

TELEMETRY-BASED IDENTIFICATION OF DECISION PATTERNS AFFECTING PRODUCTION SYSTEM PERFORMANCE DURING ORGANIZATIONAL TRANSFORMATION

Vladislav V. Tereshchenko
Elena O. Kasianenko
Sergey V. Salkutsan¹
Yury S. Klochkov

Received 13.05.2026.
Revised 19.08.2026.
Accepted 15.09.2026.

Keywords:

Organizational systems, Adverse managerial decision patterns, Simulation-based decision sessions, Telemetry, Identification of behavioral models, Transformation readiness



ABSTRACT

Organizational transformation projects fail predominantly for managerial rather than technical reasons, while established risk assessment relies on expert judgment. This paper builds a typology of adverse managerial decision patterns in production-type organizational systems from the telemetry of simulation-based decision sessions on a discrete-event model. Two corpora from one simulation platform family cover 1.13 million events of 691 teams. Candidate patterns are operationalized as predicates over timed decision sequences and validated with permutation tests under a false-discovery-rate correction, hold-out, leave-one-game-out and cross-cohort checks, and session-structure and activity-level controls. Eight of ten outcome-testable candidates passed statistical screening; three replicate across cohorts and scenario revisions. Verdicts hold on an alternative criterion (service level), and accounting for decision timing reverses the verdict on one pattern. The number of diagnosed patterns is monotonically associated with lower median profit; transfer beyond the simulation environment remains to be tested.

© 2026 Published by Faculty of Engineering

1. INTRODUCTION

Transformations of production-type organizational systems, that is, programs that change the structure of the system (the composition and topology of work centers, routes, team roles) and its control parameters (target indicators, batch sizes, buffers, operating policies), are among the most common and most failure-prone undertakings in industrial management, with outcomes that vary systematically with firm size, product innovation, and production type (Wu et al., 2024). Survey evidence indicates that while most plants deploy lean-

type tools, only a marginal fraction report full achievement of the declared goals (Pay, 2008); headline figures of this kind warrant caution, since the reproduced "70 percent of change initiatives fail" claims lack a consistent evidential basis (Hughes, 2011; Oludapo et al., 2024). What does replicate is qualitative: failure concentrates in the organizational, managerial, and behavioral domain rather than the technical one: communication and management support are decisive (Worley & Doolen, 2006), leadership, management, financial capability, culture, and skills top the critical-success-factor lists for SMEs (Achanga et al., 2006),

¹ Corresponding author: Sergey V. Salkutsan
Email: salkutsan@spbstu.ru

personal top-management involvement ranks first across plants (Netland, 2016), and the same managerial factors persist in digital transformation (Chirumalla et al., 2025). The recurring deficit is uniform: insufficient analysis of the mutual influence of concurrent changes, neglect of the behavioral properties of managerial decision making, and the absence of quantitative risk estimates at the design stage.

The instruments used in practice to assess transformation risks, namely failure mode and effect analysis (FMEA), value stream mapping, and the structured expert techniques catalogued in the international risk assessment standard IEC 31010 (International Electrotechnical Commission [IEC], 2019), share three systematic shortcomings with respect to the managerial and behavioral component of risk. First, the frequency and severity of risk realizations are estimated by expert judgment rather than by mass observation of the actual behavior of management teams. Second, the field data that could replace expertise are unusable for a principled reason: production contexts are incomparable (different systems, projects, and markets), and the decisions of management teams are not recorded in a reproducible form that would be comparable across cases. Third, the behavioral component of risk is not operationalized as a measurable quantity: the methods produce a qualitative list of "what can go wrong" but not a quantitative typology with frequencies and a measured cost to the outcome. Removing all three shortcomings requires empirical material of a different kind: the telemetry of mass simulation-based decision sessions on a controlled model, used as a complement to, not a replacement of, the established expert toolkit.

A controlled environment that removes these limitations is provided by simulation-based decision sessions: a management team operates a discrete-event simulation model over a multi-period horizon, and every decision is recorded automatically with a timestamp and period reference. In the terms of the theory of control in organizational systems (Burkov et al., 2015; Novikov, 2013; Lazić et al 2025), the team is a collective active element and the session is an identification experiment: identical initial conditions make behavior comparable, and telemetry yields reproducible evidence about the behavioral models of active elements that is unattainable by expert survey or field observation. Compared with self-reports and interviews, the implicit fixation of decisions during work substantially reduces self-report bias and strategic misrepresentation.

The purpose of this paper is to construct a typology of adverse managerial decision patterns in organizational systems (Singh et al., 2024) undergoing structural and parametric transformation, identified from the telemetry of simulation-based decision sessions, with a reproducible operationalization of classes, frequency estimates, and validation of the link between each class and the financial outcome. Throughout the paper, "error

class" is used as a defined shorthand for a telemetry-based adverse decision pattern; its interpretation as a managerial error in the full sense is a hypothesis that requires construct validation by independent experts, which is planned but not performed here (Section 5.2). The contributions are as follows: (a) fourteen candidate classes are operationalized as testable predicates over timed decision sequences, computable from telemetry without observer participation; (b) on two corpora (Section 3.1), eight of the ten fully outcome-testable classes pass the statistical screening under a false-discovery-rate correction, and a core of three classes replicating across cohorts and scenario revisions is isolated, with an explicit gradation of the evidence behind each class; (c) it is demonstrated that accounting for decision timing can reverse the verdict on a class relative to accounting for the mere fact of a decision; (d) a monotonic association between the number of diagnosed classes and median profit is established. The paper deliberately limits its claim: the typology describes the decisions of teams in a controlled simulation environment, and the replication achieved is cross-cohort and cross-scenario within one simulation platform family; transfer of the classes to production contexts is a hypothesis to be tested, and measuring the readiness of a specific team is the subject of separate work.

Section 2 reviews related work; Section 3 describes the data, formal statement, candidate classes, and validation; Section 4 presents the results; Section 5 discusses transferability, limitations, and integration into the control loop; Section 6 concludes.

2. RELATED WORK

2.1 Risk assessment and quality management methods

The mainstream of transformation risk assessment rests on structured expert techniques. IEC 31010 (IEC, 2019) catalogues several dozen methods, from FMEA and HAZOP to Delphi-type elicitation, whose common trait is that likelihoods and consequences are supplied by expert judgment. The same reliance on expert estimates and aggregated indicators marks the quality management line: Klochkov and Tveryakov (2020) show that the effectiveness of quality management methods hinges on how expert procedures are organized and standardized, and Klochkov et al. (2016) assess the speed of organizational development through standards efficiency, which is an aggregate, document-level indicator. These methods answer "how mature is the organization" at the level of aggregated assessments but do not observe the individual decision: its content, timing, or cost to the outcome. The present work is complementary: telemetry of simulation-based sessions supplies exactly the missing decision-level evidence, from which frequencies and effect sizes of specific errors can be estimated directly rather than elicited.

2.2 Behavioral evidence on managerial decision errors

The behavioral analysis of managerial decisions in controlled simulations goes back to Sterman (1989), who demonstrated systematic misperceptions of feedback, and to the gaming-simulation tradition of Shubik (1975); a recent review confirms business simulation games as an active method for developing managerial decision making (Velez & Alonso, 2025), learning analytics on simulation logs links behavioral indicators to outcomes (Hernández-Lara et al., 2019), and team cognition has been reframed for simulation-embedded measurement (Evans et al., 2025). Behavioral operations research confirmed that decision errors in operations are systematic rather than random: Croson and Donohue (2006) isolated behavioral causes of the bullwhip effect, and Bendoly et al. (2006) systematized the model assumptions that break down under actual human behavior; the field lean literature of Section 1 (Worley & Doolen, 2006; Achanga et al., 2006; Netland, 2016) converges on the same conclusion. This literature establishes the existence and direction of systematic errors but provides no reproducible typology built from mass decision-level telemetry, with class frequencies, confidence intervals, and per-class validation against outcome; to our knowledge none has been published for transformation-type decision tasks. Methodologically, mining regularities from event logs connects to process mining (van der Aalst, 2016); a recent method abstracts low-level logs into high-level ones via supervised learning (Blatt et al., 2026), a neighbor of our convolution of granular decision codes into classes, but we test a fixed nomenclature of predicates against financial outcome across corpora rather than reconstruct a control-flow model.

2.3 Theory of control in organizational systems and the identification problem

The theory of control in organizational systems develops formal models in which a principal designs mechanism (planning, incentive, monitoring) for active agents with their own goal functions and private information (Burkov et al., 2015; Novikov, 2013; Pacana & Siwiec, 2023; Kenchappa et al., 2024). Two aspects are constitutive here. First, the object is the organizational system and the behavior of its active elements, of which a production system is a special case; we treat the management team as a collective active element and its recorded decision flow as an observable realization of its behavioral model. Second, the framework separates the identification problem (reconstructing an agent's model from data) from the control synthesis that uses it. The typology below solves an identification problem in this sense: error classes are identified from retrospective telemetry as deficient realizations of specific control functions (planning, coordination, setup, monitoring; Section 3.4), and the diagnostic vector parameterizes downstream control tasks (readiness assessment and risk-reduction

design) addressed in companion work. This also explains why reduced manipulability matters: an assessment embedded in a working session, not administered as an explicit test, offers far fewer opportunities for strategic behavior, though formally guaranteeing non-manipulability requires a dedicated experiment left to future work.

Terminologically, the model here is a discrete-event model of a production-type organizational system; although digital-twin frameworks (ISO 23247-1; International Organization for Standardization [ISO], 2021) provide an adjacent context, no real-time link with a physical asset is required for this measurement task, so the digital-twin framing is not relied upon.

2.4 Research gap

In sum: (a) risk and quality methods quantify organizational behavior only through expert judgment or aggregate indicators; (b) behavioral operations research demonstrates systematic decision errors but has produced no validated typology on mass decision-level data; (c) the theory of control in organizational systems gives the formal frame (identification of behavioral models for subsequent mechanism design) but needs exactly the empirical input neither (a) nor (b) supplies. The present study fills this gap.

3. DATA AND METHOD

3.1 Data

Two independent corpora of decision-level telemetry were used, produced by the same family of discrete-event simulation models of a production-type organizational system for discrete manufacturing (two products, more than 30 technological operations, a complete decision loop: layout, personnel, logistics, procurement, warehousing, quality, operating system, finance). Telemetry records 44 decision codes in the format "Subsystem::decision" with timestamps and settlement-period references. The simulation platform used in this study is a deployed training simulator; identifying details are omitted under double-blind review.

Corpus A comprises nine expert sessions held in 2021-2026: 567 teams played (made at least one decision or reached the third settlement period), 918,058 telemetry events, 296,997 coded decisions, and a 32-week decision horizon; 522 teams reached the final period and have a financial outcome. Corpus B comes from a corporate program of utility-industry enterprises (June-July 2026), run on the second revision of the expert scenario configuration: 203 teams in the logs, 124 played, 215,195 events, 63,049 decisions; 118 teams have a financial outcome.

Teams comprise three to six participants: practicing managers and specialists of industrial enterprises in

management development programs (corpus A: open-enrollment expert sessions; corpus B: a corporate program), not undergraduate students. Within each session all teams received an identical briefing and initial conditions. Participants consented to processing of their activity data upon registration; the analysis uses anonymized team-level records only, and team identifiers cannot be mapped to natural persons.

3.2 Duplicated data source: a methodological check

At a preliminary stage a third, nominally independent source was considered: an aggregated analytics package (267 participants) anonymized under enterprises of a different industry. Cross-checking session identifiers and team-level financial indicators established that it described the same sessions as corpus B, processed by an aggregate (not decision-level) pipeline. It was excluded from the independent corpora to avoid double counting; its estimates are retained only as a consistency check under a change of measurement method on the same sample (Section 4), not as replication. The lesson (verify sample independence before any claim of independent replication) is discussed in Section 5.2.

3.3 Formal statement

Let V be a finite alphabet of decision codes ($|V| = 44$). For a team t , telemetry defines a sequence $S(t) = ((v_1, w_1), \dots, (v_m, w_m))$, where $v_i \in V$ is a decision code and w_i is the settlement period in which it was taken (timing). An error class is a predicate $K_j: S(t) \rightarrow \{0, 1\}$ defined through the structure of the transformation task (constraint management, subsystem synchronization, initial setup, project economics) rather than the mechanics of a particular simulator, computable from telemetry without observer participation. The diagnostic vector of a team is $c(t) = (K_1(S(t)), \dots, K_n(S(t))) \in \{0, 1\}^n$. A class enters the typology under four requirements: (a) reproducible operationalization from a published rule (full pseudocode with frozen thresholds is provided as online supplementary material); (b) non-zero prevalence: the lower bound of the 95% Wilson confidence interval (Wilson, 1927) is above zero; (c) outcome validation: a significant shift of median cumulative profit of class-positive versus remaining teams (two-sided permutation test, 10,000 permutations (Good, 2005), final-period teams only); (d) reproducibility of sign and significance on a hold-out sample and/or an independent corpus.

3.4 Classification basis

The basis of classification is categorical and system-level: classes differ not by symptom or severity of consequence but by the category of the transformation task structure to which the violated action belongs: constraint management, subsystem synchronization, initial setup of the system, and project economics. The

categories are conceptually distinct and jointly cover the structure of the transformation task rather than the mechanics of a particular simulator; the class predicates themselves are not mutually exclusive and may co-occur within one team. A transformation has a structural component (the composition and topology of elements: work centers, routes, team roles) and a parametric one (values and control policies of existing elements: targets, batch sizes, buffers); the task categories touch both.

In the language of Section 2.3, the four categories correspond to canonical control functions of a principal (planning under a binding constraint, coordination of subsystems, setup of the control loop before operation, and monitoring of the financial state), so each class identifies a deficient realization of one control function, and the diagnostic vector $c(t)$ is an identified fragment of the behavioral model of a collective active element, suitable as input to control synthesis. A secondary, competence-based reading follows the "competence $\rightarrow$ behavioral indicator $\rightarrow$ proficiency level" scheme (Rudskoy et al., 2018): the absence or untimeliness of a recorded action may indicate a deficit of a managerial competence rather than a psychological category, which is consistent with evidence that in-game behavior can validly indicate competency level (Bijl et al., 2024).

3.5 Candidate classes

Fourteen candidates were operationalized in two iterations that differ in epistemic status; we separate them explicitly. Iteration 1 (ten candidates) was specified from the structure of the transformation task before any outcome testing. Iteration 2 (four candidates: rigidity to demand shock, lack of changeover, planning-loop desynchronization, late personnel involvement) was added after inspecting corpus-A data, motivated by incomplete coverage and by the analysis of reactions to environmental perturbations; hence its exploratory status. All predicates are defined in Table 1.

Operationalization thresholds were fixed on the training sample (the largest session of corpus A, $n = 283$) and applied without change to all remaining data. Corpus B served as a fully closed test sample: it became available after the nomenclature and all thresholds had been fixed, and no definition was modified after its arrival. For classes admitting continuous threshold variation, a full sensitivity analysis around the base threshold was performed on both corpora (Section 4.1).

Note: K-codes come from the pipeline (gaps K10, K11, K13 in the original numbering). Three groups: (I) ten coded outcome-testable classes given the full corrected test (eight validated; K1 and K4 rejected); (II) descriptors not outcome-assessable by definition (dropout); (III) exploratory or underpowered candidates without a full test (planning-loop desynchronization, late personnel involvement, K6). Only group I enters the multiple-testing family. Threshold sensitivity: Section 4.1 and the supplement.

Table 1. Operationalization of candidate error classes

Code	Class	Predicate (base threshold)	Validated
K3	Passive start	Initial-setup decisions in periods 1-4 below the team-distribution lower quartile (0.25)	Yes (core)
K12	Rigidity to demand shock	Share of significant demand changes (>10% w/w) with no adapting operating-system decision within 2 periods ≥ 0.5	Yes (core)
K9	Formal period completion	Longest run of consecutive settlements with no decisions between them ≥ 8	Yes (core)
K2	Supply chain blind spot	Share of Suppliers + Warehouse decisions among all decisions < 0.02	Yes
K14	Lack of changeover under product-mix change	Changeover decisions (OCS::d_changeProd) under two-product demand = 0	Yes (scenario-dependent)
K7	Neglect of personnel training	Personnel training decisions = 0	Yes
K8	Tunnel concentration	Normalized entropy of decisions over eight subsystems < 0.5	Yes
K5	Borrowing under accumulated loss	Loan taken while cumulative profit is negative	Yes (scenario-dependent)
K1	Layout churn	Share of workplace-relocation decisions ≥ 0.30	No
K4	Pull activation without preparation	Pull production activated with < 10 preparatory decisions before it	No
n/a	Planning loop desynchronization	Mean relative plan - demand error in the upper quartile	No
n/a	Late personnel involvement	First training decision after mid-horizon (week 16)	No
K6	Neglect of quality	Quality decisions (target quality, incoming inspection, defect handling) = 0	Underpowered
n/a	Dropout before the final period	Final settlement period (week 32) not reached	Not assessed

3.6 A remark on the nature of frequency estimates (relation to Monte Carlo)

Unlike interval risk-assessment methods, where realization frequencies are unknown and reconstructed by stochastic simulation (Monte Carlo) or expert elicitation, here the frequencies are observed directly as empirical shares over 296,997 recorded decisions made under conditions identical for all teams; stochastic generation of the input would be redundant. Scenario generation on the model is separate work, not part of the present typology.

3.7 Validation procedure

Class frequencies are estimated with 95% Wilson confidence intervals; the link to outcome is tested by a two-sided permutation test of the difference of median cumulative profit between class-positive and remaining teams. Because candidates are screened against outcome, the decision rule is fixed ex ante: a class enters the typology if (1) its predicate is computable from the published rule and has at least five class-positive teams; (2) the test clears a Benjamini-Hochberg false-discovery-rate correction at $q \leq 0.05$ over the family of coded outcome-screened candidates; and (3) the sign and significance are preserved on a hold-out sample and/or an independent corpus. Because teams are nested in sessions, tests are additionally run stratified (labels permuted within sessions). Because many predicates are

mechanically related to how much a team decides, each validated class is further tested against activity-level controls: a median regression of profit on the class indicator with $\log(\text{total decisions})$, the number of active periods, and session fixed effects; a stratification by activity quartiles; and, as a deliberately conservative over-control, an early-performance covariate (cumulative profit at week 10), which partly reflects the same early decisions and so bounds the confound from above. Reproducibility is checked in three stages: hold-out within corpus A ($n = 284$), leave-one-game-out (LOGO) cross-validation, and replication on corpus B. Finally, the monotonicity of profit in the number of diagnosed classes is tested by the Jonckheere-Terpstra test and Spearman correlation. Results carry an explicit evidence gradation: replicated core (cross-corpus and LOGO-stable), internally replicated (within corpus A), suggestive (borderline or few teams), scenario-dependent, and rejected.

4. RESULTS

Tables 2a and 2b summarize class frequencies and effects per corpus; Table 3 reports group sizes, confidence intervals, multiple-testing corrections, and robustness checks on corpus A; Figure 1 visualizes the per-class profit shifts on corpus A. All verdicts below are additionally confirmed on an alternative outcome criterion that enters no class definition, namely the service level (Section 4.4). The main results are as follows.

Table 2a. Class frequencies and validation, corpus A (training / hold-out)

Class	Frequency, train, %	Frequency, hold-out, %	Δ profit, c.u.	p (all)	p (hold-out)
Passive start	25.8	20.8	-232,678	< 0.001	0.002
Rigidity to demand shock	14.1	6.7	-240,781	< 0.001	0.039
Formal period completion	10.6	8.1	-212,970	< 0.001	0.061
Supply chain blind spot	3.2	6.7	-211,232	0.021	0.271
Lack of changeover under product-mix change	13.1	14.4	-231,966	< 0.001	0.041
Neglect of personnel training	8.1	3.9	-430,474	< 0.001	too few obs.
Tunnel concentration	3.9	2.8	-420,586	0.003	too few obs.
Borrowing under accumulated loss	17.7	41.2	-206,953	< 0.001	< 0.001
Layout churn	4.6	5.3	-101,274	0.169	0.162
Pull activation without preparation	14.1	4.6	-11,104	0.849	0.250
Neglect of quality	3.9	2.5	too few obs.	n/a	n/a
Dropout before the final period	6.4	7.7	no outcome	n/a	n/a

Table 2b. Class frequencies and validation, corpus B

Class	Frequency, %	Δ profit, c.u.	p
Passive start	24.2	-292,892	< 0.001
Rigidity to demand shock	9.7	-302,725	0.008
Formal period completion	4.8	-346,293	0.002
Supply chain blind spot	6.5	-260,822	0.030
Lack of changeover under product-mix change	4.0	too few obs.	n/a
Neglect of personnel training	4.0	too few obs.	n/a
Tunnel concentration	2.4	too few obs.	n/a
Borrowing under accumulated loss	12.9	+50,304	0.617
Layout churn	3.2	too few obs.	n/a
Pull activation without preparation	20.2	-70,301	0.446
Neglect of quality	0.8	too few obs.	n/a
Dropout before the final period	4.8	no outcome	n/a

Note: Δ is the shift of median cumulative profit "class minus others" (conventional units); p is the two-sided permutation test; n/a means not applicable. The aggregate corpus-B pipeline (Sections 3.2, 4.7) reproduces the direction and significance on the same sample; this is a robustness check, not an independent replication.

Table 3. Uncertainty, multiple-testing correction, and robustness of the validated classes (corpus A, teams with outcome, n = 522)

Class	n+ / n-	Δ , c.u.	95% bootstrap CI	q (BH)	Evidence grade
Passive start (K3)	103 / 419	-232,678	[-268,943; -178,949]	0.0002	Replicated core
Rigidity to demand shock (K12)	49 / 473	-240,781	[-311,051; -185,491]	0.0002	Replicated core
Formal period completion (K9)	49 / 473	-212,970	[-289,562; -148,077]	0.0002	Replicated core (activity-coupled)
Lack of changeover (K14)	49 / 473	-231,966	[-297,792; -163,424]	0.0002	Scenario-dependent
Neglect of personnel training (K7)	17 / 505	-430,474	[-736,697; -281,281]	0.0002	Internally replicated
Tunnel concentration (K8)	7 / 515	-420,586	[-656,617; -204,033]	0.0040	Suggestive
Supply chain blind spot (K2)	14 / 508	-211,232	[-308,076; -118,215]	0.0260	Suggestive
Borrowing under accumulated loss (K5)	161 / 361	-206,953	[-270,350; -105,108]	0.0002	Scenario-dependent
Layout churn (K1)	21 / 501	-101,274	n/a	0.1872	Rejected
Pull activation without preparation (K4)	50 / 472	-11,104	n/a	0.8490	Rejected

Note: n+ / n- are class-positive / remaining teams among the 522 corpus-A outcome teams; Δ is the difference of median cumulative profit; CI is the bootstrap interval over 5,000 resamples; q is the Benjamini-Hochberg correction over the ten coded candidates; n/a means not applicable (no CI for rejected candidates). Hold-out, stratified, and activity-controlled results are in Section 4.4.

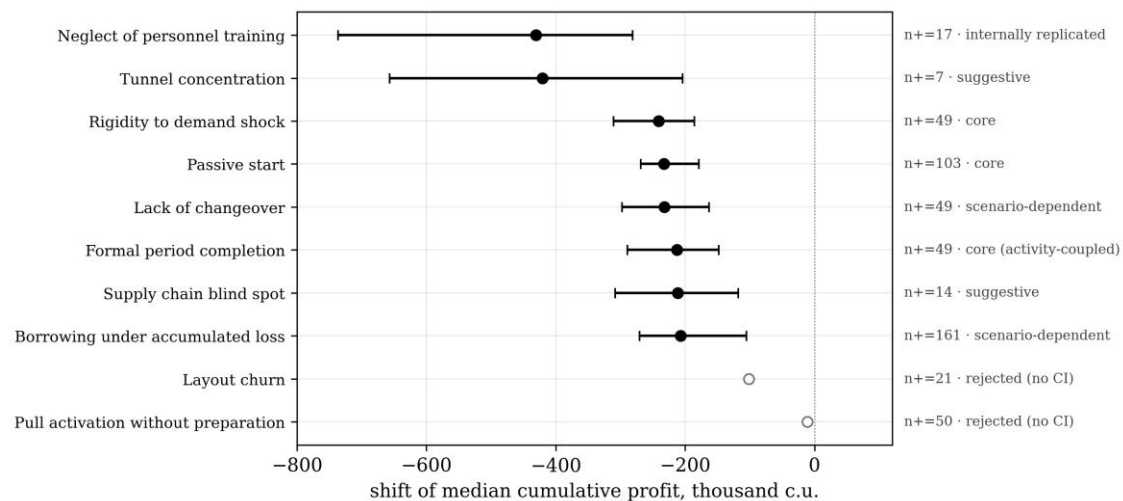


Figure 1. Per-class shift of median cumulative profit ("class-positive minus others"), corpus A: point estimate, 95% bootstrap CI, class-positive teams (n+), and evidence grade (Table 3); rejected candidates as open markers (no CI defined)

4.1 The core of reproducible classes

The passive start class exhibits exceptional cross-corpus stability: frequency 25.8% (training sample of corpus A), 20.8% (hold-out of A), 24.2% (corpus B), and the confidence intervals of all estimates overlap; the median profit shift ranges from -167 thousand conventional monetary units of the simulation economy (hereinafter c.u.; hold-out sample of corpus A) to -293 thousand c.u. (corpus B), with $p \leq 0.002$ in all checks. The aggregate corpus-B pipeline (Section 3.2) yields a consistent frequency of 24.0% (Δ -289,692; $p < 0.0001$). Substantively, the class captures comparatively low early system-design activity: the team does not perform the initial setup (layout, personnel, target parameters) in the first periods and thereafter manages in catch-up mode. The second cross-corpus class is rigidity to demand shock (added in iteration 2): the team issues no adapting operating-system decision within two periods after a significant change in market demand; frequency 14.1% (corpus A) and 9.7% (corpus B); effect -241 thousand c.u. (A, $p = 0.0001$, reproduced on hold-out with $p = 0.039$) and -303 thousand c.u. (B, $p = 0.008$).

The third core class is formal period completion (frequency 4.8-10.6%; effect -213 to -346 thousand c.u.; $p \leq 0.002$ in both corpora); as Section 4.4 shows, it is the activity-coupled member of the core. Supply chain blind spot also reproduces in both corpora (-211 to -261 thousand c.u.; $p = 0.02$ -0.03). Significant within corpus A but underpowered on corpus B (frequencies of 4% or less there) are neglect of personnel training (-430 thousand c.u., the largest effect), tunnel concentration (-421 thousand c.u.), and lack of changeover (-232 thousand c.u., $p = 0.0001$, hold-out $p = 0.041$); their corpus-B signs do not contradict corpus A.

In total, eight of the ten coded outcome-screened candidates passed the statistical screening, all clearing

the Benjamini-Hochberg correction at $q \leq 0.05$, while the rejected candidates do not (Table 3). The threshold sensitivity analysis (Table 4) confirms the robustness of the core and the fragility of the rejected candidates: the rejections are not artifacts of the chosen thresholds. Two validated classes carry interpretive caveats visible in Table 4: lack of changeover is diagnostic strictly at "zero decisions" (a binary attribute rather than a threshold on a scale), and the power of borrowing under accumulated loss depends on the depth of the loss, which places it among the scenario-dependent classes.

4.2 The role of timing: the borrowing class

The most instructive contrast is "fact versus timing" (Figure 2). The mere fact of resorting to credit does not form an error class in either corpus: the profit shift is indistinguishable from zero (corpus A: -51 thousand c.u., $p = 0.11$; corpus B: +139, $p = 0.09$), and the corporate-cohort estimates are positive: for strong teams a loan is a growth lever. Accounting for timing splits the phenomenon: a loan taken at the moment of accumulated loss forms a strong class in corpus A (17.7% of teams; -207 thousand c.u.; $p = 0.0001$; reproduced on hold-out), yet on corpus B (the second scenario revision, with modified credit mechanics) it does not reproduce (+50, $p = 0.62$).

Two conclusions follow. The first is methodological: a typology of errors fundamentally requires decision-level telemetry with timing, because aggregate fixation of the fact of an action leads to the opposite verdict on the class. The second is substantive: the classes of the typology divide into core classes (invariants reproduced across both corpora and scenario revisions) and scenario-dependent classes, whose diagnostic power is determined by the configuration of the model economy; when the typology is applied, frequency reference points are normalized by the scenario configuration.

Table 4. Full threshold sensitivity grid, 11 of 14 classes (corpus A, all teams with outcome; corpus B)

Class	Test thresholds, A (frequency, % · p)	Verdict
K2 Supply chain blind spot	0.015: 3.9%·0.042 · 0.020: 4.9%·0.022 · 0.030: 10.4%·0.033	Robust
K3 Passive start	0.1875: 18.9%·0.0001 · 0.25: 24.9%·0.0001 · 0.375: 36.9%·0.0001	Robust (core)
K7 Neglect of training	≤0: 6.0%·0.0001 · ≤2: 6.3%·0.0001 · ≤3: 6.7%·0.0001	Robust; near-binary attribute
K8 Tunnel concentration	0.50: 3.4%·0.0026 · 0.625: 5.5%·0.0019 · 0.75: 12.7%·0.0001	Robust where testable; thin at base
K9 Formal period completion	6: 15.9%·0.0001 · 8: 9.3%·0.0001 · 10: 3.9%·0.006	Robust (core)
K12 Rigidity to demand shock	0.375: 19.0%·0.0001 · 0.50: 13.1%·0.0002 · 0.75: 10.1%·0.0014	Robust (core)
K5 Borrowing under loss	-50k: 8.3%·0.0001 · 0: 29.1%·0.0001 · +100k: 38.4%·0.083	Fragile: strengthens when tightened, dies when relaxed, sign flips on B
K14 Lack of changeover	0: 13.8%·0.0001 · ≤1: 54.7%·0.779 · ≤3: 80.1%·0.917	Knife-edge: signal strictly at zero, gone at ≤1
K6 Neglect of quality	≤1: 4.2%·n<5 · ≤2: 4.6%·0.054 · ≤3: 5.1%·0.0005	Underpowered at base
K1 Layout churn (rejected)	0.20: 18.7%·0.018 · 0.30: 4.9%·0.175 · 0.40: 1.1%·0.723	Rejected; significant only when softened
K4 Pull without preparation (rejected)	5: 7.1%·0.499 · 10: 9.3%·0.856 · 15: 11.5%·0.491	Neutral; rejected in both corpora

Note: p is the two-sided permutation test (10,000 permutations, seed = 7); "n<5" means undefined. Three thresholds per class (softer · base · tighter); base-threshold frequencies and Δ are in Tables 2a and 2b, the full five-point grid in the supplement. The "= 0" classes (K6, K7, K14) can only be relaxed, not tightened; on corpus B (underpowered) the absence of significance must not be read as absence of an effect. Per-class corpus-B figures and the fixed K12 constants (10% shock, 32-week finish) are in the supplement.

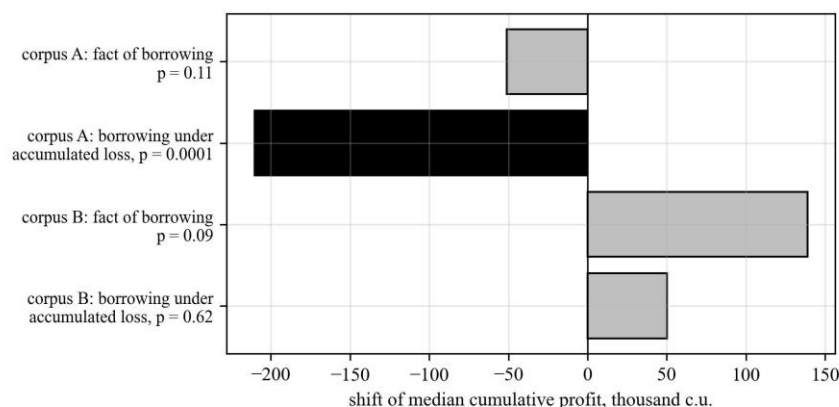


Figure 2. Borrowing: fact versus timing of the decision and the split of the class verdict across the corpora

4.3 Negative results

Four candidates failed outcome validation: layout churn ($p = 0.16$ - 0.17) and pull activation without preparation ($p = 0.24$ - 0.85) from the first iteration, planning loop desynchronization ($p = 0.32$) and late personnel involvement ($p = 0.21$; frequency 1.1%) from the second; the mere fact of activating pull production is likewise neutral with respect to outcome ($p = 0.88$). Publishing the negative results is a matter of principle here: these are typical expert judgments about "bad behavior," and their empirical rejection demonstrates that the typology is not a formalization of intuition.

4.4 Robustness: alternative criterion, session structure, and activity level

The most direct answer to the alternative explanation "the classes merely mark weak teams under one objective

function" is a check on a second outcome metric, the service level (share of demand fulfilled on time, OTIF), which enters no class definition. The two criteria are correlated across teams (Spearman $\rho = 0.73$ on corpus A), as expected for outcomes of one model; the check's value is definitional disjointness, not statistical independence. All eight validated classes are associated with lower service too: the core by 25-31 percentage points, the rest by 28-40 ($p \leq 0.0004$). The one exception is borrowing under accumulated loss, with a strong effect on profit but a weak one on service (-6.7 pp, $p = 0.04$), consistent with its financial rather than operational nature. Coincidence of verdicts across two definitionally disjoint criteria shows the classes are associated with underperformance on both metrics, not with an artifact of one objective function. Effect sizes with bootstrap confidence intervals are in Table 3; no interval of a validated class includes zero.

Because teams are nested in nine sessions, all tests were re-run with permutations stratified within sessions (Section 3.7): every validated class remains significant (core $p \leq 0.0003$; supply chain blind spot and tunnel concentration even strengthen), and the two rejections are confirmed; the effects are not an artifact of between-session heterogeneity.

The activity-level control addresses the second alternative explanation ("class-positive teams simply decide less") and differentiates the classes. Passive start is essentially insensitive to it (the median regression removes only ~8% of its effect, $p < 0.001$): its signal is specific to the timing of the initial setup, not to low activity in general. Neglect of personnel training retains ~69% of its effect; rigidity to demand shock and borrowing also remain significant. By contrast, formal period completion and lack of changeover lose independent significance under this continuous control ($p = 0.062$ and 0.115 ; ~60% of their raw shift is attributable to overall activity), while surviving the coarser stratification by activity quartiles: their incremental contribution is present but unstable, as expected for predicates close to measures of inaction. Under the deliberately conservative early-performance over-control (Section 3.7), rigidity to demand shock survives ($p = 0.039$) whereas formal period completion and borrowing lose independent significance; because early profit partly mediates the same early decisions, these estimates bound the contribution of latent preparedness from above rather than measure it.

4.5 Within-corpus replication (leave-one-game-out)

Beyond the training/hold-out split, a rolling check excludes each of the nine corpus-A sessions in turn and tests the class effect on the excluded one. Passive start is negative in all testable folds (7 of 7), formal period completion in 3 of 3, rigidity to demand shock in 2 of 2:

the session-invariant core. By contrast, borrowing keeps a negative sign in only 5 of 9 sessions and lack of changeover in 3 of 5: session-dependent, consistent with their scenario-dependent status (borrowing also failed to reproduce on corpus B). Being entirely within corpus A, the LOGO check provides replication independent of the corpus-B independence question (Section 3.2).

4.6 Cumulative effect

Median profit decreases monotonically with the number of diagnosed classes in both corpora (Figure 3): corpus A from +327 thousand c.u. (0 classes, $n = 243$) through +148, -24 and -40 to -150 (4+ classes, $n = 21$; bucket sizes 243/185/56/17/21), teams with two or more classes being loss-making at the median; corpus B from 414 to 138. The decrease is confirmed on corpus A by the Jonckheere-Terpstra test ($p < 10^{-30}$) and Spearman correlation ($\rho = -0.55$); the Pearson coefficients ($r = -0.24$ in A, -0.42 in B) understate the monotone association because of heavy tails. The highest-load bucket rests on few teams, so the tail is imprecise. The aggregate corpus-B pipeline shows the same direction (427 to 106). The association motivates the following definition, which fixes the contour of a convolution of the diagnostic vector $c(t)$ for subsequent work and is not a present result:

$$J(t) = \sum_j \alpha_j c_j(t) \quad (1)$$

where $c_j(t) = K_j(S(t)) \in \{0,1\}$ is the indicator of class j and $\alpha_j > 0$ its weight. The Δ column of Table 2a gives only an ex-ante ordering of candidate weights; because the classes are correlated and may interact, estimating the weights requires a multivariate model with regularization and independent predictive validation, which, together with the normalization of J into a readiness-indicator scale and the decision rule for a team profile, is the subject of the next work of this program (the integrated readiness indicator).

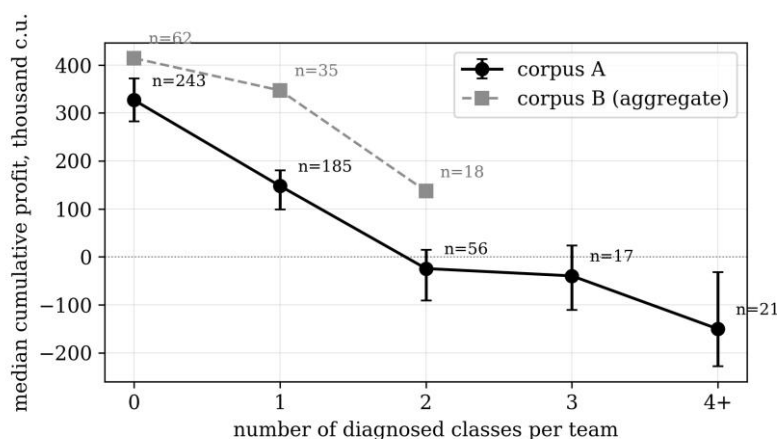


Figure 3. Monotonic association between the number of diagnosed classes and median cumulative profit across the two corpora

Note3: corpus-A buckets are computed on the canonical pipeline over all 522 outcome teams (243/185/56/17/21 for 0/1/2/3/4+ classes); error bars are 95% bootstrap CIs of the bucket median (5,000 resamples). Corpus-B buckets (0/1/2 classes; $n = 62/35/18$) come from the retained aggregate pipeline and are plotted without CIs (unavailable there); three corpus-B teams could not be assigned to the plotted buckets, and no substantive exclusion criterion was applied.

4.7 Coverage

At least one validated class is diagnosed for 53.4% of corpus-A teams [49.2; 57.7] and for 39.5% of corpus-B teams under the strict 44-code definitions [31.4; 48.3]; on the same corpus-B sample, under the soft aggregate definitions (Section 3.2), coverage reaches 81.6%: the same teams measured more strictly and more softly, not a wider sample. Iteration 2 did not raise coverage substantially: the new classes mostly flag already-flagged teams. We deliberately reject coverage as a sufficiency criterion: high coverage is achievable only by softening predicates until the link to outcome is lost; sufficiency is assessed instead by cross-corpus and rolling-check reproducibility, the monotone class-count association, and threshold robustness. A share of teams remains "clean" yet underachieves for reasons outside the "what and when" apparatus (parameter magnitudes are unrecorded; Section 5.2), which is a substantive limitation, not a coverage defect.

5. DISCUSSION

5.1 What transfers and what does not

The typology is built on team decisions in a controlled simulation. What the data support directly is replication across cohorts and scenario revisions within one simulation platform family; validation on a different platform or on field production data is future work (Section 5.3). What is hypothesized to transfer is the set of classes (invariants of decision structure) and the extraction methodology, but not the absolute frequencies, which are normalized by the scenario configuration (as the borrowing class shows), and not a prediction of a specific person's behavior in production. Outcome validation establishes association, not causality: the classes may partly reflect latent team preparedness. Three checks bound this alternative: all validated classes coincide with lower service (a disjoint criterion); the effects survive session stratification and, with the noted exceptions, activity-level controls (Section 4.4); and even the conservative early-performance over-control does not eliminate the strongest classes. A stable class-outcome association is a necessary first condition for a diagnostic instrument, not yet sufficient for diagnostic validity. Finally, this paper describes patterns and their association with outcome; predicting future performance, diagnosing specific-team competencies, and estimating intervention effects are distinct tasks it does not claim to solve.

5.2 Limitations

Telemetry records the code and moment of a decision but not the parameter value it sets; classes of the "how much" kind (as opposed to "what and when") are inaccessible. Operationalization thresholds are conventional; the sensitivity analysis (Table 4) bounds this, with two caveats carried over: the knife-edge lack of changeover and the loss-depth dependence of borrowing. Tunnel

concentration rests on fewer than 20 class-positive teams, and neglect of quality fell below the five-team threshold altogether (excluded from the validated typology); neither admits tightening without loss of power. Predicates are computed deterministically, so inter-rater agreement is not required; what remains open is construct validity, that is, whether an independent expert would agree a predicate captures a meaningful behavioral error rather than an artifact of the rule's formulation. That expert check is scheduled as the immediate next step and was not performed here. Two further limitations: the incremental contribution of formal period completion and lack of changeover beyond overall activity is unstable (Section 4.4), so these read as markers coupled with general inaction rather than fully separable mechanisms; and the typology is retrospective: most predicates span the whole horizon while the outcome is measured at its end, so whether early-window classes predict later outcomes (the requirement for early readiness diagnostics) is untested. Finally, as Section 3.2 shows, externally distinct anonymized packages may describe one sample; verification of sample independence must precede any claim of independent replication.

5.3 Integration into the control loop and further work

In the terms of Section 2.3, the results provide an initial telemetry-based implementation of the identification stage of a two-stage control task. The identified diagnostic vector $c(t)$ and the empirical ordering of class weights α_j parameterize the behavioral component of a management team's transformation readiness indicator $J(t)$; the readiness indicator, in turn, enters the synthesis stage, namely the design of a risk-reduction mechanism that allocates preparatory control actions across teams and classes subject to resource constraints. The further work program comprises three steps: (1) extension of the class nomenclature with a full threshold sensitivity analysis and an expert check of construct validity and coding reproducibility; (2) construction of the integrated readiness indicator of a management team with telemetry-based parameterization of the behavioral component using the obtained classes and frequencies; (3) formalization of the simulation-based decision session as a measurement procedure on a simulation model of the enterprise's own organizational system, including an assessment of susceptibility to strategic adaptation and manipulation-resistant measurement procedures.

6. CONCLUSION

A typology of adverse managerial decision patterns in organizational systems undergoing structural and parametric transformation has been constructed. Fourteen candidates were operationalized as reproducible predicates over decision-level telemetry; of the ten coded outcome-testable ones, eight passed the

statistical screening under a false-discovery-rate correction, of which the core (passive start, formal period completion, rigidity to demand shock) is stable under leave-one-game-out cross-validation and reproduces on corpus B. The evidence is explicitly graded: passive start survives every robustness check including session stratification and activity controls; rigidity to demand shock weakens only under the conservative early-performance over-control; the independent contribution of formal period completion beyond general activity is unstable. The verdicts hold under threshold sensitivity analysis and an alternative criterion (service level). Accounting for decision timing can reverse the verdict on a class relative to its mere occurrence, and a monotonic association between the number of diagnosed classes and

outcome was established on 640 teams with recorded outcomes. A methodological duplication of a data source was detected and eliminated, which shows the need for explicit sample-independence checks on external anonymized packages. Within the theory of control in organizational systems, the typology advances the identification stage: it turns raw decision telemetry into candidate behavioral indicators, that is, an identified fragment of the behavioral model of a collective active element, suitable for parameterizing the readiness indicator and risk-reduction mechanism developed in companion work.

Acknowledgement: This research received no external funding. The authors declare no conflict of interest.

References:

- Achanga, P., Shehab, E., Roy, R., & Nelder, G. (2006). Critical success factors for lean implementation within SMEs. *Journal of Manufacturing Technology Management*, 17(4), 460–471. <https://doi.org/10.1108/17410380610662889>
- Bendoly, E., Donohue, K., & Schultz, K. L. (2006). Behavior in operations management: Assessing recent findings and revisiting old assumptions. *Journal of Operations Management*, 24(6), 737–752. <https://doi.org/10.1016/j.jom.2005.10.001>
- Bijl, A., Veldkamp, B. P., Wools, S., & de Klerk, S. (2024). Serious games in high-stakes assessment contexts: A systematic literature review into the game design principles for valid game-based performance assessment. *Educational Technology Research and Development*, 72(4), 2041–2064. <https://doi.org/10.1007/s11423-024-10362-0>
- Blatt, J., Delfmann, P., Schubert, P., & Just, M. (2026). Event abstraction for social process mining in enterprise collaboration systems: A supervised machine learning approach for converting low-level event logs into high-level event logs. *Information Systems*, 139, Article 102703. <https://doi.org/10.1016/j.is.2026.102703>
- Burkov, V. N., Goubko, M., Korgin, N., & Novikov, D. (2015). *Introduction to theory of control in organizations*. CRC Press. <https://doi.org/10.1201/b18152>
- Chirumalla, K., Oghazi, P., Nnewuku, R. E., Tuncay, H., & Yahyapour, N. (2025). Critical factors affecting digital transformation in manufacturing companies. *International Entrepreneurship and Management Journal*, 21(1), Article 54. <https://doi.org/10.1007/s11365-024-01056-3>
- Croson, R., & Donohue, K. (2006). Behavioral causes of the bullwhip effect and the observed value of inventory information. *Management Science*, 52(3), 323–336. <https://doi.org/10.1287/mnsc.1050.0436>
- Evans, J. C., Evans, M. B., & Lingard, L. (2025). Team cognition in healthcare simulation: A framework for deliberate measurement. *Advances in Simulation*, 10(1), Article 12. <https://doi.org/10.1186/s41077-025-00333-7>
- Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). Springer. <https://doi.org/10.1007/b138696>
- Hernández-Lara, A. B., Perera-Lluna, A., & Serradell-López, E. (2019). Applying learning analytics to students' interaction in business simulation games. The usefulness of learning analytics to know what students really learn. *Computers in Human Behavior*, 92, 600–612. <https://doi.org/10.1016/j.chb.2018.03.001>
- Hughes, M. (2011). Do 70 per cent of all organizational change initiatives really fail? *Journal of Change Management*, 11(4), 451–464. <https://doi.org/10.1080/14697017.2011.630506>
- International Electrotechnical Commission. (2019). *Risk management - Risk assessment techniques* (IEC Standard No. 31010:2019).
- International Organization for Standardization. (2021). *Automation systems and integration - Digital twin framework for manufacturing - Part 1: Overview and general principles* (ISO Standard No. 23247-1:2021).
- Kenchappa, R. M., Yadav, R. K., Singh, A., et al. (2024). Utilizing a unique deep learning technique for detecting anomalies in industrial automation systems. *Proceedings on Engineering Sciences*, 6(S1), 241. <https://doi.org/10.24874/PES.SI.24.02.007>
- Klochkov, Y., Gazizulina, A., & Golovin, N. (2016). Assessment of organization development speed based on the analysis of standards efficiency. In *2016 Second International Symposium on Stochastic Models in Reliability Engineering, Life Science and Operations Management (SMRLO)* (pp. 530–532). IEEE. <https://doi.org/10.1109/SMRLO.2016.93>

- Klochkov, Y., & Tveryakov, A. (2020). Approaches to the improvement of quality management methods. *International Journal of System Assurance Engineering and Management*, 11(Suppl. 2), 163–172. <https://doi.org/10.1007/s13198-019-00939-x>
- Lazić, Z., Grujić, M., & Škorić, V. (2025). Quality in supply chain management in times of global crises. *International Journal for Quality Research*, 19(2), 627–638. <https://doi.org/10.24874/IJQR19.02-18>
- Netland, T. H. (2016). Critical success factors for implementing lean production: The effect of contingencies. *International Journal of Production Research*, 54(8), 2433–2448. <https://doi.org/10.1080/00207543.2015.1096976>
- Novikov, D. A. (2013). *Theory of control in organizations*. Nova Science Publishers.
- Oludapo, S., Carroll, N., & Helfert, M. (2024). Why do so many digital transformations fail? A bibliometric analysis and future research agenda. *Journal of Business Research*, 174, Article 114528. <https://doi.org/10.1016/j.jbusres.2024.114528>
- Pacana, A., & Siwec, D. (2023). Integrated technique searching causes of quality incompatibilities. *International Journal for Quality Research*, 17(2), 473–484 <https://doi.org/10.24874/IJQR17.02-11>
- Pay, R. (2008, March 1). *Everybody's jumping on the lean bandwagon, but many are being taken for a ride*. IndustryWeek. <https://www.industryweek.com/leadership/companies-executives/article/21946187/everybodys-jumping-on-the-lean-bandwagon-but-many-are-being-taken-for-a-ride>
- Rudskoy, A. I., Borovkov, A. I., Romanov, P. I., & Kolosova, O. V. (2018). Obshcheprofessional'nye kompetentsii sovremennogo rossiyskogo inzhenera [General professional competencies of a modern Russian engineer]. *Vysshee Obrazovanie v Rossii*, 27(2), 5–18.
- Singh, A. P., Shahid, M., Kumari, A., & Karthikeyan, M. P. (2024). Enhanced production through novel swarm-intelligent enabled virtual cell formation: multifaceted approach. *Proceedings on Engineering Sciences*, 6(1), 353–362. <https://doi.org/10.24874/PES.SI.24.02.018>
- Shubik, M. (1975). *Games for society, business and war: Towards a theory of gaming*. Elsevier.
- Sterman, J. D. (1989). Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Science*, 35(3), 321–339. <https://doi.org/10.1287/mnsc.35.3.321>
- van der Aalst, W. M. P. (2016). *Process mining: Data science in action* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-662-49851-4>
- Velez, A., & Alonso, R. K. (2025). Business simulation games for the development of decision making: Systematic review. *Education Sciences*, 15(2), Article 168. <https://doi.org/10.3390/educsci15020168>
- Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.1080/01621459.1927.10502953>
- Worley, J. M., & Doolen, T. L. (2006). The role of communication and management support in a lean manufacturing implementation. *Management Decision*, 44(2), 228–245. <https://doi.org/10.1108/00251740610650210>
- Wu, C.-H., Chou, C.-W., Chien, C.-F., & Lin, Y.-S. (2024). Digital transformation in manufacturing industries: Effects of firm size, product innovation, and production type. *Technological Forecasting and Social Change*, 207, Article 123624. <https://doi.org/10.1016/j.techfore.2024.123624>

Vladislav V. Tereshchenko

Higher School of Advanced Digital Technologies, Advanced Engineering School "Digital Engineering", Peter the Great St. Petersburg Polytechnic University, 29 Politechnicheskaya str., St. Petersburg, 195251, Russia
v.v.tereshchenko92@gmail.com
ORCID 0000-0001-9690-5434

Yury S. Klochkov

2Industrial University of Tyumen, 38 Volodarsky St., Tyumen, 625000, Russia
y.kloch@mail.ru
ORCID 0000-0002-7913-8285

Elena O. Kasianenko

Higher School of Advanced Digital Technologies, Advanced Engineering School "Digital Engineering", Peter the Great St. Petersburg Polytechnic University, 29 Politechnicheskaya str., St. Petersburg, 195251, Russia
kasyanenko_eo@spbstu.ru
ORCID 0000-0002-9988-5050

Sergey V. Salkutsan

Higher School of Advanced Digital Technologies, Advanced Engineering School "Digital Engineering", Peter the Great St. Petersburg Polytechnic University, 29 Politechnicheskaya str., St. Petersburg, 195251, Russia
salkutsan@spbstu.ru
ORCID 0000-0002-2360-1330
