behavior; deep reinforcement learning provides mechanisms for sequentially selecting actions; and quality 4.0 and digital-twin research provide a broader setting in which data, models, and operational decisions can be connected. The specific gap addressed here concerns the interface between these strands: statistical quality limits are commonly used to detect or classify abnormal behavior, whereas the present formulation directly uses them in the learning state and reward. This study combines process-variable standardization, a $\pm 3\sigma$ operating boundary, PPO-based control and compliance-sensitive reward terms.

The empirical question is whether adding the standards-aware objective improves the simulated quality-control performance relative to conventional SPC and PPO without that objective. PPO is retained as a particularly useful comparator because both PPO and S-DRL use the same optimization mechanism; therefore, the intended contrast is the effect of incorporating the statistical quality information into the objective rather than a difference between unrelated RL algorithms.

2. Methodology

2.1 Simulation Environment

The experimental environment is a synthetic manufacturing process with four monitored variables: temperature, pressure, production speed, and vibration. The simulation follows the controlled-experimentation logic described by Law (2015), whereas the verification and validation principles discussed by Sargent (2013) motivate explicit documentation of the model and its assumptions. The environment is used only to compare the controller behavior under repeatable stochastic conditions. The process state is observed at each decision epoch, a controller selects an action, the process transition and disturbance are applied, and quality-related outcomes are returned.

2.2 State representation and data standardization

Let

$$\mathbf{x}_t = [x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t}]^T \quad (1)$$

where $x_{1,t}$, $x_{2,t}$, $x_{3,t}$ and $x_{4,t}$ denote, represent temperature, pressure, production speed, and vibration.

The state observed by the DRL agent at time t is expressed as follows:

$$\mathbf{s}_t = [z_{1,t}, z_{2,t}, z_{3,t}, z_{4,t}]^T \quad (2)$$

$$z_{jt} = \frac{x_{jt} - \mu_j}{\sigma_j}, \text{ for } j = 1, \dots, 4 \quad (3)$$

where;

$\mathbf{s}_t$ state vector at time t ,

z_{jt} represent each standardization value of the process variables j ,

μ_j and σ_j denote the mean and standard deviation of the process variable j , respectively,

x_{jt} observed value of the process variable j at time t , and

where $j = 1, \dots, 4$ represents the four processing quality variables in the model.

t = current decision time, production period, or stimulation step

T = transpose, so the four process quality variables can be arranged as column vectors.

Thus, the state vector provides the agent with a common scale representation of the current process conditions.

This transformation places the four variables on a comparable scale despite their different physical units and magnitudes. The same reference μ and σ values must be retained for both training and evaluation; changing them between stages would alter the quality boundary's meaning.

2.3 Action Space and Transition

The action space defines the distinct control decision available to the S-DRL agent with the following:

$$a_t \in -1, 0, +1 \quad (4)$$

where

-1 denote a decrease,

0 denotes no change, and

+1 denotes an increase.

At each decision step t_1 the S-DRL agent selects an action a_t that modifies the current process state x_t . The simulation environment then applies the selected action and computes the resulting process state x_{t+1} .

We represent this state transition by the general function: into an updated process state by the simulated process.

$$x_{t+1} = f(x_t, a_t, \varepsilon_t) \quad (5)$$

where

x_{t+1} denote resulting state after the action is applied,

where $f(\cdot)$ denotes the implemented synthetic dynamics,

x_t denotes current state of the process environment at time t ,

a_t denotes action selected by the DRL agent, and

ε_t denotes process disturbance or uncertainty.

2.4 The quality boundary

The primary quality boundary is set at $\pm 3\sigma$, following conventional statistical process-control practice (Montgomery, 2019).

A conventional three-standard-deviation operating boundary is adopted. A process variable is considered to be within the acceptable standardized region when the following occurs:

$$|z_{j,t}| \leq 3 \quad (6)$$

The aggregate violation magnitude is represented by the following:

$$V_t = \sum_j \max(0, |z_{j,t}| - 3)^2 \quad (7)$$

Thus, a small excursion produces a limited penalty, whereas increasingly large departures from the boundary receive progressively greater weight. This makes the statistical limit part of the learning signal rather than a threshold used only after an action has been taken.

2.5 Standards-Aware Reward

A manufacturing process monitors temperature, vibration, and production speed,

$$s_t = \begin{bmatrix} z_{temperature,t} \\ z_{pressure,t} \\ z_{speed,t} \\ z_{vibration,t} \end{bmatrix}$$

The DRL agent observes this state and determines the most appropriate quality control action.

S-DRL reward is defined as

$$R_t = -w_1 D_t - w_2 V_t - w_3 C_t + w_4 Q_t \quad (8)$$

where;

D_t denotes the deficit penalty,

$-w_1 D_t$ denotes reduced reward (the agent learns to minimize deficits),

V_t is the standard violation penalty,

$-w_2 V_t$ penalty agent for violating the specified quality standard,

C_t is the control cost penalty,

$-w_3 C_t$ represents discourage the agent for marking excessive interventions

Q_t is the quality compliance reward.

$w_4 Q_t$ represents the increase in the reward of the agent, and

w_1, w_2, w_3, w_4 are weighting coefficients

The weighting coefficients determine the relative importance of these components. PPO and S-DRL share the PPO optimization procedure; the distinguishing feature of S-DRL is the additional use of standardized deviation and compliance information in the objective.

2.6 PPO learning

PPO served as the RL optimization method, selected for its clipped objective function that constrains policy updates while maintaining the foundational policy-gradient mechanics established by Schulman et al. (2017). Although the source text prescribes 300 training episodes per PPO execution, it omits critical configuration details regarding the underlying NNN architecture and optimization schedules. The exact hyperparameters and structural configurations used in the implementation have been quantified and compiled in the baseline tables below to ensure absolute experimental reproducibility.

Table 1: Experimental configuration and requirements for reproducibility

Component	Reported specification	Submission requirement
Training episodes	300	Retention of code/configuration
Controllers	SPC, PPO, S-DRL	Exact implementations
Scenarios	3	Exact disturbance parameters
Runs/cell	30	Random seeds
Evaluations/controller	90	Run-level dataset
Total observations	270	Archive dataset
State variables	4	Exact reference μ and σ
Quality boundary	$\pm 3\sigma$	Threshold implementation
Transition model	Conceptual form only	Exact equations required
Reward weights	Not supplied	Exact numerical values required
PPO architecture	Not supplied	Exact network/hyperparameters required
SPC action rule	Not supplied	Exact rule required

2.7 Baselines and experimental design of the study

To evaluate and compare the three distinct control methodologies:

- Statistical Process Control (SPC): This serves as the traditional, non-learning baseline for conventional statistical quality control.
- PPO: Configured as the standard reinforcement learning baseline, operating without the proposed standards-aware reward modifications.
- S-DRL: The proposed approach builds upon the core PPO optimization framework by integrating standardized deviation constraints and compliance-sensitive penalties directly into the objective function.

The evaluation framework consists of three operational scenarios designed to test resilience: (S_1) baseline performance under normal operations, (S_2) robustness under moderate system disturbances, and (S_3) stability under severe environmental disturbances. To ensure statistical significance and reproducibility, we implemented a balanced $3 \times 3 \times 30$ factorial experimental design. Each controller-scenario configuration was evaluated across 30 independent stochastic runs, yielding 270 run-level experimental records.

Table 2: Scenario and replication design.

Scenario	Condition	Purpose	Runs/controller
S1	Normal operation	Baseline performance	30
S2	Moderate disturbance	Adaptability assessment	30
S3	Severe disturbance	Robustness assessment	30

2.8 Statistical analysis and reproducibility

The independent controller scenario run is treated as the experimental unit. A two-factor ANOVA was used to examine the effects of the controller, scenario, and interaction. Welch tests are used for within-scenario comparisons because the cell variances are not assumed to be equal, followed by Holm adjustment for the pairwise tests. Kruskal–Wallis tests provide a distribution-free sensitivity check, and Cohen’s d is used to describe the separation between controller means. Bootstrap estimates were also reported to assess the stability of the observed differences. The bootstrap resampling count or interval-construction procedure are

not documented in the current analysis materials; these details, together with the simulation equations, reward weights, PPO settings, random seeds, software versions, and analysis scripts, should be archived with the final submission.

3. Results

3.1 Defect Rate

S-DRL produced the lowest simulated defect rate in every scenario: 1.360%, 1.923%, and 2.530% in S₁-S₃, respectively. PPO produced 2.247%, 2.988%, and 4.070%, respectively, while SPC produced 2.806%, 3.990%, and 5.368%, respectively.

Table 3: Scenario-level defect-rate performance (n = 30 per cell).

Scenario	Controller	Mean	SD	95% CI
S1	SPC	2.806	0.148	[2.750,2.861]
S1	PPO	2.247	0.184	[2.178,2.316]
S1	S-DRL	1.360	0.118	[1.316,1.405]
S2	SPC	3.990	0.236	[3.902,4.079]
S2	PPO	2.988	0.234	[2.900,3.075]
S2	S-DRL	1.923	0.201	[1.848,1.998]
S3	SPC	5.368	0.387	[5.223,5.512]
S3	PPO	4.070	0.304	[3.956,4.184]
S3	S-DRL	2.530	0.213	[2.450,2.609]

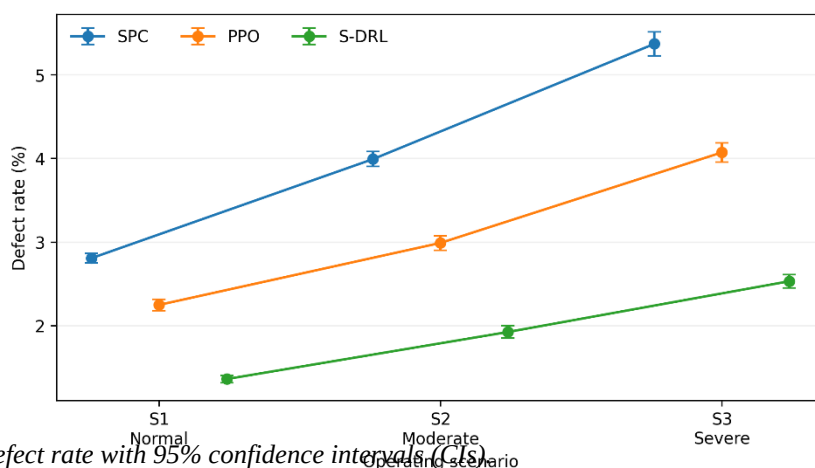


Figure 1: Simulated defect rate with 95% confidence intervals (CIs).

3.2 Standard Compliance

S-DRL achieved the highest standard compliance in every scenario: 97.085% in S₁, 96.068% in S₂, and 94.160% in S₃. PPO achieved 94.825%, 92.899%, and 88.936%, while SPC achieved 93.070%, 89.119%, and 83.022%.

Table 4: Scenario-level standard compliance (n = 30 per cell).

Scenario	Controller	Mean	SD	95% CI
S ₁	SPC	93.070	0.525	[92.874,93.266]
S ₁	PPO	94.825	0.649	[94.583,95.068]
S ₁	S-DRL	97.085	0.576	[96.870,97.301]
S ₂	SPC	89.119	0.798	[88.821,89.416]
S ₂	PPO	92.899	0.761	[92.615,93.184]
S ₂	S-DRL	96.068	0.489	[95.885,96.250]
S ₃	SPC	83.022	0.866	[82.699,83.345]
S ₃	PPO	88.936	0.936	[88.587,89.286]
S ₃	S-DRL	94.160	0.701	[93.898,94.421]

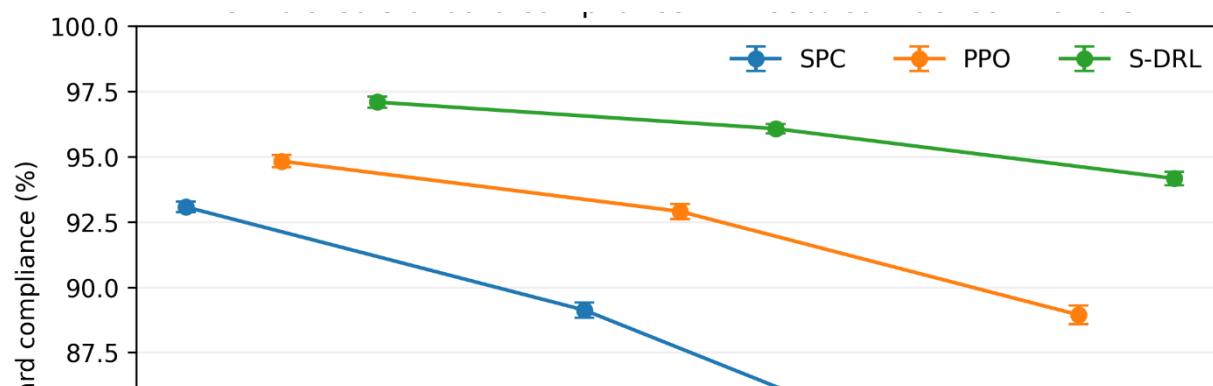


Figure 2: Simulated standard compliance with 95% confidence interval (CI).

3.3. Mean reward and overall performance

Table 5: Scenario-level mean reward.

Scenario	SPC	PPO	S-DRL
S ₁	-0.1593 ± 0.0057	-0.1531 ± 0.0054	-0.1385 ± 0.0039
S ₂	-0.1713 ± 0.0061	-0.1597 ± 0.0066	-0.1437 ± 0.0067
S ₃	-0.1917 ± 0.0087	-0.1715 ± 0.0082	-0.1487 ± 0.0081

Table 6: Overall controller performance across 90 controller evaluations

Controller	Evaluations	Defect rate (%)	Compliance (%)	Mean reward
SPC	90	4.05	88.40	-0.1741
PPO	90	3.10	92.22	-0.1614
S-DRL	90	1.94	95.77	-0.1436

Across the 90 evaluations for each controller, S-DRL recorded a simulated defect rate of 1.94%, 3.10% for PPO, and 4.05% for SPC. The mean compliance was 95.77%, compared with 92.22% for PPO and 88.40% for SPC. The corresponding reduction in defect rate was approximately 52.2% and 37.5% relative to SPC and PPO, respectively. Compliance was higher by 7.37 percentage points than SPC and by 3.55 points than PPO. These quantities describe separation within this simulation and should not be interpreted as expected industrial gains.

3.4 Inferential Validation

The factorial analysis detected statistically significant effects of controller, scenario, and controller-by-scenario interaction for both outcomes. For the defect rate, the controller effect was $F(2,261) = 1789.58$, $p < .001$, partial $\eta^2 = .932$; the scenario effect was $F(2,261) = 1369.50$, $p < .001$, partial $\eta^2 = .913$; and the interaction was $F(4,261) = 65.39$, $p < .001$, partial $\eta^2 = .501$. For compliance, the corresponding values were $F(2,261) = 2390.40$, $p < .001$, partial $\eta^2 = .948$ for the controller; $F(2,261) = 1782.33$, $p < .001$, partial $\eta^2 = .932$ for the scenario; and $F(4,261) = 190.25$, $p < .001$, partial $\eta^2 = .745$ for the interaction. The interaction terms show that the size of the controller differences changes as the severity of the disturbance increases.

Table 7: Factorial analysis of variance summary.

Outcome	Effect	Df	F	p	Partial η^2
Defect	Controller	2	1789.58	< .001	0.932
Defect	Scenario	2	1369.50	< .001	0.913
Defect	Controller × Scenario	4	65.39	< .001	0.501
Compliance	Controller	2	2390.40	< .001	0.948
Compliance	Scenario	2	1782.33	< .001	0.932
Compliance	Controller × Scenario	4	190.25	< .001	0.745

Partial η^2 is reported for the substantive model terms rather than the residual component. The magnitude of the reported effects is unusually large because the simulation produces relatively narrow within-cell variation while maintaining clear separation

between controller means. Therefore, these values should be read as measures of separation under the specified simulation conditions, not as estimates of the expected effect in an industrial population.

3.5 Heteroscedastic and pairwise confirmation

The Welch tests detected controller differences for both defect rate and compliance within each scenario (all $p < .001$). The Holm-adjusted pairwise comparisons were also significant for every controller pair. In S_3 , for example, S-DRL reduced the defect rate by 1.540 percentage points relative to PPO and by 2.838 points relative to SPC. Its compliance advantage was 5.223 points over PPO and 11.138 points over SPC. The same comparison direction was observed in S_1 and S_2 .

3.6 Effect size

Table 8: S-DRL versus baseline effect sizes

Scenario	Metric	Contrast	Mean difference	Cohen's d
S_1	Defect	S-DRL vs. PPO	-0.886	-5.74
S_1	Defect	S-DRL vs. SPC	-1.445	-10.80
S_1	Compliance	S-DRL vs. PPO	+2.260	+3.68
S_1	Compliance	S-DRL vs. SPC	+4.015	+7.28
S_2	Defect	S-DRL vs. PPO	-1.065	-4.88
S_2	Defect	S-DRL vs. SPC	-2.068	-9.44
S_2	Compliance	S-DRL vs. PPO	+3.168	+4.95
S_2	Compliance	S-DRL vs. SPC	+6.949	+10.50
S_3	Defect	S-DRL vs. PPO	-1.540	-5.86
S_3	Defect	S-DRL vs. SPC	-2.838	-9.08
S_3	Compliance	S-DRL vs. PPO	+5.223	+6.32
S_3	Compliance	S-DRL vs. SPC	+11.138	+14.14

Cohen's d estimates are large across all controller contrasts. They indicate pronounced separation among the simulated controller outcomes, but they should not be presented as direct estimates of the magnitude of the industrial effect. Manufacturing data collected under uncontrolled operating conditions would normally contain additional variability, measurement error, process drift, and unmodelled disturbances.

3.7 Sensitivity and robustness

Table 9: Stress Sensitivity from S_1 to S_3

Controller	Metric	S_1	S_3	Absolute change	Relative change
SPC	Defect	2.806	5.368	+2.562	+91.32%
PPO	Defect	2.247	4.070	+1.823	+81.14%
S-DRL	Defect	1.360	2.530	+1.169	+85.95%
SPC	Compliance	93.070	83.022	-10.048	-10.80%
PPO	Compliance	94.825	88.936	-5.889	-6.21%
S-DRL	Compliance	97.085	94.160	-2.925	-3.01%

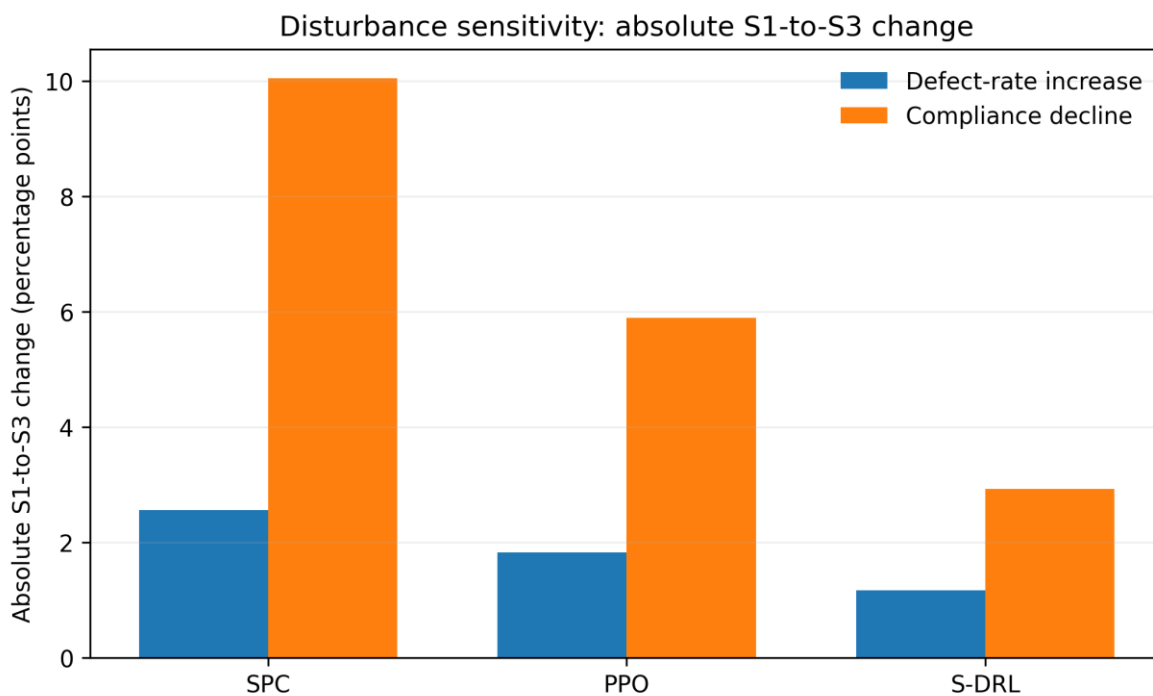


Figure 3: Absolute S1-to-S3 disturbance sensitivity. Changes are expressed as percentage points.

The disturbance-sensitivity analysis is best interpreted using the absolute changes in the two quality measures. From S1 to S3, S-DRL increased the defect rate by 1.169 percentage points, compared with increases of 1.823 points for PPO and 2.562 points for SPC. Its compliance fell by 2.925 percentage points, whereas PPO and SPC fell by 5.889 and 10.048 percentage points, respectively. Although relative percentage changes can produce a different ordering for some measures, the absolute changes show that S-DRL experienced the smallest deterioration while retaining the lowest defect rate and highest compliance. Therefore, the robustness claim is restricted to absolute deterioration within this simulation.

Table 10: 10% trimmed-mean sensitivity

Scenario	Metric	SPC	PPO	S-DRL	Ranking
S1	Defect	2.806	2.240	1.362	S-DRL > PPO > SPC
S1	Compliance	93.126	94.841	97.059	S-DRL > PPO > SPC
S2	Defect	3.986	2.991	1.918	S-DRL > PPO > SPC
S2	Compliance	89.129	92.905	96.071	S-DRL > PPO > SPC
S3	Defect	5.411	4.086	2.530	S-DRL > PPO > SPC
S3	Compliance	83.015	88.950	94.109	S-DRL > PPO > SPC

After removing the lowest and highest 10% of observations within each controller-scenario cell, the controller ordering remained unchanged. This check reduces the influence of observations at the tails of the simulated distributions and indicates that some extreme runs do not drive the reported ordering.

Table 11: Kruskal–Wallis sensitivity test results

Scenario	Metric	H	P
S ₁	Defect	78.684	< .001
S ₁	Compliance	78.163	< .001
S ₂	Defect	79.121	< .001
S ₂	Compliance	79.121	< .001

S ₃	Defect	77.739	< .001
S ₃	Compliance	79.121	< .001

All six Kruskal–Wallis tests were significant at $p < .001$. Therefore, the nonparametric results agree with the factorial and Welch analyses in indicating differences among the three controllers within each scenario.

3.8 Bootstrap assessment

The bootstrap results followed the same pattern as the main analysis. Relative to SPC, the estimated defect-rate reductions were approximately 51.5%, 51.8%, and 52.9% in S1, S2, and S3, respectively; relative to PPO, the corresponding reductions were approximately 39.5%, 35.6%, and 37.8%. The compliance gains relative to SPC were 4.31%, 7.80%, and 13.42%, whereas the gains relative to PPO were 2.38%, 3.41%, and 5.87%, respectively. The reported bootstrap intervals remained on the favorable side of zero. The final supplement should state the number of resamples and the exact interval construction used.

3.9 Threshold sensitivity boundary

The available 270-run outcome file does not contain the underlying time-step states or z-scores required to recompute performance under alternative boundaries. Consequently, $\pm 2\sigma$, $\pm 2.5\sigma$, $\pm 3.5\sigma$, and $\pm 4\sigma$ results are not estimated here. Reporting such counterfactual results from the available summary outcomes is not methodologically defensible. A proper threshold-sensitivity study should rerun the simulator at each boundary while holding controller definitions, scenario parameters, training settings, randomization procedures, and the 30-replication structure constant.

4. Discussion

4.1 Core Methodological Contribution of the Study

The main methodological feature of S-DRL is that the learning process incorporates the statistical quality boundary. Standardization converts the four process measurements to a common deviation scale; the $\pm 3\sigma$ limit defines the acceptable region; and the violation term rewards excursions outside that region. Compliance is also rewarded. Since PPO and S-DRL use the same PPO optimization mechanism, the comparison isolates the contribution of this standards-aware reward design more clearly than a comparison between unrelated learning algorithms.

4.2 Position Relative to Previous Research

The proposed formulation sits alongside, rather than replace, conventional SPC. He and Wang (2018) addressed statistical monitoring in manufacturing big-data settings; Ramos et al. (2021) and Ueda et al. (2022) reviewed multivariate SPC methods and industrial control charts; and Chang et al. (2026) examined the growing role of artificial intelligence (AI) in statistical process monitoring. The RL literature reviewed by Nian et al. (2020), Li et al. (2023), and Dogru et al. (2024) establishes the relevance of RL for industrial process decisions. This study connects these two bodies of work by allowing statistical deviation from a quality boundary to affect action selection. This is also consistent with the emphasis on using industrial data to support operational decisions in Quality 4.0 (Escobar et al., 2021).

4.3 Disturbance robustness

The controller-by-scenario interaction and the unchanged controller ranking indicate that disturbance severity affects the size of the performance differences. S-DRL did not remove the effect of disturbance: its defect rate still increased and compliance declined from S1 to S3. Its advantage was that the absolute deterioration was smaller than **that of** the two comparison controllers, while its final defect rate remained the lowest and compliance remained the highest. Therefore, the term “robustness” is used here in this limited, simulation-based sense.

4.4 Interpretation of large effect sizes

The unusually large Cohen’s d values are a consequence of the strong separation between controller means relative to the synthetic environment’s within-cell variation. They demonstrate that the controllers behave differently under the specified simulation assumptions, but they should not be interpreted as forecasts of IESs. Validation using real manufacturing data would be expected to introduce substantially more heterogeneity and, therefore, may produce smaller effects.

4.5 Implications of smart manufacturing

One possible implementation pathway would be to place the learned policy inside a digital twin or otherwise validated process simulator, with a supervisory layer checking action feasibility and safety before any physical intervention. Aniba et al. (2024) illustrated the use of deep learning with digital-twin-based quality control, and Fu et al. (2025) reviewed digital twins for machining-quality characterization and optimization. Gu et al. (2024) further emphasized the need for explicit safety constraints in RL deployment. Although these studies support the feasibility of such a future architecture, they do not demonstrate that this synthetic S-DRL controller has been validated in an industrial plant.

5. External and Real-Data Validation Strategies

The SECOM dataset is a potential external dataset because it contains manufacturing process measurements and quality labels (McCann & Johnston, 2008). However, its observational structure does not provide the intervention-to-next-state relationship required by a sequential controller. Therefore, a credible external study would need preprocessing fitted only to the training data, followed by an explicit transition model or calibrated digital twin. Then, SPC, statistical learning baselines, PPO, and S-DRL should be evaluated under the same data partition and operating assumptions. The F1 score, balanced accuracy, and Matthews correlation coefficient can complement control-oriented measures, such as defect rate, compliance, intervention frequency,

reward, and process cost, where the data support them. A second manufacturing dataset would strengthen the evidence only if its process variables and action semantics can be aligned without information leakage.

6. Limitations

- i. The environment is synthetic and cannot establish external industrial validity.
- ii. Exact transition equations, disturbance distributions, reward weights, PPO architecture/hyperparameters, random seeds, and the SPC action rule are not provided in the manuscript and must be documented from the implementation.
- iii. Additional modern adaptive control, model-based, and safe RL baselines would strengthen the benchmarking.
- iv. The safety constraints, action feasibility, uncertainty management, and human supervision are not fully represented.
- v. Numerous effect sizes are simulation-specific and should not be directly generalized to industry.
- vi. The supplied convergence figure cannot be treated as a formal statistical convergence test until its aggregation is documented across independent runs.
- vii. Alternative quality boundary experiments cannot be inferred from the available outcome records.
- viii. Real-data validation remains a future work and cannot be inferred from the synthetic experiment.

7. Conclusion

This study evaluated a standards-aware deep reinforcement learning formulation for AQC using 270 run-level observations from a controlled synthetic manufacturing environment. S-DRL produced the lowest defect rate and highest standard compliance across the three disturbance conditions, and the controller-by-scenario analysis showed that the size of these differences changed with the severity of the disturbance. The results of the trimmed-mean, nonparametric, Welch-adjusted, and bootstrap analyses were consistent with the primary controller ranking. The evidence supports the proposed reward formulation as a promising computational approach for SAQC. It does not establish industrial effectiveness. The simulation must be fully reproducible before such a conclusion can be considered. The controller should be compared with stronger adaptive and safe-RL baselines, process dynamics should be calibrated against manufacturing data, and the policy should undergo safety-constrained and prospective industrial evaluation.

8. Research Statements

8.1 Data availability

This study uses a 270-record synthetic run-level dataset with 30 observations in each of nine controller-scenario cells. The complete dataset and statistical workbook should be submitted as supplementary material or deposited in a persistent repository. They must not be regarded as real industrial measurements.

8.2 Ethics

This study uses synthetic computational data and does not involve human participants, identifiable personal data, or animal experimentation. The final ethics statement should be aligned with the requirements of the target journal.

References

- Aniba, Y., Bouhedda, M., Bachene, M., Rahim, M., Benyezza, H., & Tobbal, A. (2024). Digital twin-enabled quality control through deep learning in industry 4.0: A framework for enhancing manufacturing performance. *International Journal of modeling and Simulation*, 114, e01395899. <https://doi.org/10.1080/02286203.2024.2395899>
- Chang, S. I., Ghafariasl, P., & Aryal, B. (2026). A review of artificial intelligence impacting statistical process monitoring and future directions. *Computers and Industrial Engineering*, 216, 111989. <https://doi.org/10.1016/j.cie.2026.111989>
- Dogru, O., Xie, J., Prakash, O., Chiplunkar, R., Soesanto, J., Chen, H., Velswamy, K., Ibrahim, F., & Huang, B. (2024). Reinforcement learning in process industries: Review and perspective. *IEEE/CAA Journal of Automatica Sinica*, 11(2), 283–300. <https://doi.org/10.1109/JAS.2024.124227>
- Escobar, C. A., McGovern, M. E., & Morales-Menendez, R. (2021). Quality 4.0: A review of big data challenges in manufacturing. *Journal of Intelligent Manufacturing*, 32, 2319–2334. <https://doi.org/10.1007/s10845-021-01765-4>
- Fu, X., Li, S., Song, H., & Lu, Y. (2025). Digital twin-driven multi-scale characterization of machining quality: Current status, challenges, and future perspectives. *Robotics and Computer-Integrated Manufacturing*, 93, 102902. <https://doi.org/10.1016/j.rcim.2024.102902>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A. (2024). A review of safe reinforcement learning: Methods, theories, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12), 11216–11235. <https://doi.org/10.1109/TPAMI.2024.3457538>
- Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill Education.
- Li, C., Zheng, P., Yin, Y., Wang, B., & Wang, L. (2023). Deep reinforcement learning in smart manufacturing: A review and prospects. *CIRP Journal of Manufacturing Science and Technology*, 40, 75–101. <https://doi.org/10.1016/j.cirpj.2022.11.003>
- McCann, M., & Johnston, A. (2008). *SECOM [Dataset]*. UCI Machine Learning Repository. <https://doi.org/10.24432/C54305>
- Montgomery, D. C. (2019). *Introduction to statistical quality control* (8th ed.). Wiley.
- This publication is licensed under Creative Commons Attribution CC BY.
10.29322/IJSRP.16.09.2026.p17728

- Nievas, N., Pagès-Bernaus, A., Bonada, F., Echeverria, L., & Domingo, X. (2024). Reinforcement learning for autonomous process control in Industry 4.0: Advantages and challenges. *Applied Artificial Intelligence*, 4(1), e383101. <https://doi.org/10.1080/08839514.2024.2383101>
- Nian, R., Liu, J., & Huang, B. (2020). A review on reinforcement learning: Introduction and applications in industrial process control. *Computers and Chemical Engineering*, 139, 106886. <https://doi.org/10.1016/j.compchemeng.2020.106886>
- Paraschos, P. D., Koulinas, G. K., & Koulouriotis, D. E. (2020). Reinforcement learning for combined production-maintenance and quality control of a manufacturing system with deterioration failures. *Journal of Manufacturing Systems*, 56, 470–483. <https://doi.org/10.1016/j.jmsy.2020.07.004>
- Ramos, M., Ascencio, J., Hinojosa, M. V., Vera, F., Ruiz, O., Jimenez-Feijoó, M. I., & Galindo, P. (2021). Multivariate statistical process control methods for batch production: A review focused on applications. *Production and Manufacturing Research*, 9(1), 33–55. <https://doi.org/10.1080/21693277.2020.1871441>
- Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7(1), 12–24. <https://doi.org/10.1057/jos.2012.20>
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv. doi: 10.48550/arXiv.1707.06347
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
- Ueda, R. M., Agostino, Í. R. S., & Souza, A. M. (2022). Analysis and perspectives on multivariate statistical process control charts used in the industrial sector: A systematic literature review. *Management and Production Engineering Review*, 13(2), 48–60. <https://doi.org/10.24425/mper.2022.142054>

Supplementary Appendices

Appendix A. Descriptive statistics and confidence intervals (CIs)

Table A1. Complete the cell-level descriptive statistics.

Scenario	Controller	N	Defect mean	Defect SD	Defect 95% CI	Compliance mean	Compliance SD	Compliance 95% CI
S ₁	SPC	30	2.806	0.148	[2.750,2.861]	93.070	0.525	[92.874,93.266]
S ₁	PPO	30	2.247	0.184	[2.178,2.316]	94.825	0.649	[94.583,95.068]
S ₁	S-DRL	30	1.360	0.118	[1.316,1.405]	97.085	0.576	[96.870,97.301]
S ₂	SPC	30	3.990	0.236	[3.902,4.079]	89.119	0.798	[88.821,89.416]
S ₂	PPO	30	2.988	0.234	[2.900,3.075]	92.899	0.761	[92.615,93.184]
S ₂	S-DRL	30	1.923	0.201	[1.848,1.998]	96.068	0.489	[95.885,96.250]
S ₃	SPC	30	5.368	0.387	[5.223,5.512]	83.022	0.866	[82.699,83.345]
S ₃	PPO	30	4.070	0.304	[3.956,4.184]	88.936	0.936	[88.587,89.286]
S ₃	S-DRL	30	2.530	0.213	[2.450,2.609]	94.160	0.701	[93.898,94.421]

Appendix B. Full factorial analysis of variance

Table B1. Defect-rate ANOVA.

Effect	SS	Df	F	p	Partial η^2
Controller	202.3081	2	1789.582	< .001	0.932
Scenario	154.8184	2	1369.496	< .001	0.913
Controller × Scenario	14.7855	4	65.395	< .001	0.501
Residual	14.7527	261	—	—	—

Table B2. Compliance ANOVA.

Effect	SS	Df	F	p	Partial η^2
Controller	2443.4920	2	2390.404	< .001	0.948
Scenario	1821.9135	2	1782.330	< .001	0.932
Controller × Scenario	388.9424	4	190.246	< .001	0.745
Residual	133.3982	261	—	—	—

Appendix C. Welch- and Holm-adjusted pairwise results

Table C1. Welch tests within the scenario.

Scenario	Metric	Welch F	df1	df2	p
S ₁	Defect	901.101	2.0	56.229	< .001
S ₁	Compliance	393.352	2.0	57.579	< .001
S ₂	Defect	667.387	2.0	57.647	< .001
S ₂	Compliance	848.878	2.0	54.859	< .001
S ₃	Defect	700.854	2.0	54.681	< .001
S ₃	Compliance	1491.033	2.0	57.071	< .001

Table C2. Holm-adjusted pairwise Welch contrasts were obtained.

Scenario	Metric	Contrast	Mean difference	95% CI	df	t	p
S ₁	Defect	SPC vs.	0.559	[0.472,0.645]	55.46	12.96	< .001

		PPO					
S ₁	Defect	SPC vs. S-DRL	1.445	[1.376,1.514]	55.27	41.82	< .001
S ₁	Defect	PPO vs. S-DRL	0.886	[0.806,0.967]	49.42	22.22	< .001
S ₂	Defect	SPC vs. PPO	1.003	[0.881,1.124]	58.00	16.51	< .001
S ₂	Defect	SPC vs. S-DRL	2.068	[1.954,2.181]	56.54	36.54	< .001
S ₂	Defect	PPO vs. S-DRL	1.065	[0.952,1.178]	56.66	18.89	< .001
S ₃	Defect	SPC vs. PPO	1.298	[1.117,1.478]	54.94	14.43	< .001
S ₃	Defect	SPC vs. S-DRL	2.838	[2.675,3.000]	45.10	35.17	< .001
S ₃	Defect	PPO vs. S-DRL	1.540	[1.404,1.676]	51.92	22.70	< .001
S ₁	Compliance	SPC vs PPO	-1.755	[-2.060,-1.449]	55.56	-11.51	< .001
S ₁	Compliance	SPC vs. S-DRL	-4.015	[-4.300,-3.730]	57.50	-28.21	< .001
S ₁	Compliance	PPO vs. S-DRL	-2.260	[-2.578,-1.943]	57.20	-14.26	< .001
S ₂	Compliance	SPC vs PPO	-3.781	[-4.184,-3.378]	57.87	-18.79	< .001
S ₂	Compliance	SPC vs S-DRL	-6.949	[-7.293,-6.606]	48.12	-40.67	< .001
S ₂	Compliance	PPO vs S-DRL	-3.168	[-3.500,-2.836]	49.49	-19.18	< .001
S ₃	Compliance	SPC vs PPO	-5.915	[-6.380,-5.449]	57.65	-25.42	< .001
S ₃	Compliance	SPC vs S-DRL	-11.138	[-11.545,-10.730]	55.60	-54.77	< .001
S ₃	Compliance	PPO vs S-DRL	-5.223	[-5.651,-4.795]	53.75	-24.47	< .001

Appendix D. Sensitivity and bootstrap evidence

Table D1. Trimmed mean sensitivity.

Scenario	Metric	SPC	PPO	S-DRL	Ranking
S ₁	Defect	2.806	2.240	1.362	S-DRL > PPO > SPC
S ₁	Compliance	93.126	94.841	97.059	S-DRL > PPO > SPC

S ₂	Defect	3.986	2.991	1.918	S-DRL > PPO > SPC
S ₂	Compliance	89.129	92.905	96.071	S-DRL > PPO > SPC
S ₃	Defect	5.411	4.086	2.530	S-DRL > PPO > SPC
S ₃	Compliance	83.015	88.950	94.109	S-DRL > PPO > SPC

Table D2. The bootstrap estimates are reported in the supplied analysis.

Scenario	Metric	Baseline	Difference	Improvement	Bootstrap 95% CI
S ₁	Defect	PPO	-0.886	39.45%	[-0.966,-0.809]
S ₁	Defect	SPC	-1.445	51.51%	[-1.511,-1.377]
S ₁	Compliance	PPO	+2.260	2.38%	[1.960,2.573]
S ₁	Compliance	SPC	+4.015	4.31%	[3.738,4.294]
S ₂	Defect	PPO	-1.065	35.64%	[-1.171,-0.956]
S ₂	Defect	SPC	-2.068	51.81%	[-2.179,-1.959]
S ₂	Compliance	PPO	+3.168	3.41%	[2.850,3.482]
S ₂	Compliance	SPC	+6.949	7.80%	[6.618,7.284]
S ₃	Defect	PPO	-1.540	37.84%	[-1.669,-1.409]
S ₃	Defect	SPC	-2.838	52.87%	[-2.991,-2.680]
S ₃	Compliance	PPO	+5.223	5.87%	[4.824,5.630]
S ₃	Compliance	SPC	+11.138	13.42%	[10.743,11.534]

The bootstrap intervals above are the mean-difference scale intervals reported in the supplied analysis. The relative improvement percentages are presented separately.

Appendix E. Quality-boundary sensitivity protocol

The primary specification is $\pm 3\sigma$. Alternative thresholds cannot be estimated from the outcome records. A valid sensitivity experiment should rerun the simulator at $\pm 2\sigma$, $\pm 2.5\sigma$, $\pm 3\sigma$, $\pm 3.5\sigma$, and $\pm 4\sigma$ with identical controller definitions, scenario parameters, training settings, randomization procedure, and replication structure.

Table E1. Boundary-sensitivity plan.

Boundary	Current status	Required action
$\pm 2\sigma$	Not available	Rerun simulator
$\pm 2.5\sigma$	Not available	Rerun simulator
$\pm 3\sigma$	Primary experiment	Retain reference
$\pm 3.5\sigma$	Not available	Rerun simulator
$\pm 4\sigma$	Not available	Rerun simulator

Appendix F. Reproducibility checklist

- Archive the complete 270-run dataset.
- Provide exact state-transition equations and disturbance distributions.
- Provide standardization parameters and pre-processing rules.
- Provide numerical reward weights and definitions of reward components.
- Provide the PPO architecture and all hyperparameters.
- Provide random seeds and an independent-run generation procedure.
- Provide the exact SPC corrective action rule.
- Document the bootstrap resampling count and confidence interval construction.
- Archive the training trajectories if a convergence figure is retained.
- Archive code, environment specifications, and analysis scripts in a persistent repository.
- Keep synthetic computational evidence clearly separated from external and industrial evidence.

Appendix G. Simulation Integrity Statement

All numerical outcomes are the computational results from the synthetic evaluation dataset supplied. These are not measurements from an industrial plant, SECOM, NASA, or another real manufacturing facility. Subsequent real-data validation must use a prespecified development-evaluation split, leakage-safe preprocessing, and an explicit sequential transition mechanism.