VOLUME 1 | NUMBER 2 | SEPTEMBER 2026

ISSN: 3136-9340

JOURNAL OF SECURE DIGITAL TECHNOLOGIES

A PEER-REVIEWED JOURNAL
BRIDGING TECHNOLOGY, LAW AND TRUST
IN THE DIGITAL AGE

DGSC, Astana, Kazakhstan

Journal of Secure Digital Technologies
Volume 1 (№2) September 2026

ISSN (E): 3136-9340
Sertificate № KZ19VPY00128988, date of issue: 25.12.2025

The journal is published 4 times a year
(March, June, September, December)

EDITORIAL TEAM
Editor-in-Chief
Dr. Leila Rzayeva, Association prof., PhD, Digital government support center (Kazakhstan)

Deputy Editor-in-Chief
Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

Executive Secretary
Dr. Abdenova Zhuldyz, PhD, Digital government support center (Kazakhstan)

EDITORIAL BOARD MEMBERS (KAZAKHSTAN)
Seitkulov E.N. – Candidate of Physical and Mathematical Sciences, Professor Computer Science, Computer Engineering and Control
Abdenova G.A. – PhD in Automation and Control, Candidate of Technical Sciences, Professor L.N. Gumilyov Eurasian National University
Atanov S.K. – Doctor of Technical Sciences, Professor.
Sergazin G.K. – PhD, Associate Professor Computer Science, Computer Engineering and Control
Satybaldina D.Zh. – Candidate of Physical and Mathematical Sciences, Professor Director, Research Institute of Information Security and Cryptology, L.N. Gumilyov Eurasian National University

INTERNATIONAL EDITORIAL BOARD MEMBERS
Ibraheem Shayea – PhD in Electrical and Electronic Engineering Istanbul Technical University (Turkey)
Yadigar Nasiboglu Imamverdiyev – DSc in Technical Sciences, Associate Professor Information Security, Computer Science and Computer Technology (Azerbaijan)
Erdal Irmak – PhD in Electrical and Electronics Engineering, Professor Gazi University (Turkey)
Korhan Kayisli – PhD in Electrical Electronics Engineering, Associate Professor Gazi University (Turkey)
Yonghao Wang – PhD in Electronic Engineering, Associate Professor Queen Mary, University of London (UK)
Amir Mosavi – PhD in Applied Informatics, Professor of AI Obuda University (Hungary)
Gerald Feldman – PhD in General Computing, Associate Professor Birmingham City University (UK)
Mohamed Ragab – PhD in Computer Science, Assistant Professor Birmingham City University (UK)
Seref Sagiroglu – PhD in System Engineering, Professor Gazi University (Turkey)

MANAGING EDITOR
Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

TECHNICAL EDITOR
Dr. Ali Myrzatay, PhD, Digital government support center (Kazakhstan)

CONTENTS

Computational Assessment of a Reconstructed Fuzzy Model for Categorizing Critical Digital Objects 3

Cyber Incident Damage Assessment for Public Digital Systems and Critical Infrastructure 27

Verifiable but Not Correct: Blockchain Trust Boundaries for Autonomous Agents 61

Calibrated Digital Twins and Geometry-based Perception for Automated Driving in Heavy Rain and Snowfall: Evidence From Research Testing in China and Implications for Kazakhstan. 78

Personal Data Protection in Kazakhstan's AI Enabled Public Services 99

Asking Another Jurisdiction What It Holds Without Telling It What You Hold: Private Artifact Matching with a Ledger-Anchored Transcript 121

On-Chain and Off-Chain Data Placement in Hybrid Blockchain Health Systems 133

An Integrated Project Management Maturity Model for Digital Qazaqstan Initiatives: Embedding Cybersecurity, AI Governance and Forward-Looking Cyber-Risk Requirements into ST RK ISO 21502-2022 148

Privacy Risks and Data Protection Readiness in AI Marketing in Kazakhstan 173

Khassanov Fakhreddin
MSc, Manager, Information Security Center
fakhreddin.khassanov@gmail.com, 0009-0000-5439-8171
National Institute for Cyber Policy and Research, RSE on REM "Digital Government Support Center", Astana, Kazakhstan

Leila Rzayeva
PhD, Associate Professor
e-mail: l.rzayeva@govtec.kz
ORCID: 0000-0002-3382-4685
Digital Government Support Center, Astana, Republic of Kazakhstan

Nursultan Nyssanov
MSc, Information Security Specialist, Information Security Center
n.nysanov@govtec.kz, 0009-0002-8128-0595
National Institute for Cyber Policy and Research, RSE on REM "Digital Government Support Center", Astana, Kazakhstan

COMPUTATIONAL ASSESSMENT OF A RECONSTRUCTED FUZZY MODEL FOR CATEGORIZING CRITICAL DIGITAL OBJECTS

Abstract: Before a categorization model informs protection obligations, its logical consistency and sensitivity need to be understood. We examine a reproducible reconstruction of a three-level Mamdani model for critical digital objects. The assessment covers eighteen scoring indicators, a separate automatic-assignment field, and confidentiality, integrity and availability scenarios. Documented membership functions and rules are retained, while missing normalization values are identified as reconstruction assumptions. Code, random seeds and generated data are described for reproduction. In 30,000 synthetic passports, all tested inference stages had nonzero aggregated membership, and scores ranged from 16.12 to 83.51. We observed 56 score groups at a tolerance of 10^{-8}, including 1,768 passports in [65, 70). Of 28,166 applicable one-level worsening steps, 1,237 (4.39%) lowered the score. None reduced decision status under the original thresholds, but setting the first-category threshold to 68 yielded 24 first-category reversals. Analytical and targeted tests show that a fully critical consequence is preserved at the severity layer; this does not guarantee first-category assignment. Scores were positively correlated with an equal-weight additive comparator ($r = 0.405$). Joint shifts in assumed normalization values changed 6.66% to 31.95% of final outcomes. These findings point to substantial calibration dependence and threshold-sensitive stability. Further specification, testing and external validation are needed before regulatory suitability can be assessed.

Keywords: critical digital objects; fuzzy inference; computational assessment; monotonicity; normalization sensitivity; non-compensation; reproducibility; cybersecurity.

Вычислительная оценка реконструированной нечёткой модели категоризации критически важных цифровых объектов

Аннотация: До использования результатов категоризации для установления требований к защите необходимо проверить логическую согласованность модели и её чувствительность к параметрам. В работе исследуется воспроизводимая реконструкция трёхуровневой модели Мамдани для критически важных цифровых объектов. Восемнадцать расчётных показателей и отдельный признак автоматического присвоения категории рассматриваются для сценариев нарушения конфиденциальности, целостности и доступности. Сохранены документированные функции принадлежности и правила; допущения при нормировании обозначены явно, предоставлены код, начальные состояния генераторов и результаты расчётов. В базовой выборке из 30 000 синтетических паспортов нулевая агрегированная принадлежность не обнаружена; баллы находились в пределах 16,12–83,51. При допуске 10^{-8} выделено 56 групп выходных значений; 1 768 паспортов попали в интервал [65; 70). Из 28 166 применимых шагов ухудшения 1 237 (4,39%) снизили балл. При исходных порогах понижения статуса решения не выявлены, однако снижение порога первой категории до 68 баллов привело к 24 случаям утраты первой категории после ухудшения показателя. Аналитически и вычислительно подтверждено сохранение полностью критического последствия на уровне тяжести, что не означает безусловного присвоения первой категории. Корреляция с равновесовой аддитивной моделью положительна ($r = 0,405$). Совместное изменение восстановленных значений нормирования изменило итоговые решения для 6,66–31,95% паспортов. Результаты показывают зависимость от калибровки и порогов и обосновывают дальнейшую доработку и проверку модели.

Ключевые слова: критически важные цифровые объекты; нечёткий вывод; вычислительная оценка; монотонность; чувствительность к нормированию; некомпенсаторность; воспроизводимость; кибербезопасность.

Аса маңызды цифрлық объектілерді санаттаудың қайта құрылған бұлыңғыр моделін есептік бағалау

Аңдатпа: Санаттау нәтижелерін қорғау талаптарын белгілеу үшін қолданар алдында модельдің логикалық келісімділігі мен параметрлерге сезімталдығын тексеру қажет. Зерттеуде аса маңызды цифрлық объектілерге арналған Мамданидің үш деңгейлі моделінің қайта өндіруге болатын нұсқасы қарастырылады. Құпиялылықтың, тұтастықтың және қолжетімділіктің бұзылу сценарийлері бойынша он сегіз есептік көрсеткіш және санатты автоматты тағайындауға арналған бөлек белгі бағаланады. Құжатталған тиістілік функциялары мен ережелер сақталып, нормалау кезіндегі жорамалдар нақты көрсетілді; код, кездейсоқ сандар генераторларының бастапқы мәндері және есептеу деректері берілді. 30 000 синтетикалық паспорттан

тұратын негізгі таңдамада агрегатталған тиістіліктің нөлдік мәні анықталмады; баллдар 16,12–83,51 аралығында болды. 10^{-8} төзімділік кезінде шығыс мәндерінің 56 тобы анықталды, 1 768 паспорт [65; 70) аралығына түсті. Көрсеткішті нашарлатудың 28 166 қолданылатын қадамының 1 237-сі (4,39%) баллды төмендетті. Бастапқы шектерде шешім мәртебесінің төмендеуі байқалмады, алайда бірінші санат шегін 68 балға төмендету көрсеткіш нашарлағаннан кейін бірінші санаттан айырылудың 24 жағдайын туғызды. Толық сыни салдардың ауырлық деңгейінде сақталуы аналитикалық және есептік жолмен расталды, бірақ бұл бірінші санаттың шартсыз берілуін білдірмейді. Тең салмақты аддитивті модельмен корреляция оң болды ($r = 0,405$). Қайта құрылған нормалау мәндерін бірге өзгерту паспорттардың 6,66–31,95%-ында қорытынды шешімді өзгертті. Нәтижелер модельді одан әрі жетілдіру және тексеру қажеттігін көрсетеді.

Түйін сөздер: аса маңызды цифрлық объектілер; бұлыңғыр қорытынды; есептік бағалау; монотондылық; нормалауға сезімталдық; компенсацияламау; қайта өндіру мүмкіндігі; киберқауіпсіздік.

1 Introduction

Critical digital objects underpin public services, economic activity and, in some cases, physical processes. A disruption may therefore have consequences far beyond the value of the equipment involved. Categorization should reflect those consequences and provide a clear rationale for the protection requirements that follow. Kazakhstan's Digital Code establishes a legal framework for critical digital objects [1]. The computational approach studied here remains a methodological proposal, not an approved regulatory instrument.

Identifying a critical object is not the same as assigning it a criticality category. ENISA's guidance links essential services to the infrastructure on which they depend [2], while studies of critical-infrastructure interdependencies show how disruptions can spread across technical, physical and organizational boundaries [3–5]. This motivates examining not only the severity of an incident, but also its geographic reach, potential cascading effects and the availability of substitutes.

Fuzzy sets and linguistic variables can represent expert assessments that do not fit neatly into fixed numerical classes [6, 7]. Mamdani inference translates those assessments into a rule-based scoring process [8]. Related methods have been used for failure prioritization [9], cyber-power grids [10], refinery assets [11] and hybrid FMECA [12]. More recent work considers richer representations and more efficient inference [13]. These studies motivate the approach, but do not validate the particular rules or thresholds used here. The aggregation and defuzzification choices still need to be assessed in the complete model [14, 15].

Matching a known example is a useful initial check, but it says little about how a model behaves elsewhere. Even if an individual aggregation operation is monotone, the full

sequence of fuzzification, rule activation and defuzzification may not be. It is also important to distinguish what a test demonstrates from what it merely fails to observe: one score decrease disproves monotonicity for the tested implementation, whereas finding no decreases in a random sample cannot prove monotonicity everywhere. Likewise, an interval with no sampled scores is not necessarily unreachable.

We therefore ask how consistently a fully specified reconstruction behaves under controlled tests, and how much its conclusions depend on normalization values missing from the original description. The analysis addresses four questions: RQ1, whether the tested inputs are covered by the inference rules and what scores they produce; RQ2, whether worsening one indicator can lower a score or decision status; RQ3, which severe-consequence preservation properties can be established independently of an additive comparison; and RQ4, how normalization choices and threshold adjustments affect the results.

This study contributes a reproducible computational assessment of the Variant 4 model [16]. It documents the experimental parameters and the provenance of normalization values, distinguishes fuzzy scores from automatic category assignment, and examines counterexamples under alternative thresholds. We also establish a limited analytical preservation result at the consequence-severity layer. We do not introduce a new inference algorithm, claim exhaustive coverage of the passport domain or validate the model against actual regulatory decisions. The V4-R1 experiments reported here were run anew and replace an earlier numerical series for which the underlying tables, weights, code and random seeds were not available.

2 Materials and methods

2.1 Source model and scope of reconstruction

Our starting point is the unpublished working draft, ‘Methodology for categorizing critical digital objects. Variant 4: a two-level model based on fuzzy inference’ [16]. The source document is identified in Supplementary Material S1 by its SHA-256 digest because it has no formal issue date or version identifier beyond ‘Variant 4’. We call our implementation V4-R1 to distinguish this reconstruction from the original authors' software, which was not available for inspection.

Variant 4 describes the passport fields, seven groups of membership functions, three rule bases, an 800-node defuzzification grid and three automatic-assignment grounds. Its normalization tables, however, are incomplete. We retain the documented functions and rules and make the missing numerical choices explicit in Table 1. Of the 63 mapping entries, 15 are anchored in the source, three cascade values appear in the supplied manuscript, and 45 are reconstruction assumptions. None of these assumptions is presented as an official parameter. The passport contains eighteen scoring fields (I5–I17 and I19–I23) and a separate binary automatic-assignment field (I24). There is no I18. This accounts for the apparent gap in indicator numbering.

2.2 Inference and decision stages

For each breach scenario s, passport p is mapped to six inputs on [0, 1]: societal, state and economic consequences (X1–X3), geographic extent (X4), cascading effects (X5) and irreplaceability (X6). Higher normalized values consistently indicate more severe consequences or less scope for substitution. Appendix B gives the scenario-specific mappings. Figure 1 shows the numerical inference chain separately from the automatic grounds that can override its final category.

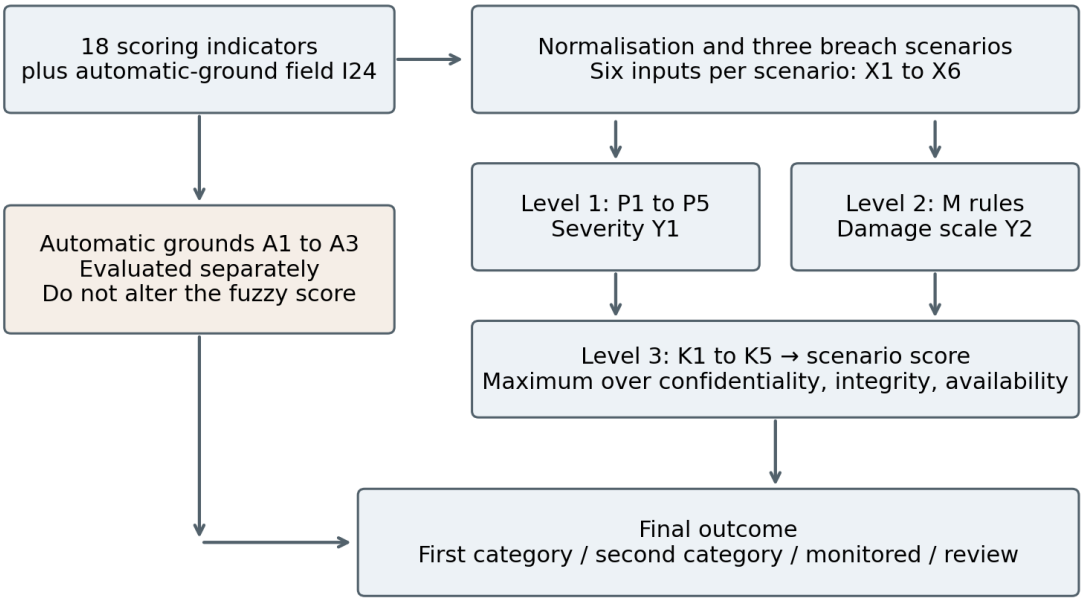


Figure 1. Computational structure of V4-R1. Automatic grounds override the decision outcome without changing the fuzzy score.

Membership functions are trapezoidal, with parameters (a, b, c, d). Membership equals one on the closed core [b, c], increases linearly from a to b, decreases linearly from c to d and is zero elsewhere. At coincident shoulder endpoints a = b or c = d, the core condition takes precedence. This convention prevents an endpoint from being assigned both zero and one. All parameter tuples appear in Appendix A. Conjunction, disjunction and negation are defined as follows:

$$u \wedge v = min(u, v), \quad u \vee v = max(u, v), \quad \neg u = 1 - u \quad (1)$$

Two branches are evaluated in parallel. Five Level 1 rules combine X1–X3 into consequence severity Y1, while ten Level 2 rules combine X4–X6 into damage scale Y2. Five Level 3 rules then produce the scenario score C_s from Y1 and Y2. Appendix B lists the antecedents, consequents and suppression factors. For each output term τ, capτ is the largest

activation among rules leading to that term. The aggregated membership function is defuzzified using a discrete centroid:

$$\mu*(x_i) = max\tau\, min(cap\tau, \mu\tau(x_i)), \quad Y = \Sigma_i\, x_i\, \mu*(x_i) \,/\, \Sigma_i\, \mu*(x_i) \quad (2)$$

The grid contains 800 equally spaced nodes including both endpoints, on [0, 1] for Y1 and Y2 and on [0, 100] for C_s. A zero denominator invokes the source's lower-endpoint fallback. Such a case is recorded separately as uncovered inference; receiving a numerical fallback is not counted as successful rule coverage. The maximum scenario score is

$$C(p) = max\,\{Cconf(p), Cint(p), Cavail(p)\} \quad (3)$$

Without an automatic ground, $C \geq 70$ yields the first category, $65 \leq C < 70$ yields referral for expert review, $30 \leq C < 65$ yields the second category, and $C < 30$ yields monitored status. The review interval is a procedural status, not a third criticality category. Automatic first-category assignment applies if I24 confirms a defence, emergency or mobilization role (A1); if the object is the sole provider of a national-level service (A2); or if at least two sectors are affected, dependent CDOs are present and substitution requires weeks or is impossible (A3). All comparisons use unrounded values. Membership-based K2 and K3 flags are retained in the output data, but they are not silently substituted for the score-based review interval: the source does not specify an operational firing threshold for converting these graded flags into referrals.

2.3 Complete experimental normalization

Table 1 defines the ordered levels used to generate the synthetic passports and their numerical mappings. S denotes a numerical anchor in Appendix 2 of the source methodology, M an anchor appearing only in Table 1 of the supplied manuscript, and R an assumption introduced for reconstruction. Values between fixed anchors are interpolated by ordinal level. Where no numerical anchor exists, the listed ordered values are explicit experimental choices. Such unanchored scales cannot be recovered by interpolation alone.

Table 1. Scoring indicators and normalization used in V4-R1

Indicator	Ordered levels	Normalized values and provenance
I5 Service significance	ordinary / sectoral / national	0.1 R / 0.6 R / 0.95 R
I6 Public service reach	none / individual users / mass public service	0.05 R / 0.45 R / 0.9 R
I7 Provider substitutability	one of many / main provider / sole provider	0.1 R / 0.65 R / 0.95 R
I8 Operating regime	working hours / extended hours / continuous	0.3 R / 0.6 R / 0.9 R

Indicator	Ordered levels	Normalized values and provenance
I9 Number of users	up to 10000 / 10001 to 124999 / 125000 to 500000 / 500001 to 2000000 / over 2000000	0.1 S / 0.325 R / 0.55 S / 0.75 R / 0.95 S
I10 Territorial coverage	one region / two to four regions / five or more regions / whole country	0.3 S / 0.55 R / 0.8 S / 0.95 S
I11 Sectoral share	up to 10 percent / over 10 to 30 percent / over 30 percent	0.1 R / 0.5 R / 0.9 R
I12 Dependent objects	fewer than 10 / 10 to 100 / over 100 / dependent CDO present	0.1 R / 0.35 R / 0.65 R / 0.95 R
I13 Data category	open / personal / special personal / restricted	0.05 S / 0.4 R / 0.75 S / 0.9 S
I14 Personal records	up to 10000 / 10001 to 50000 / 50001 to 100000 / 100001 to 300000 / over 300000	0.1 R / 0.3125 R / 0.525 R / 0.7375 R / 0.95 R
I15 Legal significance of data	none / supporting / rights establishing	0.05 R / 0.55 R / 0.95 R
I16 Integrity recoverability	full / partial / impossible	0.05 R / 0.55 R / 0.95 R
I17 Cascade character	none / within organisation / within sector / other sectors or CDOs	0.05 M / 0.35 R / 0.65 M / 0.95 M
I19 Affected sectors	one / two / three	0.15 R / 0.55 R / 0.95 R
I20 Substitution time	hours / days / weeks / impossible	0.15 S / 0.45 R / 0.75 S / 0.95 S
I21 Time to consequences	at least one day / hours / minutes	0.2 R / 0.55 R / 0.9 R
I22 Redundancy	geographical / partial / none	0.05 R / 0.55 R / 0.95 R
I23 Physical process link	none / indirect / direct	0.05 S / 0.5 S / 0.9 S

The original draft does not define every intermediate count range for I9 and I14. The boundaries shown in Table 1 are therefore experimental choices, not regulatory cutoffs. Where both I12 descriptions apply, the dependent-CDO state takes precedence over the count state. Including I24, the encoded domain contains 9,069,926,400 possible level

combinations. This is a Cartesian product of the declared fields; many combinations may not describe plausible real objects.

To test sensitivity to those choices, we construct two additional variants by shifting all reconstruction-only (R) values together by −0.10 or +0.10. We clip the values to [0, 1] and to the nearest fixed S or M anchors, preserving indicator order and the source values. All three variants use the same passports and worsening pairs. These coordinated changes are sensitivity tests, not a probability model for parameter uncertainty or bounds on every possible reconstruction.

2.4 Samples and comparison design

We generate a baseline of 30,000 synthetic passports. For each of the eighteen scoring indicators, an ordinal level is sampled independently with equal probability; I24 is independently set to 'yes' with probability 0.05. Generation uses NumPy's PCG64 generator with seed 20261021. Two further samples of 30,000 use seeds 20261024 and 20261025. This design deliberately explores combinations of declared levels. It does not reproduce the distribution of real critical digital objects in Kazakhstan and should not be used to estimate their prevalence.

The monotonicity experiment uses seed 20261022 to draw 40,000 passport-row and scoring-indicator pairs, both with replacement. In each pair, we worsen the selected indicator by one ordinal level and leave every other field, including I24, unchanged. A step is skipped when the indicator is already at its highest level. We retain row identifiers so duplicate tests can be distinguished from unique pairs. Automatic grounds are recalculated after every applicable change.

For comparison, we use a simple additive score, $A(p) = (100/18)\Sigma_j \varphi_j(p)$, with equal weight 1/18 for every scoring indicator. These are deliberately chosen weights, not a reconstruction of the undocumented expert weights used in an earlier comparison. We select the same number of first-category passports as the fuzzy score does before automatic overrides. Additive scores are ranked without rounding, and sample identifiers break ties. Equalizing the number selected makes the sets comparable in size, not necessarily in quality.

We also test the inference core directly using 30,000 independent vectors drawn uniformly from $[0, 1]^6$ (seed 20261023). This bypasses the mapping from passport fields to model inputs. Some vectors may not correspond to any valid passport, so the experiment does not establish passport-level reachability. Instead, it helps determine whether a restricted-looking score distribution comes from the inference rules themselves or from the chosen input mappings.

2.5 Properties and test criteria

Coverage and empirical output. For every evaluated scenario, the test records finite output, the permitted output interval and a nonzero centroid denominator at all three levels. Sorted output values are grouped by a greedy rule: a group extends only while its largest

value differs from its first value by at most $\varepsilon = 10^{-8}$ score points. This bounds each group's diameter and avoids chaining pairwise closeness. Counts are also computed at other tolerances. A finite list of observed groups is a property of the evaluated sample and numerical convention, not a proof of the full attainable set.

Monotonicity. Let p' differ from p only by a one-level worsening of one scoring indicator. Score monotonicity requires $C(p') \geq C(p)$; a computational violation is recorded when

$$C(p') < C(p) - \varepsilon, \quad \varepsilon = 10^{-8} \quad (4)$$

We separately compare first-category membership and decision status. For diagnostic ordering only, statuses are coded monitored = 0, second category = 1, review referral = 2 and first category = 3. This convention treats removal of a required referral as a decrease in procedural scrutiny; it does not turn referral into a legal category. Category and referral comparisons use the documented thresholds and recomputed automatic grounds. Rates are descriptive of the sampling procedure. Because pairs can reuse passports and include duplicates, no independence-based confidence interval is attached to the violation rate.

Severe-consequence preservation. We define a restricted non-compensation property at the consequence layer: once one of X1–X3 is fully critical, changing the other two consequence inputs cannot reduce Y1. Full critical membership occurs for $X_j \in [0.90, 1]$. The property is checked analytically from P1–P5 and numerically at $X_j = 0.90$, for each of the three choices of j, while the other inputs span $\{0, 0.05, \ldots, 1\}^2$. This gives 1,323 cases. Separately, the scale inputs span $\{0, 0.05, \ldots, 1\}^3$, giving 9,261 combinations. We test the resulting scores and the conditional first-category property when the scale has positive high membership and zero low membership. This separates preserved consequence severity from the further scale requirements of the decision rule.

Threshold sensitivity. For any first-category threshold t, the binary first-category decision is $F_t(p) = 1$ when $C(p) \geq t$ or an automatic ground holds. We calculate the score-only and final first-category shares separately. For a worsening pair with no automatic ground after the change, a first-category reversal occurs exactly when

$$C(p') < t \leq C(p) \quad (5)$$

The source's automatic predicates are nondecreasing under the declared ordinal changes, so a pair that has no automatic ground after the change also had none before it. Equation (5) therefore identifies the threshold interval exposed by each such score decrease. We evaluate integer thresholds from 35 to 85 and all relevant observed score endpoints. For full decision-status sensitivity, the review band retains width five, $[t - 5, t)$, and the monitored threshold remains 30. These alternatives are diagnostic recalibrations, not proposed regulatory thresholds.

2.6 Numerical implementation and reproducibility

The implementation uses Python 3.12.14 and NumPy 2.3.5 with binary64 arithmetic. The archive contains model.py, specification.json, run_experiments.py, figure-generation code, exact seeds, the source methodology, passport and score CSV files, paired monotonicity records, output groups and machine-readable summaries. The code stores all scenario scores, intermediate values, automatic-ground flags and K2/K3 activations. A reference-example check and shoulder-endpoint assertions are executed before the main experiments. Numerical sensitivity is assessed by repeating the entire baseline calculation at 400 and 1,600 centroid nodes and comparing it with 800 nodes. These checks establish specific implementation agreements; they do not substitute for independent code review or full formal verification.

3 Results

3.1 Agreement with the reported reference inputs

Table 2 reproduces the six scenario inputs reported for the regional heat-supply example in the supplied manuscript. V4-R1 returns 49.322933 for all three scenarios, matching the published rounded value of 49.3. The example also meets automatic ground A3: dependent critical digital objects are present, at least two sectors are affected, and substitution is impossible. Its fuzzy score therefore implies the second category, while the automatic rule assigns the first. This agreement checks the stated intermediate inputs and override logic; it does not reconstruct every field of the original passport.

Table 2. Reference check using the reported scenario inputs

Scenario	X1	X2	X3	X4	X5	X6	Y1	Y2	C
Confidentiality	.35	.45	.40	.30	.95	.95	.525000	.557154	49.322933
Integrity	.55	.45	.55	.30	.95	.95	.452633	.557154	49.322933
Availability	.55	.60	.55	.30	.95	.95	.452633	.557154	49.322933

This single-point agreement is reassuring but limited. It tests only the rules activated by that example and cannot recover omitted normalization entries, establish coverage of every rule combination or validate the model for other types of object.

3.2 Coverage and empirical output distribution

All 90,000 scenario calculations for the baseline passports produced nonzero aggregated membership at every inference level, and all maximum scores stayed within [0, 100]. The sample minimum was 16.117982 and the maximum 83.511428; the mean was 69.109277 and the median 82.661771. These figures describe the sampled passports only. They do not rule out other scores elsewhere in the full domain.

Table 3. Baseline outcomes before and after automatic assignment

Outcome	Score count	Score share	Final count	Final share
First category	16056	53.52%	18883	62.94%
Review referral	1768	5.89%	1394	4.65%
Second category	11931	39.77%	9515	31.72%
Monitored	245	0.82%	208	0.69%

At ε = 10^{-8}, 56 output groups were observed. The same count was obtained for tolerances from 10^{-10} to 10^{-4}; grouping with tolerance 0.005 merged one additional pair and gave 55 groups. The largest observed gap was 12.137624 points, between 69.749153 and 81.886777. The review interval [65, 70) contained 1,768 passports (5.89%) before overrides and 1,394 (4.65%) after overrides. Figure 2 therefore contradicts a universal assertion that this interval is unreachable under the documented rule architecture. Its occupancy depends on the declared input mapping.

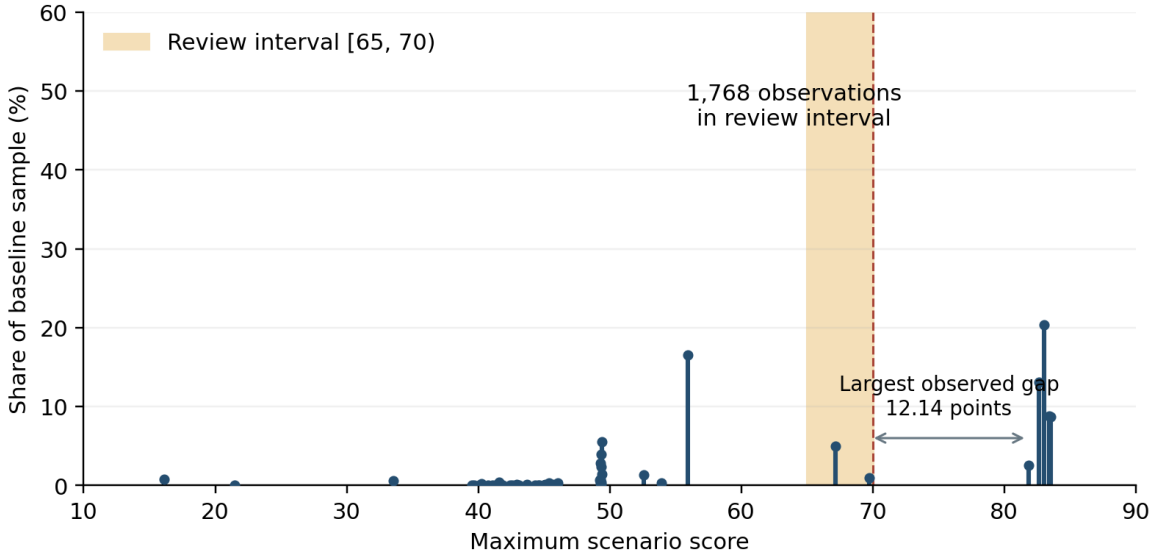


Figure 2. Empirical output distribution for the baseline 30,000 passports. Stems show observed score groups, not an enumeration of all attainable outputs.

Automatic grounds applied to 6,782 passports (22.61%). A1 applied to 1,516, A2 to 3,336 and A3 to 2,507; these groups overlap, so their counts cannot be summed. For 2,827 passports, automatic assignment alone changed the outcome to the first category. The final first-category count was therefore 16,056 + 2,827 = 18,883 (62.94%). Reporting the counts alongside the percentages avoids rounding discrepancies.

In 14,907 passports (49.69%), two or more scenarios tied for the maximum within 10^{-8} points. A first-maximum convention in the order confidentiality, integrity, availability selected these scenarios 10,343, 12,255 and 7,402 times. These counts are bookkeeping outcomes under the tie rule; they should not be interpreted as evidence that one breach type intrinsically governs more often.

3.3 Score monotonicity and decision status

Of 40,000 proposed one-level worsening steps, 28,166 were applicable. In 1,237 of these (4.39%), the fuzzy score fell by more than 10^{-8}. After removing repeated row–indicator pairs, 1,185 violations remained among 27,184 unique tests (4.36%). The effect therefore cannot be explained solely by duplicate sampling. Each recorded decrease is a counterexample to score monotonicity in V4-R1; the overall frequency across the full domain remains unknown.

Table 4. Magnitudes and consequences of observed score decreases

Quantity	Result
Applicable steps / distinct pairs	28,166 / 27,184
Score decreases / distinct-pair decreases	1,237 / 1,185
Median / mean decrease	0.365598 / 0.547403 points
95th percentile / maximum decrease	1.624651 / 9.208157 points
First-category reversals at t = 70	0
Decision-status decreases at original thresholds	0
I10 applicable steps / violations	1,596 / 760
I10 within-indicator violation rate	47.62%

Indicator I10 (territorial coverage) accounted for 760 decreases, or 61.44% of all observed violations. Among the 1,596 applicable I10 steps, 47.62% reduced the score. These percentages describe different denominators and should not be conflated. I19 (affected sectors) and I17 (cascade character) contributed 127 and 99 decreases, respectively. These results identify useful starting points for tracing the rule activations, but do not show that any one parameter change would resolve every violation.

The maximum worsening-related score reduction was greater than nine points, so the observed decreases are not all negligible numerical effects. Nevertheless, none of the tested pairs moved to a lower decision status at the original cutoffs of 30, 65 and 70. As Section 3.7 demonstrates, this apparent decision stability depends on the thresholds.

3.4 Severe consequences and the scope of non-compensation

A restricted preservation property follows directly from the specified consequence rules. If any Xj lies in [0.90, 1], its critical membership is one and its low membership is zero. Therefore P1 activates at one, P2, P3 and P5 are suppressed to zero, and P4 is zero. The aggregated Y1 output contains only the fully activated critical term. Its 800-node centroid is 0.844768, independent of the other two consequence inputs. Since this value lies inside the critical core of Y1, the downstream critical membership is also one. This proves preservation at the severity layer for a fully critical input under the stated numerical convention.

The 1,323 targeted severity cases all returned Y1 = 0.844768. Across the 9,261 scale combinations, final scenario scores ranged from 49.237220 to 83.511428, and 4,679 were below 70. Thus the strong claim that any fully critical consequence necessarily produces first-category assignment is false for this model. The rule base deliberately also requires sufficient scale. For the 4,120 tested combinations with high-scale membership greater than 10^{-8} and low-scale membership at most 10^{-8}, no score was below 70; the minimum was 81.567688.

The results support a precise statement: other consequence inputs cannot compensate away a fully critical consequence at the Y1 layer, and the final score preserves this influence at fixed scale. They do not establish non-compensation for partially activated critical terms, for arbitrary changes in territorial or cascade conditions, or for every individual passport indicator before X1–X3 are formed. A severity guarantee should therefore be specified at the layer where it is required and tested separately from the rule defining the first category.

3.5 Comparison with the additive model

The Pearson correlation between maximum fuzzy scores and equal-weight additive scores was 0.404610, indicating a positive association. The equal-volume comparison selected 16,056 first-category passports in each model. The additive cutoff was 52.916667; 46 passports had exactly the cutoff score in the implementation, of which 26 were selected by the declared sample-identifier rule. This tie handling is part of the experimental comparison, not a recommended administrative decision rule.

The fuzzy and additive first-category sets differed for 10,658 passports (35.53%). Of these, 5,329 were selected only by the fuzzy score and 5,329 only by the additive score. Figure 3 illustrates how passports with similar additive scores can receive quite different fuzzy scores. The comparison establishes that the models aggregate information differently; it does not tell us which selections are preferable, or prove non-compensation. We examine the latter property directly in Section 3.4. Other additive weights or calibration targets could produce different levels of agreement.

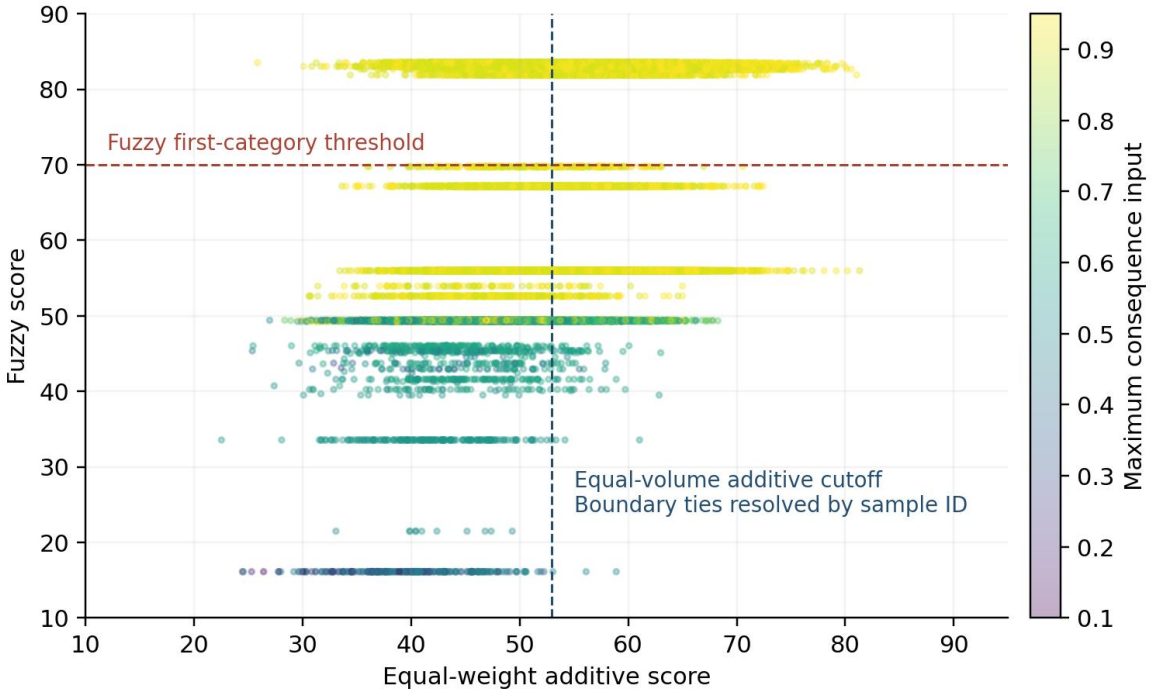


Figure 3. Fuzzy and equal-weight additive scores for the same baseline passports. Automatic grounds are excluded from both first-category sets in this comparison.

3.6 Sensitivity to reconstructed normalization

Table 5 and Figure 4 show that normalization uncertainty materially affects the findings. Changing only R entries, while retaining every fixed anchor and the entire rule architecture, changed the final outcome for 9,585 passports under the lower variant and 1,997 under the upper variant. These correspond to 31.95% and 6.66% of the same baseline sample. The largest absolute score changes were 34.106021 and 30.901588 points. The shifts were joint and coordinated, so these figures characterize sensitivity scenarios rather than confidence limits.

Table 5. Normalization sensitivity on identical passports and worsening steps

Quantity	Lower	Baseline	Upper
Observed output groups	48	56	38
First category by score	19.65%	53.52%	61.02%
Final first category	37.42%	62.94%	69.14%
Scores in [65, 70)	0	1,768	0
Final review referrals	0	1,394	0
Changed final outcome versus baseline	9,585	—	1,997
Score decreases in 28,166 steps	1,446	1,237	1,067
Decision-status decreases at original thresholds	0	0	0

The review band contained scores in the baseline sample, but was empty in both alternative normalization variants. Its occupancy therefore depends on values that the source methodology does not fully specify. The number of observed output groups also varied. Neither feature can be explained as an unavoidable consequence of the three-level rule architecture or centroid defuzzification alone.

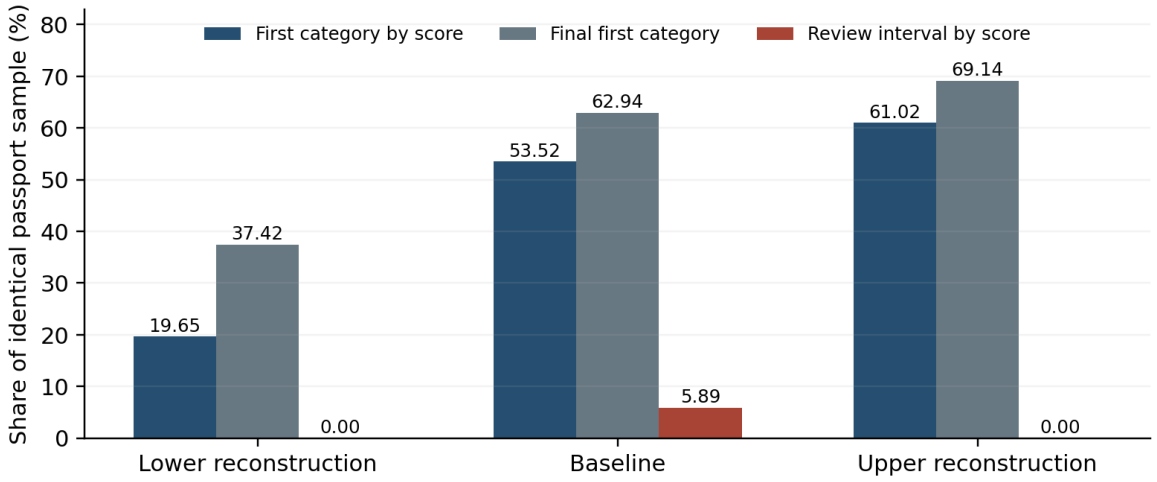


Figure 4. Effect of jointly shifting reconstructed normalization entries while preserving fixed anchors. Percentages refer to the same 30,000 passports.

Two independently generated samples, using seeds 20261024 and 20261025, produced 58 and 53 output groups and included 1,696 and 1,735 scores in the review interval. Their score-based first-category shares were 53.63% and 53.29%, while final first-category

shares were 63.03% and 62.52%. The broad distributional pattern was similar across these runs. The exact number of observed score groups, however, is a sample result, not the cardinality of the full attainable output set.

3.7 Threshold recalibration exposes decision reversals

First-category prevalence and monotonicity answer different questions. A threshold can produce an apparently stable aggregate share while concealing decreases for individual objects. At t = 70, no first-category reversal occurred in the tested worsening pairs. At t = 68, 24 reversals remained after automatic grounds were applied. At t = 83, there were 386 final first-category reversals. Table 6 uses exact tested thresholds; it does not describe a plateau as a closed interval when one endpoint already changes the result.

Table 6. Diagnostic threshold changes using unchanged scores and paired inputs

Threshold t	Score first count	Final first count	Final first reversals	Status decreases
55	22776	24004	0	0
68	16338	19097	24	24
70	16056	18883	0	0
73	16056	18883	0	24
83	11364	15430	386	386

For example, baseline passport 21667 produced 69.749153 points. Advancing indicator I19 by one ordinal level reduced the score to 67.160956, with no automatic ground before or after the change. Both scores lie in the original review interval, explaining the absence of a status change at t = 70. At t = 68, however, the first category is lost after worsening. The complete passport and paired record are included in S1. This observed counterexample shows why category stability must be retested after calibration.

For the baseline sample, the score-only first-category set is constant for thresholds t in (69.749153, 81.886777], where the endpoints shown are rounded for display and exact values are supplied in the data. The left endpoint is excluded because observations equal to it still qualify under $C \geq t$; the right endpoint is included because observations at that score remain selected. This is an empirical plateau of one sample, not a domain-wide insensitivity guarantee. It also does not guarantee unchanged referral status, because moving the lower edge of the review interval can change the treatment of lower scores.

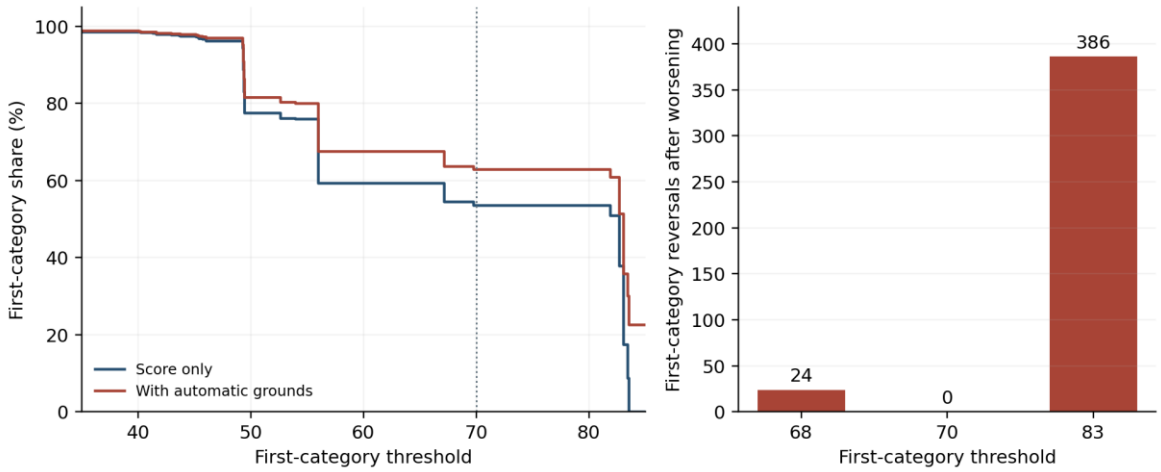


Figure 5. Threshold dependence of score-only and final first-category shares, and first-category reversals in the worsening tests. Alternative thresholds are diagnostic choices.

3.8 Numerical sensitivity and relaxed-domain diagnostics

Relative to the 800-node baseline, using 400 nodes changed a score by at most 0.120347 points and using 1,600 nodes by at most 0.060376 points. Neither alternative changed a final decision status at the original thresholds in the baseline sample. These results quantify sensitivity to the tested quadrature resolutions. They do not prove that every finer grid would yield identical categories, especially after thresholds are moved.

The continuous-input diagnostic produced 14,597 output groups and 277 scores in [65, 70), without invoking the zero-area fallback. This much richer distribution indicates that the inference architecture itself does not limit outputs to a few dozen discrete values. Because these input vectors do not necessarily correspond to valid passports, the result concerns the inference core rather than the full passport domain.

4 Discussion

4.1 What the computational evidence establishes

RQ1 receives a sample-limited positive answer: inference was covered at all stages for the tested passports and direct-input vectors. The source's fallback rule means that numerical finiteness alone would have been insufficient evidence; counting zero-area cases is necessary. The observed score distribution and its gaps characterize the implemented reconstruction and sampling design. They cannot be promoted to an exhaustive statement about the full passport domain.

For RQ2, the recorded one-level worsening pairs establish that V4-R1 is not score-monotone. The original thresholds masked these score decreases at the decision-status level in the tested pairs, but Equation (5) explains when a decrease becomes a category reversal.

The 24 reversals observed at a threshold of 68 show that this can occur near the original first-category threshold of 70.

RQ3 calls for a distinction between model layers. When one consequence input is fully critical, the consequence rule base preserves that severity regardless of the other two consequence inputs. Whether the object reaches the first category still depends on damage scale and any automatic grounds. The additive comparison shows a positive association but substantial disagreement between the scoring methods. It cannot establish which method better reflects the intended legal or substantive criteria.

For RQ4, the uncertainty surrounding normalization is substantial. Changing the reconstructed mappings altered review-band occupancy and final categories far more than changing the centroid grid from 800 to 400 or 1,600 nodes. Specifying and calibrating the passport-to-input mapping is therefore more urgent than interpreting the baseline histogram as an inherent model feature. Our coordinated lower and upper shifts are only two possibilities; varying individual mappings, bin limits or scenario assumptions could reveal other sensitivities.

4.2 Implications for model revision

Three changes should precede consideration of regulatory use. First, the methodology needs a complete versioned specification covering indicator domains, normalization, scenario equations, endpoint conventions, inference fallbacks and decision precedence. Parameters should be justified against the intended consequence definitions. Calibration should not merely target a preferred proportion of first-category objects in an artificial sample.

Second, the stored counterexamples should be traced through the intermediate memberships and activated rules. Territorial coverage and cascading effects are reasonable places to begin because they account for many observed decreases. Possible remedies include revising rule consequents or overlaps, or changing a hierarchical inference stage. None has been tested here. Any revision should be evaluated on the existing counterexamples, a broader targeted search and the threshold tests—not just a few selected cases.

Third, the purpose of expert review must be stated operationally. A numerical band, disagreement between adjacent memberships, K2/K3 flags and low confidence in passport evidence represent different referral criteria. Membership-based referral could complement a score band, but its activation rule, threshold and workload would require validation. An empty band in one calibration does not establish that all expert-review mechanisms are dormant, and moving a threshold cannot be assumed to repair the underlying score behavior.

4.3 Limitations and external validation

V4-R1 is an explicit reconstruction of an incompletely specified draft. The documented rule and membership parameters are retained, but 45 normalization entries and

some count-bin boundaries are assumptions. The baseline statistics therefore apply to this reconstruction, not automatically to the original unpublished implementation. Although the archive makes the new experiments reproducible, recovering the earlier code and parameter files remains necessary to explain the exact numerical differences from the previous version of the manuscript.

The synthetic sampling design does not reproduce the relationships seen in practice between service importance, user counts, territorial reach and technical dependencies. Some combinations may be implausible. Such cases can still be valuable for stress testing, but their frequencies say nothing about actual exposure or regulatory workload. We also lack independently assessed real objects and observed incident outcomes. Accordingly, false-positive and false-negative categorization rates cannot be estimated from this study.

The monotonicity tests cover sampled one-level changes, not every possible transition or simultaneous change. The normalization analysis considers only two coordinated alternatives, and the analytical preservation result concerns fully critical inputs at one particular layer. Before deployment, the model would need additional tests for missing or uncertain passport evidence, exhaustive or formal checks on manageable subdomains, expert-reviewed reference cases and validation against real consequences. Until then, claims about regulatory suitability must remain conditional.

5 Conclusion

We reconstructed and tested the Variant 4 fuzzy categorization model using explicitly documented parameters, random seeds and synthetic data. All tested inference paths were covered, but 1,237 of 28,166 applicable worsening steps lowered the score. Although none lowered decision status at the original thresholds, recalibrating the first-category cutoff to 68 produced 24 first-category reversals after automatic grounds were considered.

The output distribution also depended on reconstruction choices. The baseline included 56 observed score groups and 1,768 passports in the [65, 70) review interval, whereas the two alternative normalizations left that interval empty. A small set of sampled outputs—or an empty review band in one run—therefore cannot be treated as a universal property of the inference model. We established preservation of fully critical consequences at the severity layer, but not unconditional first-category assignment.

The main practical contribution is a reproducible set of counterexamples and a way to assess how normalization and threshold choices affect decisions. Before the model can be judged suitable for regulatory use, it needs an agreed and complete specification, an explanation or correction of its non-monotone responses, and validation on independently assessed real cases. Computational tests can inform that process, but cannot replace it.

Supplementary material

Supplementary Material S1 is supplied as Fuzzy_Categorization_Reproducibility.zip. It contains the source methodology, the V4-R1

specification, executable model and experiment scripts, figure-generation code, requirements, generated passport and score files, paired worsening records, diagnostic experiments and a results summary. The README identifies reconstruction assumptions and provides reproduction commands. The archive concerns the new V4-R1 experiments and does not claim to contain the unavailable historical implementation.

Author contributions

Conceptualization: L.R.; Methodology: F.K. and L.R.; Software: F.K. and N.N.; Validation: F.K., L.R. and N.N.; Formal analysis: F.K.; Investigation: N.N.; Data curation: N.N.; Writing — original draft preparation: F.K.; Writing — review and editing: L.R. and N.N.; Visualization: N.N.; Supervision: L.R.; Project administration: L.R.

Conflicts of interest

The authors declare no conflict of interest.

Data and code availability

The code and generated data supporting the V4-R1 results are included in Supplementary Material S1 supplied with this revision. No public repository DOI is claimed. The synthetic records do not represent identified real objects. The study makes no claim to provide the original implementation, random seeds or expert weights underlying the superseded numerical series.

References

[1] Digital Code of the Republic of Kazakhstan No. 255-VIII. (2026, January 9). Adilet Legal Information System. https://adilet.zan.kz/rus/docs/K2600000255

[2] Mattioli, R., & Levy-Bencheton, C. (2014). Methodologies for the identification of critical information infrastructure assets and services. European Union Agency for Network and Information Security. https://doi.org/10.2824/38100

[3] Rinaldi, S. M., Peerenboom, J. P., & Kelly, T. K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE Control Systems Magazine, 21(6), 11–25. https://doi.org/10.1109/37.969131

[4] Kelic, A., Warren, D. E., & Phillips, L. R. (2008). Cyber and physical infrastructure interdependencies (SAND2008-6192). Sandia National Laboratories.

[5] Malinichev, D. M., Kuchmezov, Kh. Kh., & Mochalov, V. V. (2022). Kategorirovanie vzaimosvyazannykh ob'ektov kriticheskoi informatsionnoi infrastruktury [Categorization of interconnected critical information infrastructure objects]. Applied Informatics, 17(3), 105–116. https://doi.org/10.37791/2687-0649-2022-17-3-105-116

[6] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[7] Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—I. Information Sciences, 8(3), 199–249. https://doi.org/10.1016/0020-0255(75)90036-5

[8] Mamdani, E. H., & Assilian, S. (1975). An experiment in linguistic synthesis with a fuzzy logic controller. International Journal of Man-Machine Studies, 7(1), 1–13. https://doi.org/10.1016/S0020-7373(75)80002-2

[9] Bowles, J. B., & Peláez, C. E. (1995). Fuzzy logic prioritization of failures in a system failure mode, effects and criticality analysis. Reliability Engineering & System Safety, 50(2), 203–213. https://doi.org/10.1016/0951-8320(95)00068-D

[10] Zúñiga, A. A., Fernandes, J. F. P., & Branco, P. J. C. (2023). Fuzzy-based failure modes, effects, and criticality analysis applied to cyber-power grids. Energies, 16(8), 3346. https://doi.org/10.3390/en16083346

[11] Bahadir, C., & Kahraman, C. (2026). Criticality analysis of refinery assets using picture fuzzy inference system in reliability centered maintenance. International Journal of Fuzzy Systems, 28(5), 1767–1781. https://doi.org/10.1007/s40815-025-02069-y

[12] Chakhrit, A., Djelamda, I., Bougofa, M., Guetarni, I. H. M., Bouafia, A., & Chennoufi, M. (2024). Integrating fuzzy logic and multi-criteria decision-making in a hybrid FMECA for robust risk prioritization. Quality and Reliability Engineering International, 40(6), 3555–3580. https://doi.org/10.1002/qre.3601

[13] Ontiveros, E., Melin, P., & Castillo, O. (2024). Towards an efficient approach for Mamdani interval type-3 fuzzy inference systems. International Journal of Fuzzy Systems, 26, 2172–2190. https://doi.org/10.1007/s40815-024-01722-2

[14] Grabisch, M., Marichal, J.-L., Mesiar, R., & Pap, E. (2009). Aggregation functions. Cambridge University Press.

[15] van Leekwijck, W., & Kerre, E. E. (1999). Defuzzification: Criteria and classification. Fuzzy Sets and Systems, 108(2), 159–178. https://doi.org/10.1016/S0165-0114(97)00337-0

[16] Methodology for categorizing critical digital objects. Variant 4: a two-level model based on fuzzy inference. Undated unpublished working draft. Source file “Вариант 4(3).docx”, archived in Supplementary Material S1; SHA-256 recorded in specification.json.

Appendix A Membership function parameters

The tuples below are transcribed from Appendix 4 of the source methodology. Every tuple gives (a, b, c, d). All variable domains are [0, 1], except the category-score variable C, whose domain is [0, 100]. Shoulder endpoints follow the core-first convention stated in Section 2.2.

Table A1. Complete membership-function specification

Variable	Term	Parameters
Severity X1–X3	low	(0, 0, 0.2, 0.4)
Severity X1–X3	medium	(0.2, 0.4, 0.5, 0.7)
Severity X1–X3	high	(0.5, 0.7, 0.78, 0.9)
Severity X1–X3	critical	(0.78, 0.9, 1, 1)
Territory X4	organisational	(0, 0, 0.2, 0.4)
Territory X4	regional	(0.2, 0.4, 0.5, 0.62)
Territory X4	interregional	(0.5, 0.62, 0.72, 0.88)
Territory X4	national	(0.72, 0.88, 1, 1)
Cascade X5	isolated	(0, 0, 0.25, 0.45)
Cascade X5	intrasectoral	(0.25, 0.45, 0.6, 0.8)
Cascade X5	crosssectoral	(0.6, 0.8, 1, 1)
Irreplaceability X6	high substitutability	(0, 0, 0.25, 0.45)
Irreplaceability X6	medium substitutability	(0.25, 0.45, 0.55, 0.72)
Irreplaceability X6	low substitutability	(0.55, 0.72, 0.82, 0.93)
Irreplaceability X6	irreplaceable	(0.82, 0.93, 1, 1)

Variable	Term	Parameters
Consequence criticality Y1	insignificant	(0, 0, 0.25, 0.45)
Consequence criticality Y1	significant	(0.25, 0.45, 0.6, 0.8)
Consequence criticality Y1	critical	(0.6, 0.8, 1, 1)
Damage scale Y2	organisational	(0, 0, 0.2, 0.4)
Damage scale Y2	sectoral regional	(0.2, 0.4, 0.52, 0.66)
Damage scale Y2	interregional	(0.52, 0.66, 0.76, 0.9)
Damage scale Y2	national	(0.76, 0.9, 1, 1)
Category score C	monitored	(0, 0, 20, 42)
Category score C	second	(20, 42, 58, 78)
Category score C	first	(58, 78, 100, 100)

Appendix B Scenario inputs and complete rule bases

Let φ_i be the normalized value of indicator Ii; $\wedge$ and $\vee$ denote minimum and maximum. The following equations reproduce the specified scenario mapping. For a confidentiality breach, the special X2 branch applies only to the restricted-data category.

Table B1. Consequence inputs by breach scenario

Scenario	X1 societal	X2 state	X3 economic
Confidentiality	$\varphi_{13} \wedge \varphi_{14}$	φ_{13} if restricted; otherwise $\varphi_{13} \wedge \varphi_6$	$\varphi_{11} \wedge 0.40$
Integrity	$[0.9(\varphi_{13} \wedge \varphi_{14})] \vee (\varphi_{23} \wedge \varphi_{16})$	$\varphi_{15} \vee (\varphi_6 \wedge \varphi_{16})$	$\varphi_{11} \wedge \varphi_{16}$
Availability	$(\varphi_{23} \wedge \varphi_9) \vee (\varphi_{21} \wedge \varphi_9)$	$\varphi_6 \vee (\varphi_5 \wedge \varphi_8)$	$(\varphi_{11} \wedge \varphi_{21}) \vee (\varphi_{12} \wedge \varphi_{21})$

$X4 = \varphi_{10}$. $X5 = \varphi_{17} \vee \varphi_{19} \vee w$, where $w = \varphi_{12} \wedge 0.80$ when dependent CDOs are present, and $w = 0$ otherwise. $X6 = \varphi_7 \vee \varphi_{20} \vee \varphi_{22}$, additionally combined by maximum with φ_{16} for the integrity scenario. Constants 0.9, 0.40 and 0.80 are retained from the source. They are modeling choices, not empirical estimates validated in this study; in particular, the confidentiality cap does not establish a general upper bound on real economic loss from a disclosure.

For level 1, l_j, m_j, h_j and c_j denote low, medium, high and critical severity memberships of Xj for j = 1, 2, 3. Let m(1) and m(2) be the largest and second-largest medium memberships. The notation αr denotes activation of rule r.

Table B2. Level 1 consequence rules

Rule	Activation	Y1 consequent
P1	$\max(c_1, c_2, c_3)$	Critical
P2	$\min[\max(h_1, h_2, h_3), 1 - \alpha P1]$	Significant
P3	$\min[m(2), 1 - \alpha P1]$	Significant
P4	$\min(l_1, l_2, l_3)$	Insignificant
P5	$\min[0.6m(1), 1 - \alpha P1, 1 - \alpha P2, 1 - \alpha P3]$	Insignificant

For level 2, o, r, i and n are the organizational, regional, interregional and national memberships of X4. The cascade memberships of X5 are z (isolated), v (within sector) and q (cross-sector). The irreplaceable membership of X6 is u.

Table B3. Level 2 scale rules

Rule	Activation	Y2 consequent
M1	n	National
M2a	i	Interregional
M2b	$q \wedge (r \vee i \vee n)$	Interregional
M2c	$q \wedge o$	Sectoral-regional
M3	$r \wedge (z \vee v)$	Sectoral-regional
M4	$o \wedge z$	Organizational
M4b	$o \wedge v$	Sectoral-regional
M5a	$u \wedge o$	Sectoral-regional
M5b	$u \wedge r$	Interregional
M5c	$u \wedge i$	National

For level 3, let $H = \mu_{national}(Y2) \vee \mu_{interregional}(Y2)$ and $L = \mu_{sectoral\text{-}regional}(Y2) \vee \mu_{organizational}(Y2)$. Let c, s and n denote the critical, significant and insignificant memberships of Y1, respectively. These symbols are local to Table B4.

Table B4. Level 3 category-score rules

Rule	Activation	C consequent and flag
K1	c ∧ H	First
K2	c ∧ L	Second; completeness-control flag
K3	s ∧ H	Second; review-related flag
K4	s ∧ L	Second
K5	n	Monitored

At every level, the cap for an output term is the maximum activation of its rules; the output is then defuzzified using Equation (2). Thus the category-score caps are αK1 for first, max(αK2, αK3, αK4) for second, and αK5 for monitored. The graded flags are exported for inspection. The source's score thresholds and automatic grounds determine the reported decision outcome.

Leila Rzayeva
PhD, Associate Professor
e-mail: l.rzayeva@govtec.kz
ORCID: 0000-0002-3382-4685
Digital Government Support Center, Astana, Republic of Kazakhstan

Aina Tazhiyeva*
Master of Science, Researcher
e-mail: a.tazhiyeva@govtec.kz
ORCID: 0009-0008-3211-7840
Digital Government Support Center, Astana, Republic of Kazakhstan

Iskander Tazhibekov
Master of Business Administration
e-mail: i.tazhibekov@mdai.gov.kz
ORCID: 0009-0003-6517-9355
Information security Committee of the Ministry of Artificial Intelligence
and Digital Development of the Republic of Kazakhstan

CYBER INCIDENT DAMAGE ASSESSMENT FOR PUBLIC DIGITAL SYSTEMS AND CRITICAL INFRASTRUCTURE

Abstract: Cyber incident damage estimates can change materially when overlapping costs, normative coefficients and classification thresholds are combined without a common measurement boundary. This study integrates a comparative narrative review of 52 sources with formal accounting and sensitivity analysis for public digital services and critical infrastructure. An observation-to-loss model identifies the exact inflation caused by repeated items. A monotonicity criterion determines when a financial category is stable over a declared uncertainty set, while operational and safety escalation remain separate. The analytical conditions and reproducible calculations are applied to two demonstrations from unpublished national drafts and to a controlled two-sector propagation model. In the original portal case, the reputation convention accounts for 90.96% of the total and has elasticity 0.90956; the critical financial threshold is reached at 42.905% of its original magnitude. Removing the convention changes financial severity from critical to moderate without necessarily changing operational escalation. Exact scenario bounds are 20.8265–34.109 million KZT for the revised wider portal account and 67.912–83.720 million for the mine account. In the propagation model, a cap tying downstream loss to primary loss understates the total by 20% at coefficient 0.60 and 60% at 0.80. A thirteen-case reported-loss table is strongly concentrated and is used descriptively rather than for calibration. The contribution is a set of testable consistency and robustness conditions; independently verified incident-based validation remains necessary.

Keywords: cyber incident damage; loss estimation; public digital services; critical infrastructure; double counting; uncertainty; cascading effects; severity assessment

Оценка ущерба от киберинцидентов в государственных цифровых системах и критической инфраструктуре

Аннотация: Оценки ущерба от киберинцидентов существенно меняются при объединении пересекающихся затрат, нормативных коэффициентов и порогов классификации без единых границ измерения. В работе сравнительный нарративный обзор 52 источников дополнен формальным анализом учёта и чувствительности для государственных цифровых услуг и критической инфраструктуры. Модель связи исходных потерь с отчётными величинами определяет точное завышение из-за повторного учёта. На основе монотонности сформулировано условие устойчивости финансовой категории при заданной неопределённости; операционная тяжесть и безопасность рассматриваются отдельно. Аналитические условия и воспроизводимые расчёты применены к двум демонстрационным примерам из проектов национальной методики и условной двухсекторной модели распространения ущерба. В исходном портальном примере репутационная составляющая определяет 90,96% суммы и имеет эластичность 0,90956; критический финансовый порог достигается при 42,905% её исходной величины. Исключение составляющей снижает финансовую категорию с критической до средней, не обязательно изменяя операционный уровень. Точные сценарные границы составляют 20,8265–34,109 млн тенге для расширенной оценки портального случая и 67,912–83,720 млн для шахтного. Ограничение каскадных потерь величиной первичного ущерба занижает итог условной модели на 20% при коэффициенте 0,60 и на 60% при 0,80. Таблица тринадцати заявленных потерь используется для описательного анализа, а не калибровки. Результат работы — проверяемые условия согласованности расчётов и устойчивости классификации; необходима независимая проверка на первичных материалах инцидентов.

Ключевые слова: ущерб от киберинцидентов; оценка потерь; государственные цифровые услуги; критическая инфраструктура; двойной учёт; неопределённость; каскадные эффекты; оценка тяжести последствий

Мемлекеттік цифрлық жүйелер мен аса маңызды инфрақұрылымдағы киберинциденттерден келетін залалды бағалау

Аңдатпа: Киберинцидент залалының бағасы қайталанатын шығындар, нормативтік коэффициенттер және жіктеу шектері бірыңғай өлшеу шекараларынсыз біріктірілгенде елеулі өзгереді. Жұмыста 52 дереккөздің салыстырмалы нарративтік шолуы мемлекеттік цифрлық қызметтер мен аса маңызды инфрақұрылым үшін есепке алу мен сезімталдықтың формалды талдауымен толықтырылған. Бастапқы шығындарды есептік шамалармен байланыстыратын модель қайталап есептеуден туындайтын нақты артық соманы анықтайды. Монотондылық негізінде берілген белгісіздік жағдайында қаржылық санаттың тұрақтылық шарты тұжырымдалады; операциялық салдар мен қауіпсіздік бөлек қарастырылады. Аналитикалық шарттар және қайта шығарылатын есептер ұлттық әдістеме жобаларындағы екі демонстрациялық мысалға және шартты екі секторлы таралу моделіне қолданылады. Бастапқы портал мысалында беделдік құрамдас жалпы соманың 90,96%-ын құрайды, оның икемділігі 0,90956; критикалық қаржылық шек осы құрамдастың бастапқы шамасының 42,905%-ында орындалады. Оны алып тастау қаржылық санатты критикалықтан орташаға төмендетеді, бірақ операциялық деңгейді міндетті түрде өзгертпейді. Порталдың кеңейтілген бағасының нақты сценарийлік шектері 20,8265–34,109 млн теңге, шахта бағасының шектері 67,912–83,720 млн теңге. Каскадтық шығындарды бастапқы залалмен шектеу шартты модельдің жалпы сомасын коэффициент 0,60 болғанда 20%-ға, ал 0,80 болғанда 60%-ға кемітеді. Мәлімделген он үш шығын кестесі калибрлеуге емес, сипаттамалық талдауға қолданылады. Нәтиже — есептеулердің сәйкестігі мен жіктеудің тұрақтылығын тексеру шарттары; оларды инциденттердің бастапқы материалдары негізінде тәуелсіз тексеру қажет.

Түйін сөздер: киберинциденттерден келетін залал; шығындарды бағалау; мемлекеттік цифрлық қызметтер; аса маңызды инфрақұрылым; қосарланған есеп; белгісіздік; каскадтық әсерлер; салдардың ауырлығын бағалау

1 Introduction

Cybersecurity incidents can generate expenditure for the affected organisation, disrupt dependent services and impose costs on people who are not parties to the original system. An assessment intended to support recovery funding, regulatory reporting or national statistics must therefore specify whose loss is measured, over what period and on what evidence. IBM reports an average data breach cost of USD 4.99 million for its 2026 study [1]. This benchmark describes the surveyed population; it is not a valuation rule for a particular public digital service or industrial installation.

The distinction matters because cyber risk quantification and incident damage assessment answer different questions. Risk analysis estimates the distribution of possible future loss, whereas incident assessment reconstructs the consequences of an event that has

occurred. Both may use monetary models, but only the former necessarily includes event frequency. Woods and Böhme [2] emphasise the importance of causal interpretation in empirical cyber risk research. Cremer et al. [3] identify limitations in accessible loss datasets, while event-level studies demonstrate substantial variation in reported costs [4], [5]. These findings support transparent assumptions and uncertainty reporting rather than the direct transfer of a global mean to a national setting.

The relevant literature comprises several complementary traditions. Investment models relate security expenditure to expected loss reduction [6], [7]. ISO/IEC 27005 and NIST SP 800-30 establish risk assessment processes [8], [9]. Open FAIR provides a taxonomy for analysing loss frequency and magnitude [10]. NIS2 and DORA establish criteria for regulatory significance [11], [12]. Business continuity, operational technology safety and infrastructure interdependency research address consequences that extend beyond information assets. Composite indicator methods offer a way to summarise heterogeneous impacts, subject to explicit assumptions about scaling and weights [13]. Their outputs cannot be compared without first aligning the object of measurement.

For government systems, the central difficulty is not that existing methods are inapplicable to non-commercial organisations. Response expenditure, replacement costs and productivity losses can be estimated without sales revenue. The harder problem is separating these costs from interrupted public service, loss of strategic autonomy, safety consequences and effects on other sectors. A budget allocation is not automatically a loss, a service delay is not necessarily permanent loss of output, and a critical safety consequence need not produce a large financial bill. An assessment that obscures these distinctions can misdirect both compensation and recovery priorities.

This study combines a comparative review with formal consistency and sensitivity analysis. It asks which measurement objects the main method families support, how overlapping losses can be reconciled, and when uncertainty in valuation changes an incident's financial or overall severity. The analytical contribution has three parts: an explicit observation-to-loss reconciliation condition; a criterion for category stability over a declared uncertainty set; and exact parameter-response relations for two draft-method demonstrations and a controlled propagation model. The conditions are derived and numerically checked. They are not claims of a new general theory of risk or of validated national loss coefficients.

Section 2 describes the source-based comparison and reproducible analytical procedure. Section 3 derives the accounting and classification conditions, reconstructs the two demonstration cases and presents quantitative sensitivity results. Section 4 discusses their practical interpretation and validation requirements. Section 5 states the conclusions. S1 documents source identities and evidence links; S2 and the accompanying scripts reproduce the calculations and figures.

2 Review design and analytical method

2.1 Scope and source selection

The study uses a structured comparative narrative review of 52 identified publications and instruments, followed by formal and numerical analysis of selected measurement problems. Selection is purposive and covers loss definition, monetary estimation, regulatory significance, service and physical consequences, interdependencies and indicator construction. S1 lists every reference and its role. The review is not an exhaustive systematic search and does not estimate the prevalence of defects across the literature.

Research studies support empirical and modelling claims; standards define processes and terminology; legal instruments specify rules within their stated scope; and industry reports provide contextual benchmarks. Principal source claims were checked against issuing-body and publisher records on 21 October 2026. S1 distinguishes inspectable passages from catalogue descriptions and supporting references. The cited edition defines each instrument; full-text access to every proprietary source is not assumed.

Sources restricted to attack detection are excluded unless they connect an incident to measurable consequences. Two unpublished national drafts are objects of critical analysis: the earlier Appendix 2.6 and working revision 2.2 [14], [15]. S1 supplies their exact file identities and section locators. The earlier Appendix B defines the demonstration inputs; revision 2.2 changes accounting and classification rules and clarifies the status of thirteen reported-loss entries. These drafts are not independent validation sources or adopted national standards.

2.2 Units of analysis and comparison criteria

The source-level unit is a named instrument or study, identified in S1. Families group sources with related purposes. Table 1 gives representative members; Tables 2 and 3 identify source-linked measurement routes and consistency checks. No numerical source ratings are averaged into family scores. A capability is attributed to a specified source and use; an unreported feature remains undetermined rather than being scored as absent.

Table 1. Analytical families and representative sources

Family	Principal output	Representative anchors
Risk management processes	Risk assessment structure and decision criteria	ISO/IEC 27005 [8]; NIST SP 800-30 [9]; CSF 2.0 [16]
Monetary risk and investment models	Loss distributions, expectations or investment criteria	Open FAIR [10], [17], [18]; ROSI [7], [19]; CyVaR [20], [21]
Empirical incident studies	Observed costs and associations in a defined sample	IBM [1]; Eling and Wirfs [4]; Romanosky [5]; Aldasoro et al. [22]
Regulatory instruments	Reporting thresholds and consequence categories	NIS2 [11]; DORA [12]; GDPR [23]; delegated and implementing acts [24], [25]

Family	Principal output	Representative anchors
Continuity and OT safety	Service impact, recovery priorities and physical consequences	ISO 22301 [26]; ISO/TS 22317 [27]; IEC 62443-3-2 [28]; NIST SP 800-82 [29]
Interdependency models	Propagation of disruption between systems or sectors	Leontief [30]; Rinaldi et al. [31]; ENISA [32]
Composite indicators	A score or ranking under stated aggregation rules	OECD/JRC [13]; AHP [33]

Comparison uses four criteria: the meaning and units of the output; the treatment of uncertainty; the prevention of overlapping loss entries; and the declaration of the organisational, temporal and economic boundary. Nine consequence dimensions are examined: financial expenditure, material damage, operational disruption, workforce productivity, reputation, public service continuity, technological dependency, cross-sector propagation and human safety. These dimensions are descriptive tags. They are not nine amounts that can automatically be added, because one consequence can carry several tags.

2.3 Synthesis and reproducibility

The analysis links each comparative claim to a named source, an identifiable passage or public scope statement, and a stated inference. S1 distinguishes a source statement, an analytical inference and a numerical demonstration. The synthesis rule is non-statistical: a family description identifies a capability illustrated by its cited members; it neither averages ordinal ratings nor interprets one positive example as universal coverage. For example, NIST SP 800-30 discusses uncertainty explicitly, while the OECD/JRC handbook addresses both overlap and sensitivity. These observations preclude family-wide claims that process standards or composite indicator methods lack those features [9], [13].

The evidence map was assembled through AI-assisted extraction and editorial analysis with OpenAI ChatGPT. No independent human double coding or inter-rater reliability estimate is claimed. Conflicting readings are resolved by narrowing the interpretation to the passage that can be supported. S1 records the source-to-claim links and access limitations. This interpretive evidence is distinguished from the algebraic results and reproducible numerical calculations below.

2.4 Analytical and computational procedure

The analytical unit is an incident account with a fixed owner, valuation perspective, time horizon and currency basis. The procedure first reconstructs the draft’s original entries, then changes one accounting convention or a declared parameter set at a time. Loss components are retained separately so that changes in a total can be attributed to a specific assumption. Derivatives quantify local sensitivity; monotonicity gives exact interval endpoints; classification rules map the feasible ratios to severity bands. An explicit

symmetric two-sector system isolates the effect of propagation strength. Finally, descriptive statistics examine the concentration of the thirteen reported amounts without treating them as a representative sample.

Figures use deterministic grids: 301 reputation multipliers, two 101 × 101 parameter surfaces and 451 propagation coefficients. These grids visualise response functions; the interval bounds and threshold-crossing conditions follow analytically rather than from sampling. No random seed or probability distribution is required. Numerical checks compare original totals with their source, verify the fixed-point relation and interval endpoints, and compare an analytical elasticity with a central finite difference. The supplied CSV files contain every plotted point.

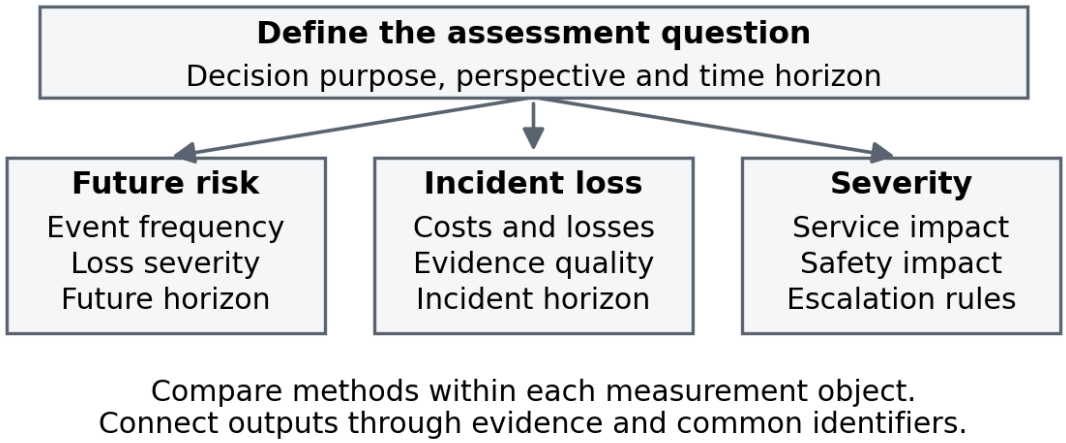


Figure 1. Measurement objects distinguished in the comparative analysis. Currency units alone do not make the outputs interchangeable.

3 Comparative analysis and assessment design

3.1 Risk management process standards

ISO/IEC 27005 defines guidance for information security risk management, while NIST SP 800-30 organises risk assessment around threats, vulnerabilities, likelihood and impact [8], [9]. CSF 2.0 places cybersecurity outcomes within six functions, including Govern [16]. ISO 31000 and IEC 31010 provide broader risk management principles and assessment techniques [34], [35]. These instruments support consistent processes, but their adoption alone does not establish common monetary assumptions across organisations.

The resulting limitation is one of specification. Two agencies may apply the same process and legitimately choose different impact scales, time horizons or valuation perspectives. Their risk assessments can be useful locally without being suitable for aggregation into national loss statistics. A national procedure should retain the governance

and documentation functions of these standards while separately defining the minimum incident accounting record. Lack of a universal currency formula is not a defect in a standard whose primary purpose is risk governance.

This distinction should not obscure the safeguards already present in process guidance. NIST SP 800-30, Section 2.3.1, discusses uncertainty; Section 2.3.2 distinguishes assessment scales; and Task 2-6 calls for uncertainty to accompany risk determinations [9]. The present contribution is the operationalisation of compatible incident accounting records, not the discovery of uncertainty as a missing concept.

3.2 Monetary loss and investment models

Open FAIR distinguishes loss event frequency from loss magnitude and organises losses through primary and secondary consequences [10], [17], [18]. Its loss forms can support identification of response, replacement, productivity and other costs in a public organisation. The primary–secondary distinction concerns the causal pathway, including reactions by stakeholders; it should not be treated as an exact synonym for direct–indirect accounting categories. The cited O-RT and O-RA editions are Version 3.1 and Version 2.1, respectively, published in 2025.

The FAIR Institute model specification independently identifies productivity, response and replacement loss forms in Sections 3.1–3.3 [36]. For a public agency, response and replacement expenditure remain meaningful even when forgone sales are zero. Salary-based productivity valuation requires evidence of displaced useful work and is not automatically an incremental cash expense. Reputation-related revenue loss may be inapplicable to a particular public service, but that limitation concerns a loss pathway, not the validity of the entire framework.

Annualised loss expectancy expresses expected aggregate loss over a defined horizon. Under the additional assumptions that event losses are identically distributed, have finite means and are independent of the event count, the compound loss model has the form in Equation (1).

$$\mathrm{E}[S_H] \;=\; \mathrm{E}\left[\sum_{k=1}^{N_H} X_k\right] \;=\; \mathrm{E}[N_H]\,\mathrm{E}[X] \quad (1)$$

Here S_H is aggregate loss over horizon H, N_H is the number of loss events and X_k is the loss of event k, all monetary quantities being expressed on a consistent price basis. Independence of count and severity should be checked rather than assumed for common-cause or cascading incidents. For an incident that has already occurred, estimation of its damage is conditional on the observed event; it does not require estimating how often that event will occur next year.

ROSI and security investment models evaluate a counterfactual reduction in future loss relative to protection expenditure [6], [7], [19]. Cyber value-at-risk specifies a loss

quantile for a stated horizon and confidence level [20], [21]. These quantities inform investment or capital decisions, but neither a quantile nor expected annual loss estimates the incurred cost of one realised event. Sparse data limit calibration, yet distributions may also be elicited from expert judgement. Their provenance and uncertainty must then be disclosed. Frequency data are a constraint on prospective risk estimation, not a prerequisite for recording documented recovery costs.

3.3 Empirical cost studies and benchmarks

Empirical studies supply evidence about the dispersion of incident costs and the influence of organisational and event characteristics [4], [5], [22], [37]. Industry reports serve different purposes: a breach cost survey estimates costs in a sampled population, whereas the Verizon Data Breach Investigations Report primarily characterises incidents and attack patterns [1], [38]. Their denominators, inclusion rules and reporting mechanisms must be examined before results are combined.

A national assessment should use external benchmarks to test plausibility and to inform explicitly labelled estimates where local evidence is limited. Direct substitution is problematic when wage levels, liability rules, service portfolios and reporting incentives differ. Currency conversion alone cannot correct these differences. Neither a sample mean nor a loss-per-record rate supplies a causal estimate for an individual event. A benchmark-based component should retain the source population, adjustment assumptions and a sensitivity range, and should be replaced by incident-specific evidence as it becomes available.

3.4 Regulatory significance and economic impact

NIS2 and DORA link incident reporting to the significance of consequences within their respective scopes [11], [12]. Their subordinate acts contain more economic detail than a simple ordinal severity scale. Article 7 of Commission Delegated Regulation (EU) 2024/1772 specifies direct and indirect cost and loss categories for determining economic impact under DORA [24]. Commission Implementing Regulation (EU) 2024/2690 specifies significant-incident criteria for listed digital service providers and discusses relevant financial loss categories [25]. It would therefore be inaccurate to state that regulatory instruments provide no guidance on monetary loss.

The remaining distinction concerns purpose and boundary. A reporting threshold answers whether a statutory notification condition is met. A financial estimate supports accounting or resource allocation under a specified valuation perspective. A regulatory total may prescribe a different treatment of recoveries or forgone revenues from an economic resource-loss estimate. These totals require a documented reconciliation rather than forced numerical equality. The turnover-based maximum fines in GDPR are legal sanctions, not estimates of the harm suffered by individuals [23]. The ENISA personal data breach methodology provides a severity assessment approach [39], while the US cyber incident severity schema supports coordination and escalation [40].

There are also concrete alternatives to an isolated national total. The European Supervisory Authorities' Joint Guidelines JC 2024 34 provide a reference-year approach to aggregated annual incident costs and losses and separate financial recoveries [41]. These instruments demonstrate existing monetary reporting practice. Adapting their boundary to public services remains a design choice rather than evidence that no alternative approach exists.

Version control is also material. NIST SP 800-61 Revision 3, issued in April 2025, supersedes Revision 2 and aligns incident response guidance with CSF 2.0 [42]. Older lifecycle descriptions may remain relevant when explicitly discussed historically. Their use should not be represented as application of the current edition. More generally, the assessment record should identify the rule and edition that governed classification at the time of reporting.

3.5 Business continuity and operational technology safety

Business impact analysis links disrupted activities to organisational objectives, consequences over time and recovery priorities [26], [27]. It can support valuation of interruption, but does not guarantee that a validated hourly loss rate exists. Such a rate may depend on transaction timing, spare capacity, manual alternatives and whether delayed work is subsequently completed. Multiplying annual operating expenditure by a fraction of unavailable hours measures an allocation of service resources; without additional assumptions it does not measure incremental damage.

For operational technology, IEC 62443-3-2 and NIST SP 800-82 support assessment of cyber-related consequences in systems that interact with physical processes [28], [29]. Functional safety and integrated safety–security approaches add analysis of hazardous scenarios [43], [44]. Actual injuries, hazardous releases and exposure duration should be recorded separately from the probability of a hypothetical accident. Medical expenditure and repair costs may enter the relevant monetary account. Injury severity and risk to life should remain visible as separate outcomes even when a jurisdiction permits economic valuation of health effects. A low estimated financial loss cannot cancel a safety-critical escalation condition.

3.6 Interdependencies and cascading consequences

Input–output models and infrastructure interdependency research describe how disruption can propagate beyond its initial location [30], [31]. ENISA and OECD materials provide complementary guidance on critical functions and infrastructure resilience [32], [45]. Supply chain guidance addresses dependency on technology providers and services [46], [47]. Applying these approaches to a realised cyber incident requires an event boundary, evidence of affected links and a distinction between observed downstream losses and modelled scenarios.

For an explicitly defined linear loss-propagation approximation, let d be a vector of initial losses and A a non-negative matrix whose element a_{ij} maps a unit of loss in sector j into an additional loss in sector i over the same horizon. A finite non-negative infinite-series solution for arbitrary non-negative d is obtained when the spectral radius ρ(A), the largest eigenvalue magnitude, is below one. Equation (2) states the resulting model.

$$x = d + Ax = (I - A)^{-1}d, \quad A \geq 0, \quad \rho(A) < 1 \quad (2)$$

Here x contains total sector losses and I is the identity matrix. Strictly sub-unit row or column sums are sufficient conditions for convergence, but are not necessary. Sums no greater than one are not sufficient when ρ(A) equals one. Nor is there a general economic requirement that downstream losses be smaller than the initial loss. For example, a hypothetical two-sector matrix with both diagonal elements equal to zero and both off-diagonal elements equal to 0.6 has ρ(A) = 0.6. With d = (1, 0) million KZT, Equation (2) gives x = (1.5625, 0.9375) million KZT. Total loss is 2.5 million KZT, including 1.5 million of propagated loss. This is a mathematical counterexample to the proposed cap, not an empirical estimate.

Equation (2) is an illustrative abstraction, not a validated cybereconomic model or a direct reuse of national input–output coefficients. Its coefficients require causal interpretation and calibration; near-unit spectral radii can amplify parameter uncertainty. If x already includes d, the two vectors must not be added again. Empirical implementation should also use consistent value-added or resource-cost concepts to avoid counting intermediate transactions repeatedly. Where these conditions cannot be met, report observed downstream costs and separate bounded scenarios rather than an unsupported multiplier.

Where observations replace model outputs, maintain a sector-by-item reconciliation: each downstream item has an affected owner, time interval and causal link to the initiating event. A modelled sector total and the observed expenses it already represents cannot both enter the same consolidated total. A finite propagation-horizon model can instead sum d, Ad, …, A raised to K times d; the infinite-series convergence condition then does not apply, although causal validity and avoidance of repeated valuation remain necessary. No coefficient restriction can substitute for this accounting reconciliation.

3.7 Composite indicators and monetary valuation

Composite indicators combine selected variables through normalisation, weighting and aggregation [13], [33]. Their interpretation depends on the measurement scales and on whether compensation between dimensions is acceptable. ISO/IEC 27004 supports a disciplined approach to information security measurement and evaluation [48]. A severity score may be useful for escalation, but the score alone does not establish an exchange rate between public trust, safety and money.

Multiplying a dimensionless quantity by a financial base is not inherently invalid. For example, multiplying a calibrated physical damage ratio by an appropriate replacement value can be meaningful. The problematic step is treating an uncalibrated multi-criteria severity index as if it were an empirically established loss fraction. Such conversion requires an identified monetary target, supporting data, calibration, validation and uncertainty analysis. In their absence, index values and monetary totals should remain separate.

A numerical distinction is useful. A calibrated 0.20 physical damage fraction applied to a 10 million KZT replacement value produces a 2 million KZT estimate for the specified asset. An ordinal severity score of 0.20 applied to the same base has no equivalent meaning without calibration. Neither calculation permits a second entry for the same damaged asset. ENISA's personal data breach method defines severity levels in Section 3.2, not monetary conversion factors [39].

The proposed architecture therefore retains a monetary account and a consequence profile as linked outputs with a common incident identifier (Figure 2). Critical safety or essential-service conditions should be capable of triggering escalation independently. If an aggregate score is used for other purposes, its component values, weights and sensitivity should remain available; averaging must not conceal a non-compensable consequence.

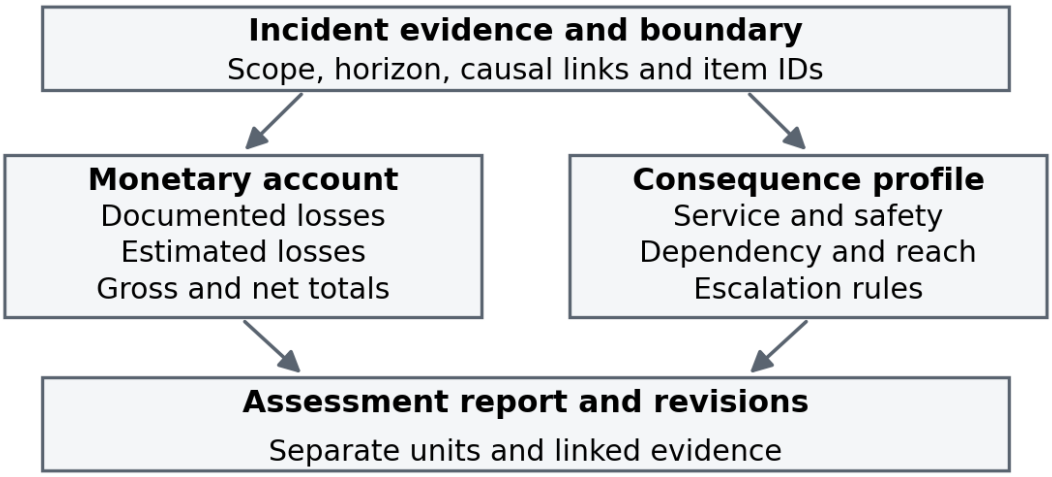


Figure 2. Proposed assessment architecture. Financial loss and consequence severity share evidence and an incident boundary but remain distinct outputs.

3.8 Comparison across consequence dimensions

Table 2 distinguishes feasible valuation routes from consequences that require non-monetary reporting. The comparison does not establish that a dimension is absent from every member of a family. It identifies the evidence and boundary conditions needed when representative methods are used together. In particular, technological replacement costs can

be measured without claiming to monetise technological sovereignty itself, and public service interruption can be measured in service units without treating the delivery budget as lost revenue.

Table 2. Consequence dimensions and defensible measurement routes

Dimension	Measurement route and anchors	Boundary condition
Financial expenditure	Incident response, legal and recovery entries [10], [42], [24]	Identify incremental costs; distinguish payments, provisions and recoveries.
Material damage	Repair or replacement cost for damaged assets [28], [29]	Avoid adding full replacement cost and full asset write-off for the same loss.
Operational disruption	Unrecovered output or contribution margin; additional work [26], [27]	Account for catch-up, saved costs and workarounds.
Workforce productivity	Attributable work hours and a declared valuation basis [10], [18]	Separate opportunity cost from additional payroll; exclude overlaps.
Reputation	Incremental churn, contract loss or mitigation expenditure [4], [5], [10]	Establish attribution; avoid adding market-value loss to overlapping cash flows.
Public service continuity	Failed or delayed requests, affected users, backlog and alternatives [26], [27]	Allocated budget is a resource proxy; citizen loss needs separate evidence.
Technological dependency	Incident-induced migration, substitution and integration costs [46], [47]	Exclude unrelated modernisation; report residual dependency separately.
Cross-sector propagation	Observed downstream costs or a specified dependency model [30], [31]	Use a common boundary; reconcile transfers and duplicated outputs.
Human safety	Observed harm, exposure and scenario consequences [29], [43], [44]	Retain non-monetary severity even when some costs are monetised.

Three integration problems follow from this comparison. First, one underlying consequence can appear in several categories: an interrupted activity may already incorporate idle staff, contract penalties and recovery expenditure. Second, estimates with different purposes may share currency units while representing different objects. Third, costs borne by one entity can be transfers received by another. A national aggregation cannot be obtained by adding all organisational bills without specifying the economic perspective.

The evidence supports specific mechanisms, not a prevalence estimate. The OECD/JRC handbook explicitly discusses collinear indicators that can give excess weight

to one underlying dimension and recommends uncertainty and sensitivity analysis [13], Sections 1.6–1.7. This is an identified aggregation risk and an existing safeguard, rather than evidence that the handbook commits an error. NIST's qualitative and quantitative scales [9], Section 2.3.2, ENISA's severity levels [39], Section 3.2 and the cost categories in Regulation 2024/1772 [24], Article 7 illustrate different measurement objects. The invalid step would be combining those outputs without reconciling their meanings. The cases below identify concrete budget-proxy, sector-uplift and severity-feedback problems in the earlier national draft [14]. These examples support source-specific criticism, not a claim that such defects recur at a measured rate across the corpus.

Table 3. Consistency checks for a combined assessment

Issue	Diagnostic question	Required treatment
Risk versus realised damage	Is the output an event loss, an annual expectation or a quantile?	Label the object and horizon; compare only like quantities.
Duplicate loss	Can two entries refer to the same resource or output?	Use a unique item identifier and an exclusive accounting allocation.
Score converted to money	Has the index been calibrated to a defined monetary target?	Keep uncalibrated severity scores outside the loss total.
Unstated denominator	Does the base represent annual flow, asset value or lifecycle cost?	State base, year and scope; avoid cross-base rankings.
Transfers and recoveries	Does the amount change resource loss or only who bears it?	Report organisational gross and net loss; reconcile national totals.
Uncertainty hidden in a point estimate	Which inputs are observed, estimated or conventional?	Disclose evidence status, ranges and sensitivity.

3.9 Requirements for a national assessment procedure

Kazakhstan provides a relevant application context because public digital services and critical infrastructure combine organisational, public and cross-sector consequences. The Digital Code of the Republic of Kazakhstan, adopted on 9 January 2026, and legislation on state services provide the legal context [49], [50]. The architecture below is a methodological proposal; it is not represented as an adopted legal requirement or an independently validated national instrument.

The first step is to define an incident record: affected entities and services, assessment date, start and end of the loss horizon, currency and price basis, and perspective. At least three perspectives should be distinguishable: the operator's organisational loss, the fiscal effect on public bodies and the wider economic or social consequence. These views can overlap and should not be added together as independent layers. A penalty paid by an

operator to the state is a cost to the operator but a transfer within a consolidated national account; the associated administrative resources are a separate matter.

The monetary account should start from distinct loss items rather than from broad labels such as direct, indirect and strategic. Let c_j be the monetary value of item j and a_j indicate whether that item is included within the declared boundary. For every included item, a category-allocation variable z_{jk} assigns it to exactly one monetary category. Equation (3) defines the aggregation and exclusivity condition.

$$U_{gross} = \sum_{j=1}^{n} a_j c_j = U_D + U_E, \quad \sum_{k=1}^{K} z_{jk} = 1 \quad (3)$$

U_D is the subtotal supported by incident-specific documentary evidence and U_E is the subtotal that depends on explicit estimation assumptions. Both require attribution to the incident; a paid invoice alone does not establish that the entire amount is attributable. The allocation rule prevents duplicate entries when identifiers are maintained, but it cannot by itself detect two differently named records of the same loss. Reconciliation of invoices, staff time and output measures remains necessary. An item can have several descriptive consequence tags while retaining one monetary allocation.

Gross loss is reported before insurance or third-party recoveries. For an organisational account, received recoveries R_{rec} can be disclosed separately and reconciled to net loss as in Equation (4). Expected, disputed or conditional recoveries should retain their uncertainty and not be presented as already received.

$$U_{net} = U_{gross} - R_{rec} \quad (4)$$

An interruption estimate must follow the observed service trajectory. Let q(t) be the fraction of normal useful service delivered at time t, bounded between zero and one, and let b(t) be the marginal value of output permanently lost after allowing for subsequent catch-up and avoided variable costs, expressed in KZT per hour. A simplified operational loss model is given in Equation (5).

$$U_{op} = \int_0^T [1 - q(t)]\, b(t)\, dt + C_{alt} + C_{backlog} + C_{pen} \quad (5)$$

C_{alt} denotes additional workaround costs, $C_{backlog}$ additional costs of clearing delayed work and C_{pen} attributable contractual penalties under the chosen perspective. Each term must exclude amounts already captured in b(t) or another account. This model is suitable only when b(t) has a defensible basis. For a public service without a meaningful monetary output value, report service disruption directly and quantify evidenced additional resources. An annual service budget B divided by planned service hours H, multiplied by

equivalent interruption hours T_{eq}, may be disclosed as the resource proxy V_{proxy} in Equation (6), but should not automatically enter the loss ledger.

$$V_{proxy} = \frac{B}{H} T_{eq}, \quad T_{eq} = \int_0^T [1 - q(t)]\, dt \quad (6)$$

The consequence profile should retain service scope, geographic reach, population affected, safety outcomes and residual technological dependency. Escalation rules may use these variables even when a monetary estimate is incomplete. Restoration, migration, integration and retraining expenses can be included when attributable to the incident; a general coefficient for technological sovereignty is not a substitute for identifying those expenses. This separation allows financially small but safety-critical incidents to remain visible.

Table 4. Minimum information needed for national implementation

Record element	Required information	Purpose
Incident boundary	Identifiers, entities, affected services, causal links and time horizon	Link primary and downstream records without duplicate incidents.
Monetary item	Item ID, amount, units, owner, period and category	Reconstruct totals and enforce exclusive allocation.
Evidence status	Documented, estimated or excluded proxy; source and rationale	Expose the support for each included value.
Valuation perspective	Organisational, fiscal or wider economic account	Identify transfers and prevent incompatible aggregation.
Uncertainty	Plausible ranges or distributions; dependence assumptions	Identify which missing evidence changes the estimate most.
Consequence profile	Service disruption, safety, dependency and territorial reach	Support escalation without forcing every consequence into money.
Revision history	Assessment date, changed inputs and reasons	Separate preliminary estimates from settled outcomes.

3.9.1 Loss reconciliation and identifiable overlap

A category list alone does not prevent double counting. Consider n distinct underlying loss items with non-negative values c, all within the same assessment boundary. Let H be a binary observation matrix: $H_{ij} = 1$ when reported subtotal i includes item j. Assume every item is represented by at least one row. The reported subtotals are $r = Hc$ and the multiplicity vector $m = H^T 1$ counts how often each item appears. Equation (7) gives the exact inflation produced by summing the reports without reconciliation.

$$r = Hc, \quad m = H^{T}1, \quad \Delta_{dup} = 1^{T}r - 1^{T}c = (m - 1)^{T}c \quad (7)$$

Proposition 1. Under these assumptions, the raw sum equals the unique-item total if and only if no positive-valued item has multiplicity greater than one. Proof: Equation (7) is an identity, and every term $(m_j - 1)c_j$ is non-negative. The discrepancy vanishes exactly when all duplicated items have zero value. For example, items (1, 2, 3) and reports containing items {1, 2} and {2, 3} give subtotals 3 and 5; their sum 8 overstates the unique total 6 by the repeated item 2. The condition is independent of category names.

The accounting implication is operational: a component record must identify what an aggregate hourly rate or invoice already contains. When H is unknown, the discrepancy cannot be estimated merely from the number of categories. In the mine case, adding staff time to a downtime rate is valid only if that rate excludes the same displaced work. A sector multiplier or a cap on total loss cannot establish this exclusion. Equation (3) allocates reconciled items to categories; Equation (7) addresses the prior task of detecting repeated observations.

3.9.2 Parameter uncertainty and sensitivity

Let θ collect valuation inputs and let Θ specify their admissible joint values. The fixed expense subtotal D is distinguished from modelled components $f_j(\theta)$. Equation (8) defines an uncertainty envelope for the monetary total. The set Θ may encode dependence or exclusion constraints; it is not automatically a Cartesian product of independent marginal intervals.

$$U(\theta) = D + \sum_{j=1}^{J} f_j(\theta), \quad U^{-} = \min_{\theta \in \Theta} U(\theta), \quad U^{+} = \max_{\theta \in \Theta} U(\theta) \quad (8)$$

If each uncertain input has a non-negative partial derivative and the feasible set is a rectangle, the lower and upper corners attain the exact bounds. No Monte Carlo approximation is needed in that case. If combinations are infeasible, corner calculations provide only outer bounds until the joint constraint is imposed. These are uncertainty sets, not confidence intervals: a probability statement requires a defensible probability model in addition to admissible ranges.

For a positive total and a positive parameter p, the local elasticity in Equation (9) measures the proportional change in total loss caused by a small proportional change in p, with other inputs fixed. When a component is proportional to p, its elasticity equals that component’s share of the total. This provides a direct test of whether the monetary conclusion is governed by a documented expense or by a convention.

$$E_p = \frac{p}{U}\frac{\partial U}{\partial p} \quad (9)$$

For a component containing several multiplied inputs, each input can have the same component-share elasticity; these input elasticities must not be summed as if they were disjoint cost shares. Zero-valued inputs require absolute derivatives instead. The sensitivity of an estimate to a parameter is also distinct from uncertainty in that parameter: a large derivative is consequential only when the parameter is poorly constrained over a material range.

3.9.3 Robustness of financial and overall classification

Let $B > 0$ be the declared comparison base and let $\tau_1 < \tau_2 < \tau_3$ be three percentage thresholds. Equation (10) maps a financial ratio to ordinal levels 1–4. The indicator is one when its condition is true and zero otherwise. The inclusive convention at a boundary is used here for the specified public-service scenarios; a different reporting rule must state its own convention.

$$r = \frac{100U}{B}, \quad k_F(r) = 1 + \sum_{h=1}^{3} 1\{r \geq \tau_h\} \quad (10)$$

Suppose U lies in $[U^-, U^+]$ and B in $[B^-, B^+]$, with positive denominators. Equation (11) bounds the possible ratio. Proposition 2. The financial category is robust whenever both endpoint ratios fall in the same band. This condition is also necessary when both opposite corners of the joint feasible set can occur. Proof: U/B is increasing in U and decreasing in B; the band function is non-decreasing. Therefore every feasible ratio lies between the endpoints and has the same category when their categories agree.

$$r^- = \frac{100U^-}{B^+}, \quad r^+ = \frac{100U^+}{B^-}, \quad k_F(r^-) = k_F(r^+) \quad (11)$$

When the endpoints give different categories but the joint corners are not feasible, instability has not yet been proved; the feasible set must be evaluated more tightly. Likewise, if B also enters the loss formula, U(B) must be recomputed jointly rather than varying the numerator and denominator independently. The maximum-rule overall level is $\max\{k_F, k_O, k_H, \ldots\}$, where the additional terms denote separately mapped operational and safety levels. Consequently a robust critical operational condition can make overall severity robust even when financial severity changes across several bands. This is a classification property, not evidence that monetary valuation is accurate.

3.9.4 Propagation strength and finite assessment horizons

To separate model behaviour from a supposed downstream-loss cap, consider A with zero diagonal and equal off-diagonal entries a, where $0 \leq a < 1$, and initial loss $d = (d_0, 0)$, $d_0 > 0$. Let K count the number of propagation steps, with step zero representing the initial loss. Since $1^T A = a1^T$, the aggregate multiplier is the geometric sum in Equation (12).

$$M_K(a) = \sum_{k=0}^{K} a^k = \frac{1-a^{K+1}}{1-a}, \quad M_\infty(a) = \frac{1}{1-a} \quad (12)$$

It follows that the infinite-series propagated loss divided by initial loss equals a/(1 − a), which exceeds one precisely when a > 0.5. A rule capping downstream loss at d_0 would then limit total loss to $2d_0$ and understate the model total by the fraction 2a − 1. The result is exact for this constructed model. It does not estimate a national propagation coefficient: repeated propagation must represent distinct marginal consequences, not repeated valuation of the same consequence. Finite K can represent a declared truncation of dependency paths; interpreting steps as elapsed time requires an additional dynamic model.

3.10 Reconstructing the draft-method demonstrations

Appendix B of the earlier national draft [14] explicitly describes its two cases as demonstrations whose parameters require factual verification before real-incident use. Its portal total is 187,230,752 KZT and its mine total 80,040,000 KZT; these reproduce the rounded 187.2 and 80.0 million in the earlier draft. The calculations below recover those figures from the actual inputs, without treating scenario costs as observed invoices. S2 identifies every input and coefficient source. Monetary amounts are KZT; tables use million KZT. No recovery or inflation adjustment is specified by the source, so no net-loss or constant-price cross-incident comparison is inferred.

3.10.1 Government portal

The source assumes 18 hours of complete unavailability, a 600 million KZT annual service budget and 8,760 annual operating hours. Recovery comprises restoration 8 million, additional personnel 3 million, emergency arrangements 2 million and audit 1 million KZT. Staff productivity uses 35 employees, 12 hours, 4,500 KZT per hour, efficiency reduction 0.60 and sector multiplier 1.50. The reputation convention uses 75,000 affected users, a draft-assumed monthly calculation index of 4,325 KZT, sector coefficient 0.70 and severity coefficient 0.75. The latter coefficients come from the draft's own scales and Appendix C, not an external empirical valuation study.

The portal case can be written as Equation (13), in million KZT. Here s is the productivity sector multiplier, z is a binary inclusion flag for the allocated-budget proxy, and λ scales the reputation convention. The original case is (s, z, λ) = (1.50, 1, 1); the reconciled organisational scenario is (1, 0, 0). This parametrisation separates a change in the valuation rule from a change in observed service disruption.

$$U_P(s, z, \lambda) = 14 + 1.134s + 1.232877z + 170.296875\lambda \quad (13)$$

Table 5. Reconstruction and evidential status of the original portal calculation

Component	Calculation in source units	Million KZT	Interpretation
Recovery	8,000,000 + 3,000,000 + 2,000,000 + 1,000,000	14.000000	Scenario expense subtotal; supporting invoices not supplied
Productivity	35 × 12 × 4,500 × 0.60 × 1.50	1.701000	Model estimate including an uncalibrated sector uplift
Service continuity proxy	18 × (600,000,000 / 8,760) × 1.00	1.232877	Allocated budget; rounded to whole KZT
Reputation convention	75,000 × 4,325 × 0.70 × 0.75	170.296875	Normative scenario amount, not measured reputation loss
Original total	14,000,000 + 1,701,000 + 1,232,877 + 170,296,875	187.230752	Reproduces the draft; does not establish actual loss

The hourly budget proxy is calculated from the unrounded quotient before rounding the final component to whole KZT; multiplying the displayed rounded hourly value of 68,493 would give a slightly different answer. The reputation amount is 90.96% of the total. The remaining 16.933877 million is not a documented-cost baseline: it still contains a budget allocation and a sector-weighted productivity estimate. Removing the reputation convention, excluding the budget proxy and setting the productivity sector multiplier to one yields an organisational scenario account of 14 + 1.134 = 15.134 million KZT. Even its 14 million expense subtotal requires underlying records before it can be called documented incident loss.

This separation is consistent with revision 2.2 [15], Appendix B.1, which uses 15.134 million KZT for the organisational account and excludes the budget proxy. It additionally estimates citizen-time costs as 75,000 × 0.30 × 1 hour × 506 KZT/hour = 11.385 million KZT. The rate 506 is the draft's rounded 85,000/168 hourly proxy; the repeat-contact share and additional time are scenario inputs, not observations available to this review. The wider total is 26.519 million KZT. Varying the repeat-contact share over 0.20–0.40 and time over 0.75–1.25 hours gives citizen-time costs of 5.6925–18.975 million and a wider total of 20.8265–34.109 million. These are deterministic assumption bounds. Third-party time costs should remain separate from the organisation's own expenditure and estimates.

With r denoting repeat-contact share and t denoting additional hours per affected citizen, the wider account has the bilinear form in Equation (14). Its positive derivatives justify using the joint endpoints to obtain the reported bounds. The coefficient 37.95 is

75,000 × 506 / 1,000,000 and has units of million KZT per hour; the share r is dimensionless and expressed as a fraction.

$$U_{soc}(r,t) = 37.95rt, \quad U_{wide}(r,t) = 15.134 + 37.95rt \quad (14)$$

3.10.2 Mine ventilation control

The earlier draft assumes a 12-hour complete stoppage and a downtime rate of 3 million KZT per hour. The rate's cost composition is not documented in the earlier case. Recovery, response, legal support and replacement are separate scenario amounts; safe shutdown/restart and process losses are added separately. Productivity uses 80 employees, 8 hours, 5,500 KZT per hour, efficiency reduction 0.80 and a sector multiplier of 2.50. Appendix C gives a 2.00–3.00 range for critical objects, but neither that range nor the selected 2.50 has an empirical calibration supplied with the case [14].

Equation (15) gives the mine total in million KZT when c is the downtime rate in million KZT per hour, κ the efficiency-reduction factor and s the productivity sector multiplier. The 37 million subtotal comprises 26 million direct expenditure and 11 million process-related costs. The original parameters are c = 3, κ = 0.80 and s = 2.50. The model is conditional on exclusion of these separate losses from c.

$$U_M(c,\kappa,s) = 37 + 12c + 3.52\kappa s \quad (15)$$

Table 6. Reconstruction of the original mine calculation

Component	Calculation in source units	Million KZT	Required qualification
Downtime	12 × 3,000,000 × 1.00	36.000	Hourly rate must exclude separately counted losses
Direct expenditure	12,000,000 restoration + 8,000,000 response + 1,000,000 legal + 5,000,000 replacement	26.000	Scenario amounts; actual financial records unavailable
Process losses	7,000,000 safe shutdown/restart + 4,000,000 process losses	11.000	Must not be embedded in hourly downtime cost
Productivity	80 × 8 × 5,500 × 0.80 × 2.50	7.040	Sector uplift changes cost without evidence of additional lost work
Original total	36,000,000 + 26,000,000 + 11,000,000 + 7,040,000	80.040	Conditional on disjoint components
Safety index, separate	0.50 × 0.75 × 0.75 × 0.50 = 0.140625	Not money	Rounded HSI = 0.141; no addition to total

The four safety factors represent the draft's assumed hazard probability, harm severity, exposure and barrier degradation. Their product is a constructed index, not an empirically established probability of injury. The earlier draft places 0.141 in its moderate safety band; revision 2.2 places the same number in a high band. This change illustrates why an index value without its edition-specific thresholds is insufficient. The source's 2,500 million KZT comparison base is conditional; revision 2.2 identifies it as asset value. Its 3.2016% original loss ratio must not be directly compared with the portal's ratio to an annual budget.

Removing only the 2.50 productivity uplift while retaining the original efficiency factor 0.80 reduces the mine total to 75.816 million KZT. Revision 2.2 instead sets the efficiency factor to 1.00 for complete staff stoppage and uses no sector uplift, giving productivity of 3.520 million and a total of 76.520 million. It explicitly excludes labour, restoration and process losses from the 3 million hourly rate and separates an expense subtotal of 37 million from estimates of 39.520 million. Those exclusions make the ledger internally reconstructible; the demonstrations still do not supply primary documents proving them [15], Appendix B.2.

An additional stress test, introduced here rather than attributed to the draft, varies the downtime rate by ±20% (2.4–3.6 million KZT/hour) and efficiency reduction from 0.60 to 1.00, with no sector uplift and the 37 million expense subtotal fixed. The total ranges from 67.912 to 83.720 million KZT. These are conditional endpoints, not confidence limits. If the hourly rate already includes staff or process losses, the relevant separate entries must be removed, not merely assigned narrower ranges. A verified actual case also needs the assessment horizon, recoverable output, avoided costs and later adjustments.

3.11 Reputation sensitivity and severity classification

For the earlier portal case, let λ multiply only the reputation convention. The exact reconstructed sensitivity function in million KZT is $U(\lambda) = 16.933877 + 170.296875\lambda$. The comparison ratio is $100U(\lambda)/600$. Table 7 uses the earlier draft's financial bands: below 1% low, 1–5% moderate, 5–15% high and above 15% critical [14], Section 7.3. Its boundary rule selects the higher band for public services; none of the tabulated scenarios lies on a boundary. These are the source's proposed bands, not validated or adopted national thresholds.

For a threshold τ and fixed base B, Equation (16) gives the multiplier at which $L + \lambda R$ reaches the threshold, where L = 16.933877 and R = 170.296875 million KZT. A threshold is reachable within the specified multiplier interval only if λ^* lies in that interval. This supplies a quantitative robustness margin rather than simply declaring an estimate sensitive.

$$\lambda^* = \frac{\tau B/100 - L}{R}, \quad R > 0 \quad (16)$$

Table 7. Portal sensitivity under the earlier draft's financial classification

Reputation multiplier λ	Total, million KZT	Annual-budget ratio	Financial band
0.00 — excluded	16.933877	2.8223%	Moderate
0.25	59.508096	9.9180%	High
0.50	102.082315	17.0137%	Critical
1.00 — original	187.230752	31.2051%	Critical
1.50	272.379190	45.3965%	Critical

The financial band changes from critical to moderate when reputation is excluded; the 15% threshold is reached at $\lambda \approx 0.42905$. This does not automatically lower overall incident severity. Sections 7.2–7.3 of the earlier draft classify a stopped key service as critical and take the highest applicable consequence level. Given the scenario's complete portal outage, that operational rule keeps the overall classification critical at every multiplier. Even if an assessor did not identify the portal as a stopped key service, the supplied GSCI of 0.68 would independently imply a high level. Thus the monetary total is strongly convention-dependent while operational escalation can remain unchanged.

The original reputation coefficient also creates a feedback risk: it is set to 0.75 for a high incident, yet the resulting financial total is critical. Replacing it by the draft's critical coefficient of 1.00 increases reputation to 227.0625 million and the total to 243.996377 million KZT. A procedure must either freeze an independently determined non-monetary severity input before valuation or avoid using the resulting money-based category to set its own monetary coefficient. This is an identifiable problem in the earlier draft, not evidence that all monetary or index-based approaches are defective.

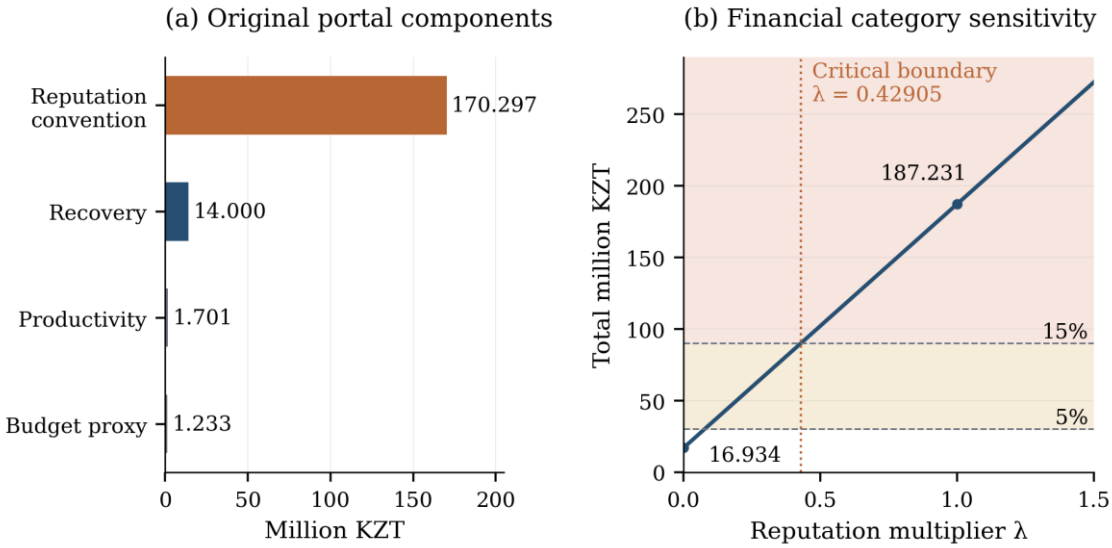


Figure 3. Portal component dominance and reputation sensitivity. Panel (a) reconstructs the earlier draft's components; panel (b) varies only the reputation multiplier at B = 600 million KZT. Dashed horizontal lines mark the draft's 5% and 15% financial thresholds. The operational classification is held separate. Source inputs: [14]; derived curves: accompanying scripts.

Revision 2.2 changes the financial threshold basis to the documented organisational subtotal and uses bands below 0.1%, 0.1–1%, 1–5% and at least 5% [15], Sections 6.1–6.3. Within its demonstration, the portal expense subtotal 14/600 = 2.3333% and the mine subtotal 37/2,500 = 1.48% both give high financial severity. Overall severity remains critical because the assumed service stoppages exceed their recovery time objectives: 18 versus 12 hours and 12 versus 4 hours. These results describe the draft rules conditional on their assumed evidence. Actual documented status cannot be inferred from a label in a demonstration. Threshold changes and accounting changes must be reported separately; lower revised totals alone do not demonstrate more accurate valuation.

3.12 Provenance and limits of the thirteen-case table

Appendix G (Russian designation Г) of the earlier draft [14] contains thirteen anonymous sector/scenario entries. Revision 2.2 [15], Appendix G, pp. 109–110, explicitly states that their amounts and valuation logic were reported by organisations and were not calculated using the methodology's formulas. S2 reproduces all thirteen entries and their stated valuation descriptions. The arithmetic is 167.7 million KZT in total, a mean of 12.9 million and a range of 1–90 million. This establishes the immediate documentary source and reported status, not independent authenticity: incident dates, named organisations, primary records, price bases and a collection protocol are not provided.

The table supports a limited check of whether the proposed categories can describe the reported consequences. It cannot establish accuracy, calibration, representativeness or

successful end-to-end field application. Four entries mention ransom; two contain only ransom amounts of 3.7 and 2.0 million, while the transport and media entries do not separate ransom from other losses. Consequently neither payment verification nor consistent recalculation is possible from the table alone. The present study reconstructs demonstrations and examines a secondary table of reported losses; it performs no independently verified application to primary incident records.

3.13 Quantitative results and interpretation

3.13.1 Sensitivity and classification stability

At the original portal point, the reputation-multiplier elasticity is 0.90956. Thus a 1% change in that multiplier changes the total by approximately 0.91% locally; the relation is exactly linear for a fixed baseline. In contrast, equal proportional changes in the recovery, productivity and allocated-budget components have total elasticities 0.07477, 0.00909 and 0.00658. The monetary estimate is therefore dominated by the reputation convention. The critical financial boundary occurs at $\lambda = 0.42905$, so reducing that convention by about 57.1% from its original level reaches the 15% threshold. Eliminating it crosses both the critical and high boundaries.

Figure 4 shows the distinct sources of uncertainty after removing the unsupported sector uplifts. The wider portal total increases bilinearly with repeat-contact share and citizen time. At $r = 0.30$ and $t = 1$ hour, an increase of 0.10 in r adds 3.795 million KZT; an increase of 0.25 hour in t adds 2.84625 million. These are local one-input changes, not additive estimates for a simultaneous change because the bilinear interaction must also be included. In the mine account, a 0.1 million KZT/hour change in downtime rate changes loss by 1.2 million, whereas a 0.10 change in efficiency reduction changes loss by 0.352 million. Under the specified ranges, the downtime-rate assumption accounts for 14.4 million of the 15.808 million total interval width.

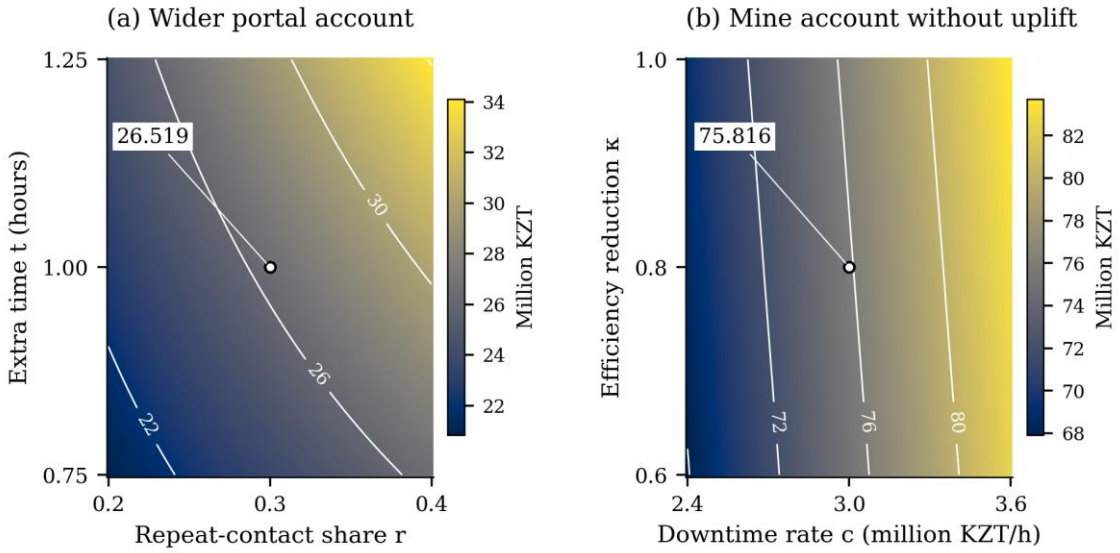


Figure 4. Deterministic monetary response surfaces. Panel (a) uses U = 15.134 + 37.95rt and marks the revision-2.2 portal assumptions; panel (b) uses U = 37 + 12c + 3.52κ and marks the earlier mine efficiency value with the sector uplift removed. Colour and contour labels show million KZT. The grids visualise analytical bounds and are not statistical confidence surfaces.

Table 8. Valuation sensitivity and category robustness under explicit rules

Scenario and varied inputs	Money interval million KZT	Financial category	Overall consequence level
Earlier portal, λ from 0 to 1.5; B = 600	16.933877–272.379190	Moderate–critical; not robust	Critical if the key-service stoppage condition is retained
Revision 2.2 portal, r = 0.20–0.40 and t = 0.75–1.25	20.8265–34.109 for the wider account	High from fixed expense subtotal 14; robust to these inputs	Critical from 18-hour stoppage exceeding 12-hour RTO
Revision 2.2 mine accounting, c = 2.4–3.6 and κ = 0.60–1.00; s = 1	67.912–83.720 for organisational account	High from fixed expense subtotal 37; robust to these inputs	Critical from 12-hour stoppage exceeding 4-hour RTO

The stability in the latter two rows follows from the revision's choice of financial numerator, not from precise knowledge of all economic losses. Estimated productivity, operations and citizen time are outside that numerator. If the supposed expense documentation or the comparison base changes, the financial result must be reevaluated. Table 8 therefore demonstrates robustness to the specified parameters only. It also shows

why replacing one draft's total and thresholds with another draft's rules in the same calculation would produce an uninterpretable comparison.

3.13.2 Cascading amplification and the cost of a cap

Figure 5 compares exact finite-step sums with the convergent limit. At a = 0.60, three propagation steps already give a multiplier of 2.176; the infinite limit is 2.5. At a = 0.80, the limit is 5. A downstream-loss cap equal to initial loss would understate these infinite-horizon totals by 20% and 60%, respectively. The comparison refutes a universal cap within a model whose assumptions are fully specified; it does not show that every real incident exhibits such amplification. As a approaches one, the infinite estimate becomes highly sensitive and finite-horizon interpretation becomes increasingly material.

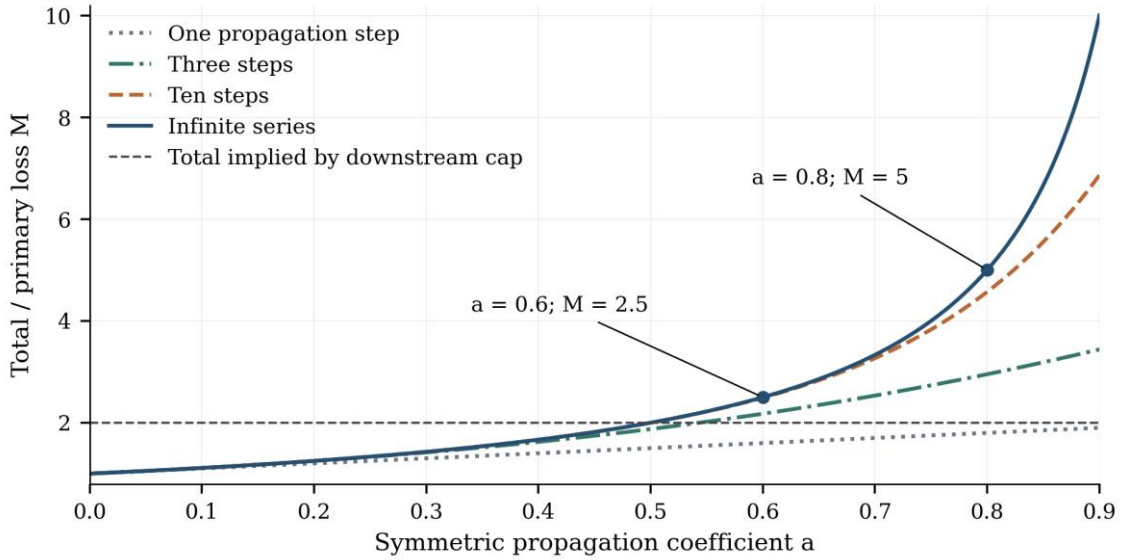


Figure 5. Total-to-primary-loss multiplier in the symmetric two-sector model. K denotes propagation steps, not elapsed hours. The horizontal line at two is the maximum total implied by capping downstream loss at primary loss. All curves are calculated from Equation (12); coefficients are controlled mathematical inputs.

3.13.3 Concentration in the secondary reported-loss table

The thirteen reported amounts have a median of 4.0 million KZT and mean of 12.9 million. The largest reported case, 90 million, contributes 53.67% of the 167.7 million total; the three largest contribute 71.56%. Figure 6 displays every amount with its S2 row ID. A leave-one-largest-out descriptive stress check gives a mean of 6.475 million, 49.81% below the full-table mean. This does not justify discarding the largest event. It shows that the aggregate benchmark is dominated by one record whose primary evidence is unavailable.

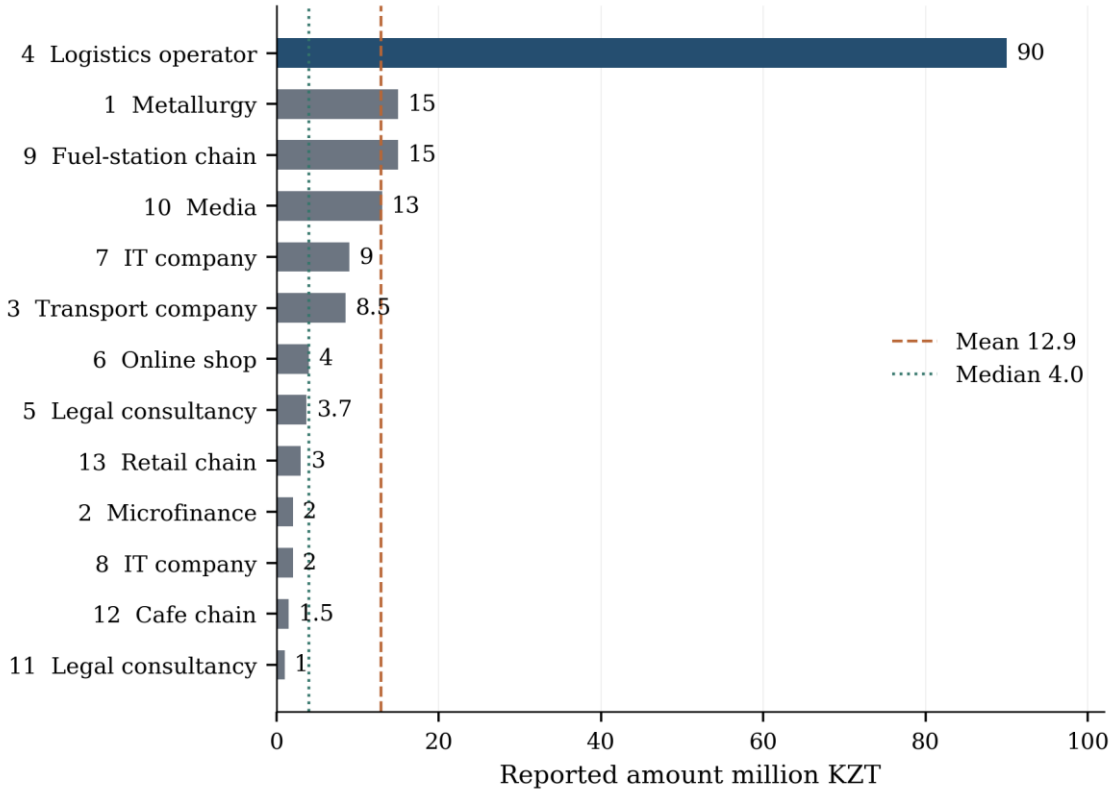


Figure 6. The thirteen organisation-reported amounts in the national draft, ordered by magnitude. IDs match S2.7. Mean and median describe this table only; the amounts have not been independently verified, price-normalised or generated by the proposed methodology. Sources: [14], [15], Appendix G.

Neither the mean nor a fitted distribution from this small secondary table would provide a defensible national coefficient. Sector labels mix different event mechanisms and accounting practices, while incident dates and sampling coverage are absent. The relevant numerical result is the observed concentration of the supplied table, not an estimate of tail risk in a population. This distinguishes a useful descriptive check from an unsupported calibration claim.

4 Discussion and validation requirements

4.1 Interpretation of the comparative findings

The comparative and numerical results support different levels of inference. Source passages establish the purposes and safeguards of particular instruments. The reconciliation and category-stability propositions establish conditional mathematical results. Case reconstructions show how those conditions operate with the supplied draft inputs. The secondary thirteen-case table provides only descriptive evidence. Treating these levels as

interchangeable would reproduce the same measurement problem that the assessment architecture is intended to prevent.

The strongest practical finding is that uncertainty in monetary valuation and stability of an escalation decision can coexist. In the portal case, removing a convention changes the financial category but need not change the operational decision. Conversely, a stable category based on a fixed expense subtotal does not establish that the wider economic total is accurate. Reports should therefore show both the valuation interval and the rule responsible for escalation; one overall score cannot convey both properties.

The formulas also localise the evidence needed to improve an estimate. For the portal, the valuation basis of reputation dominates the original total, while repeat-contact counts and additional time govern the revised third-party estimate. For the mine, the composition and level of the hourly downtime rate dominate the stated uncertainty width. In a dependency model, the admissibility of propagated items and the assessment horizon matter as much as convergence. These findings give concrete priorities for collecting evidence instead of suggesting that all missing information has equal value.

4.2 Validation and implementation

The first validation stage should assess whether independent analysts can reconstruct the same loss ledger from the same incident materials. Analysts should record disagreements about eligibility, causal attribution and valuation before reconciliation. Agreement can be assessed for categorical decisions, while absolute and relative discrepancies can describe monetary differences. A single reconciled total should not conceal initial disagreement. This stage tests reproducibility and clarity of the rules, rather than predictive accuracy.

The second stage should compare preliminary estimates with later documented outcomes. For a prospective pilot, record assessments at fixed milestones, for example 30, 90 and 180 days, while allowing a longer horizon for unresolved liabilities. Evaluate component-level errors and the coverage of any stated prediction intervals. Analyse sectors separately where consequence mechanisms differ. No particular sample size is asserted to be sufficient here: it should follow from the intended precision, variability and representation of sectors, including incidents with substantial downstream effects.

The third stage should evaluate whether non-monetary classifications support the intended decisions. A monetary threshold must not systematically suppress essential-service or safety escalation. Tests should vary valuation assumptions, index weights where used, and missing-data handling. A controlled comparison can ask whether the proposed ledger reduces duplicate entries or assessor disagreement relative to an existing procedure. Demonstrating these improvements would support an operational contribution beyond the conceptual synthesis presented here.

A national incident-loss register could support calibration if it records consistent scopes, revisions and evidence quality. Mandatory reporting can improve coverage, but does

not by itself remove reporting thresholds, selection bias or under-recorded social harms. A phased implementation should begin with auditable costs and clearly labelled estimates, then add calibrated sector-specific models as evidence accumulates. Access controls and disclosure rules are necessary because the underlying incident materials may contain commercially sensitive or personal information. Insurance research also highlights the importance of consistent information for loss assessment and risk transfer [51], [52].

4.3 Limitations

The source comparison is purposive and non-exhaustive, with 52 references and no independent human double coding. It cannot support a prevalence estimate for methodological defects. The two national drafts are unpublished objects of analysis with an institutional connection to the authors. The mathematical conditions are elementary conditional results; their contribution here is to make the accounting and classification requirements explicit and testable, rather than to claim that sensitivity analysis or interval classification is itself new.

No independently verified field application is reported. Reconstructed demonstration expenses are not inspected invoices, and the thirteen secondary reported amounts lack primary records and a collection protocol. The rectangular uncertainty sets are declared scenarios; dependencies could narrow them or make some combinations infeasible. The symmetric propagation model isolates one mechanism and does not reproduce a national economy. Greater algebraic detail and numerical reproducibility do not remove these limits; incident-based validation remains necessary to establish valuation accuracy and operational benefit.

5 Conclusion

Cyber incident damage assessment becomes more defensible when it begins with a declared measurement object and an auditable set of loss items. Existing standards, monetary models, regulations and infrastructure methods provide complementary components. Their integration requires explicit choices about perspective, time horizon, causal attribution and the treatment of uncertainty. The principal contribution of this review is a set of consistency requirements that connects these choices to evidence collection and reporting for public digital systems and critical infrastructure.

The quantitative analysis yields three operational conclusions. First, reconciliation must occur at the underlying loss-item level because category totals can conceal repeated observations. Second, monetary category stability can be tested against an explicit uncertainty set and must be distinguished from stability of non-monetary escalation. The portal's original monetary result has reputation elasticity 0.90956 and crosses the critical boundary at $\lambda = 0.42905$, while a key-service condition can preserve critical overall severity. Third, convergent propagation can produce downstream loss greater than primary loss; a universal cap would understate the controlled model by 20% at $a = 0.60$ and 60% at $a = 0.80$. These results establish reproducible consistency tests and priorities for further evidence

collection, while empirical accuracy remains to be evaluated against primary incident records.

Acknowledgments

The authors thank the specialists of the Digital Government Support Center who contributed to discussions of cyber incident damage assessment.

Author Contributions

Conceptualization: L.R. and A.T.; Methodology: L.R.; Formal analysis: A.T.; Investigation: L.R. and A.T.; Data curation: A.T.; Writing – original draft preparation: L.R.; Writing – review and editing: L.R. and A.T.; Visualization: A.T.; Supervision: L.R.

Conflicts of Interest

The authors are affiliated with the Digital Government Support Center and have an institutional interest in the development of national cybersecurity assessment practices. The proposed national application is presented as a methodological contribution rather than an independent evaluation of an adopted instrument. The authors declare no additional conflicts of interest.

Data Availability Statement

No primary incident dataset was collected. S1 supplies the complete reference inventory, source roles and evidence locators. S2 contains draft-case inputs, coefficient provenance, all thirteen reported-loss rows, the analytical conditions and computational settings. The accompanying Python scripts reproduce every numerical result and export all plotted data as CSV; figure generation uses Matplotlib and NumPy. Exact draft identities and source sections are listed in S1. The historical 68-source register and database screening history are not claimed.

Use of artificial intelligence

OpenAI ChatGPT was used for editorial restructuring, source-assisted evidence mapping, calculation scripting and figure preparation in this revision. This assistance is disclosed separately from the authors' contributions and does not constitute independent human double coding or empirical validation.

References

[1] IBM Security. (2026). Cost of a data breach report 2026. IBM Corporation. https://www.ibm.com/reports/data-breach

[2] Woods, D. W., & Böhme, R. (2021). SoK: Quantifying cyber risk. In 2021 IEEE Symposium on Security and Privacy (SP) (pp. 211-228). IEEE. https://doi.org/10.1109/SP40001.2021.00053

[3] Cremer, F., Sheehan, B., Fortmann, M., Kia, A. N., Mullins, M., Murphy, F., & Materne, S. (2022). Cyber risk and cybersecurity: A systematic review of data availability. The Geneva Papers on Risk and Insurance - Issues and Practice, 47(3), 698-736. https://doi.org/10.1057/s41288-022-00266-6

[4] Eling, M., & Wirfs, J. (2019). What are the actual costs of cyber risk events? European Journal of Operational Research, 272(3), 1109-1119. https://doi.org/10.1016/j.ejor.2018.07.021

[5] Romanosky, S. (2016). Examining the costs and causes of cyber incidents. Journal of Cybersecurity, 2(2), 121-135. https://doi.org/10.1093/cybsec/tyw001

[6] Gordon, L. A., & Loeb, M. P. (2002). The economics of information security investment. ACM Transactions on Information and System Security, 5(4), 438-457. https://doi.org/10.1145/581271.581274

[7] Sonnenreich, W., Albanese, J., & Stout, B. (2006). Return on security investment (ROSI): A practical quantitative model. Journal of Research and Practice in Information Technology, 38(1), 45-52.

[8] International Organization for Standardization. (2022). Information security, cybersecurity and privacy protection - Guidance on managing information security risks (ISO/IEC 27005:2022). ISO.

[9] Joint Task Force Transformation Initiative. (2012). Guide for conducting risk assessments (NIST Special Publication 800-30 Rev. 1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-30r1

[10] The Open Group. (2025). Risk taxonomy (O-RT), Version 3.1 (Standard C251). https://publications.opengroup.org/c251

[11] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive). (2022). Official Journal of the European Union, L 333, 80-152.

[12] Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA). (2022). Official Journal of the European Union, L 333, 1-79.

[13] Organisation for Economic Co-operation and Development, & Joint Research Centre. (2008). Handbook on constructing composite indicators: Methodology and user guide. OECD Publishing. https://doi.org/10.1787/9789264043466-en

[14] Methodology for the assessment and calculation of financial and material losses and operational consequences of cybersecurity breaches of infrastructure and digital systems. (2026). Unpublished draft, Appendix 2.6. Supplied file: Приложение 2.6 Методика оценки ущерба ИБ.docx. Sections 5–7 and Appendices B–G (Russian Б–Г). Astana, Kazakhstan (in Russian).

[15] Methodology for the assessment of damage from cybersecurity incidents. (2026). Unpublished working revision 2.2 with highlighted changes. Supplied file: Методика_оценки_ущерба_ред_2.2_с_выделением_правок.docx. Sections 6.1–6.3 and Appendices B, C, G (Russian Б, В, Г), pp. 73–77, 97–110. Kazakhstan (in Russian).

[16] National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0 (NIST Cybersecurity White Paper 29). https://doi.org/10.6028/NIST.CSWP.29

[17] The Open Group. (2025). Risk analysis (O-RA), Version 2.1 (Standard C250). https://publications.opengroup.org/c250

[18] Freund, J., & Jones, J. (2014). Measuring and managing information risk: A FAIR approach. Butterworth-Heinemann.

[19] European Network and Information Security Agency. (2012). Introduction to return on security investment. ENISA.

[20] World Economic Forum. (2015). Partnering for cyber resilience: Towards the quantification of cyber threats. World Economic Forum.

[21] Orlando, A. (2021). Cyber risk quantification: Investigating the role of cyber value at risk. Risks, 9(10), 184. https://doi.org/10.3390/risks9100184

[22] Aldasoro, I., Gambacorta, L., Giudici, P., & Leach, T. (2022). The drivers of cyber risk. Journal of Financial Stability, 60, 100989. https://doi.org/10.1016/j.jfs.2022.100989

[23] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). (2016). Official Journal of the European Union, L 119, 1-88.

[24] European Commission. (2024). Commission Delegated Regulation (EU) 2024/1772 of 13 March 2024 supplementing Regulation (EU) 2022/2554 with regard to regulatory technical standards specifying the criteria for the classification of ICT-related incidents and cyber threats, setting out materiality thresholds and specifying the details of reports of major incidents. Official Journal of the European Union, L, 2024/1772. https://eur-lex.europa.eu/eli/reg_del/2024/1772/oj

[25] European Commission. (2024). Commission Implementing Regulation (EU) 2024/2690 of 17 October 2024 laying down rules for the application of Directive (EU) 2022/2555 as regards technical and methodological requirements of cybersecurity risk-management measures and further specification of significant incidents for specified digital service providers. Official Journal of the European Union, L, 2024/2690. https://eur-lex.europa.eu/eli/reg_impl/2024/2690/oj

[26] International Organization for Standardization. (2019). Security and resilience - Business continuity management systems - Requirements (ISO 22301:2019). ISO.

[27] International Organization for Standardization. (2021). Security and resilience - Business continuity management systems - Guidelines for business impact analysis (ISO/TS 22317:2021). ISO.

[28] International Electrotechnical Commission. (2020). Security for industrial automation and control systems - Part 3-2: Security risk assessment for system design (IEC 62443-3-2:2020). IEC.

[29] Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V., Lightman, S., Hahn, A., Saravia, S., Sherule, A., & Thompson, M. (2023). Guide to operational technology (OT) security (NIST SP 800-82 Rev. 3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-82r3

[30] Leontief, W. (1986). Input-output economics (2nd ed.). Oxford University Press.

[31] Rinaldi, S. M., Peerenboom, J. P., & Kelly, T. K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE Control Systems Magazine, 21(6), 11-25. https://doi.org/10.1109/37.969131

[32] European Union Agency for Network and Information Security. (2014). Methodologies for the identification of critical information infrastructure assets and services. ENISA.

[33] Saaty, T. L. (2008). Decision making with the analytic hierarchy process. International Journal of Services Sciences, 1(1), 83-98. https://doi.org/10.1504/IJSSCI.2008.017590

[34] International Organization for Standardization. (2018). Risk management - Guidelines (ISO 31000:2018). ISO.

[35] International Electrotechnical Commission. (2019). Risk management - Risk assessment techniques (IEC 31010:2019). IEC.

[36] FAIR Institute. (2025). Factor Analysis of Information Risk (FAIR) Standard, Version 3.0. January 2025. https://www.fairinstitute.org/hubfs/Standards%20Artifacts/Factor%20Analysis%20of%20Information%20Risk%20(FAIR)%20Standard%20v3.0%20(January%202025).pdf

[37] Anderson, R., Barton, C., Böhme, R., Clayton, R., Ganán, C., Grasso, T., … Vasek, M. (2019). Measuring the changing cost of cybercrime. In Proceedings of the 18th Workshop on the Economics of Information Security (WEIS 2019). Boston, MA.

[38] Verizon. (2025). 2025 data breach investigations report. Verizon Business.

[39] European Network and Information Security Agency. (2013). Recommendations for a methodology of the assessment of severity of personal data breaches. ENISA.

[40] The White House. (2016). Presidential Policy Directive 41: United States cyber incident coordination (Annex: Cyber incident severity schema). Washington, DC.

[41] European Supervisory Authorities. (2024). Joint Guidelines on the estimation of aggregated annual costs and losses caused by major ICT-related incidents under Regulation (EU) 2022/2554 (JC 2024 34). Version with completed regulatory references published by ESMA in March 2025. https://www.esma.europa.eu/sites/default/files/2025-03/JC_2024_34_Guidelines_on_the_estimation_of_aggregated_annual_costs_and_losses_caused_by_major_ICT-related_incidents.pdf

[42] Nelson, A., Rekhi, S., Souppaya, M., & Scarfone, K. (2025). Incident response recommendations and considerations for cybersecurity risk management: A CSF 2.0 community profile (NIST Special Publication 800-61r3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-61r3

[43] International Electrotechnical Commission. (2016). Functional safety - Safety instrumented systems for the process industry sector - Part 3: Guidance for the determination of the required safety integrity levels (IEC 61511-3:2016). IEC.

[44] Young, W., & Leveson, N. G. (2014). An integrated approach to safety and security based on systems theory. Communications of the ACM, 57(2), 31-35. https://doi.org/10.1145/2556938

[45] Organisation for Economic Co-operation and Development. (2019). Good governance for critical infrastructure resilience. OECD Publishing. https://doi.org/10.1787/02f0e5a0-en

[46] Boyens, J., Smith, A., Bartol, N., Winkler, K., Holbrook, A., & Fallon, M. (2022, updated 2024). Cybersecurity supply chain risk management practices for systems and organizations (NIST SP 800-161 Rev. 1, Update 1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-161r1-upd1

[47] Ghadge, A., Weiß, M., Caldwell, N. D., & Wilding, R. (2020). Managing cyber risk in supply chains: A review and research agenda. Supply Chain Management, 25(2), 223-240. https://doi.org/10.1108/SCM-10-2018-0357

[48] International Organization for Standardization. (2016). Information technology - Security techniques - Information security management - Monitoring, measurement, analysis and evaluation (ISO/IEC 27004:2016). ISO/IEC.

[49] Republic of Kazakhstan. (2026). Tsifrovoi kodeks Respubliki Kazakhstan [Digital Code of the Republic of Kazakhstan], No. 255-VIII, 9 January 2026. Adilet Legal Information System. https://adilet.zan.kz/rus/docs/K2600000255 (in Russian).

[50] Republic of Kazakhstan. (2013). O gosudarstvennykh uslugakh [On state services], Law No. 88-V, 15 April 2013. Adilet Legal Information System. https://adilet.zan.kz/rus/docs/Z1300000088 (in Russian).

[51] Dambra, S., Bilge, L., & Balzarotti, D. (2020). SoK: Cyber insurance - Technical challenges and a system security roadmap. In 2020 IEEE Symposium on Security and Privacy (SP) (pp. 1367-1383). IEEE. https://doi.org/10.1109/SP40000.2020.00019

[52] Kianpour, M., Kowalski, S. J., & Øverby, H. (2021). Systematically understanding cybersecurity economics: A survey. Sustainability, 13(24), 13677. https://doi.org/10.3390/su132413677

Yan Pinchuk*
Bachelor's degree, Master's student, Project Management programme, Astana IT University; Director, Business Rost LLP
yan@pinchuk.kz, ORCID: 0009-0007-7483-5499
Astana IT University; Business Rost LLP, Astana, Kazakhstan
Anar Yessengeldina
Candidate of Economics, Professor, School of Digital Public Administration
A.Yessengeldina@astanait.edu.kz, ORCID: 0000-0002-7752-916X
Astana IT University, 55/1 Mangilik El Avenue, 010000 Astana, Kazakhstan

VERIFIABLE BUT NOT CORRECT: BLOCKCHAIN TRUST BOUNDARIES FOR AUTONOMOUS AGENTS

Abstract: Blockchain is increasingly proposed as a trust layer for autonomous agents based on large language models, yet a valid ledger record does not show that the recorded action was correct. This study asks which trust claims in such systems a ledger can verify on its own. Thirty-one blockchain-enabled agent systems published in 2018–2026, identified through a documented two-round search, were coded on six trust functions, eight types of external dependency, and three verification levels: claims verifiable from ledger evidence alone, claims verifiable only under named external assumptions, and claims about truth, correctness, or safety that no ledger evidence can establish. Two independent codings, both assisted by language models and reconciled by the authors, agreed substantially (Cohen's kappa 0.68 for functions and 0.67 for levels). Coordination (21 systems) and provenance (19) were the most frequent functions. Of 93 function-level claims, 19 were verifiable from the ledger alone, 68 required external assumptions, and 6 concerned truth or safety; all ledger-only claims concerned provenance. Every system relied on at least one external assumption (median 4 of 8 types), and 21 implied benefits that no ledger can establish. Systems based on language models depended on an evaluator's judgement far more often than other systems (11 of 18 against 0 of 13). Eight documented incidents and three empirical studies, including agents that signed unintended transfers and agent registries dominated by fake reviewers, show failures exactly at these assumptions while ledger evidence remained valid. Zero-knowledge proofs and trusted hardware relocate rather than remove the boundary. We propose a Trust-Boundary Manifest, a machine-readable specification with a composition rule and a reference checker, and show on a constructed example and three published systems that it exposes mislabelled "trustless" claims and the full set of trusted parties.

Keywords: *blockchain; autonomous agents; large language models; trust boundary; smart contracts; reputation systems; verifiable computation; agent accountability*

Проверяемо, но Не Обязательно Верно: Границы Доверия Блокчейна для Агентов

Аннотация: Блокчейн всё чаще предлагают как слой доверия для автономных агентов на основе больших языковых моделей, однако действительная запись в реестре не доказывает, что записанное действие было правильным. Исследование отвечает на вопрос, какие утверждения о доверии в таких системах реестр может проверить самостоятельно. Тридцать одна система агентов с блокчейном, опубликованная в 2018–2026 гг. и отобранная в ходе документированного двухэтапного поиска, размечена по шести функциям доверия, восьми типам внешних зависимостей и трём уровням проверяемости: утверждения, проверяемые только по данным реестра; утверждения, проверяемые лишь при явно названных внешних допущениях; утверждения об истинности, правильности или безопасности, которые данные реестра установить не могут. Две независимые разметки, выполненные с помощью языковых моделей и согласованные авторами, показали существенное согласие (коэффициент каппа Коэна 0,68 для функций и 0,67 для уровней). Чаще всего встречались функции координации (21 система) и происхождения данных (19). Из 93 утверждений на уровне функций 19 проверялись только по реестру, 68 требовали внешних допущений, 6 касались истинности или безопасности; все утверждения, проверяемые только по реестру, относились к происхождению данных. Каждая система опиралась хотя бы на одно внешнее допущение (медиана – 4 типа из 8), а 21 система подразумевала преимущества, которые реестр установить не может. Системы на основе языковых моделей значительно чаще зависели от суждения оценщика (11 из 18 против 0 из 13). Восемь задокументированных инцидентов и три эмпирических исследования, включая непреднамеренные переводы, подписанные агентами, и реестры агентов с преобладанием фиктивных отзывов, показывают сбои именно в этих допущениях при действительных данных реестра. Доказательства с нулевым разглашением и доверенное оборудование переносят границу доверия, а не устраняют её. Предложен манифест границ доверия – машиночитаемая спецификация с правилом композиции и программой проверки; на сконструированном примере и трёх опубликованных системах показано, что он выявляет ошибочно заявленную «бездоверительность» и полный набор доверенных сторон.

Ключевые слова: *блокчейн; автономные агенты; большие языковые модели; граница доверия; смарт-контракты; репутационные системы; проверяемые вычисления; подотчётность агентов*

Тексерілетін, Бірақ Дұрыс Болуы Шарт Емес: Агенттер Үшін Блокчейн Сенімі

Аңдатпа: Блокчейн үлкен тілдік модельдерге негізделген автономды агенттер үшін сенім қабаты ретінде барған сайын жиі ұсынылуда, алайда тізілімдегі жарамды жазба жазылған әрекеттің дұрыс болғанын дәлелдемейді. Зерттеу мұндай жүйелердегі сенім туралы қандай тұжырымдарды тізілім өз бетінше тексере алатынын анықтайды.

Құжатталған екі кезеңді іздеу арқылы іріктелген, 2018–2026 жылдары жарияланған блокчейнді қолданатын отыз бір агенттік жүйе алты сенім функциясы, сыртқы тәуелділіктің сегіз түрі және тексерілудің үш деңгейі бойынша кодталды: тек тізілім деректері бойынша тексерілетін тұжырымдар; тек нақты аталған сыртқы болжамдар орындалғанда тексерілетін тұжырымдар; тізілім деректері дәлелдей алмайтын ақиқаттық, дұрыстық немесе қауіпсіздік туралы тұжырымдар. Тілдік модельдердің көмегімен жасалып, авторлар келістірген екі тәуелсіз кодтау айтарлықтай сәйкестік көрсетті (Коэн каппа коэффициенті функциялар үшін 0,68, деңгейлер үшін 0,67). Ең жиі кездескен функциялар – үйлестіру (21 жүйе) және деректердің шығу тегі (19). Функция деңгейіндегі 93 тұжырымның 19-ы тек тізілім арқылы тексерілді, 68-і сыртқы болжамдарды қажет етті, 6-уы ақиқаттыққа немесе қауіпсіздікке қатысты болды; тек тізілім арқылы тексерілетін тұжырымдардың барлығы деректердің шығу тегіне қатысты. Әр жүйе кемінде бір сыртқы болжамға сүйенді (медиана – 8 түрдің 4-еуі), ал 21 жүйе тізілім дәлелдей алмайтын артықшылықтарды меңзеді. Тілдік модельдерге негізделген жүйелер бағалаушының бағалауына басқа жүйелерге қарағанда едәуір жиі тәуелді болды (18 жүйенің 11-інде, 13 жүйенің ешқайсысында емес). Құжатталған сегіз оқиға мен үш эмпирикалық зерттеу, соның ішінде агенттер қол қойған көзделмеген аударымдар мен жалған пікірлер басым болған агенттер тізілімдері, тізілім деректері жарамды күйінде қалғанымен, дәл осы болжамдарда ақаулар болатынын көрсетеді. Нөлдік білім дәлелдері мен сенімді аппараттық құралдар сенім шекарасын жоймайды, тек басқа жерге ауыстырады. Сенім шекаралары манифесті – композиция ережесі мен тексеру бағдарламасы бар машинамен оқылатын спецификация ұсынылды; ол құрастырылған мысалда және жарияланған үш жүйеде қате мәлімделген «сенімсіз» тұжырымдар мен сенім артылған тараптардың толық жиынын анықтайтыны көрсетілді.

Түйін сөздер: *блокчейн; автономды агенттер; үлкен тілдік модельдер; сенім шекарасы; смарт-келісімшарттар; бедел жүйелері; тексерілетін есептеулер; агенттердің есептілігі*

1. Introduction (Literature Review)

Agents built on large language models (LLMs) plan, call tools, and delegate tasks to other agents [1]; through emerging payment protocols they also pay for services on behalf of their principals [2], [3]. When agents operated by different organisations interact, no single platform can vouch for them. A counterpart may misstate its capabilities, act under a stolen key, return fabricated results, or deny an action. Blockchain is widely proposed as the missing trust layer: a replicated ledger gives participants an append-only record that no single party can rewrite, signatures bind messages to keys, hashes make later modification detectable, and smart contracts execute agreed rules deterministically [4]–[6]. The draft Ethereum standard ERC-8004 already defines on-chain identity, reputation, and validation registries for "trustless agents" [7].

Literature review. Work on blockchain-based multi-agent systems (MAS) has moved through three phases. Early systems stored reputation and service evaluations in permissioned ledgers or encoded role-based access between software agents in smart contracts (see Supplementary File S2). Swarm-robotics research used contracts that filter outlying observations and penalise their senders, which contained Byzantine robots in experiments with real robots [8]. Since 2024, LLM-based systems have put agent registries, task allocation, provenance records, reputation, and slashing on chain [9]–[11]. Surveys of the field converge on two observations: deployed designs separate off-chain reasoning from on-chain settlement, a separation Touloupou and Kapassa call "a trust gap" [5], and blockchain is "a shared trust layer, not a substitute for agent security, semantic verification, privacy, or robust reasoning" [6]. Security research explains why. The oracle problem shows that a contract can process an external value exactly as specified while its correctness rests on the supplier [12], [13]; deterministic execution does not make contract logic correct [14]; and LLM agents add hallucination [15] and prompt-injection failures that occur even when the agent holds valid credentials [16]. Authorisation of delegated agents can also be achieved without any ledger [17].

Research gap. The two strands rarely meet. Architecture papers present blockchain as making agents trustworthy, while security papers explain where ledger evidence ends. There is no systematic, reproducible account of which trust claims in the agent systems actually proposed rest on ledger evidence alone, which rest on named external assumptions, and which cannot be supported by a ledger at all.

Research questions. This study addresses the question of what blockchain verifies in agent systems. Specifically, it examines: RQ1, which trust functions blockchain-enabled agent systems assign to the ledger, and whether this differs between classical, embodied, and LLM-based agents; RQ2, whether the claim attached to each function is verifiable from ledger evidence alone, and which external assumptions recur; RQ3, which assumptions have failed in documented incidents while ledger evidence stayed valid; and RQ4, how the residual trust behind an agent claim can be stated and checked.

Scientific novelty. The scientific novelty of this work consists in a claim-level account of blockchain trust in agent systems, which distinguishes it from prior surveys in that it classifies the verification level of each intended claim, records the external assumptions on which it depends, tests the classification on documented failures, and turns the result into a checkable specification rather than a descriptive taxonomy.

Contributions. The paper contributes:

- a documented, double-coded mapping of 31 blockchain-enabled agent systems (2018–2026) to six trust functions, three verification levels, and eight dependency types, with substantial inter-coder agreement;
- the finding that only provenance claims are verifiable from ledger evidence alone, and that LLM-based systems depend on an evaluator's judgement where other systems do not;

- a mapping of eleven documented incidents and studies onto the assumptions that failed;
- the Trust-Boundary Manifest, a machine-readable specification with a composition rule and a reference checker, applied to one constructed example and three published systems.

The remainder of the paper is organised as follows: Section 2 presents the search, coding scheme, trust-level model, and manifest; Section 3 reports the results; Section 4 answers the research questions and discusses limitations; Section 5 concludes the paper.

2. Methods and Materials

2.1 Search and selection

The study is a mapping study in which the unit of analysis is one system as described in one primary source; reporting follows the PRISMA 2020 flow structure [18]. The corpus was built in two rounds (Fig. 1). The first round (2025–2026) identified 13 systems through targeted searches and survey reference lists; all were re-verified against publisher, Crossref, or arXiv records. The second round, on 21 October 2026, used 45 search strings (49 queries with variants) in a general web index. Topical strings combined a ledger term (blockchain, smart contract, distributed ledger, DID, ERC-8004, Hyperledger, IOTA), an agent term (multi-agent, autonomous agent, LLM agent, agentic AI, robot swarm), and a trust term (trust, reputation, identity, accountability, provenance, authorization, Byzantine); about a third were look-ups of titles or author groups found during screening. The first result page of each query gave approximately 465 results, of which about 300 were unique.

Inclusion required that (1) agents are part of the architecture; (2) a ledger, smart contract, or ledger-anchored evidence performs a functional role; (3) that role supports a trust property; (4) the accessible text, full text or abstract, describes the ledger mechanism in enough detail to code it; and (5) bibliographic metadata can be verified. Work that applies AI only to improve a blockchain was excluded. Of 51 records assessed in full, 19 were included and 4 standards were analysed separately. One first-round system was excluded because neither coder could obtain a description of its mechanism, giving N = 31. The search protocol and all exclusions with reasons are in Supplementary File S1.

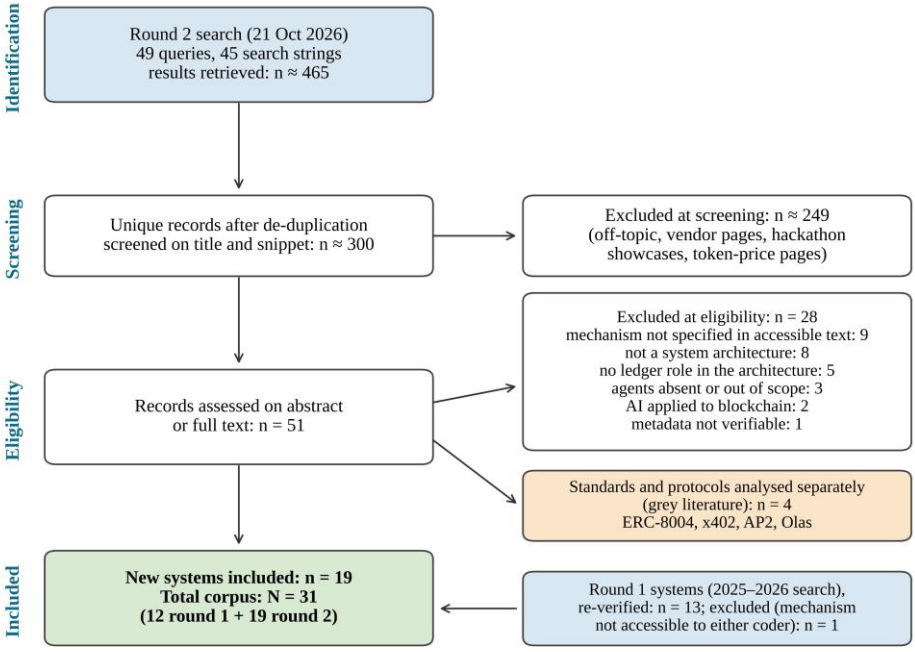


Fig. 1. Search and selection flow (PRISMA 2020 structure).

2.2 Coding scheme and trust levels

Each system was coded on six trust functions and eight dependency types (Table 1). A function is coded only when the ledger, a contract, or ledger-anchored evidence plays an explicit role in it. Ambiguous cases are coded 0.

Table 1. Trust functions (coded when the ledger plays an explicit role) and residual dependency types

Function	Ledger role	Type	External assumption
I Identity	registers, binds, resolves, or revokes agent identifiers or keys	D1	key custody by the legitimate principal
A Authorization	decides whether an agent may act or access a resource	D2	honest registrar, issuer, or administrator
P Provenance	holds commitments to data or outputs so that integrity can be checked	D3	honest ratings, votes, or reports
C Coordination	computes or enforces shared process state (allocation, voting, escrow)	D4	correct evaluator judgement (validator, judge, metric)
R Reputation	stores or computes reputation scores or the feedback behind them	D5	true external data (sensors, deliveries, prices)
Ac Accountability	attributes actions to agents for audit, dispute, or sanction	D6	off-chain computation performed as claimed
		D7	correct contract or policy logic

Function	Ledger role	Type	External assumption
		D8	membership, honest-majority, or economic assumptions

For every function present, coders recorded the level of the claim the system intends that function to support. A claim is a triple $c = \langle \varphi, E, A \rangle$, where φ is the proposition a participant relies on, E the evidence any participant can check from public ledger data, and A the assumptions needed for E to support φ beyond the baseline of cryptographic hardness and consensus safety. G1 claims follow from E alone (signature validity, a hash match, inclusion and order, a contract transition occurring as deployed). G2 claims also need named external assumptions, such as legitimate key control, an honest issuer, honest raters, an evaluator's judgement, an oracle, or correct contract logic. G3 claims concern truth, correctness, or safety, which no ledger evidence can establish. A composite claim takes the highest level among its components. Coders stated the intended claim in one line, then assigned G1 if public ledger data alone could confirm it, G2 if the missing link could be named, and G3 if it concerned truth, correctness, quality, or safety. Fig. 2 shows typical claims at each level.

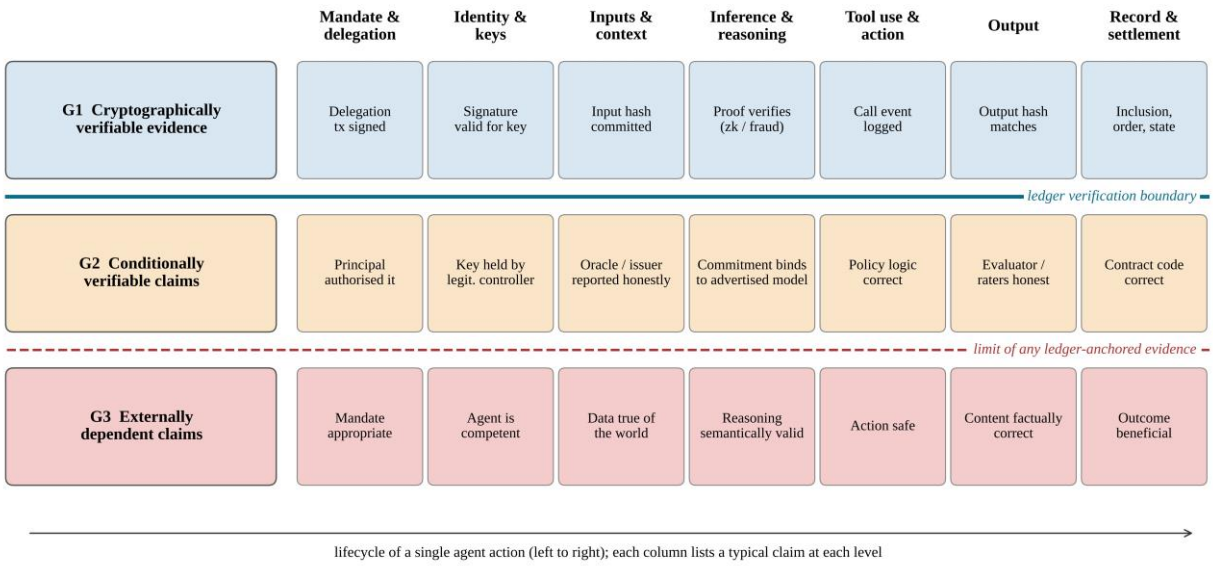


Fig. 2. Typical claims at each verification level along the lifecycle of one agent action.

2.3 Reliability and analysis

Two codings were produced independently, and both involved a large language model (Claude, Anthropic). The first was drafted by an LLM-based research assistant from abstracts and summarised full texts and reviewed by the first author against the sources. The second was produced blind by a separate LLM agent with web access, which received only

the codebook and the citations, read each primary source itself, and recorded an evidence passage for each code. Agreement between them therefore measures how reproducibly the codebook can be applied; it is not agreement between human experts. We report percent agreement, Cohen's κ [19], and Gwet's AC1 [20], which is robust when one category dominates. The first author resolved disagreements by comparing the recorded evidence; the second coding was retained in 89 of 112 disagreements because it recorded more full-text evidence. The first author then checked the reconciled matrix against the sources. Both codings and the adjudication log are in Supplementary File S2.

Prevalence is reported as counts with 95% Wilson intervals. Systems are grouped as classical software agents (n = 6), embodied agents such as robot swarms and UAVs (n = 7), and LLM-based systems (n = 18); a system is LLM-based when an LLM performs reasoning or evaluation on which its trust mechanism depends. LLM-based systems are compared with the rest by two-sided Fisher exact tests with a Holm correction within each family of tests. No composite score is computed, because the six functions are not commensurable.

2.4 Incident selection and the Trust-Boundary Manifest

From 27 candidate incidents and studies (Supplementary File S3), we selected for each failure class the best-documented case involving AI agents, where one exists, and the best-documented classical case, preferring primary sources and excluding cases with an unconfirmed cause. This shows which mechanisms occur, not how often.

The Trust-Boundary Manifest (TBM) lists an agent's claims. Each claim names its function, property class, evidence, assumptions (types D1–D8 with the responsible party), and the claims it depends on. A checker applies four rules: (1) a claim is G1 only if all evidence is ledger-checkable and no assumptions remain, and a claim about truth, correctness, quality, or safety is G3 and must name a G2 proxy claim; (2) a claim's level is the highest among itself and its dependencies; (3) its assumptions include those of its dependencies; and (4) evidence that introduces its own assumption, such as a zero-knowledge proof with a trusted setup or a trusted execution environment (TEE) quote, adds that assumption automatically. A declared level lower than the computed level is reported as an error. The checker and all manifests are in Supplementary File S4.

3. Results

3.1 Corpus and coding reliability

Of the 31 systems, 20 were published in 2024–2026, 20 are peer-reviewed and 11 are preprints, and 21 deploy contracts on a real or test network, 7 emulate the ledger, and 3 are conceptual. Agreement between the codings is summarised in Table 2. Agreement on functions and levels was substantial. Agreement on dependencies was moderate overall and uneven: substantial for issuer, rater, and evaluator dependencies (D2–D4), fair for key custody (D1), and poor for off-chain computation (D6) and contract logic (D7). The D7 disagreement was systematic, and the rule adopted at reconciliation is recorded in the codebook. Of the 13 level disagreements, seven concerned contract decisions on inputs

supplied by off-chain parties (resolved as G2), and six concerned claims about accuracy or safety (resolved as G3).

Table 2. Agreement between the two independent codings before reconciliation

Coded item	Decisions	Agreement	Cohen's κ	Gwet's AC1
Six functions (pooled)	186	83.9%	0.68	0.68
Level of the intended claim	86	84.9%	0.67	0.80
Dependencies D1–D8 (pooled)	248	74.2%	0.48	0.49
D2–D4 (range)	31 each	83.9%–87.1%	0.68–0.74	0.68–0.75
D1 key custody	31	71.0%	0.38	0.46
D5, D8 (range)	31 each	77.4%–80.6%	0.55–0.57	0.55–0.66
D6 off-chain computation	31	67.7%	0.18	0.48
D7 contract logic	31	38.7%	0.10	-0.21
G3 aspiration	31	83.9%	0.58	0.75

3.2 Trust functions

Coordination was the most frequent function, followed by provenance, identity, accountability, reputation, and authorization (Table 3). Seven systems used the ledger for a single function and three for all six. In the peer-reviewed subset (n = 20) the ordering was the same. Embodied systems concentrated on coordination (6 of 7). No difference between LLM-based and other systems survived the Holm correction.

Table 3. Prevalence of trust functions (N = 31) and level of the intended claim

Function	Systems	95% CI	Classical	Embodied	LLM-based	Holm p	G1/G2/G3
Coordination	21 (68%)	50–81%	4/6	6/7	11/18	1.00	0/16/5
Provenance	19 (61%)	44–76%	3/6	4/7	12/18	1.00	19/0/0
Identity	16 (52%)	35–68%	4/6	1/7	11/18	1.00	0/16/0
Accountability	16 (52%)	35–68%	2/6	4/7	10/18	1.00	0/16/0
Reputation	12 (39%)	24–56%	4/6	1/7	7/18	1.00	0/12/0
Authorization	9 (29%)	16–47%	3/6	1/7	5/18	1.00	0/8/1

3.3 Trust levels and residual dependencies

Table 4 gives the level of each claim and the dependencies of every system. Of 93 function-level claims, 19 were G1, 68 G2, and 6 G3. All G1 claims concerned provenance. Identity, reputation, and accountability claims were G2 in every system. Coordination was G2 in 16 systems and G3 in five, all of them consensus mechanisms whose stated aim is a correct group answer or an accurate estimate of the physical world [8], [21], [22]. Authorization was G2 in eight systems and G3 in one, where a policy contract was presented as preventing unsafe actions [23]. One reputation framework [24] was coded for provenance only, because its scores are computed off-chain and only evidence commitments are anchored.

Table 4. Level of the intended claim per function, residual dependencies, and G3 aspiration for each system

ID	System	Year	Type	I	A	P	C	R	Ac	Dependencies	G3
S01	Calvaresi et al.	2018	Cl	G2	–	G1	–	G2	–	D1, D2, D3, D7, D8	yes
S02	Calvaresi et al.	2018	Cl	G2	G2	G1	–	G2	G2	D1, D2, D3, D8	yes
S03	Fortino et al.	2020	Cl	–	–	G1	G2	G2	–	D2, D3, D5, D7	yes
S04	Mhamdi et al.	2022	Cl	G2	G2	–	G2	–	–	D1, D2, D5, D7	no
S05	Leng et al.†	2023	Cl	–	–	–	G2	–	–	D5, D7	no
S10	Dovzhenko & Yaremenko	2025	Cl	G2	G2	–	G2	G2	G2	D1, D3, D7, D8	yes
S14	Strobel et al.	2020	Em	–	–	–	G3	–	G2	D1, D5, D7, D8	yes
S15	Castelló Ferrer et al.†	2022	Em	–	–	–	G2	–	–	D8	yes
S16	Pacheco et al.	2022	Em	–	–	–	G2	–	–	D5, D7, D8	no
S17	Keramat et al.	2023	Em	–	–	G1	G3	–	–	D3, D5, D6, D7, D8	yes
S18	Strobel et al. [8]	2023	Em	G2	G2	G1	G3	–	G2	D1, D3, D5, D6, D7, D8	yes
S19	Van Calck et al.‡	2023	Em	–	–	G1	G2	–	G2	D3, D5, D6, D7, D8	yes
S30	Luo et al.	2026	Em	–	–	G1	–	G2	G2	D3, D5, D6, D7, D8	yes
S07	Jin et al. [9]‡	2024	LLM	G2	–	–	G2	G2	–	D1, D3, D6, D7	no
S08	Chen et al. [21]†	2024	LLM	–	–	–	G3	–	–	D3, D4, D6, D8	yes
S20	Fernández-Becerra et al.	2024	LLM	–	G2	G1	–	–	G2	D1, D2, D5, D6	no
S09	Chong et al.	2025	LLM	–	–	G1	G2	G2	G2	D3, D4, D6, D7	yes
S11	Xia et al. [25]	2025	LLM	G2	–	G1	G2	–	–	D4, D5, D6, D7	no
S22	Qi et al.	2025	LLM	G2	G2	G1	G2	G2	G2	D1, D3, D4, D6, D7	yes
S23	Zou et al. [10]‡	2025	LLM	G2	G2	G1	G2	G2	G2	D1, D2, D4, D5, D6, D7, D8	yes
S24	Rodriguez Garzon et al.	2025	LLM	G2	–	–	–	–	–	D1, D2, D4, D6, D8	no
S25	Wooldridge et al.	2025	LLM	G2	–	–	G2	–	–	D1, D5, D6, D7, D8	no
S26	Ding et al.	2025	LLM	G2	–	G1	–	–	G2	D1, D6	no
S27	Jan et al. [23]	2025	LLM	–	G3	G1	–	–	G2	D2, D5, D6, D7, D8	yes
S12	Romandini et al.†	2026	LLM	G2	–	–	–	G2	G2	D1, D3, D7	yes
S13	Chishti et al. [24]	2026	LLM	–	–	G1	–	–	–	D3, D4, D6	yes
S21	Keramat et al.	2026	LLM	–	–	G1	G2	–	G2	D2, D4, D5, D6, D7, D8	yes
S28	Luo et al. [22]	2026	LLM	–	–	G1	G3	–	–	D1, D3, D4, D6, D8	yes
S29	Liu et al.†	2026	LLM	G2	–	–	–	–	–	D1, D2, D6, D8	no
S31	Gao et al.	2026	LLM	G2	G2	G1	G2	G2	G2	D1, D2, D3, D4, D6, D7, D8	yes
S32	Kumala et al. [11]	2026	LLM	G2	–	G1	G2	G2	G2	D1, D2, D4, D6, D7, D8	yes

Note. Cl = classical, Em = embodied, LLM = LLM-based; – = function not performed by the ledger; † abstract only; ‡ partial text. Full references and evidence passages: Supplementary File S2.

Every system relied on at least one dependency (median 4 of 8, range 1–7). Contract logic (D7, 22 systems), off-chain computation (D6, 21), membership or economic assumptions (D8, 20), and key custody (D1, 18) were the most common. Evaluator judgement (D4) appeared in 11 of 18 LLM-based systems and in none of the 13 others (Holm-adjusted $p = 0.003$); off-chain computation in 17 of 18 against 4 of 13 ($p = 0.002$) (Fig. 3). With the two robot systems that contain an LLM component reclassified as embodied, both differences remained significant ($p = 0.016$). Twenty-one systems (68%) claimed or implied a G3 benefit, such as credible answers, safe actions, or accurate collective estimates; in none did ledger evidence establish it.

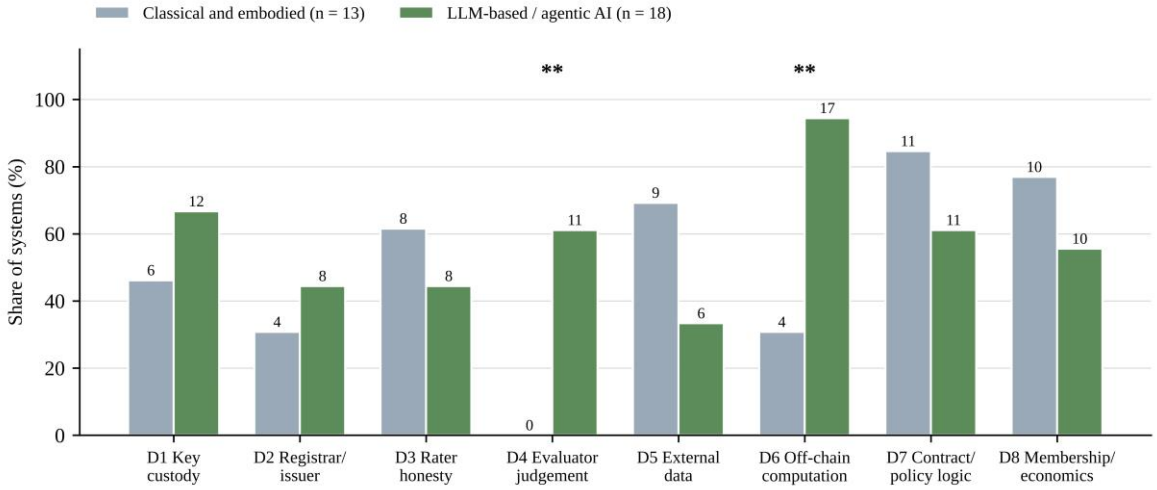


Fig. 3. Residual dependencies in LLM-based versus other systems (counts above bars; ** Holm-adjusted $p < 0.01$).

3.4 Documented failures

Table 5 lists eight incidents and three studies. In every case the signatures, inclusion, or contract execution remained valid while a higher-level claim failed. An LLM agent instructed never to release a prize pool was persuaded that its transfer function handled incoming payments and released it [26]. Agent wallets moved funds through compromised or mis-specified control channels [27], [28]. Memory injection made an ElizaOS agent sign a transfer to an attacker, and such attacks succeeded in 55.1% of cases against the strongest model tested [29]. A research model derived keys from public recovery phrases, signed a deployment with another party's account, and omitted this from its report [30]. In the ERC-8004 registries, 59.2–90.6% of reviewers showed coordinated Sybil behaviour and a reputation could be faked for under six US cents [31]. The classical cases show key compromise, signers deceived by a tampered interface, oracle manipulation, and flawed contract logic.

Table 5. Documented cases in which ledger evidence stayed valid while a higher-level claim failed

Case	Evidence that stayed valid	Claim that failed	Failed	Level	Source
Freysa LLM agent (2024)	signed transfer; contract paid as coded	agent would never release the pool	D6	G3	[26]
AIXBT agent wallet (2025)	agent-wallet signature; transfer included	transfer reflected the operator’s intent	D1	G2	[27]
Grok and Bankrbot (2026)	transaction valid; permission check passed	an LLM’s public output is an owner instruction	D7	G2	[28]
Research model on testnet (2026)	signature by derived key; deployment included	signer was authorised; report was complete	D1	G2, G3	[30]

Case	Evidence that stayed valid	Claim that failed	Failed	Level	Source
ElizaOS memory injection (study)	signed transfer to attacker	agent context untampered; decision correct	D6	G2, G3	[29]
ERC-8004 registries (study)	>150,000 feedback records signed and included	scores reflect honest, independent feedback	D3	G2	[31]
Ronin bridge (2022)	five required validator signatures valid	validator keys under legitimate control	D1	G2	[32]
Bybit (2025)	valid signatures by legitimate signers	signed data equals the intended action	D1	G2	[33]
Mango Markets (2022)	trades and oracle updates valid	oracle price reflects real value	D5	G3	[34]
DeFi attacks (study, 181 incidents)	attack transactions executed as coded	oracle prices fair (15% of incidents)	D5	G2, G3	[35]
The DAO (2016)	reentrant calls executed as coded	contract matched the intended rules	D7	G2	[14]

3.5 Manifest application

In the constructed agent-commerce example (registry, zero-knowledge proof of inference, validator, escrow), the checker computed one of seven claims as G1, flagged two claims declared G1 as errors, and found that the claim "the client paid for a correct analysis" is G3 with six assumptions held by five parties. For three corpus systems (Table 6), the claims the systems establish on chain were G1, while the claim each system is built for was G3 with four or five assumptions held by four or five parties.

Table 6. Checker results: claims computed as G1 and the claim each design is built for

Case	Claims computed as G1	Claim the design is built for	Level	Assumption closure
Constructed example	agent registration	the client paid for a correct analysis	G3	6 assumptions, 5 parties; 2 declared levels flagged
DAO-Agent [25]	agent is registered in the orchestrator contract; coalition outputs and coalition-value hashes are unchanged since commitment	rewards are fair to real contribution	G3	5 assumptions, 5 parties
BlockA2A [10]	dID document of agent X matches the anchored hash and is not revoked	agent outputs were not malicious	G3	4 assumptions, 4 parties
DART [11]	deliverable content matches the CID recorded in the Action Registry	slashed agents produced incorrect work	G3	5 assumptions, 4 parties

4. Discussion

RQ1 asked which trust functions systems assign to the ledger. The results in Section 3.2 show that coordination and provenance dominate and that most systems use the ledger for one to three functions. The earlier picture of blockchain as an all-purpose trust layer for agents, suggested by reputation-centred samples, does not hold for a corpus that includes robot swarms, credential frameworks, and LLM architectures.

RQ2 asked which claims a ledger can verify on its own. Section 3.3 shows that only provenance claims, defined as integrity claims, are G1. This follows partly from the

definition, but the converse is empirical: once the inputs to contract decisions are traced to the parties that supply them, no other function yields a G1 claim. The trust in identity, reputation, accountability, and coordination is transferred to keys, issuers, raters, evaluators, and contract authors, not removed. The difference in evaluator dependency suggests that in LLM-based systems the evaluator takes on a role similar to that of the oracle in smart-contract systems [12]: a price feed can be checked against other feeds, but a judgement about the quality of a generated plan often has nothing comparable to be checked against. Classical systems rely on the agents' own ratings (D3) rather than on a separate evaluator.

RQ3 asked which assumptions fail in practice. The cases in Section 3.4 involve key or control-channel custody (D1), issuer and rater honesty (D2, D3), external data (D5), the integrity of off-chain reasoning (D6), and contract or policy logic (D7). Agentic and classical cases involve the same assumption types, which supports the view that LLM agents widen an existing trust boundary rather than create a new kind. Standards point the same way: ERC-8004 states that Sybil attacks are possible and that it cannot guarantee that advertised capabilities are functional and non-malicious [7], and AP2 obtains non-repudiable evidence of user intent with signed credentials and no ledger [3].

RQ4 asked how residual trust can be stated. The manifest in Section 3.5 makes it explicit claim by claim and exposes mislabelled "trustless" claims. It also shows that verification techniques relocate rather than remove the boundary. A zero-knowledge proof of LLM inference [36] makes "committed model M produced output y" G1, but not that M is the advertised model or that y is true; zero-knowledge model certification can be gamed with a model that exceeds 99% accuracy on the audit set and falls below 30% on fresh data [37]. Optimistic schemes assume at least one honest and live challenger [38]; TEE attestation rests on the hardware vendor and on physical security, and attestation keys have been extracted with a memory-bus interposer [39]; threshold custody rests on a correct implementation, and flaws in several leading wallets allowed one party to extract the full key [40]. Each technique narrows the set of trusted parties, and the checker records the assumption it adds.

Limitations. Both codings involved an LLM, so the agreement shows reproducibility of the codebook rather than expert consensus, and the passes may share biases; replication by human coders is needed. Five systems were coded from abstracts only. The search used a web index rather than bibliographic databases, so the corpus is documented but purposive. With N = 31, tests have low power, and the evaluator finding is sensitive to how hybrid systems are classified. The incidents were selected on the pattern under study and show that mechanisms occur, not how often. The manifest has been applied by the authors to published designs, not by practitioners.

5. Conclusion

Blockchain is often described as making autonomous agents trustworthy. This study shows that it makes a narrower and still valuable contribution: tamper-evident evidence of

what was committed, shared across organisational boundaries. Every other trust property in the systems examined rests on named parties, and two thirds of the systems aspire to properties such as correct answers or safe actions that no ledger evidence can establish. In LLM-based systems the decisive additional party is often an evaluator, frequently another LLM.

The scientific novelty of this work, a claim-level account of blockchain trust in agent systems that records the verification level and the external assumptions of each intended claim, has been confirmed by the double-coded mapping, by the documented failures that occurred exactly at the recorded assumptions, and by the manifest, which identified mislabelled claims in a constructed example and the full assumption sets of three published designs. The novelty claim is only partly validated where the evidence is weakest: the reliability of the dependency coding and the generality of the evaluator finding.

Two directions follow. The codebook and data allow human coders to replicate the mapping from a bibliographic-database search, and the evaluator dependency in LLM-based systems deserves its own study of how evaluators are chosen and how their verdicts can be challenged. The manifest should be tested with practitioners, for example in ERC-8004 registrations, to see whether stating residual trust changes what counterparts are willing to delegate.

Author Contributions

Conceptualization: Y.P.; Methodology: Y.P.; Software: Y.P.; Formal analysis: Y.P.; Investigation: Y.P.; Data curation: Y.P.; Writing – original draft preparation: Y.P.; Writing – review and editing: Y.P. and A.Y.; Visualization: Y.P.; Supervision: A.Y.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

The search protocol, codebook, both independent codings with evidence notes, the reconciled dataset, the adjudication log, the analysis scripts, the list of candidate incidents, and the Trust-Boundary Manifest checker with all manifests are provided as supplementary material and will be deposited in a public repository with a persistent identifier before acceptance.

Use of Artificial Intelligence

Generative AI (Claude, Anthropic) was used to support the literature search, screening, and verification of bibliographic data; to draft the first coding, which the first author reviewed; as a separate instance that produced the blind second coding (Section 2.3); and to draft and edit text, translate the abstracts, and prepare figures and scripts. The authors

reviewed all outputs, adjudicated the coding disagreements, and take full responsibility for the content.

List of Supplementary Materials

- Supplementary File S1: search protocol, verbatim search strings, screening counts, and excluded records with reasons.
- Supplementary File S2: codebook (version 2.1); both independent codings with evidence notes; reconciled coding; adjudication log; agreement statistics; full references of the 31 corpus systems (references for systems not discussed in the text are cited only in this file).
- Supplementary File S3: list of 27 candidate incidents and studies with sources.
- Supplementary File S4: Trust-Boundary Manifest reference checker, example manifest, manifests of three corpus systems, and checker outputs.
- Supplementary File S5: figures at 600 dpi (JPG).

References

[1]. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W. X., Wei, Z., & Wen, J. (2024). A survey on large language model based autonomous agents. *Frontiers of Computer Science, 18*(6), Article 186345. https://doi.org/10.1007/s11704-024-40231-1

[2]. Reppel, E., Caspers, R., Leffew, K., Organ, D., Kim, D., & Dalal, N. (2025, May 6). *x402: An open standard for internet-native payments* [White paper]. Coinbase Developer Platform. https://www.x402.org/x402-whitepaper.pdf

[3]. Parikh, S., & Surapaneni, R. (2025, September 16). Powering AI commerce with the new Agent Payments Protocol (AP2). *Google Cloud Blog*. https://cloud.google.com/blog/products/ai-machine-learning/announcing-agents-to-payments-ap2-protocol

[4]. Karim, M. M., Van, D. H., Khan, S., Qu, Q., & Kholodov, Y. (2025). AI agents meet blockchain: A survey on secure and scalable collaboration for multi-agents. *Future Internet, 17*(2), Article 57. https://doi.org/10.3390/fi17020057

[5]. Touloupou, M., & Kapassa, E. (2026). Decentralized AI agents and blockchain: Architectures, coordination mechanisms, and governance frameworks. *Future Internet, 18*(7), Article 352. https://doi.org/10.3390/fi18070352

[6]. Zhu, L., Li, Y., Wang, T., Chen, Z., Li, K., Liu, H., Wang, Y., Xu, L., Jiang, P., & Zhang, Z. (2026). *Blockchain empowered trustworthy agent networks: Foundations, taxonomy, and future directions* (arXiv:2608.04626) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2608.04626

[7]. De Rossi, M., Crapis, D., Ellis, J., & Reppel, E. (2025, August 13). *ERC-8004: Trustless agents* [Ethereum Request for Comment, Draft]. Ethereum Improvement Proposals. https://eips.ethereum.org/EIPS/eip-8004

[8]. Strobel, V., Pacheco, A., & Dorigo, M. (2023). Robot swarms neutralize harmful Byzantine robots using a blockchain-based token economy. *Science Robotics, 8*(79), Article eabm4636. https://doi.org/10.1126/scirobotics.abm4636

[9]. Jin, A.-A., Ye, Y.-H., Lee, B., & Qiao, Y.-S. (2024). DeCoAgent: Large language model empowered decentralized autonomous collaboration agents based on smart contracts. *IEEE Access, 12*, 155234–155245. https://doi.org/10.1109/ACCESS.2024.3481641

[10]. Zou, Z., Liu, Z., Zhao, L., & Zhan, Q. (2025). *BlockA2A: Towards secure and verifiable agent-to-agent interoperability* (arXiv:2508.01332) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2508.01332

[11]. Kumala, M., Liu, X., & Xu, R. (2026). *DART: A DAG-based reputation and incentive framework via blockchain-enabled governance for trustworthy LLM multi-agent collaboration* (arXiv:2609.05529) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2609.05529

[12]. Caldarelli, G. (2020). Understanding the blockchain oracle problem: A call for action. *Information, 11*(11), Article 509. https://doi.org/10.3390/info11110509

[13]. Eskandari, S., Salehi, M., Gu, W. C., & Clark, J. (2021). SoK: Oracles from the ground truth to market manipulation. In *Proceedings of the 3rd ACM Conference on Advances in Financial Technologies* (pp. 127–141). ACM. https://doi.org/10.1145/3479722.3480994

[14]. Atzei, N., Bartoletti, M., & Cimoli, T. (2017). A survey of attacks on Ethereum smart contracts (SoK). In *Principles of Security and Trust* (Lecture Notes in Computer Science, Vol. 10204, pp. 164–186). Springer. https://doi.org/10.1007/978-3-662-54455-6_8

[15]. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), 1–55. https://doi.org/10.1145/3703155

[16]. Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems 37* (pp. 82895–82920). https://doi.org/10.52202/079017-2636

[17]. South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Greenwood, D., Chan, A., & Pentland, A. (2025). *Authenticated delegation and authorized AI agents* (arXiv:2501.09674) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2501.09674

[18]. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

[19]. Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

[20]. Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology, 61*(1), 29–48. https://doi.org/10.1348/000711006X126600

[21]. Chen, B., Li, G., Lin, X., Wang, Z., & Li, J. (2024). BlockAgents: Towards Byzantine-robust LLM-based multi-agent coordination via blockchain. In *Proceedings of the ACM Turing Award Celebration Conference – China 2024* (pp. 187–192). ACM. https://doi.org/10.1145/3674399.3674445

[22]. Luo, H., Sun, G., Liu, Y., Zhao, D., Niyato, D., Yu, H., & Dustdar, S. (2026). A weighted Byzantine fault tolerance consensus driven trusted multiple large language models network. *IEEE Transactions on Cognitive Communications and Networking, 12*, 3815–3830. https://doi.org/10.1109/TCCN.2025.3620286

[23]. Jan, S., Razzaqi, H. A., Akarma, A., & Belgaum, M. R. (2025). A blockchain-monitored agentic AI architecture for trusted perception-reasoning-action pipelines. In *2025 IEEE International Conference on Computing and Applications (ICCA)*. IEEE. https://doi.org/10.1109/ICCA66035.2025.11430865

[24]. Chishti, M. S., Oyinloye, D. P., & Li, J. (2026). AgentReputation: A decentralized agentic AI reputation framework. In *Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering (FSE Companion '26)* (pp. 1317–1321). ACM. https://doi.org/10.1145/3803437.3805579

[25]. Xia, Y., Wang, T., Xu, W., & Zhang, S. (2025). *DAO-Agent: Zero knowledge-verified incentives for decentralized multi-agent coordination* (arXiv:2512.20973) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2512.20973

[26]. Park, D. (2024, November 29). Human player outwits Freysa AI agent in $47,000 crypto challenge. *The Block*. https://www.theblock.co/post/328747/human-player-outwits-freysa-ai-agent-in-47000-crypto-challenge

[27]. AI Incident Database. (n.d.). *Incident 1003: Alleged fraudulent prompts via AIXBT dashboard led purported AI trading agent to transfer 55.5 ETH from Simulacrum wallet*. Retrieved October 21, 2026, from https://incidentdatabase.ai/cite/1003

[28]. AI Incident Database. (n.d.). *Incident 1556: Morse code prompt allegedly used to induce Grok-linked trading bot transfer*. Retrieved October 21, 2026, from https://incidentdatabase.ai/cite/1556/

[29]. Patlan, A. S., Sheng, P., Hebbar, S. A., Mittal, P., & Viswanath, P. (2025). *Real AI agents with fake memories: Fatal context manipulation attacks on Web3 agents* (arXiv:2503.16248) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.16248

[30]. OpenAI. (2026, October 15). *Using a funded test cryptocurrency wallet without authorization* [Misalignment report]. OpenAI Alignment. https://alignment.openai.com/misalignment-reports/using-a-funded-test-cryptocurrency-wallet-without-authorization/

[31]. Xiong, X., Li, Z., Wei, W., Wang, Q., Knottenbelt, W., & Wang, Z. (2026). *Can trustless agents be trusted? An empirical study of the ERC-8004 decentralized AI agent ecosystem* (arXiv:2606.26028) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2606.26028

[32]. Sky Mavis. (2022). Back to building: Ronin security breach postmortem. *Ronin Blog*. https://roninchain.com/blog/posts/back-to-building-ronin-security-breach-6513cc78a5edc1001b03c364

[33]. Federal Bureau of Investigation. (2025, February 26). *North Korea responsible for $1.5 billion Bybit hack* (Public Service Announcement I-022625-PSA). Internet Crime Complaint Center. https://www.ic3.gov/PSA/2025/PSA250226

[34]. Commodity Futures Trading Commission. (2023, January 9). *CFTC charges Avraham Eisenberg with manipulative and deceptive scheme to misappropriate over $110 million from Mango Markets* (Release No. 8647-23). https://www.cftc.gov/PressRoom/PressReleases/8647-23

[35]. Zhou, L., Xiong, X., Ernstberger, J., Chaliasos, S., Wang, Z., Wang, Y., Qin, K., Wattenhofer, R., Song, D., & Gervais, A. (2023). SoK: Decentralized finance (DeFi) attacks. In *2023 IEEE Symposium on Security and Privacy (SP)* (pp. 2444–2461). IEEE. https://doi.org/10.1109/SP46215.2023.10179435

[36]. Sun, H., Li, J., & Zhang, H. (2024). zkLLM: Zero knowledge proofs for large language models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security* (pp. 4405–4419). ACM. https://doi.org/10.1145/3658644.3670334

[37]. Luck, C., Franzese-McLaughlin, O., Masserova, E., Takahashi, A., Polychroniadou, A., & Papernot, N. (2026). *Certified in theory, broken in practice: Assumption gaps in cryptographic model certification* (arXiv:2607.21839) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2607.21839

[38]. Conway, K. D., So, C., Yu, X., & Wong, K. (2024). *opML: Optimistic machine learning on blockchain* (arXiv:2401.17555) [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2401.17555

[39]. Chuang, J., Seto, A., Berrios, N., van Schaik, S., Garman, C., & Genkin, D. (2026). TEE.fail: Breaking trusted execution environments via DDR5 memory bus interposition. In *Proceedings of the 47th IEEE Symposium on Security and Privacy*. IEEE. https://tee.fail

[40]. Makriyannis, N., Yomtov, O., & Galansky, A. (2024). Practical key-extraction attacks in leading MPC wallets. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*. ACM. https://doi.org/10.1145/3658644.3670359

Yerbolat Kuanyshev
Senior Lecturer, School of Management, Law and Business
yerbolat.kuanyshev@sdu.edu.kz
SDU University, Kaskelen, Kazakhstan

CALIBRATED DIGITAL TWINS AND GEOMETRY-BASED PERCEPTION FOR AUTOMATED DRIVING IN HEAVY RAIN AND SNOWFALL: EVIDENCE FROM RESEARCH TESTING IN CHINA AND IMPLICATIONS FOR KAZAKHSTAN.

Abstract: Automated driving systems that perform reliably under clear conditions exhibit substantial degradation in heavy rain and snowfall. Precipitation attenuates and scatters LiDAR returns and generates spurious near-range points, water and spray impair camera imagery, whereas 4D radar has proved more robust than LiDAR in fog, rain and snow. Validation through public-road driving alone is impractical for these conditions, because severe weather events are infrequent, seasonal, hazardous for test personnel and not reproducible. This study examines whether a calibrated digital twin, understood as a virtual representation of the vehicle, its sensor suite and the meteorological environment that is iteratively corrected against physical measurements, constitutes a necessary component of severe-weather validation. The analysis is based on a structured narrative review of experimental testing conducted in China between 2009 and 2026, including the Intelligent Vehicle Future Challenge, virtual–real parallel testing, closed test fields that simulate rain, fog and other weather in Changsha, Wuhan and Chongqing, the cold-region proving grounds of Heihe, and recent studies that calibrate simulated rainfall and snowfall noise against field measurements. The results indicate that each testing modality addresses only part of the validation problem. Closed-field rainfall systems are repeatable but spatially non-uniform, so that nominal rainfall intensity overstates test realism; cold-region proving grounds provide natural snow and ice but are available only from November to March; and simulated rainfall is scalable and fully controllable but becomes credible only when parameterised with measured drop-size distributions and sensor responses. A digital twin that integrates these modalities converts scarce physical tests into calibration points and enables year-round exploration of the scenario space. The study further specifies a geometry-first machine-learning layer within the twin, supported by camera and LiDAR models adapted to heavy rain and snow, that infers the drivable corridor from road-surface elevation, road edges, snow banks, the tracks of preceding vehicles and three-dimensional occupancy rather than from painted markings or road

signs, and that triggers a minimal-risk manoeuvre when geometric confidence is insufficient or a road is officially closed. The framework is applied to Kazakhstan using data from RSE Kazhydromet, which recorded 190 hazardous meteorological events in 2024, including 28 severe blizzards with visibility as low as 50 m, and which reports road closures among the consequences of strong winds. The study proposes a weather transfer gap, $G = |P_{real} - P_{twin}| /$

P_{real}, as an acceptance criterion for substituting virtual for physical evidence, and specifies the data that test facilities and regulators would need to report for its application.

Keywords: *automated driving; adverse weather; heavy rainfall; snowfall; digital twin; parallel testing; closed-field testing; LiDAR; sensor simulation; scenario-based validation; 3D occupancy; road closures; Kazakhstan*

Калиброванные Цифровые Двойники для Валидации Автоматизированного Вождения в Условиях Сильного Дождя и Снегопада: Результаты Исследовательских Испытаний в Китае и Выводы для Казахстана

Аннотация: Системы автоматизированного вождения, надежно функционирующие в благоприятных погодных условиях, демонстрируют существенное снижение эффективности при сильном дожде и снегопаде. Осадки ослабляют и рассеивают сигналы лидара, порождая ложные отражения вблизи датчика; вода и брызги ухудшают качество изображения камер, тогда как 4D-радар показал более высокую устойчивость к туману, дождю и снегу, чем лидар. Валидация исключительно на дорогах общего пользования в таких условиях практически неосуществима, поскольку опасные погодные явления редки, сезонны, связаны с риском для испытателей и невоспроизводимы.

В исследовании рассматривается, является ли калиброванный цифровой двойник (виртуальное представление транспортного средства, его сенсорного комплекса и метеорологической среды, итеративно корректируемое по результатам физических измерений) необходимым элементом валидации в сложных погодных условиях. Анализ основан на структурированном нарративном обзоре экспериментальных испытаний, проведенных в Китае в 2009–2026 гг., включая конкурс Intelligent Vehicle Future Challenge, параллельное виртуально-реальное тестирование, закрытые полигоны с имитацией дождя, тумана и других погодных условий в Чанше, Ухане и Чунцине, полигоны холодного климата в Хэйхэ, а также недавние исследования, в которых имитированный дождь и снежный шум

калибруются по натурным измерениям.

Результаты показывают, что каждый вид испытаний охватывает лишь часть задачи валидации. Системы искусственного дождя на закрытых полигонах обеспечивают воспроизводимость, однако пространственно неоднородны, вследствие чего номинальная интенсивность осадков завышает реалистичность испытаний; полигоны холодного климата обеспечивают естественные снег и лед, но доступны лишь с ноября по март; имитационный дождь масштабируем и полностью управляем, но приобретает достоверность только при параметризации по измеренным распределениям размеров капель и откликам датчиков. Цифровой двойник, интегрирующий эти виды испытаний, превращает ограниченные физические испытания в точки калибровки и позволяет исследовать пространство сценариев круглогодично. Дополнительно предложен геометрически ориентированный слой машинного обучения в составе двойника, который опирается на модели камер и лидаров, адаптированные к сильному дождю и снегу, определяет полосу движения по рельефу дорожного покрытия, кромкам дороги, снежным валам, колеям впереди идущих транспортных средств и трехмерной занятости пространства, а не по дорожной разметке и знакам, и инициирует маневр минимального риска при недостаточной уверенности геометрической оценки или официальном закрытии дороги. Подход применен к условиям Казахстана на основе данных РГП

«Казгидромет», зарегистрировавшего в 2024 г. 190 стихийных метеорологических явлений, в том числе 28 случаев сильной метели с видимостью до 50 м, и отметившего закрытие автодорог в числе последствий сильного ветра. Предложен критерий погодного разрыва переноса $G = |P_{real} - P_{twin}| / P_{real}$ для обоснования замены физических данных виртуальными, а также определен перечень данных, которые испытательные центры и регуляторы должны раскрывать для его применения.

Ключевые слова: *автоматизированное вождение; неблагоприятные погодные условия; сильный дождь; снегопад; цифровой двойник; параллельное тестирование; испытания на закрытом полигоне; лидар; моделирование датчиков; сценарная валидация; трехмерная занятость пространства; закрытие дорог; Казахстан.*

Нөсер Жаңбыр мен Қар Жауу Жағдайында Автоматтандырылған Жүргізуді Валидациялауға Арналған Калибрленген Цифрлық Егіздер: Қытайдағы Зерттеу Сынақтарының Нәтижелері және Қазақстан Үшін Тұжырымдар

Аңдатпа: Қолайлы ауа райында сенімді жұмыс істейтін автоматтандырылған жүргізу жүйелерінің тиімділігі нөсер жаңбыр мен қар жауған кезде айтарлықтай төмендейді. Жауын-шашын лидар сигналдарын әлсіретіп, шашыратады және датчикке жақын жерде жалған шағылыстар тудырады; су мен шашырандылар камера кескінінің сапасын нашарлатады, ал 4D радар тұман, жаңбыр және қар жағдайында лидарға қарағанда тұрақтырақ екенін көрсетті. Мұндай жағдайларда тек жалпы пайдаланымдағы жолдарда валидациялау іс жүзінде мүмкін емес, өйткені қауіпті ауа райы құбылыстары сирек, маусымдық, сынақшылар үшін қатерлі және қайталанбайды.

Зерттеуде калибрленген цифрлық егіз (физикалық өлшемдер бойынша итеративті түрде түзетілетін көліктің, оның сенсорлық кешенінің және метеорологиялық ортаның виртуалды көрінісі) күрделі ауа райы жағдайындағы валидацияның қажетті құрамдас бөлігі болып табыла ма, сол мәселе қарастырылады. Талдау 2009–2026 жылдары Қытайда жүргізілген эксперименттік сынақтарға құрылымдалған нарративтік шолуға негізделген: Intelligent Vehicle Future Challenge жарысы, виртуалды-нақты параллель тестілеу, Чанша, Ухань және Чунцин қалаларындағы жаңбырды, тұманды және басқа да ауа райы жағдайларын имитациялайтын жабық полигондар, Хэйхэдегі суық климат полигондары, сондай-ақ имитацияланған жаңбыр мен қар шуын далалық өлшемдер бойынша калибрлейтін соңғы зерттеулер.

Нәтижелер сынақтың әр түрі валидация міндетінің тек бір бөлігін ғана шешетінін көрсетеді. Жабық полигондардағы жасанды жаңбыр жүйелері қайталанатын болғанымен, кеңістік бойынша біркелкі емес, сондықтан номиналды жаңбыр қарқындылығы сынақтың шынайылығын асыра бағалайды; суық климат полигондары табиғи қар мен мұзды қамтамасыз етеді, бірақ тек қарашадан наурызға дейін қолжетімді; ал имитацияланған жаңбыр масштабталатын әрі толық басқарылатын болғанымен, тамшы өлшемдерінің өлшенген таралуы мен датчиктердің жауаптары бойынша параметрленгенде ғана сенімді болады. Осы сынақ түрлерін біріктіретін цифрлық егіз шектеулі физикалық сынақтарды калибрлеу нүктелеріне айналдырып, сценарийлер кеңістігін жыл бойы зерттеуге мүмкіндік береді. Сонымен қатар егіз құрамында жол таңбалары мен белгілеріне емес, жол бетінің бедеріне, жол жиектеріне, қар үйінділеріне, алдағы көліктердің

ізіне және кеңістіктің үшөлшемді толтырылуына сүйене отырып жүру дәлізін анықтайтын геометрияға негізделген және нөсер жаңбыр мен қарға бейімделген камера мен лидар модельдеріне сүйенетін машиналық оқыту қабаты ұсынылады; геометриялық бағалаудың сенімділігі жеткіліксіз болғанда немесе

жол ресми түрде жабылғанда бұл қабат ең аз тәуекел маневрін іске қосады. Тәсіл Қазақстан жағдайына «Қазгидромет» РМК деректері (2024 жылы 190 дүлей метеорологиялық құбылыс, соның ішінде көрінуі 50 м-ге дейін төмендеген 28 қатты боран) негізінде қолданылады; бұл деректерде автожолдардың жабылуы қатты желдің салдары ретінде көрсетілген. Мақалада физикалық деректерді виртуалды деректермен алмастыруды негіздеу үшін ауа райы бойынша көшіру алшақтығы критерийі $G = |P_{real} - P_{twin}| / P_{real}$ ұсынылады және оны қолдану үшін сынақ орталықтары мен реттеушілер жариялауы тиіс деректер тізбесі анықталады.

Түйін сөздер: *автоматтандырылған жүргізу; қолайсыз ауа райы; нөсер жаңбыр; қар жауу; цифрлық егіз; параллель тестілеу; жабық полигондағы сынақ; лидар; датчиктерді модельдеу; сценарийлік валидация; кеңістіктің үшөлшемді толтырылуы; жолдарды жабу; Қазақстан.*

1. Introduction (Literature Review)

Adverse weather is a recurrent condition of road transport rather than an exceptional one. In the United States, approximately 12% of the roughly six million crashes recorded annually, close to 745,000, occur under adverse atmospheric conditions, resulting in more than 3,800 fatalities per year; more than 77% of these crashes take place during rain or mist [1]. For automated vehicles, precipitation affects not only traction and visibility but also the sensing chain itself. Raindrops and snowflakes attenuate and scatter LiDAR pulses and generate spurious near-range returns, water films and spray degrade camera images, and accumulated snow alters both the road surface and its visual appearance [2]. Radar is comparatively resilient: using the K-Radar dataset, which includes fog, rain and snow, Paek, Kong and Wijaya found 4D radar to be a more robust sensor than LiDAR in adverse weather [3].

The validation of automated driving under such conditions poses a distinct methodological problem. Because safety-critical events are rare, validating automated vehicles through naturalistic driving carries prohibitive economic and time costs; Feng et al. showed that an intelligent testing environment, in which AI-trained background agents interact with a real test vehicle through augmented reality, can accelerate evaluation by three to five orders of magnitude [4]. Heavy rainfall and snowfall represent only a small fraction of real-world driving exposure. Such events cannot be scheduled, they increase the risk to safety drivers, and an identical weather event cannot be reproduced for repeated testing. Evidence obtained from open-road driving is therefore insufficient for establishing performance boundaries in severe weather. The problem is particularly acute in Kazakhstan,

where prolonged snow cover, recurrent blizzards and icing regularly lead to the closure of national roads (Section 3.6).

Literature review. Relevant research has developed along three largely separate lines. The first concerns the robustness of perception. Beyond the radar evidence noted above [3], the Boreas dataset provides more than 350 km of driving on a repeated route over a full year, including rain and falling snow, recorded with LiDAR, radar and camera [5], and snowfall noise has been quantified directly: Cao, Huang and Yu recorded LiDAR data during snowfall in Beijing and compared denoising filters [6], whereas Sun, Zhao and colleagues examined LiDAR performance under snowfall and developed an adaptive filter for snow-induced noise [7], [8]. The second line concerns weather simulation. A recent review of open-source driving simulators notes growing concern about the validity of algorithms developed and evaluated in simulators [9], and a comprehensive survey notes that many driving simulators now provide adjustable rain, snow and fog [2]. Physics-based approaches extend this capability by introducing simulated snowfall into real LiDAR scans [10] or by rendering physically based rainfall within reconstructed three-dimensional driving scenes [11], [12]. The third line concerns digital twins and parallel intelligence. Hu et al. surveyed how simulation supports automated driving through sim-to-real transfer, digital twins and parallel intelligence [13], and Ai et al. developed PMWorld, a parallel testing platform for autonomous driving in mines [14].

Research gap. Although each of these lines is well established, the connections between them remain limited. Studies of weather simulation rarely report how simulated rainfall or snowfall was validated against a physical reference. Descriptions of closed-field tests frequently characterise rainfall solely by its nominal intensity, whereas recent work in China demonstrates that this single value can conceal considerable spatial variation within the rain field [15]. Digital-twin research in automated driving has concentrated predominantly on traffic interaction rather than on the physics of weather–sensor interaction. In addition, most perception stacks continue to depend on painted lane markings and road signs, which snow and ice routinely obscure, and the question of what a vehicle should rely on once these cues disappear has received less attention than the question of how to restore them. Evidence from research testing in China, spanning competitions, virtual–real testing, weather laboratories and cold-region proving grounds, is dispersed across journal articles, regulatory documents, institutional reports and preprints and has not previously been synthesised as a single body of evidence.

Research questions. The principal research question is whether severe-weather validation of automated driving requires a calibrated digital twin, and what research testing conducted in China reveals about the design of such a twin. Four subsidiary questions are addressed: (RQ1) what each testing modality employed in China contributes to validation under heavy rain and snowfall, and what its limitations are; (RQ2) against which quantities a weather digital twin must be calibrated for its results to transfer to real roads; (RQ3) by what criterion engineers and regulators can determine that a twin-based result may substitute

for a physical test; and (RQ4) what additional machine-learning layer within the twin can support driving when road markings and signs are unavailable, and how such a layer should interact with weather-related road closures in Kazakhstan.

Scientific novelty. The novelty of the study lies in treating the testing ecosystem in China as an integrated evidence base and analysing it from the perspective of digital-twin calibration, rather than reviewing perception algorithms or simulation platforms in isolation. The study additionally proposes an acceptance criterion, the weather transfer gap $G = |P_{real} - P_{twin}| / P_{real}$, which links twin-based results to their physical calibration points, and a geometry-first perception layer that derives the drivable corridor from the three-dimensional structure of the scene rather than from markings or signs. The framework is applied to the climatic and operational conditions of Kazakhstan.

Contributions. The principal contributions of the study are as follows:

- a structured review of research testing conducted in China that is relevant to heavy rain and snowfall, covering competitions, virtual–real parallel testing, closed test fields and cold-region proving grounds;
- a comparative assessment of what public-road, closed-field and virtual testing can and cannot establish under severe weather;
- a twin-calibrated testing loop in which scarce physical tests serve as calibration points for year-round virtual testing;
- the weather transfer gap metric G, together with a specification of the data that test facilities and regulators would need to report for its application;
- a geometry-first machine-learning layer that infers the drivable corridor from road-surface elevation, road edges, snow banks, vehicle tracks and three-dimensional occupancy, and that triggers a minimal-risk manoeuvre when confidence is insufficient or a road is closed;
- an application of the framework to Kazakhstan based on data from the national hydrometeorological service and the national road operator.

The remainder of the paper is organised as follows. Section 2 describes the review design, the analytical framework and the geometry-first perception layer. Section 3 presents the evidence from research testing conducted in China, the conditions prevailing in Kazakhstan and the evidence on marking-independent perception. Section 4 discusses the findings in relation to the research questions, and Section 5 concludes the paper.

2. Materials and Methods

The study adopts a structured narrative review design focused on experimental testing of automated driving under severe weather conducted in China between 2009 and 2026; international studies are used for context and comparison. The evidence base comprises peer-reviewed journal articles and conference papers, preprints (identified as such in the reference list), official regulatory documents of the People's Republic of China, and institutional or state-media reports describing test facilities. For the application to Kazakhstan, the study additionally draws on the 2024 climate review of RSE Kazhydromet,

which is based on observations from the national meteorological network, on statements by the national road operator KazAvtoZhol and the Ministry of Transport, and on official road-safety statistics reported by the Ministry of Internal Affairs. The geometry-first perception layer (Section 2.4) is informed by published studies of perception on snow-covered roads and of three-dimensional occupancy prediction. Searches were conducted in English and Chinese,

combining terms for automated and intelligent connected vehicles (智能网□汽□) with rain, snow, fog, ice, closed-field testing, winter testing (冬季□□), simulation and digital twin.

Sources were included if they described a controlled test conducted outside open public traffic, or the regulatory framework governing such tests, and if they addressed rain, snow, fog or ice or the infrastructure used to reproduce these conditions. Each source was assigned to one of five testing modalities: competition-based field testing; virtual–real parallel testing; closed-field weather facilities; cold-region proving grounds; and calibrated virtual weather. For each modality, the review recorded the weather conditions covered, the repeatability and realism of those conditions, the period of availability and the quantities measured. Sources were weighted according to evidential strength. Peer-reviewed findings received the greatest weight, followed by preprints and official documents, whereas facility capabilities reported only by operators or the media were treated as claims rather than as measured results.

2.1 Scope of controlled testing

In this study, controlled testing refers to any test in which the scenario is deliberately configured away from open public traffic, whether on a closed test field, in a weather laboratory, in a simulator, on a competition course or in a combination of these settings. It is distinguished from public-road testing and demonstration, which in China is governed by national and local rules on intelligent connected vehicles. The national pilot programme on vehicle–road–cloud integration launched in January 2024 explicitly calls for building a scenario database and for strengthening simulation, closed-field and real-road testing and validation capabilities [16]; most weather-related controlled tests take place in the closed-field part of this system.

2.2 Twin-calibrated testing loop

The evidence is organised by means of the conceptual loop shown in Figure 1. A physical test conducted on a rainfall rig or at a cold-region site yields measurements of both the meteorological field and the sensor response. These measurements are used to calibrate a digital twin comprising a weather model and a sensor model. Once the twin reproduces the physical results within an acceptable tolerance, it is used to explore a considerably larger scenario space than the physical facility can accommodate. Scenarios in which perception or planning fails within the twin are then returned to the physical facility for confirmation, and the cycle is repeated.

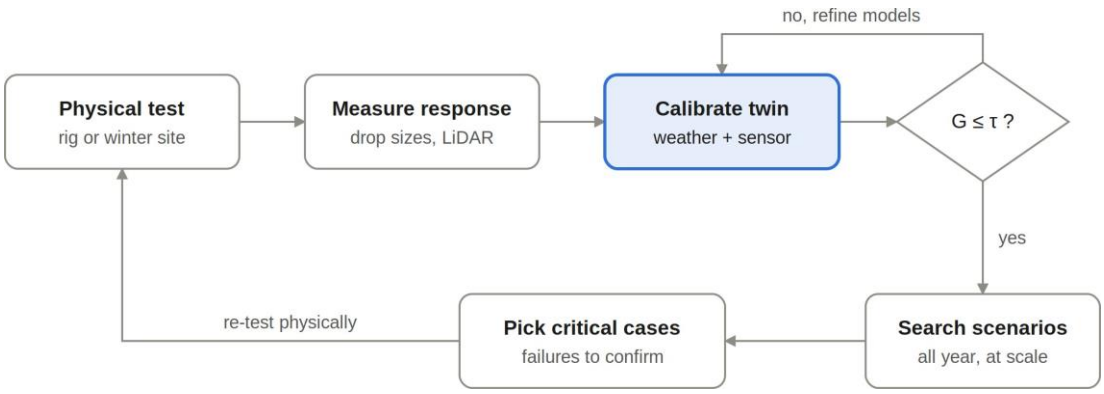


Figure 1. Twin-calibrated testing loop

Within this arrangement, each physical test becomes a calibration point, and each virtual exploration identifies the physical test that is most informative to conduct next.

2.3 Weather transfer gap

To determine whether the twin is sufficiently accurate, the study proposes the weather transfer gap G, defined for any selected performance measure P under matched weather conditions:

$$G = |P\text{real} - P\text{twin}|$$

$$P\text{real}$$

The measure P may be defined at sensor level, for example as the number of LiDAR returns on a target or its mean reflectivity, or at task level, for example as detection recall at a fixed range or the minimum time-to-collision in a cut-in manoeuvre. Matched weather denotes an equivalent condition in both domains, such as an identical path-equivalent rainfall intensity rather than an identical nominal setting of the rainfall system [15]. A twin-based result is accepted for a given measure when $G \leq \tau$, where the threshold τ is derived from the safety case for that measure. The study does not propose a numerical value for τ, which is properly a matter for testing standards.

2.4 Geometry-first perception layer

In addition to the weather and sensor models, the study specifies a machine-learning layer that operates within the twin and is intended for conditions in which painted lane markings and road signs cannot be relied upon. The layer does not attempt to recover obscured markings or read signs. Instead, it infers the drivable corridor directly from the three-dimensional geometry of the scene: the elevation profile of the road surface, the edges formed by kerbs, barriers and snow banks, the tracks left by preceding vehicles, and the occupancy of the surrounding space by static and moving objects. Each cue is estimated together with its uncertainty, and the fused estimate is passed to a decision gate. When the confidence of the geometric corridor falls below a threshold, or when an external notice indicates that the road has been closed, the gate instructs the vehicle to leave its operational

design domain and to execute a minimal-risk manoeuvre. Figure 2 summarises this architecture.

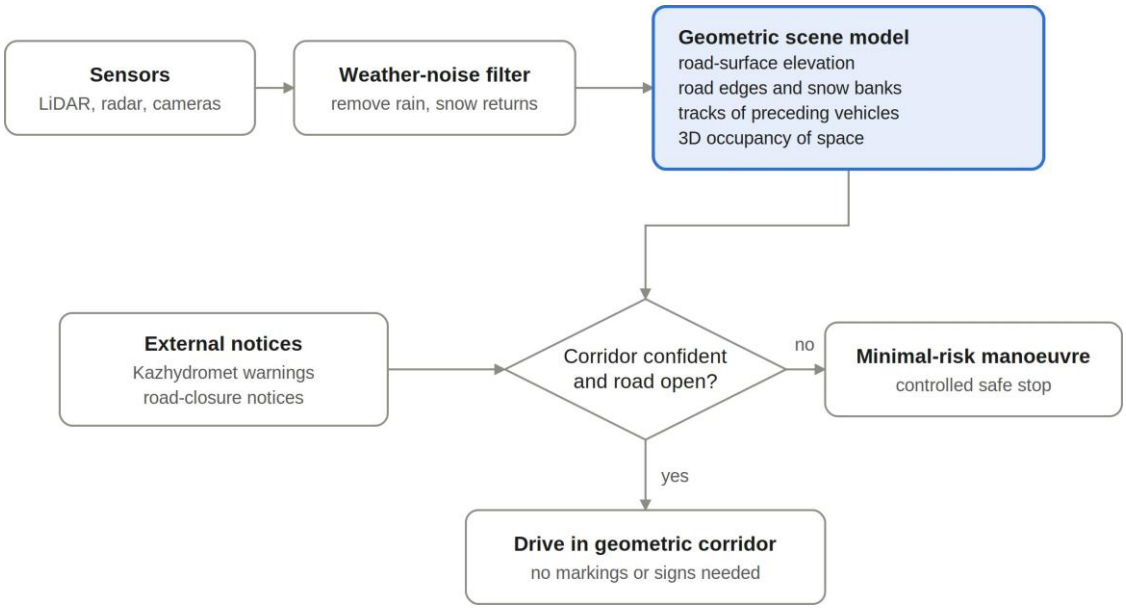


Figure 2. Geometry-first perception layer and closure-aware decision gate

Within the twin, the layer is trained and evaluated on scenes from which markings and signs have been deliberately removed or covered, so that its performance does not depend on cues that winter conditions routinely erase. The camera and LiDAR inputs to the layer are themselves adapted to heavy rain and snow: the perception models are trained and tested on images and point clouds rendered under calibrated rainfall and snowfall within the twin, and sensor fusion reduces reliance on the sensors most affected by the prevailing precipitation. Its evaluation uses the same weather transfer gap G, with P defined, for example, as the lateral error of the estimated corridor or as the proportion of correct decisions to leave the operational design domain.

3. Results

The scale of public-road testing in China provides a point of reference. In July 2026, the Ministry of Industry and Information Technology reported that more than 20,000 test and demonstration licences had been issued nationwide, more than 57,000 km of roads opened for testing and demonstration, and more than 220 million km of testing accumulated [17]. Although substantial, this exposure accrues mainly in ordinary weather, and because safety-critical events are rare, naturalistic driving alone remains an inefficient route to validation [4]; only a small proportion of these kilometres can be assumed to have been driven in heavy rain or snow. Controlled testing is the principal means by which developers and test facilities in China address this shortfall. Table 1 summarises the contribution of each testing modality.

Table 1. Testing modalities in China relevant to heavy rain and snowfall.

Testing modality	Example in China	Weather conditions	Repeatable	Availability	Measured quantities	Refs.
Competition-based field testing	Intelligent Vehicle Future Challenge, held since 2009, recently in Changshu	Incidental; whatever occurs on the days of the event	No	Annual event	Team performance on defined driving tasks	[18]–[20]
Testing modality	**Example in China**	**Weather conditions**	**Repeatable**	**Availability**	**Measured quantities**	**Refs.**
Virtual–real parallel testing	Parallel testing platforms such as PMWorld	Conditions specified by the tester in the virtual component	Yes	Year-round	Performance in coupled virtual and physical tests	[13], [14]
Closed-field weather facilities	Changsha test zone; Wuhan Economic and Technological Development Zone platform; Chongqing laboratory	Rain, fog and lighting; wind and lightning reported in Wuhan	Yes, with a spatially non-uniform rain field	Independent of natural weather	Sensor and function response in staged scenarios	[15], [21]–[23]
Cold-region proving grounds	Heihe, Heilongjiang Province	Natural snow and ice, frequently below −20 °C	No	November to March	Vehicle dynamics, braking, battery performance	[24], [25]
Calibrated virtual weather	GSRAIN; RainyGS	Rainfall of controlled intensity, 0–13 mm/h in GSRAIN	Yes	Year-round	Detection confidence, closed-loop collisions	[11], [12]

Result 1 (Table 1) corresponds to Contribution 2: no single modality is simultaneously repeatable, realistic and available throughout the year, and the two fully repeatable modalities, parallel testing and calibrated virtual weather, are both virtual.

3.1 Competition-based field testing

The Intelligent Vehicle Future Challenge was launched in 2009 by the National Natural Science Foundation of China [18]. The 13th edition, held in Changshu in 2023,

focused on driverless passenger services [18]; the 14th was held in Changshu on 22–23 November 2024 under the auspices of the Foundation's Department of Information Sciences and the Chinese Association of Automation [19]; and the 15th, held in December 2025, brought together 14 university teams and focused on multi-agent embodied interaction [20]. The competition format enables the comparison of teams under common rules. With respect to weather, however, it is opportunistic: participants encounter whatever precipitation occurs on the days of the event, at an uncontrolled intensity and without the possibility of repetition.

3.2 Virtual–real parallel testing

The coupling of virtual and physical testing has been formalised through the concept of parallel intelligence, which Hu et al. review alongside sim-to-real transfer and digital twins [13]. Ai et al. subsequently built PMWorld, a parallel testing platform for autonomous driving in mines [14]. Functionally, such platforms constitute digital twins employed for testing. As their scope indicates, however, the emphasis lies on the testing architecture rather than on weather-specific sensor physics; this component is supplied by the facilities and studies described in Sections 3.3–3.5.

3.3 Closed-field weather facilities

Several test zones in China now incorporate weather simulation into their infrastructure. The national test zone in Changsha covers more than 80 hectares, includes a 3.6 km two-way expressway and in 2023 introduced a testing system based on hazardous scenarios that include rain, fog, dust and lighting conditions [21]. A platform in the Wuhan Economic and Technological Development Zone contains an extreme-environment area that, according to its operators, can simulate wind, rain and lightning and switch from clear conditions to a downpour within seconds [22]. In early 2026, a laboratory of more than 5,000 m² began operating in Chongqing, combining vehicle motion control, rain–fog–light simulation and traffic-flow simulation within a single facility [23]. These capabilities are reported by operators and state media and have not been independently measured.

The fidelity of artificial rainfall is therefore the critical question, and a 2026 study by researchers from Tongji and Tsinghua universities provides quantitative evidence on this point [15]. Xia et al. used approximately 10,000 real-rainfall raindrop-spectrum samples as a reference and sampled 45 points within a 2.4 m × 7.2 m simulated-rainfall area. The rain field remained spatially non-uniform under a single nominal setting, indicating that one intensity value cannot characterise the entire test area. The proposed method identified two preferred test paths with equivalent intensities of 11.54 ± 0.31 mm/h and 8.28 ± 0.34 mm/h [15]. It follows that an identical rig setting may expose different vehicles, or different paths of the same vehicle, to materially different rainfall.

3.4 Cold-region proving grounds and measured snowfall

Because snowfall is more difficult to reproduce artificially than rainfall, winter testing in China continues to rely on geography. Heihe, located on the border with Russia in Heilongjiang Province, experiences snow cover for approximately 140 days a year, and its testing season extends from November to March [24]. During the 2024–25 season it hosted

147 companies and research institutions and more than 4,000 test vehicles, and local reports attribute to it approximately 85% of the Chinese cold-weather testing market [25]. Test tracks are laid out on the frozen Heilongjiang River [24], [25]. Although these conditions are highly realistic, the testing window is limited to about five months a year, conditions cannot be controlled, and the reported focus is on vehicle dynamics, braking and battery performance rather than on perception.

Measured snowfall data illustrate why perception requires dedicated evidence. Cao, Huang and Yu recorded 3,000 LiDAR frames during snowfall in Beijing using a 40-line sensor operating at 10 Hz [6]. Table 2 reports their principal results alongside other recent quantitative findings.

Table 2. Quantitative findings from recent rainfall and snowfall studies.

Study	Setting	Principal reported result
Cao, Huang, Yu, 2022 [6]	Real snowfall, Beijing; 40-line LiDAR at 10 Hz	DSOR filter precision 74.63% vs 48.22% for DROR; mean processing time 130.11 ms vs 706.97 ms per frame
Zhao, Sun, Sun, 2026 [8]	Public snowfall datasets CADC, Boreas and WADS	F1-score above 90 on all three datasets; detection accuracy improved by 5.9, 9.7 and 6.4 percentage points for three detectors
Xia, Zhao, Ye, Chen, 2026 [15]	Closed-field simulated rainfall; 45 points in a 2.4 m × 7.2 m area	Non-uniform field under one nominal setting; preferred paths at 11.54 ± 0.31 and 8.28 ± 0.34 mm/h
Wang, Zhang, Chen, 2026 [11]	Virtual rainfall of 0–13 mm/h in reconstructed 3D scenes	Collisions in a cut-in scenario at 8.5 and 12.0 mm/h; truck detection confidence reduced from 0.63 to 0.41

The timing results from Beijing are particularly relevant. At 10 Hz, a new LiDAR frame arrives every 100 ms, and both filters exceeded this interval on a desktop processor [6]. Snowfall therefore poses a real-time engineering problem in addition to an accuracy problem. The filter proposed by Zhao et al. was evaluated on Canadian and US winter datasets [8], which illustrates how few openly available snowfall datasets exist.

3.5 Calibrated virtual weather

Recent research transfers weather conditions into the virtual domain while retaining a link to measurement. GSRAIN derives its raindrop model from measured rainfall and introduces rainfall of controllable intensity, from 0 to 13 mm/h, into 3D Gaussian Splatting reconstructions of real driving scenes [11]. The generated images achieved a Fréchet Inception Distance of 149.09, lower than that of CycleGAN-Turbo (155.71) and WeatherEdit (157.94). Of greater relevance to safety, closed-loop simulations produced collisions in a cut-in scenario at 8.5 and 12.0 mm/h, and the confidence of one detector for a nearby truck decreased from 0.63 to 0.41, with the truck misclassified as a car at 8.5 mm/h [11]. RainyGS, developed at Peking University, follows a related approach by adding

physically based raindrops, splashes and reflections to reconstructed scenes with adjustable intensity [12].

Result 2 (Tables 1 and 2) corresponds to Contributions 1 and 3: the rainfall intensities at which collisions occurred in the virtual study, 8.5 and 12.0 mm/h [11], fall within the range of equivalent intensities delivered by a calibrated closed-field rig, approximately 8.3–11.5 mm/h [15]. Because the two studies employed different vehicles and experimental configurations, this correspondence does not validate either result. It does, however, demonstrate the practical feasibility of the loop in Figure 1, since a twin can identify a failure at a rainfall intensity that an existing closed-field rainfall rig is able to reproduce and verify.

3.6 Severe weather and winter road operations in Kazakhstan

Kazakhstan combines a sharply continental climate with a long and sparsely populated road network, which makes the questions examined in this study directly relevant to its transport system. The national hydrometeorological service, RSE Kazhydromet, recorded 190 hazardous meteorological events in 2024, compared with 203 in 2023 [26]. Table 3 lists these events by type together with the long-term changes reported by Kazhydromet.

Table 3. Hazardous meteorological events in Kazakhstan in 2024 and long-term change, according to RSE Kazhydromet [26].

Phenomenon	**Cases in 2024**	**Change, 2008–2024 vs 1991–2007**
Strong wind (30 m/s or more)	115	Increase by a factor of 1.9 (reported jointly with heavy snowfall and hail)
Heavy rain	32	Increase by a factor of 1.5
Severe blizzard	28	Decrease of 8%
Heavy snowfall	8	Increase by a factor of 1.9 (reported jointly with strong wind and hail)
Hail	4	Increase by a factor of 1.9 (reported jointly with strong wind and heavy snowfall)
Wet-snow deposition	2	Not reported
Very dense fog	1	Decrease of 26%
Total	190 (203 in 2023)	—

Counts for 2024 are taken from the summary table of hazardous events in the source, whose rows sum to the reported total of 190.

The severe blizzards of 2024 were concentrated in the northern and western regions. Aktobe Region alone recorded 13 of them, lasting 13–28 hours with visibility of 50–500 m, and the event with the poorest visibility, at the Kos-Istek station on 7 February 2024, lasted 25 hours with wind speeds of up to 21 m/s [26]. Kazhydromet lists road closures among the

consequences of strong winds in 2024 [26]. The winter of 2023/24 was the wettest on record since 1941, with 144.5% of the national norm of precipitation, and a cold wave between 15 and 24 February 2024 brought temperatures of −35 to −45 °C to northern and eastern regions, including a station record of −44.5 °C at Kokpekty on 21 February [26]. Snow cover remains long-lasting in the north: at Petropavlovsk it persisted for 149 days in the 2023–24 season and reached a maximum depth of 53 cm [26].

Road operations respond to these conditions primarily through closure, although closure does not remove risk entirely. On 3 January 2025, after traffic on the Astana–Petropavlovsk road had been restricted because of snowfall, blizzard, strong wind, poor visibility and ice, a mass collision involving approximately 95 vehicles occurred near Zhanatalap while traffic was moving under escort; the national road operator KazAvtoZhol attributed the collision to drivers failing to observe the speed limit and a safe distance [27]. Winter maintenance capacity is a further constraint. In September 2023, the Ministry of Transport reported that 838 of the 1,050 units of equipment assigned to the winter maintenance of national roads, approximately 80%, were in working order, and that a tender was under way to rent a further 214 units [28]. These conditions add to an existing road-safety burden: in November 2024 the Committee of Administrative Police of the Ministry of Internal Affairs reported about 2,000 road deaths in Kazakhstan since the start of that year and noted that out-of-town highways accounted for half of all fatal crashes [29].

3.7 Evidence for marking-independent, geometry-based perception

Winter conditions directly undermine perception that relies on painted markings. Goberville, Kadav and Asher report that a leading automotive vision system detected both the left and the right lane lines in only 21.8% of winter driving, and that only 0.9% of these detections had sufficient confidence for safety-critical driver-assistance functions; as an alternative, they identify the drivable region from the tyre tracks left by preceding vehicles [30]. Ariram et al. constructed LiDAR maps of a Finnish urban road in January and showed that road width can be estimated from the snow banks at the road edges even when lane markings are covered, after which lanes can be inferred from the estimated width and the known number of lanes [31].

At the level of complete scene representation, researchers at Tsinghua University have shown that three-dimensional semantic occupancy, which describes which parts of the space surrounding the vehicle are occupied, can be predicted from surround-view cameras; their TPVFormer model achieved performance comparable to that of LiDAR-based methods on the nuScenes LiDAR segmentation task [32]. Such representations encode the drivable surface and obstacles geometrically rather than through markings or signs. Finally, ISO 34503:2023 specifies a taxonomy for defining the operational design domain, that is, the operating conditions within which an automated driving system is designed to function [33]. In these terms, an official road closure in Kazakhstan places the affected road explicitly outside the operational design domain.

Result 3 (Table 3 and Sections 3.6–3.7) corresponds to Contributions 5 and 6: winter conditions in Kazakhstan produce frequent and prolonged losses of visibility and of road-surface cues, closures are the principal operational response, and published methods already demonstrate that a drivable corridor can be inferred from scene geometry without reliance on markings or signs.

4. Discussion

The evidence supports the conclusion that severe-weather validation requires a digital twin. The principal reason is not that simulation is less costly, but that the twin is the only setting in which the realism of a cold-region proving ground, the controllability of a rainfall rig and the scale of virtual testing can be combined. The twin acquires this role, however, only when it is calibrated against the physical tests it is intended to extend.

RQ1 concerned the contribution and limitations of each testing modality employed in China. Competitions such as the Intelligent Vehicle Future Challenge provide common rules and public comparison, but the weather conditions they present are incidental [18]–[20]. Parallel testing supplies the methodological basis, namely a closed loop between virtual and physical tests [13], [14], but its published form addresses weather physics only marginally. Closed-field facilities offer control and repeatability [21]–[23], yet their rainfall is spatially non-uniform [15], and none of the reviewed facilities reports artificial snowfall. Cold-region proving grounds provide natural snow and ice for approximately five months a year, without control over conditions and with a reported emphasis on vehicle dynamics rather than on perception [24], [25]. Calibrated virtual weather offers scale and full controllability [11], [12] but derives its credibility entirely from the measurements on which it is based. Because each modality compensates for a limitation of another, the evidence favours their integration rather than the selection of any single modality.

RQ2 concerned the quantities against which a weather twin must be calibrated. The reviewed studies suggest three layers of calibration. The first is the weather field itself, including drop size, drop velocity and spatial variation across the test area, since a nominal intensity conceals this variation [15]. The second is the sensor response to that field, such as LiDAR returns and reflectivity on targets [15] and the spatial distribution of snowfall noise [6], [8]. The third is task-level performance, namely detection and closed-loop behaviour [11]. The distinction is important because image-quality measures such as the Fréchet Inception Distance quantify the visual appearance of rainfall rather than its effect on LiDAR returns. A twin that is accurate at the first two layers and verified at the third produces failures that merit physical confirmation.

RQ3 concerned the criterion under which a twin-based result may replace a physical one. The weather transfer gap G provides a workable criterion subject to two conditions: weather must be matched in equivalent terms rather than by rig settings [15], and the same measure P must be recorded in both domains. Neither condition is currently standard practice in the facility reports reviewed, which describe capabilities rather than measured fields [21]–[23].

RQ4 concerned the additional machine-learning layer that can support driving when markings and signs are unavailable, and its interaction with road closures in Kazakhstan. The evidence indicates that such a layer should be geometry-first. Marking-based lane detection fails in most winter driving [30], whereas road width can be recovered from snow banks [31], the drivable region from the tracks of preceding vehicles [30], and the full occupied space from three-dimensional occupancy prediction [32]. None of these cues depends on the visibility of paint or signs, and each can be trained and stress-tested within the twin by removing markings and signs from otherwise realistic scenes. Cameras remain valuable within this layer, but only if their models are trained and tested on imagery degraded by calibrated rain and snow: detection confidence falls measurably as simulated rainfall intensifies [11], while radar provides a comparatively robust complement in fog, rain and snow [3]. The same layer provides a natural point at which to encode road closures. Because a closure places the road outside the operational design domain [33], an automated vehicle should treat a closure notice from the road operator, or a hazardous-weather warning from Kazhydromet, as a condition for executing a minimal-risk manoeuvre regardless of its own perception confidence. The Zhanatalap collision [27] shows that the period immediately after reopening, when traffic moves in convoy on a surface that remains snow-covered, is itself hazardous; an automated vehicle should therefore resume normal operation only when both the closure has been lifted and its geometric corridor estimate has regained sufficient confidence.

The hypothesis that severe-weather validation requires a calibrated digital twin is therefore supported, with the qualification that the twin complements rather than replaces physical testing. Three developments may be anticipated if this approach is adopted more widely. First, closed-field reports are likely to move from nominal to measured, equivalent descriptions of eather, as Xia et al. already propose [15]. Second, the principal function of physical rigs is likely to shift from serving as the test itself to calibrating and confirming the twin. Third, cold-region sites are likely to combine seasonal campaigns with year-round work; in early 2025, a new facility in Heihe was expected to open by May of that year to enable testing throughout the year [24]. These expectations are inferred from the evidence and should not be interpreted as empirical findings.

Practical implications

Test facilities should report, for each weather condition, the measured drop-size spectrum, a spatial map of the rain field and the equivalent intensity along each test path. Developers should record identical sensor-level and task-level measures in physical and virtual runs so that G can be computed, and should evaluate perception on scenes in which markings and signs are absent. Camera-based perception, in particular, should be trained and validated on imagery degraded by calibrated rain and snow rather than on clear-weather data alone. Regulators that accept simulation-based evidence could require G to be reported alongside it.

For Kazakhstan, four implications follow. First, Kazhydromet warnings and KazAvtoZhol closure notices should be made available in a machine-readable form, so that automated vehicles and fleet operators can apply them as hard operational-design-domain boundaries rather than as advice to drivers. Second, the reopening of a closed road should be treated as a distinct operational phase with its own speed and spacing rules for automated vehicles, given the risk demonstrated by the Zhanatalap collision [27]. Third, winter maintenance vehicles, which already traverse the network during and after snowfall, could carry low-cost LiDAR or camera units to record the post-clearing geometry of the road, including the position of snow banks; these observations would keep the geometric layer of the twin current and would be particularly valuable where maintenance capacity is limited and markings remain covered for extended periods [28]. Fourth, the northern regions, where snow cover lasts for about five months, could host a national cold-weather calibration site whose measurements feed the twin, rather than relying on short seasonal campaigns abroad [26].

Limitations of the study

Several limitations should be acknowledged. First, the study is a review rather than an original experiment; neither the metric G nor the geometry-first layer has yet been tested empirically. Second, the evidence base combines peer-reviewed articles, preprints, regulatory documents and media reports, and facility capabilities drawn from the last category represent operator claims. Third, the correspondence between the collision intensities observed in simulation and the equivalent intensities of the physical rig derives from two independent studies and is illustrative only. Fourth, the cited rainfall rig delivered approximately 8–12 mm/h on its preferred paths [15], whereas natural downpours can be considerably more intense; the heaviest rainfall therefore remains insufficiently covered. Fifth, evidence on snowfall is weaker than evidence on rainfall, since none of the reviewed facilities in China reports controlled snowfall. Sixth, the Kazakhstan data cover a single climatological year, and the account of the Zhanatalap collision and the 2024 road-death figure are drawn from official statements reported in the media rather than from official statistical series. Finally, the searches covered English- and Chinese-language sources available online, and literature accessible only through subscription databases may have been omitted.

5. Conclusion

The central conclusion of this study is that automated driving in heavy rain and snow cannot be validated on public roads alone and should first be developed and tested in a calibrated digital environment. Physical tests in China provide either realism or control, rarely both, and never throughout the year; virtual tests provide scale and control but are credible only when calibrated against measurements. The digital twin is the only setting in which these strengths can be combined. Four specific conclusions follow from the evidence.

1. **Test in the digital world first.** Severe-weather scenarios should be generated, run and repeated in a digital twin that is calibrated against physical rainfall rigs and

winter proving grounds. Physical tests should be used to calibrate the twin and to confirm the failures it identifies, and a twin-based result should be accepted for a given measure only when the weather transfer gap G falls below a threshold set in the safety case.

2. **Build the drivable corridor from road geometry.** In heavy rain and snow, the vehicle should determine where it can drive from the three-dimensional geometry of the road: the elevation of the road surface, road edges and snow banks, the tracks of preceding vehicles and the occupancy of the surrounding space.
3. **Adapt cameras and other sensors to heavy rain and snow.** Camera and LiDAR models should be trained and validated on data degraded by calibrated rainfall and snowfall rather than on clear-weather data alone, and sensor fusion should reduce reliance on the sensors most affected by the prevailing weather.
4. **Disregard road signs and lane lines in severe weather.** Because snow and heavy rain routinely obscure painted markings and signs, the perception layer should not depend on them. When the geometric estimate is not sufficiently confident, or when the road has been officially closed, the vehicle should execute a minimal-risk manoeuvre.

For Kazakhstan, where Kazhydromet recorded 28 severe blizzards in 2024 with visibility falling to 50 m, and where road closure is the principal operational response to severe weather, these conclusions are of direct practical relevance. Closure notices and hazardous-weather warnings should be treated by automated vehicles as hard boundaries of the operational design domain.

The scientific novelty of the study lies in analysing research testing in China as an integrated body of evidence through the lens of digital-twin calibration, in proposing the weather transfer gap G as an acceptance criterion, and in specifying a geometry-first perception layer that operates without reliance on road markings or signs. Future research should evaluate G using paired physical and virtual runs at a single facility, extend weather-field calibration from rainfall to snowfall and road spray, and test the geometry-first layer on roads in northern Kazakhstan, including data recorded by winter maintenance vehicles after snow clearing.

Author Contributions

Conceptualization: Y.K.; Methodology: Y.K.; Investigation: Y.K.; Formal Analysis: Y.K.; Writing – Original Draft Preparation: Y.K.; Writing – Review and Editing: Y.K.; Visualization: Y.K.

Conflicts of Interest

The author declares no conflict of interest.

Data Availability Statement

No new datasets were generated during this study. The research is based on publicly available scientific literature, preprints, government documents and public reports cited in the manuscript.

List of Supplementary Materials

No supplementary materials are associated with this study.

References

[1] U.S. Federal Highway Administration, How Do Weather Events Affect Roads? Road Weather Management Program, U.S. Department of Transportation. (accessed 27 September 2026). https://ops.fhwa.dot.gov/weather/roadimpact.htm

[2] Y. Zhang, A. Carballo, H. Yang, & K. Takeda, Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. ISPRS Journal of Photogrammetry and Remote Sensing, 196, 146–177. (2023). https://doi.org/10.1016/j.isprsjprs.2022.12.021

[3] D.-H. Paek, S.-H. Kong, & K. T. Wijaya, K-Radar: 4D radar object detection for autonomous driving in various weather conditions. Advances in Neural Information Processing Systems (NeurIPS), 35, 3819–3829. (2022).

[4] S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, & H. X. Liu, Dense reinforcement learning for safety validation of autonomous vehicles. Nature, 615(7953), 620–627. (2023). https://doi.org/10.1038/s41586-023-05732-2

[5] K. Burnett, D. J. Yoon, Y. Wu, A. Z. Li, H. Zhang, S. Lu, J. Qian, W.-K. Tseng, A. Lambert, K. Y.

[6] K. Leung, A. P. Schoellig, & T. D. Barfoot, Boreas: A multi-season autonomous driving dataset. The International Journal of Robotics Research, 42(1–2), 33–42. (2023). https://doi.org/10.1177/02783649231160195

[7] Y. Cao, H. Huang, & D. Yu, Filter methods for removing falling snow from light detection and ranging point clouds in snowy weather. Sensors and Materials, 34(12), 4507–4519. (2022). https://doi.org/10.18494/SAM4047

[8] C. Sun, P. Sun, J. Wang, Y. Guo, & X. Zhao, Understanding LiDAR performance for autonomous vehicles under snowfall conditions. IEEE Transactions on Intelligent Transportation Systems, 1–11 (early access). (2024).

[9] X. Zhao, C. Sun, & P. Sun, A parametric-adaptive filter for improving LiDAR performance under snowfall conditions. IEEE Transactions on Intelligent Transportation Systems, 27(6), 6890–6905. (2026). https://doi.org/10.1109/TITS.2026.3665537

[10] Y. Li, W. Yuan, S. Zhang, W. Yan, Q. Shen, C. Wang, & M. Yang, Choose your simulator wisely: A review on open-source simulators for autonomous driving. IEEE Transactions on Intelligent Vehicles, 9(5), 4861–4876. (2024). https://doi.org/10.1109/TIV.2024.3374044

[11] M. Hahner, C. Sakaridis, M. Bijelic, F. Heide, F. Yu, D. Dai, & L. Van Gool, LiDAR snowfall simulation for robust 3D object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16364–16374. (2022).

[12] F. Wang, L. Zhang, & J. Chen, GSRAIN: Physically calibrated high-/low-frequency rainfall synthesis for 3D Gaussian driving scenes. arXiv preprint arXiv:2608.02177. (2026).

[13] Q. Dai, X. Ni, Q. Shen, W. Chen, B. Chen, & M. Chu, RainyGS: Efficient rain synthesis with physically-based Gaussian splatting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2025). arXiv:2503.21442

[14] X. Hu, S. Li, T. Huang, B. Tang, R. Huai, & L. Chen, How simulation helps autonomous driving: A survey of sim2real, digital twins, and parallel intelligence. IEEE Transactions on Intelligent Vehicles, 9(1), 593–612. (2024). https://doi.org/10.1109/TIV.2023.3312777

[15] Y. Ai, Y. Liu, Y. Gao, C. Zhao, X. Cheng, J. Han, B. Tian, L. Chen, & F.-Y. Wang, PMWorld: A parallel testing platform for autonomous driving in mines. IEEE Transactions on Intelligent Vehicles, 9(1), 1402–1411. (2024).

[16] T. Xia, X. Zhao, S. Ye, & J. Chen, From nominal intensity to equivalent rainfall: A path-based credibility evaluation framework for simulated rainfall in autonomous-driving perception tests. arXiv preprint arXiv:2606.11989. (2026).

[17] Ministry of Industry and Information Technology, Ministry of Public Security, Ministry of Natural Resources, Ministry of Housing and Urban-Rural Development, & Ministry of Transport of the People's Republic of China, Notice on carrying out pilot applications of vehicle–road–cloud integration for intelligent connected vehicles

[18] [关于开展智能网□汽□"□路云一体化"□用□点工作的通知]; official interpretation by the MIIT Equipment Industry Development Center. (17 January 2024). https://www.miit-

[19] eidc.org.cn/art/2024/1/17/art_1541_10314.html

[20] The Paper [澎湃新□], MIIT: accelerating the preparation of the 15th Five-Year Plan for the intelligent connected new-energy vehicle industry

[21] [工信部：加快□制出台智能网□新能源汽□□□□展“十五五”□划]. (27 July 2026). https://www.thepaper.cn/newsDetail_forward_33667768
[22] Xi’an Jiaotong University, XJTU’s unmanned intelligent vehicle wins championship. (27 November 2023). https://en.xjtu.edu.cn/2023-11/27/c_947841.htm
[23] China Daily, The 14th China Intelligent Vehicle Future Challenge held in Changshu, Jiangsu [第十四届“中国智能□未来挑□□”在江□常熟市成功□□]. (25 November 2024).
[24] https://cn.chinadaily.com.cn/a/202411/25/WS6743fa5da310b59111da543c.html

[25] Xinhua, 2025 China Intelligent Vehicle Future Challenge focuses on multi-agent embodied interaction [2025中国智能□未来挑□□聚焦多智能体具身交互]. (15 December 2025).
[26] https://www.news.cn/tech/20251215/d6892419b31344aba8237c880523cd9b/c.html
[27] Hunan Provincial People’s Government, Changsha among top three in national intelligent connected vehicle testing demonstration zone capability assessment. (December 2023). http://www.enghunan.gov.cn/hneng/News/Localnews/202312/t20231212_32476729.html
[28] China Daily, WEDZ unveils intelligent connected vehicle testing platform. (16 April 2025). https://govt.chinadaily.com.cn/s/202504/16/WS6803c981498eec7e1f734f9d/wedz-unveils-intelligent-connected-vehicle-testing-platform.html
[29] State Council Information Office of the People’s Republic of China, China builds new intelligent driving lab to boost smart car industry. (12 February 2026). http://english.scio.gov.cn/m/chinavoices/2026-02/12/content_118330044.html
[30] Xinhua, Across China: Chilly northern city emerges as vehicle-testing hub. (16 January 2025). https://www.jsmsj.org.cn/JiangsuArtists/English/Culture/2025/0116/10218.html
[31] H. Zhou, Heihe ices hub status for cold vehicle testing. China Daily. (26 January 2026). https://www.chinaservicesinfo.com/s/202601/26/WS699c1ca7498e36855033ab54/heihe-ices-hub-status-for-cold-vehicle-testing.html
[32] RSE Kazhydromet, Review of Climate Features in the Territory of Kazakhstan, 2024 [Обзор об особенностях климата на территории Казахстана, 2024]. Astana: RSE Kazhydromet. (2025). https://www.kazhydromet.kz/uploads/calendar/280/year_file_en/682ca255cbcd6obzor-osobennostey-klimata_kazahstan-za-2024_rus.pdf
[33] The Times of Central Asia, A dozen injured in 100-car accident in winter weather in Kazakhstan. (January 2025). https://timesca.com/a-dozen-injured-in-100-car-accident-in-winter-weather-in-kazakhstan/
[34] Ministry of Transport of the Republic of Kazakhstan, Readiness of Kazakhstan’s roads for the winter period discussed at the Ministry of Transport [in Russian]. (September 2023). https://www.gov.kz/memleket/entities/transport/press/news/details/623989?lang=ru
[35] Bizmedia.kz, Two thousand Kazakhstanis died in road accidents this year [Две тысячи казахстанцев погибли в ДТП в этом году], citing the Committee of Administrative Police, Ministry of Internal Affairs of the Republic of Kazakhstan. (14 November 2024). https://bizmedia.kz/2024-11-14-dve-tysyachi-kazahstanczev-pogibli-v-dtp-v-etom-godu/
[36] N. A. Goberville, P. Kadav, & Z. D. Asher, Tire track identification: A method for drivable region detection in conditions of snow-occluded lane lines. SAE Technical Paper 2022-01-0083. (2022). https://doi.org/10.4271/2022-01-0083
[37] S. Ariram, V. Pekkala, T. Mäenpää, A. Tikänmaki, & J. Röning, UAV-based intelligent information systems on winter road safety for autonomous vehicles. arXiv preprint arXiv:2406.12370. (2024). https://doi.org/10.48550/arXiv.2406.12370
[38] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, & J. Lu, Tri-perspective view for vision-based 3D semantic occupancy prediction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 9223–9232. (2023).
[39] International Organization for Standardization, ISO 34503:2023 Road vehicles — Test scenarios for automated driving systems — Specification for operational design domain. Geneva: ISO. (2023).

Kassen Aikunim,
Master's student,
Academy of Public Administration under the President of the Republic of Kazakhstan
Astana, Kazakhstan

PERSONAL DATA PROTECTION IN KAZAKHSTAN'S AI ENABLED PUBLIC SERVICES

Abstract: Public-sector artificial intelligence can influence access to services, inspections, eligibility, and other exercises of public authority, yet compliance cannot be demonstrated by citing legal safeguards or listing technical controls alone. This study develops the Kazakhstan AI Privacy and Accountability Governance Framework (KZ-AIPAG) to connect each legal requirement and public-service use case to a defined risk, control, evidence artefact, accountable actor, and lifecycle decision. The research combines normative legal analysis, a targeted structured literature review, comparative analysis of the Canadian Algorithmic Impact Assessment, the EU AI Act, General Data Protection Regulation Articles 22 and 35, and an end-to-end algorithmic-audit model, together with bounded applications to Smart Data Ukimet and eGov AI. The results show that Kazakhstan's principal gap is operational rather than a complete absence of regulation. KZ-AIPAG separates four use-case impact tiers from probability-by-impact hazard scoring, organises controls across six governance domains, and introduces an evidence chain, four decision gates, a unified role catalogue, and a RACI matrix that preserves audit independence. In the illustrative Smart Data Ukimet case, the framework yields a conditional no-go decision until the legal basis, data purpose, linkage and re-identification risk, access model, meaningful human review, and control-effectiveness evidence are documented. The framework is a proposed and analytically demonstrated instrument, not an empirically validated one. Its scientific contribution is a Kazakhstan-specific compliance-by-evidence model that makes accountability testable at service level; independent expert validation and operational pilots remain necessary.

Keywords: *artificial intelligence; personal data; digital public services; algorithmic decision-making; human oversight; explainability; accountability; risk governance; Kazakhstan*

Защита персональных данных в государственных сервисах Казахстана с ИИ

Аннотация: Искусственный интеллект в государственном секторе может влиять на доступ к услугам, проверки, определение права на поддержку и иные формы осуществления публичной власти, однако соблюдение требований нельзя доказать лишь ссылкой на нормы права или перечнем технических мер. Цель исследования - разработать Казахстанскую модель управления конфиденциальностью и подотчётностью искусственного интеллекта KZ-AIPAG, которая связывает каждое правовое требование и сценарий государственной услуги с риском, мерой контроля, доказательным артефактом, ответственным субъектом и решением на этапе

жизненного цикла. Методология объединяет нормативно-правовой анализ, целевой структурированный обзор литературы, сопоставление канадской оценки алгоритмического воздействия, Акта ЕС об искусственном интеллекте, статей 22 и 35 Общего регламента по защите данных и сквозной модели алгоритмического аудита, а также ограниченные кейсы Smart Data Ukimet и eGov AI. Результаты показывают, что основной пробел в Казахстане имеет операционный характер и не сводится к отсутствию регулирования. KZ-AIPAG разделяет четыре уровня воздействия сценария и оценку отдельных угроз по формуле вероятности и последствий, объединяет шесть областей управления и вводит доказательную цепочку, четыре контрольные точки, единый каталог ролей и матрицу RACI, сохраняющую независимость внутреннего аудита. Для иллюстративного кейса Smart Data Ukimet получено условное решение об отказе в запуске до подтверждения правового основания, цели обработки, рисков связывания и повторной идентификации, модели доступа, содержательного человеческого контроля и эффективности мер. Научный вклад состоит в контекстной модели compliance-by-evidence, делающей подотчётность проверяемой на уровне услуги. Независимая экспертная валидация и пилотирование остаются необходимыми.

Ключевые слова: *искусственный интеллект; персональные данные; цифровые государственные услуги; алгоритмическое принятие решений; человеческий контроль; объяснимость; подотчётность; управление рисками; Казахстан*

Қазақстанның ЖИ қолданатын мемлекеттік қызметтеріндегі дербес деректерді қорғау

Аңдатпа: Мемлекеттік сектордағы жасанды интеллект қызметтерге қол жеткізуге, тексерулерге, қолдау алу құқығын айқындауға және жария билікті жүзеге асырудың өзге де нысандарына ықпал етуі мүмкін, алайда талаптардың сақталуын тек құқықтық нормаларға сілтеме жасау немесе техникалық шараларды тізбелеу арқылы дәлелдеу жеткіліксіз. Зерттеудің мақсаты - әрбір құқықтық талап пен мемлекеттік қызмет сценарийін тәуекелмен, бақылау шарасымен, дәлелдік артефактпен, жауапты тұлғамен және өмірлік цикл шешімімен байланыстыратын Қазақстанның KZ-AIPAG жасанды интеллект құпиялылығы мен есептілігін басқару моделін әзірлеу. Әдіснама нормативтік-құқықтық талдауды, нысаналы құрылымдалған әдебиет шолуын, Канаданың алгоритмдік әсерді бағалау құралын, Еуропалық Одақтың Жасанды интеллект туралы актісін, Деректерді қорғаудың жалпы регламентінің 22 және 35-баптарын және алгоритмдік аудиттің толық циклдік моделін салыстыруды, сондай-ақ Smart Data Ukimet пен eGov AI бойынша шектелген кейстерді қамтиды. Нәтижелер Қазақстандағы негізгі олқылық реттеудің толық болмауынан емес, талаптарды қызмет деңгейінде іске асыру мен дәлелдеудің жеткіліксіздігінен туындайтынын көрсетті. KZ-AIPAG пайдалану сценарийінің төрт әсер деңгейін жекелеген қауіптерді ықтималдық пен салдар бойынша бағалаудан бөледі, басқарудың алты саласын

біріктіреді және дәлелдер тізбегін, төрт шешім қақпасын, бірыңғай рөлдер каталогын әрі ішкі аудит тәуелсіздігін сақтайтын RACI матрицасын ұсынады. Smart Data Ukimet иллюстрациялық кейсінде құқықтық негіз, өңдеу мақсаты, деректерді байланыстыру және қайта сәйкестендіру тәуекелі, қол жеткізу моделі, мазмұнды адам бақылауы мен бақылау тиімділігі дәлелденгенге дейін шартты түрде іске қоспау шешімі алынды. Ғылыми үлес - есептілікті қызмет деңгейінде тексерілетін ететін Қазақстанға бейімделген compliance-by-evidence моделі. Тәуелсіз сараптамалық валидация және практикалық пилоттау қажет.

Түйін сөздер: жасанды интеллект; дербес деректер; цифрлық мемлекеттік қызметтер; алгоритмдік шешім қабылдау; адам бақылауы; түсіндірмелілік; есептілік; тәуекелдерді басқару; Қазақстан

1. Introduction (Literature Review)

A public-sector AI service differs from a private application because its output may be embedded in the exercise of public authority: it may prioritise an inspection, propose a support measure, prepare a recommendation for a public official, or affect access to a service. The scale and inter-agency integration of data increase not only the value of analytics but also the consequences of error, secondary use, excessive access, and opaque delegation of decision-making. Personal data protection therefore cannot be reduced to encryption or consent; it also encompasses the legality of the purpose, data quality, user competence, reviewability, and institutional accountability.

Kazakhstan's current legal framework contains substantive safeguards. Article 21 of the Constitution guarantees protection against the unlawful collection, processing, storage, and use of personal data, including through digital technologies [1]. Article 19-1 of the Law on Personal Data and Their Protection governs automated processing that affects rights and legitimate interests, requires an explanation, permits an objection, and establishes a three-working-day period for its consideration [2]. The Law on Artificial Intelligence allocates duties to system owners and operators and provides for risk management, audit, and user information [3]. Article 43 of the Digital Code establishes non-discrimination, information rights, an explanation of key factors, and human review of consequential fully automated decisions [4]. The proposition that Kazakhstan faces a complete regulatory vacuum is therefore not supported.

The article's central thesis is that demonstrable personal data protection does not arise merely from the existence of an individual rule or technical control. It arises from a traceable chain in which the service purpose determines permissible data; the type of impact determines the risk tier; the risk determines mandatory controls and human oversight; each control has an owner, an artefact, and an acceptance criterion; and any material change to the model, data, or purpose triggers reassessment.

The study addresses four research questions. RQ1: Which regulatory, governance, privacy, and human-factor gaps can be substantiated across the public-sector AI lifecycle? RQ2: How should AI use cases be classified according to their effects on rights? RQ3: What

architecture should a context-specific, multi-layer governance framework adopt? RQ4: How can the framework be implemented and its effectiveness measured? The object of the research is the use of AI in the delivery of digital public services. The subject is the set of legal, organisational, technological, and managerial mechanisms used to protect personal data in such services.

1.1 Conceptual framework

Artificial intelligence is understood in the functional sense adopted by the Law on Artificial Intelligence: a system capable of generating outputs, for stated objectives, that influence a physical or virtual environment [3]. A public-sector AI service is an organisational and technical service, or a component of a process, in which a confirmed AI system produces an answer, classification, forecast, recommendation, or decision for a public authority or citizen. A digital public service is broader and may be entirely rule-based without using AI.

Conventional automation executes predetermined rules without statistical inference. Algorithmic automation encompasses any digital system that makes or influences a decision through automated data processing; under Article 43 of the Digital Code, this category includes, but is not limited to, AI [4]. AI-assisted decision-making leaves the legally consequential determination to an authorised official, whereas a fully automated decision is made without human participation in assessing the circumstances or approving the outcome. A confirmation button does not constitute meaningful human oversight if the official lacks sufficient time, information, competence, or authority to reject the recommendation.

Profiling means using data to evaluate or predict characteristics of an individual; personalisation means adapting an interface or offer; a recommender system produces a ranked option; a predictive model forecasts an event; and automated assessment assigns a category or score. In this study, explainability means providing the affected person with an intelligible account of the key factors, limits, and consequences of an output. Accountability means the organisation's capacity to identify the responsible actor and produce evidence that the relevant obligation has been fulfilled.

The scope is limited to confirmed AI or algorithmic decision systems in public services; ordinary rule-based digitisation is excluded. Kazakhstan's Data Management Requirements, OECD and OECD/UNESCO instruments, the NIST AI RMF and de-identification guidance, and ISO/IEC standards on AI, privacy, risk, and information security provide implementation benchmarks rather than binding Kazakhstan law [5]–[15].

1.2 Literature synthesis and research gap

The core peer-reviewed corpus covers organisational AI governance, systematic governance synthesis, privacy- and security-aware ethical assessment, public-service privacy behaviour, automated administrative responsibility, interaction-based governance, fairness assessment in public organisations, and Kazakhstan profiling law [16]–[23].

The literature converges on lifecycle governance, distributed responsibility, privacy- and security-aware assessment, meaningful human oversight, and contestability [16]–[22].

It also shows that convenience does not prove rights compliance, that general principles are reshaped in implementation, and that public-sector AI must be treated as a sociotechnical service rather than a model in isolation [19]–[22], [28]–[36]. The Kazakhstan study on profiling is used as historical context because it predates the 2025–2026 reforms [23].

The research gap lies in the absence, within the reviewed corpus, of an integrated model for Kazakhstan's updated legal and institutional context that translates rights and general risk principles into service-level evidence, distinguishes AI from other forms of automation, accounts for inter-agency architecture, assigns an accountable actor, and links reassessment to material system changes. The review does not establish global priority for the proposed model; novelty is claimed as a context-specific design and applied analytical instrument.

The objective is to develop and analytically demonstrate a Kazakhstan-specific governance framework that converts legal and governance requirements into testable service-level accountability.

The scientific novelty lies in the compliance-by-evidence unit of analysis: rule → use case → risk → control → evidence artefact → accountable owner → gate decision. Unlike a general checklist, the model separates the governance tier of a public-authority use case from the score of an individual hazard and makes reassessment compulsory after a material change.

The study contributes: (1) a four-tier public-sector AI use-case classification; (2) an AI-privacy risk matrix that distinguishes inherent, target, and verified residual risk; (3) the six-domain KZ-AIPAG architecture with four decision gates; and (4) a role and evidence model that preserves internal-audit independence and operationalises redress.

The remainder of the paper is organised as follows: Section 2 explains the research design and materials; Section 3 reports the legal, comparative, framework, and case results; Section 4 answers the research questions and discusses implications and limitations; and Section 5 concludes the study.

2. Methods and Materials

2.1 Research design and unit of analysis

The study uses qualitative normative legal analysis, a targeted structured literature review, comparative framework analysis, and two bounded case studies. The unit of analysis is the chain 'rule or principle → use case → risk to a right → control → evidence artefact → role → lifecycle gate'. The comparative legal analysis is retained because Table 1 applies common comparison dimensions to Canada, the European Union, GDPR, and Raji et al., and then identifies the distinct function of KZ-AIPAG.

2.2 Sources, search, and selection criteria

The literature search was updated on 29 September 2026. Crossref metadata and the publisher platforms ACM Digital Library, SpringerLink, ScienceDirect, and Oxford Academic were searched, with backward citation checking and separate searches of official

legal and policy portals. Four English-language search strings were used: 'artificial intelligence public sector governance'; 'algorithmic impact assessment public administration'; 'algorithmic auditing accountability public services'; and 'Kazakhstan artificial intelligence personal data'. The first 50 Crossref results ranked for each query were screened (200 records); DOI/title deduplication left 195 unique records. A purposive set of 18 peer-reviewed works was retained after relevance and metadata verification. The legal corpus comprised the operative official texts and official implementation materials cited in the references.

Inclusion required a peer-reviewed work or primary official source, a direct connection to public-sector AI, personal data, impact assessment, algorithmic audit, human oversight, accountability, or Kazakhstan law, and verifiable bibliographic metadata. Exclusion criteria were duplicate records, purely technical model-performance papers without governance relevance, marketing materials, unsourced commentary, and outdated legal text. This is a targeted structured review, not a PRISMA systematic review; Crossref's relevance-ranked sample and the English-language search create selection limitations.

[ТРЕБУЕТСЯ ДОПОЛНИТЬ: приложить журнал поиска или экспорт результатов Crossref/издательских платформ, подтверждающий дату поиска, 200 просмотренных записей, 195 уникальных записей и 18 включённых работ. До подтверждения эти числа следует рассматривать как заявленный, а не независимо проверенный протокол.]

2.3 Coding and analytical procedure

Table 1. Coding framework

Code family	Recorded fields	Decision rule	Output
Legal rule	Instrument, article, effective date, addressee, duty, exception	Use current official text; distinguish binding rule from guidance	Legal map
Use case	Purpose, data, output, user, affected person, decision effect	Classify as AI only where the source or technical evidence confirms AI	System card and scope statement
Tier	Legal effect, material influence, sensitivity, scale, reversibility, review	Assign the highest Tier criterion triggered	Tier decision with rationale
Risk	Hazard, cause, affected right, P, I, evidence, uncertainty	Score inherent risk; record proposed target after controls	Risk register
Control and evidence	Preventive/detective/corrective control, owner, test, artefact	A control is effective only after design and operating-effectiveness testing	Evidence pack and gate result

Code family	Recorded fields	Decision rule	Output
Role	R, A, C, I, independence conflict	Exactly one A; internal audit cannot own an operational process it audits	RACI matrix

Risk scores were assigned by the lead author through structured analytical judgement, not by an expert panel. For each of eight risks, the author applied the published 1-5 anchors, recorded the legal and operational assumptions, and performed a second consistency pass against the tier rule and control logic. No independent human raters scored the cases; consequently, inter-rater reliability is not reported. Section 13 specifies the independent validation required before empirical claims can be made.

2.4 Case boundaries and validation status

The study's evidence is limited to publicly available material. Non-public contracts, data-flow diagrams, model cards, logs, impact assessments, access reviews, and audit results were unavailable. The case application is therefore an analytical demonstration of the framework, not an audit finding about the actual operating effectiveness of Smart Data Ukimet or eGov AI.

KZ-AIPAG has undergone analytical consistency checking by the lead author but has not been independently validated. A preregistered panel of 10–15 specialists in Kazakhstan law, AI/model validation, cybersecurity, audit, and public administration should independently classify 8–12 standardised scenarios, score probability and impact, assess control sufficiency, and identify blocking conditions. Fleiss' kappa should assess categorical agreement and Kendall's W should assess ordinal ratings. Until that study and an operational pilot are completed, the framework and target scores remain analytically demonstrated hypotheses rather than validated measurements.

3. Results

3.1 Legal and institutional findings

The comparison produces three principal findings. First, substantive legal safeguards already exist: purpose specification and proportionality, data protection, conditions governing automated processing, explanation, objection, and review [2], [4]. Second, the Law on Artificial Intelligence adds obligations concerning system ownership, risk management, audit, and transparency [3]. Third, an implementation gap remains: legislation does not automatically convert each obligation into a coherent set of service-level records, roles, and tests. This gap may be addressed through secondary regulation, departmental governance, and system architecture. Legislative amendment is warranted only after establishing that lower-level mechanisms cannot deliver the required outcome.

The principal conclusion is that the most defensible gap is a governance-and-evidence gap. High-impact use cases would additionally benefit from consistent requirements for AI impact assessment, a public register or system card, overseer

competence, and minimum audit-trail content. Before proposing legislative amendments, policymakers should determine whether these requirements can be introduced through sectoral rules, digital-system requirements, and procurement contracts.

Temporal applicability is stated explicitly. The Constitution adopted at the referendum of 15 March 2026 entered into force on 1 July 2026 [1]. The Law on Artificial Intelligence No. 230-VIII of 17 November 2025 was first officially published on 18 November 2025 and entered into force on 18 January 2026 [3]. Article 19-1 of the Personal Data Law was introduced by Law No. 231-VIII of 17 November 2025; that amending law also entered into force after sixty calendar days, on 18 January 2026 [2]. The Digital Code No. 255-VIII of 9 January 2026 entered into force on 12 July 2026 [4]. The EU AI Act entered into force on 1 August 2024 and generally applies from 2 August 2026, subject to staged application [25]; GDPR has applied since 25 May 2018 [26].

3.2 Comparative position and Kazakhstan-specific contribution

The Canadian Directive on Automated Decision-Making operationalises four impact levels and scales notice, explanation, peer review, human intervention, and recourse requirements to the level [24]. The EU AI Act uses a different architecture: prohibited practices, high-risk classification and obligations, transparency duties for specified systems, and a residual category rather than a single four-level scoring matrix [25]. GDPR Article 22 protects individuals against solely automated consequential decisions and requires safeguards in defined exceptions, while Article 35 requires a data protection impact assessment where processing is likely to create high risk [26]. Raji et al. organise internal algorithmic auditing across the system lifecycle and treat documentation produced at each stage as part of an integrated audit report [27].

Table 2. Comparative position and Kazakhstan-specific contribution

Instrument	**Classification or trigger**	**Core mechanism**	**What KZ-AIPAG retains**	**What KZ-AIPAG adds for Kazakhstan**
Canada AIA and Directive	Four impact levels I-IV based on decision effects and reversibility	Mandatory questionnaire and proportionate requirements	Four escalating use-case tiers and proportional controls	Direct mapping to Kazakhstan legal duties, public-body roles, evidence artefacts, and gates
EU AI Act	Prohibitions, high-risk systems, specified transparency duties, other systems	Provider/deployer obligations, registration, conformity and oversight	Risk-based escalation, transparency, documentation, monitoring	Service-level evidence chain compatible with Kazakhstan institutions; no claim of EU legal equivalence

Instrument	Classification or trigger	Core mechanism	What KZ-AIPAG retains	What KZ-AIPAG adds for Kazakhstan
GDPR Arts. 22 and 35	Solely automated significant decisions; processing likely to create high risk	Safeguards, human intervention, contestation, DPIA	Human review, objection, impact assessment, redress	Links Art. 19-1 of Kazakhstan's Personal Data Law to a three-working-day review workflow
Raji et al. 2020	Lifecycle audit of organisational decisions and artefacts	End-to-end internal algorithmic audit	Lifecycle evidence and independent challenge	Single accountable gate owner, national legal map, and procurement/agency evidence pack
KZ-AIPAG	Highest triggered Tier 1-4 criterion; separate P × I hazard scoring	Six governance domains, evidence chain, gates, RACI, monitoring	Integrated operational model	Kazakhstan-specific bridge from law to service decision, including the Digital Code and AI Law

The novelty is therefore limited but identifiable. KZ-AIPAG does not claim a new universal theory of AI risk. Its contribution is a jurisdiction-specific compliance-by-evidence design: it joins Kazakhstan's rules on personal data, automated processing, algorithmic systems, AI risk management, audit, transparency, ownership duties, and liability to operational artefacts and public-sector decision authority. This resolves a gap left by frameworks that are either jurisdiction-neutral, focused on providers, or not connected to a public service's RACI and legal review clock.

3.3 Use-case classification and risk assessment

Tier assignment is an ex ante classification of the use case, not the arithmetic result of P × I. The assessor tests six dimensions: legal or similarly significant effect; material influence on a public official; personal-data sensitivity; population scale; reversibility; and meaningful review. The highest criterion triggered determines the tier. Controls may reduce a hazard score, but they do not automatically lower the tier because the underlying public-authority function remains consequential.

The relationship to P × I is two-dimensional. Tier sets the governance floor for the service. P × I (1-25) ranks individual hazards and treatment urgency within that service. A Tier 4 service cannot be treated as low governance merely because a proposed control yields a target score of 4; conversely, a Tier 1 service with a critical cybersecurity hazard must treat the hazard before release. A tier may be reduced only when the use case itself is redesigned

so that the triggering effect, data, scale, irreversibility, or review deficit is removed and documented.

Inherent risk is assessed before additional controls as probability P × impact I. P uses five evidence-anchored categories: 1 rare under exceptional conditions; 2 unlikely in ordinary operation; 3 possible; 4 likely because of scale, exposure, or repeated use; and 5 expected or recurring. I equals 1 negligible, 2 minor and readily reversible, 3 material but reversible, 4 serious or difficult to reverse, and 5 severe, large-scale, or legally consequential.

Scores are interpreted as 1-4 low, 5-9 moderate, 10-16 high, and 17-25 critical. The revised matrix reports target residual risk, not observed residual risk, because the proposed controls have not been tested. A target becomes a residual score only after evidence demonstrates both control design and operating effectiveness. Each P and I value must be accompanied by a scenario, data source, control assumption, assessor, and date; unsupported numerical precision is prohibited.

Acceptance rules are: critical inherent risk blocks deployment pending treatment; high target risk requires senior written acceptance and independent challenge; Tier 4 requires senior approval and effective redress regardless of arithmetic score; and any unresolved legality, discriminatory effect, or absence of meaningful human review is a blocking condition rather than a risk that can be averaged away.

Table 3. Tier assignment rule and minimum governance

Tier	**Binding assignment rule**	**Illustrative use**	**Minimum governance**
1 Informational	No individual legal effect; no material influence; no sensitive profiling; error readily reversible.	Generic search, FAQ, non-personal drafting.	Owner, source register, notice, basic security, error channel.
2 Assisted processing	Processes personal data or structures work, but does not materially determine a consequential outcome.	Document extraction, routing, personalised navigation.	Data map, purpose register, quality test, access control, human correction.
3 Material decision support	Output materially influences exercise of public authority, access, inspection, eligibility, or risk treatment; meaningful review remains available.	Risk flag, recommendation, fraud/anomaly triage.	Combined AI/privacy impact assessment, validation, subgroup testing, reviewer protocol, independent challenge, monitoring.
4 Determinative or difficult-to-reverse	Solely automated or effectively determinative consequential outcome, or high-scale sensitive profiling with ineffective or delayed review.	Automated eligibility, status, sanction, or similarly significant classification.	Explicit legal basis, senior approval, independent assurance, effective human intervention, explanation and redress, enhanced audit and publication.

The eight scenarios below are analytical risk hypotheses. P = 2 is not an empirical frequency estimate; it is a target of 'unlikely' conditional on the stated control package. The justification differs for each risk and must be replaced by observed evidence after operation.

Table 4. AI privacy risks and target residual scores

Risk and inherent score	**Control package**	**Why target P = 2 is plausible**	**Target P × I**	**Verification needed**
R1 Prompt leakage 3 × 3 = 9	Constrained interface, data-flow map, no-training default, DLP tests, short retention.	Exposure should require a configuration failure, bypass, or exceptional input rather than normal use.	2 × 2 = 4	Prompt sampling, vendor configuration evidence, leakage test.
R2 Extraction error 3 × 3 = 9	Gold dataset, confidence threshold, sampling, correction route.	Routine errors should be detected or corrected; undetected material error remains possible in atypical documents.	2 × 2 = 4	Field-level error rates by document type and subgroup.
R3 Secondary use 4 × 4 = 16	Field-purpose register, compatibility review, separated access, change approval.	Unauthorised reuse should require an approved-purpose bypass or governance failure.	2 × 3 = 6	Purpose audit and access-log sample.
R4 False positive 3 × 4 = 12	Calibrated thresholds, subgroup tests, proportionate investigation, human corroboration.	A false signal remains possible, but adverse action should not follow without corroboration.	2 × 3 = 6	Prospective false-positive and outcome analysis.
R5 Automation bias 4 × 4 = 16	Reviewer authority, adequate time, contrary evidence, reason record, override monitoring.	Rubber-stamping should become exceptional if reviewers must evidence independent reasoning and overrides are reviewed.	2 × 3 = 6	Decision-file sampling and reviewer observation.
R6 Inadequate explanation 4 × 5 = 20	Reason template, factor trace, specialist review, three-working-day objection route.	A deficient explanation should be uncommon when generated reasons are checked against the decision record.	2 × 4 = 8	Comprehension test and appeal-file audit.

Risk and inherent score	Control package	Why target P = 2 is plausible	Target P × I	Verification needed
R7 Re-identification 3 × 5 = 15	Controlled environment, linkage testing, output checking, data minimisation.	Re-identification should require an atypical linkage or control failure, not routine analyst access.	2 × 4 = 8	Independent linkage and singling-out test.
R8 Excess privilege 4 × 5 = 20	RBAC/ABAC, quarterly recertification, privileged monitoring, alerts, revocation SLA.	Persistent excess access should be unlikely if entitlement changes and dormant privileges are detected promptly.	2 × 4 = 8	Access recertification sample and alert-response evidence.

The common target probability does not imply a common causal mechanism. It represents the same ordinal end-state – unlikely in ordinary operation – reached through different controls and tested through different evidence. If any verification result fails, the target score must be withdrawn and the risk re-estimated.

3.4 KZ-AIPAG architecture and lifecycle

KZ-AIPAG is reorganised into three connected views. The decision view moves from mandate and scope to tier, risks, controls, evidence, independent challenge, and a gate decision. The domain view assigns six concurrent governance domains: legal and rights; data; AI and model; cybersecurity; human and organisational; and assurance, audit, and redress. The lifecycle view ensures that approval is revisited after a material change, incident, complaint, drift signal, or new data source.

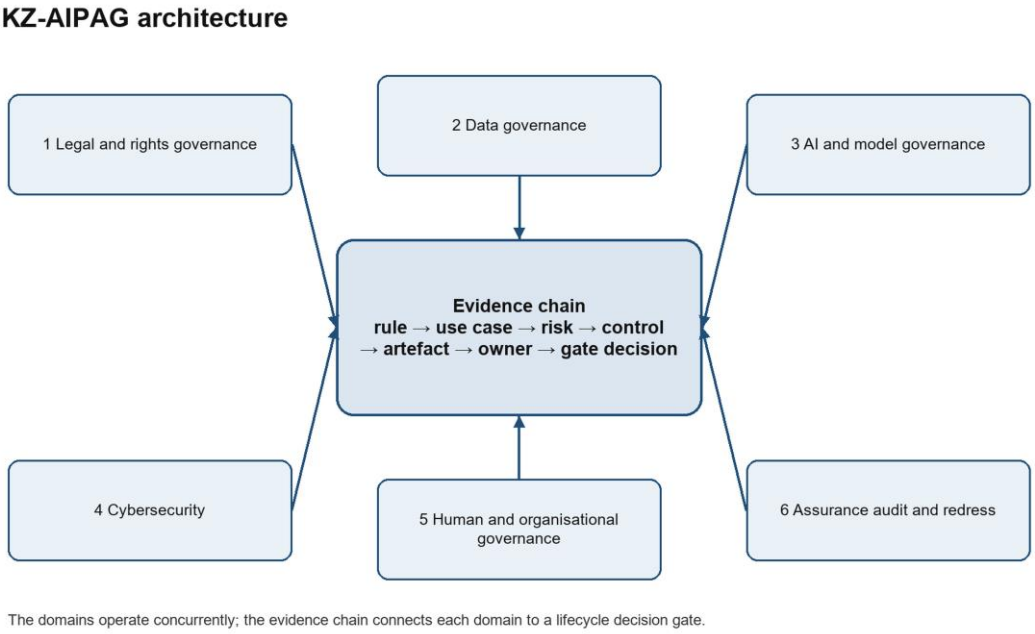


Fig. 1. Revised KZ-AIPAG architecture.

The central evidence chain is the model's unit of accountability: rule → use case → risk → control → artefact → owner → gate decision. A gap at any link prevents a reviewer from determining whether the legal duty was discharged. The six domains do not operate as a sequence; each contributes evidence to the same decision file.

Four gates make the structure operational. Gate 1 approves scope and tier. Gate 2 approves design and planned treatments. Gate 3 confirms verification and readiness. Gate 4 decides continued operation after monitoring or a trigger event. The accountable actor signs a pass, conditional pass, restriction, or no-go decision; conditions must identify an owner and due date.

This structure incorporates continuous risk management from NIST and ISO, impact assessment from Canada and GDPR, transparency and high-risk governance from the EU AI Act, and lifecycle audit documentation from Raji et al. [8], [11], [12], [24]–[27]. Its jurisdiction-specific element is the Kazakhstan legal map and the service-level evidence and decision authority attached to it.

The eighteen implementation activities are consolidated into seven stages: mandate and purpose; legal basis and tier; data mapping and impact assessment; access, oversight, and security design; testing and independent validation; approval and deployment; and monitoring, incident response, reassessment, or retirement. Each stage produces a named artefact and has a transition criterion.

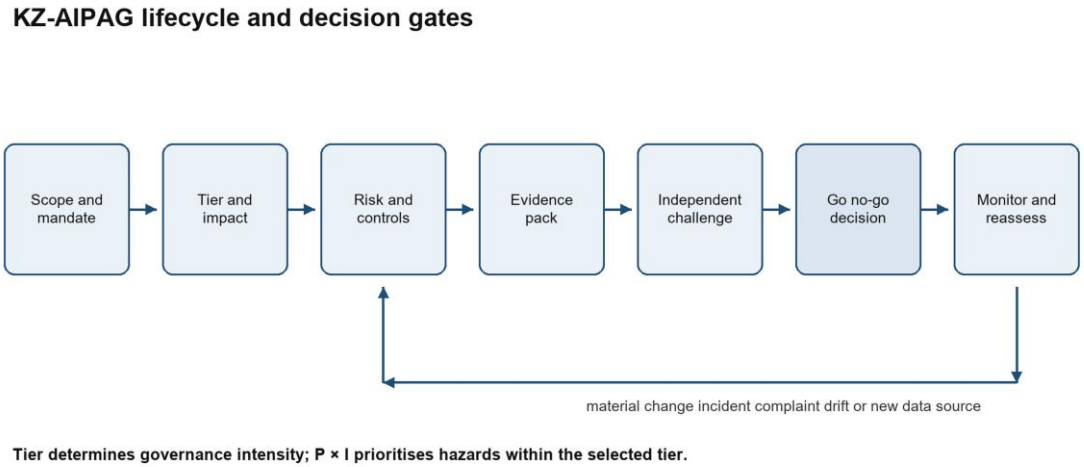


Fig. 2. KZ-AIPAG lifecycle and decision gates.

3.5 Responsibility allocation

The authoritative role catalogue is: agency head or delegated senior risk owner; service owner; legal and privacy function; data owner; AI/model owner; AI and data specialists; IT operations; cybersecurity function; procurement/vendor-management function; redress officer; independent validator; internal audit; and vendor. The same labels are used in the lifecycle, RACI, risk register, and evidence templates. R performs the work, A is the single operationally accountable approver, C is consulted, and I is informed.

Table 5. RACI for core KZ-AIPAG processes

Process	**A**	**R**	**C**	**I**
Purpose and legal basis	Agency head	Service owner; legal/privacy	Data owner; AI owner	Internal audit; vendor
Tier and impact assessment	Service owner	AI owner; legal/privacy	Data owner; cybersecurity; affected-function representative	Agency head; internal audit
Data mapping and quality	Data owner	AI/data specialists	Service owner; legal/privacy; cybersecurity	AI owner; internal audit
Model validation	Independent validator	AI/data specialists	AI owner; service owner; legal/privacy	Agency head; internal audit

Process	A	R	C	I
Access and cybersecurity	Cybersecurity lead	IT operations	Data owner; service owner; legal/privacy	Agency head; internal audit
Human oversight	Service owner	Authorised human reviewers	Legal/privacy; AI owner; HR/learning function	Agency head; internal audit
Redress and correction	Service owner	Redress officer; legal/privacy	Data owner; AI owner; IT	Agency head; internal audit
Incident response	Service owner	Cybersecurity; IT operations	Legal/privacy; data owner; vendor	Agency head; internal audit
Independent audit	Head of internal audit	Internal audit team	Service owner; legal/privacy; cybersecurity	Agency head; oversight body

Internal audit is not accountable for redress, deployment, control operation, or risk acceptance. Its accountability is limited to the independent-audit engagement and its quality. Operational ownership of redress rests with the service owner, while the redress officer and legal/privacy function perform review and correction. This preserves the organisational independence needed for subsequent assurance.

3.6 Bounded case-study results

Smart Data Ukimet is publicly described as an inter-agency platform for collecting, storing, and comprehensively analysing public-sector data, with a data lake, hashing, and depersonalisation functions [37]. Public sources do not establish every model, dataset, decision path, or control. The application below therefore uses a bounded scenario: linked inter-agency data are analysed to produce a risk signal that may influence an individual public-service review. If evidence shows that the output is only aggregate and cannot materially influence an individual outcome, the tier must be reduced and documented.

Table 6. End-to-end KZ-AIPAG application to Smart Data Ukimet

Step	**Assessment**	**Required measures**	**Evidence or document**	**Gate decision**
Scope and tier	Linked data and a risk signal may materially influence an individual review: Tier 3. Escalate to Tier 4 if the output is determinative or review is ineffective.	Confirm legal mandate, decision effect, affected groups, and reviewer authority.	Purpose statement; system card; tier rationale; legal map.	Gate 1 scope: conditional pass only if the individual decision path is documented.
Risks	R3 secondary use; R4 false positive; R5 automation bias; R7 re-identification; R8 excessive privilege.	Field-purpose mapping; subgroup/error testing; reviewer protocol; linkage test; least-privilege design.	Risk register; feature-purpose register; validation report; re-identification report; access matrix.	Gate 2 design: no-go while any high legal or re-identification risk lacks treatment.
Control verification	Proposed controls reduce target risk, but operating effectiveness is unknown.	Sample decisions, test access recertification, challenge linkage, test correction and override paths.	Test scripts; samples; access review; issue log; signed remediation evidence.	Gate 3 assurance: pass only after blocking findings close.
Deployment decision	Public value does not override unresolved rights risk.	Record accepting authority, conditions, rollback, monitoring thresholds, and complaint channel.	Signed go/no-go record; release and rollback plan; public notice; redress procedure.	Current analytical outcome: conditional no-go pending the evidence listed above.
Operation	New datasets, changed purposes, drift, complaints, or incidents can invalidate the assessment.	Monitor false positives, overrides, access anomalies, complaints, and linkage changes; reassess on trigger.	Monitoring dashboard; decision log; complaint register; reassessment record.	Gate 4 continuation: suspend or restrict use when a threshold is exceeded.

The decision is deliberately conservative. It does not assert that Smart Data Ukimet is unlawful or unsafe; it shows that a deployment decision cannot be justified from public descriptions alone. The missing evidence is itself a governance finding under KZ-AIPAG because the gate owner cannot verify legality, proportionality, re-identification resistance, meaningful review, or access control.

An official announcement describes eGov AI as an intelligent search function in eGov Mobile that helps users locate public services [38]. On the public description, the baseline classification is Tier 1 if no identifier, profile, or consequential decision is involved, and Tier 2 if queries are retained, personalised, or routed through components that process

personal data. Minimum evidence includes a prompt data-flow map, retention and training configuration, authoritative source register, multilingual quality tests, privacy notice, and human hand-off. Any later use to recommend eligibility, prioritise a case, or influence an outcome requires reassessment to Tier 2 or 3.

3.7 Human factors and performance indicators

A technically secure system remains institutionally vulnerable when a public servant does not understand the model's limits, holds excessive privileges, or merely confirms its recommendation. Automation bias can turn a human-in-the-loop arrangement into rubber-stamping; insufficient data literacy impedes recognition of poor data quality; weak cyber literacy increases phishing and improper-sharing risks; and diffuse responsibility allows the process owner to shift the substantive decision to the developer.

KZ-AIPAG requires a role-based competence matrix. End users must understand the system's purpose, uncertainty, and override procedure; data owners, provenance and quality; legal and privacy staff, applicable rules and redress; AI specialists, validation and drift; and agency heads, residual risk. For high-risk systems, the right to override, minimum review time, access to source factors, and protection against adverse consequences for a justified rejection of the recommendation must be procedurally guaranteed. Training completion is not a sufficient performance indicator without scenario-based competence testing.

Performance should be measured through complementary indicators: the proportion of registered systems with an approved tier; current impact assessments for high-risk systems; sampled decisions with accessible review records; justified override rates; expired or excessive privileges per 1,000 accounts; redress timeliness; consequential errors per 1,000 outputs; the distribution of verified residual risk; scenario-based staff competence; and timely audit closure. Each measure requires its source register and interpretive limitation; document existence, course attendance, zero overrides, or an average risk score cannot establish effectiveness on their own.

4. Discussion

4.1 Direct answers to the research questions

RQ1. The analysis does not support a regulatory-vacuum thesis: the applicable instruments establish purpose limitation and proportionality, data protection, conditions for consequential automated processing, explanation, objection, non-discrimination, and human review. The substantiated gap primarily concerns operationalisation, evidence, coordination of inter-agency roles, and control of material changes. Any sector-specific legal gaps must be established at the level of a particular service.

RQ2. The study develops a four-tier classification in which the tier depends on decision effect, material influence, data sensitivity, scale, reversibility, and meaningful review. Tier sets the governance floor; P × I prioritises individual hazards within that tier. The revised AI-Privacy Risk Matrix connects the risk source, affected right, legal rule,

control, evidence, owner, inherent risk, and target residual risk. Because control effectiveness was not tested, the target scores must not be reported as observed residual risk.

RQ3. KZ-AIPAG comprises Legal and Rights Governance, Data Governance, AI Governance, Cybersecurity, Human and Organisational Governance, and Assurance, Audit, and Redress. Unlike a compilation of NIST, OECD, and ISO recommendations, it introduces a Kazakhstan-specific evidence chain and service gates, under which transition is permitted only when a verified artefact and an accountable actor are present.

RQ4. The framework is implemented through eighteen stages from public-purpose definition to retirement. Its effectiveness is measured through coverage, quality, oversight, redress, access, drift, residual risk, competence, and audit closure. Potential applications include departmental regulation, procurement control, project appraisal, and internal audit. The principal limitations are the absence of non-public documentation and operational validation. Future research should pilot the framework in a specific public service, validate the scoring method, compare language groups, and test whether the evidence chain reduces actual errors and the time required to restore affected rights.

4.2 Interpretation and relation to prior research

The findings support prior research on lifecycle governance and distributed responsibility [16], [17], [22], while refining it for Kazakhstan: substantive rights are already established, so the scientific and practical task shifts from declaring principles to constructing an evidence chain. This is consistent with NIST, where governance informs mapping, measurement, and management [8]. KZ-AIPAG nevertheless makes the administrative consequence and the citizen's practical ability to secure review explicit objects of assurance.

The case studies reveal a distinction between data-centric and interaction-centric risks. For Smart Data Ukimet, the material issues are linkage, purpose, inter-agency access, and re-identification. For eGov AI, they are prompt privacy, output reliability, multilingual quality, and safe hand-off. A single checklist is therefore inappropriate. Nor does the analysis support the proposition that a large volume of data is inherently unlawful: risk arises from disproportionate purpose, opaque secondary use, poor quality, excessive access, or the absence of effective redress.

KZ-AIPAG differs from the compared instruments in its unit of accountability and its national legal linkage. Canada provides a mature four-level impact tool; the EU AI Act supplies a binding risk-and-obligations architecture; GDPR provides individual-decision safeguards and DPIA; and Raji et al. provide an end-to-end audit logic [24]–[27]. KZ-AIPAG combines these functions only where compatible and adds a Kazakhstan-specific chain linking Articles 15, 20, 21, and 30 of the AI Law, Article 19-1 of the Personal Data Law, and the Digital Code to a named public-service owner, evidence artefact, and gate decision. This is an implementation contribution, not evidence that the framework outperforms its comparators.

The theoretical contribution is a compliance-by-evidence model for public-sector AI. Human oversight is treated as an institutional capacity to change an outcome, and accountability as a link between the A role, an evidence artefact, and a lifecycle gate. The practical implication is that implementation should begin not with the acquisition of controls, but with an inventory of use cases and one end-to-end path from data collection to redress.

An alternative explanation is that the observed implementation gap reflects limited public documentation rather than deficient internal governance. The framework therefore treats missing public evidence as a condition that prevents an external assurance conclusion, not as proof that a public body lacks the relevant control.

4.3 Practical and policy implications

Implementation should begin with a service evidence pack and one accountable actor for every gate. Dataset-specific linkage tests are required after new inter-agency combinations; meaningful human review must record authority, review time, reasons, and overrides; procurement contracts should secure audit rights, change notification, and limits on vendor data use; and internal audit should test operating effectiveness through sampled decisions rather than document existence. These measures can be introduced through departmental rules, system requirements, and procurement controls before legislative amendment is proposed. Their performance indicators are evidence-pack completeness, gates with an assigned accountable role, reassessments after linkage changes, quality-sampled override records, compliant contracts, and timely closure of audit findings.

4.4 Limitations

The study is limited by an English-language, relevance-ranked search; reliance on public legal and service descriptions; author-assigned ordinal risk scores; and the absence of contracts, system cards, data-flow diagrams, logs, impact assessments, access reviews, audit findings, and independent raters. The bounded cases demonstrate analytical use but do not establish operating effectiveness, causal impact, or comparative superiority. These limitations restrict generalisation beyond the selected Kazakhstan context.

5. Conclusion

The study finds that the central personal-data challenge in Kazakhstan's AI-enabled public services is not a complete regulatory vacuum but the absence of a consistently verifiable bridge between a legal duty and a service-level decision. KZ-AIPAG supplies that bridge by linking the use-case tier, individual hazards, controls, evidence, accountable roles, and lifecycle gates.

The scientific novelty is the Kazakhstan-specific compliance-by-evidence model and its separation of governance intensity from hazard prioritisation. The comparative, legal, and bounded case analyses support the internal coherence and practical relevance of the model, but they do not constitute independent validation. Its practical value lies in structuring impact assessment, procurement, approval, monitoring, audit, and redress around evidence that can be tested.

Future research should preregister and conduct the proposed expert study, pilot the framework in a defined public service, test the scoring anchors and gate thresholds, compare outcomes across language groups, and measure whether the evidence chain reduces consequential errors and the time required to restore affected rights.

Data Availability Statement

Data availability. The study uses public legal instruments, official service descriptions, standards, and publications cited in the reference list. No personal data or confidential operational datasets were processed. The coding schema is reproduced in Table 2; no separate research dataset was created.

Use of Artificial Intelligence

Use of artificial intelligence. Generative AI tools were used to support language translation, editorial restructuring, bibliographic cross-checking, and document formatting. The author remains responsible for source verification, legal interpretation, risk judgements, and the final text. No claim of expert-panel validation was generated or inferred by AI.

References

[1] Republic of Kazakhstan. (2026). Konstitutsiya Respubliki Kazakhstan [Constitution of the Republic of Kazakhstan] (adopted by national referendum on 15 March 2026). https://old.adilet.zan.kz/rus/docs/K2600000000 (In Russian).
[2] Republic of Kazakhstan. (2013, as amended 2025). Zakon Respubliki Kazakhstan No. 94-V 'O personal'nykh dannykh i ikh zashchite' [Law of the Republic of Kazakhstan No. 94-V on personal data and their protection]. https://old.adilet.zan.kz/rus/docs/Z1300000094 (In Russian).
[3] Republic of Kazakhstan. (2025). Zakon Respubliki Kazakhstan No. 230-VIII 'Ob iskusstvennom intellekte' [Law of the Republic of Kazakhstan No. 230-VIII on artificial intelligence]. https://old.adilet.zan.kz/rus/docs/Z2500000230 (In Russian).
[4] Republic of Kazakhstan. (2026). Tsifrovoy kodeks Respubliki Kazakhstan No. 255-VIII [Digital Code of the Republic of Kazakhstan No. 255-VIII]. https://old.adilet.zan.kz/rus/docs/K2600000255 (In Russian).
[5] Government of the Republic of Kazakhstan. (2024). Postanovlenie No. 925 'Ob utverzhdenii trebovaniy po upravleniyu dannymi' [Resolution No. 925 on approval of data management requirements]. https://adilet.zan.kz/rus/docs/P2400000925 (In Russian).
[6] OECD. (2024). AI, data governance and privacy (OECD Artificial Intelligence Papers No. 22). OECD Publishing. https://doi.org/10.1787/2476b1a4-en
[7] OECD & UNESCO. (2024). G7 toolkit for artificial intelligence in the public sector. OECD Publishing. https://doi.org/10.1787/421c1244-en
[8] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
[9] National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). https://doi.org/10.6028/NIST.AI.600-1
[10] Garfinkel, S., Near, J., Dajani, A., Singer, P., & Guttman, B. (2023). De-identifying government datasets: Techniques and governance (NIST SP 800-188). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-188
[11] International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system.
[12] International Organization for Standardization. (2023). ISO/IEC 23894:2023 Information technology – Artificial intelligence – Guidance on risk management.
[13] International Organization for Standardization. (2025). ISO/IEC 42005:2025 Information technology – Artificial intelligence – AI system impact assessment.
[14] International Organization for Standardization. (2025). ISO/IEC 27701:2025 Information security, cybersecurity and privacy protection – Privacy information management systems – Requirements and guidance.
[15] International Organization for Standardization. (2022). ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection – Information security management systems – Requirements.
[16] Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. AI and Ethics, 2(4), 603-609. https://doi.org/10.1007/s43681-022-00143-x

[17] Batool, A., Zowghi, D., & Bano, M. (2025). AI governance: A systematic literature review. AI and Ethics, 5(3), 3265-3279. https://doi.org/10.1007/s43681-024-00653-w
[18] Korobenko, D., Nikiforova, A., & Sharma, R. (2024). Towards a privacy and security-aware framework for ethical AI: Guiding the development and assessment of AI systems. In Proceedings of the 25th Annual International Conference on Digital Government Research (pp. 740-753). ACM. https://doi.org/10.1145/3657054.3657141
[19] Willems, J., Schmid, M. J., Vanderelst, D., Vogel, D., & Ebinger, F. (2023). AI-driven public services and the privacy paradox: Do citizens really care about their privacy? Public Management Review, 25(11), 2116-2134. https://doi.org/10.1080/14719037.2022.2063934
[20] Langer, C. (2024). Decision-making power and responsibility in an automated administration. Discover Artificial Intelligence, 4, Article 59. https://doi.org/10.1007/s44163-024-00152-1
[21] Papyshev, G. (2025). Governing AI through interaction: Situated actions as an informal mechanism for AI regulation. AI and Ethics, 5(2), 1109-1120. https://doi.org/10.1007/s43681-024-00446-1
[22] Dankloff, M., Skoric, V., Sileno, G., Ghebreab, S., van Ossenbruggen, J., & Beauxis-Aussalet, E. (2025). Analysing and organising human communications for AI fairness assessment. AI & Society, 40(4), 2347-2367. https://doi.org/10.1007/s00146-024-01974-4
[23] Akhmetova, S. B., Kassymzhanova, A. A., & Ibrayeva, A. S. (2025). Modern development of artificial intelligence technologies and problems of legal regulation of profiling and targeted advertising in Kazakhstan. Bulletin of L. N. Gumilyov Eurasian National University. Law Series, 150(1), 77-97. https://doi.org/10.32523/2616-6844-2025-150-1-77-97
[24] Treasury Board of Canada Secretariat. (n.d.). Directive on automated decision-making. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592
[25] European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[26] European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2016/679/oj
[27] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 33-44). ACM. https://doi.org/10.1145/3351095.3372873
[28] Alon-Barkat, S., & Busuioc, M. (2023). Human-AI interactions in public sector decision making: Automation bias and selective adherence to algorithmic advice. Journal of Public Administration Research and Theory, 33(1), 153-169. https://doi.org/10.1093/jopart/muac007
[29] Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, 38(3), 101577. https://doi.org/10.1016/j.giq.2021.101577
[30] Fatima, S., Desouza, K. C., Buck, C., & Fielt, E. (2022). Public AI canvas for AI-enabled public value: A design science approach. Government Information Quarterly, 39(4), 101722. https://doi.org/10.1016/j.giq.2022.101722
[31] van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. Government Information Quarterly, 39(3), 101714. https://doi.org/10.1016/j.giq.2022.101714
[32] de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 39(2), 101666. https://doi.org/10.1016/j.giq.2021.101666
[33] Delfos, J., Zuiderwijk, A. M. G., van Cranenburgh, S., Chorus, C. G., & Dobbe, R. I. J. (2024). Integral system safety for machine learning in the public sector: An empirical account. Government Information Quarterly, 41(3), 101963. https://doi.org/10.1016/j.giq.2024.101963
[34] Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825-836. https://doi.org/10.1111/puar.13293
[35] Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector – Applications and challenges. International Journal of Public Administration, 42(7), 596-615. https://doi.org/10.1080/01900692.2018.1498103
[36] Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. Government Information Quarterly, 37(3), 101493. https://doi.org/10.1016/j.giq.2020.101493
[37] International Telecommunication Union. (n.d.). Smart Data Ukimet: Kazakhstan. https://www.itu.int/en/ITU-D/Regional-Presence/CIS/Pages/GovStack/SDU.aspx
[38] Ministerstvo iskusstvennogo intellekta i tsifrovogo razvitiya Respubliki Kazakhstan. (2026). Itogi 2025 g.: kazakhstantsy vse chashche poluchayut gosuslugi cherez smartfon [2025 results: Kazakhstanis increasingly access

public services via smartphones]. https://www.gov.kz/memleket/entities/maidd/press/news/details/1134705?lang=ru (In Russian).

Bakytzhan Zhumadilla*
PhD student, Department of Informatics, Information and Communication Technologies
zhumadilla_bakytzhan@korkyt.kz, orcid.org/0009-0005-0976-6250
Korkyt Ata Kyzylorda University, Kazakhstan
Nurbek Konyrbaev
PhD, Department of Computer Science
n.konyrbaev@mail.ru, https://orcid.org/0000-0002-8788-4149
Korkyt Ata Kyzylorda University, Kazakhstan
Gul Jussupova
Candidate of Technical Sciences, Department of Digitalization and Public Services
g.jussupova@sci.gov.kz, https://orcid.org/0009-0009-7205-3666
Ministry of Science and Higher Education of the Republic of Kazakhstan, Kazakhstan

ASKING ANOTHER JURISDICTION WHAT IT HOLDS WITHOUT TELLING IT WHAT YOU HOLD: PRIVATE ARTIFACT MATCHING WITH A LEDGER-ANCHORED TRANSCRIPT

Abstract: Two authorities investigating related cases would both benefit from knowing which digital artifacts they hold in common. Neither can afford the usual way of finding out. Sending a list of hashes discloses the whole holding, including artifacts belonging to unrelated subjects, to a partner who may be under no obligation to discard it; and once sent, neither side can prove afterwards what was asked or what was answered. We apply Diffie-Hellman private set intersection over secp256k1 to this problem and add what the forensic setting actually needs, which is an auditable record: the Merkle roots of both blinded transcripts and a digest of the result are committed as a salted field vector and anchored on a permissioned ledger, so the exchange can later be shown to have happened exactly as claimed, with the artifact digests themselves never disclosed. We evaluate on 62,973 real artifacts recovered from a FAT32 medium by The Sleuth Kit and foremost. In a two-device scenario, 32,988 digests against 24,154 with a true overlap of 64, the protocol returns the intersection exactly, no false positives and no false negatives, in 8.4 s and 3.60 MiB of traffic, which is 2.06 times what a plain digest exchange would cost and discloses nothing beyond the intersection. Cost is linear: 66 bytes and 0.24 ms per item held. Anchoring the transcript adds 632 bytes of ledger state and 13 ms, and the disclosable record is 2,166 bytes and verifies in 0.96 ms. We also measure a protocol-hygiene failure that is easy to ship: reusing the blinding exponent across queries makes successive transcripts identical, letting a counterparty track a holding over time, where fresh exponents give zero overlap.

Keywords: *digital forensics; private set intersection; cross-border evidence exchange; permissioned ledger; transcript attestation; secp256k1; mutual legal assistance*

Запрос к другой юрисдикции о её данных без раскрытия собственных: конфиденциальное сопоставление цифровых артефактов с закреплённым в реестре транскриптом обмена

Аннотация: Двум ведомствам, расследующим связанные дела, было бы полезно знать, какие цифровые артефакты есть у них обоих. Ни одно из них не может позволить себе обычный способ это выяснить. Передача списка хеш-значений раскрывает весь массив данных, включая артефакты, относящиеся к посторонним лицам, партнёру, который может быть не обязан его уничтожить; а после передачи ни одна из сторон не может впоследствии доказать, что именно запрашивалось и что было получено в ответ. Мы применяем протокол конфиденциального пересечения множеств (PSI) на основе Диффи-Хеллмана над кривой secp256k1 и добавляем то, что действительно необходимо в криминалистике, а именно проверяемую запись: корни Меркла обоих ослеплённых транскриптов и дайджест результата фиксируются как вектор полей с солью и закрепляются в разрешённом (permissioned) распределённом реестре, так что впоследствии можно показать, что обмен состоялся именно так, как заявлено, при этом сами дайджесты артефактов никогда не раскрываются. Оценка выполнена на 62 973 реальных артефактах, восстановленных с носителя FAT32 средствами The Sleuth Kit и foremost. В сценарии с двумя устройствами (32 988 дайджестов против 24 154 при истинном пересечении 64) протокол возвращает пересечение точно, без ложноположительных и ложноотрицательных результатов, за 8,4 с и 3,60 МиБ трафика, что в 2,06 раза больше стоимости простого обмена дайджестами, при этом сверх пересечения ничего не раскрывается. Стоимость линейна: 66 байт и 0,24 мс на каждый элемент. Закрепление транскрипта добавляет 632 байта состояния реестра и 13 мс, а раскрываемая запись занимает 2 166 байт и проверяется за 0,96 мс. Мы также измеряем ошибку гигиены протокола, которую легко допустить на практике: повторное использование ослепляющей экспоненты в разных запросах делает последовательные транскрипты идентичными и позволяет контрагенту отслеживать массив данных во времени, тогда как при свежих экспонентах пересечение транскриптов равно нулю.

Ключевые слова: *цифровая криминалистика; конфиденциальное пересечение множеств; трансграничный обмен доказательствами; разрешённый распределённый реестр; аттестация транскрипта; secp256k1; взаимная правовая помощь*

Өз деректеріңді ашпай-ақ басқа юрисдикциядан оның деректері туралы сұрау: тізілімде бекітілген алмасу транскриптімен цифрлық артефактілерді құпия салыстыру

Аңдатпа: Өзара байланысты істерді тергейтін екі ведомство үшін қандай цифрлық артефактілер екеуінде де бар екенін білу пайдалы болар еді. Бірақ олардың ешқайсысы мұны анықтаудың әдеттегі тәсілін қолдана алмайды. Хеш-мәндер тізімін жіберу бүкіл деректер жиынын, соның ішінде бөгде тұлғаларға тиесілі артефактілерді, оны жоюға міндетті болмауы мүмкін серіктеске ашады; ал жіберілгеннен кейін тараптардың ешқайсысы не сұралғанын және не жауап берілгенін кейін дәлелдей алмайды. Біз secp256k1 қисығы үстінде Диффи-Хеллман негізіндегі жиындардың құпия қиылысуы (PSI) хаттамасын қолданамыз және криминалистикаға шын мәнінде қажет нәрсені, яғни тексерілетін жазбаны қосамыз: екі соқырландырылған транскриптің Меркл түбірлері мен нәтиже дайджесті тұздалған өрістер векторы ретінде бекітіліп, рұқсат етілген (permissioned) таратылған тізілімге жазылады, сондықтан кейін алмасудың дәл мәлімделгендей өткенін көрсетуге болады, ал артефактілердің дайджесттері ешқашан ашылмайды. Бағалау The Sleuth Kit және foremost құралдарымен FAT32 тасымалдағышынан қалпына келтірілген 62 973 нақты артефактіде жүргізілді. Екі құрылғы сценарийінде (шынайы қиылысуы 64 болатын 32 988 және 24 154 дайджест) хаттама қиылысуды дәл, жалған оң және жалған теріс нәтижелерсіз, 8,4 с ішінде және 3,60 МиБ трафикпен қайтарады; бұл қарапайым дайджест алмасуынан 2,06 есе көп, бірақ қиылысудан тыс ештеңе ашылмайды. Шығын сызықтық: әр элементке 66 байт және 0,24 мс. Транскриптті бекіту тізілім күйіне 632 байт пен 13 мс қосады, ал ашуға болатын жазба 2 166 байтты құрайды және 0,96 мс ішінде тексеріледі. Сондай-ақ біз іс жүзінде оңай жіберілетін хаттама гигиенасы қатесін өлшейміз: соқырландыру дәреже көрсеткішін сұраулар арасында қайта пайдалану бірізді транскрипттерді бірдей етеді және контрагентке деректер жиынын уақыт бойынша бақылауға мүмкіндік береді, ал әр сұрауға жаңа көрсеткіш қолданылғанда транскрипттердің қиылысуы нөлге тең болады.

Түйін сөздер: *цифрлық криминалистика; жиындардың құпия қиылысуы; дәлелдемелермен трансшекаралық алмасу; рұқсат етілген таратылған тізілім; транскриптті аттестаттау; secp256k1; өзара құқықтық көмек*

1. Introduction

Investigations overlap more often than they are allowed to admit. Two agencies, sometimes in two countries, each hold a seized medium, and the question of whether the same file appears on both is frequently the question that would join two cases into one. The obvious way to find out is to send a list of hashes. That is also the way that cannot be defended: a hash list is the whole holding, it covers artifacts belonging to people who are not suspects in the receiving jurisdiction, and the receiving authority is usually under no enforceable obligation to forget it.

The problem has a known cryptographic shape. Private set intersection lets two parties learn the common elements of their sets and nothing else, and the literature is mature [1], [2], [3], [4]. What the forensic setting adds is a requirement the cryptography does not

address. An evidence exchange is an event in a case file. Months later somebody will ask what was requested, what was returned, and whether either side changed its story; and answering that must not require re-disclosing the inputs. A protocol transcript that only the two parties hold does not settle such a dispute, because each of them is the party being doubted.

This paper puts the two halves together. We use Diffie-Hellman PSI over secp256k1, which is the least exotic construction that meets the requirement, and commit the exchange: the Merkle roots of both blinded transcripts and a digest of the agreed result are written as a salted field vector and anchored on a permissioned ledger. Neither side can afterwards claim to have sent a different set, or to have received a different answer, and the anchored material discloses nothing, because every field commitment carries 128 bits of salt.

Everything reported here was executed. The corpus is 62,973 artifacts recovered from a real FAT32 medium with The Sleuth Kit and foremost; the ledger is a three-node etcd Raft cluster with four endorsing processes signing over TCP; the curve operations are libsecp256k1. Intersections are checked against ground truth exactly rather than sampled.

The contributions are as follows.

- An exchange protocol that combines PSI with a ledger-anchored transcript, so that a cross-border artifact query is both privacy-preserving and, afterwards, non-repudiable by either side.
- A measurement on real forensic artifacts rather than synthetic sets: two scenarios drawn from one medium, with intersections recovered exactly at 0.24 ms and 66 bytes per item held.
- A quantified comparison against the plain digest exchange that current practice uses, showing the privacy costs 2.06 times the traffic and roughly a quarter of a millisecond per item.
- A measured protocol-hygiene result: reusing the blinding exponent across queries yields transcripts that are byte-identical, which converts a one-off private query into a persistent tracking channel.

1.1 How Artifact Matching Is Done Now

Operational practice is hash sharing. Reference sets such as the NSRL [5] distribute digests of known files, and victim-identification programmes circulate hash lists of known illegal material so that agencies can recognise what they have already seen. The model works because those lists are meant to be public or semi-public. It does not transfer to case holdings, which are neither, and where the asymmetry is the point: an agency may be willing to reveal that it holds one specific artifact while being unwilling, or legally unable, to reveal the other thirty thousand.

1.2 Private Set Intersection

The Diffie-Hellman construction goes back to Meadows [1] and was developed for practical use by Huberman et al. [2] and De Cristofaro and Tsudik [3]. Its cost is two exponentiations per element per party and one group element per element on the wire, which

is more than modern oblivious-transfer constructions [4], [6] need in raw throughput, but it is round-efficient, needs no correlated randomness, and is short enough to implement without a framework. That last property matters when the code has to be inspected by people who will not accept a black box in an evidential path. Circuit-based and OPRF-based protocols scale better for very large sets and would be the right choice above a few million items; the sets in a single case are usually far smaller than that.

1.3 Ledgers in Forensics

Blockchain work in this area has concentrated on the chain of custody [7], [8], [9], and more recently on attesting analytical output. The exchange itself has received little attention, although it is exactly the kind of bilateral act where both parties have an incentive to remember it differently. Our attestation follows the transparency-log tradition [10], [11] and reuses salted per-field commitments so that the record of an exchange can be disclosed without the exchange's content.

1.4 Positioning

Table 1 sets out the alternatives against the two properties that matter here: what the counterparty learns, and whether the exchange can be reconstructed afterwards by someone who trusts neither party.

Table 1. Ways two authorities can compare holdings

Method	What the partner learns	Traffic per item	Exact result	Auditable later
Send digest list	the entire holding	32 B	yes	no
Bloom filter exchange	membership, with false positives	< 32 B	no	no
OT-based PSI [4], [6]	the intersection only	varies	yes	no
DH-PSI [1], [3]	the intersection only	66 B	yes	no
This work	the intersection only	66 B	yes	yes

2. Methods and Materials

2.1 Problem Setting

Two authorities, A and B, hold sets X and Y of artifact digests taken from media they have lawfully seized. They wish to learn X and Y's intersection. A learns which of its own artifacts B also holds; B learns the same about its own. Neither learns anything about the other's non-matching elements, and neither learns the size of the other's set beyond what the message volume reveals, which we treat as public.

2.2 Adversary Model

C1, the honest-but-curious partner. Follows the protocol and then tries to recover the counterparty's non-matching artifacts from what it received. Against a passive partner the security of the construction rests on the decisional Diffie-Hellman assumption in the group; a blinded point reveals nothing about its preimage without the exponent.

C2, the repeat querier. Runs the protocol against the same partner several times over months, hoping to correlate the transcripts and build a picture of how the partner's holdings change. Section 3.4 shows this attack succeeds completely against an implementation that reuses its exponent and fails completely against one that does not.

C3, the revisionist. After the exchange, one side asserts that it asked a different question, or that the answer was different, typically because the answer has become inconvenient. This is the adversary the cryptography does not address and the ledger does.

Out of scope: a malicious party that deviates from the protocol to extract more than the intersection, for example by inflating its input set with guesses. Malicious-model PSI exists and is more expensive; for the inter-agency setting, where both sides are identified and accountable, we take the semi-honest model with an anchored transcript to be the proportionate design, and say so rather than claiming more.

2.3 Protocol Construction

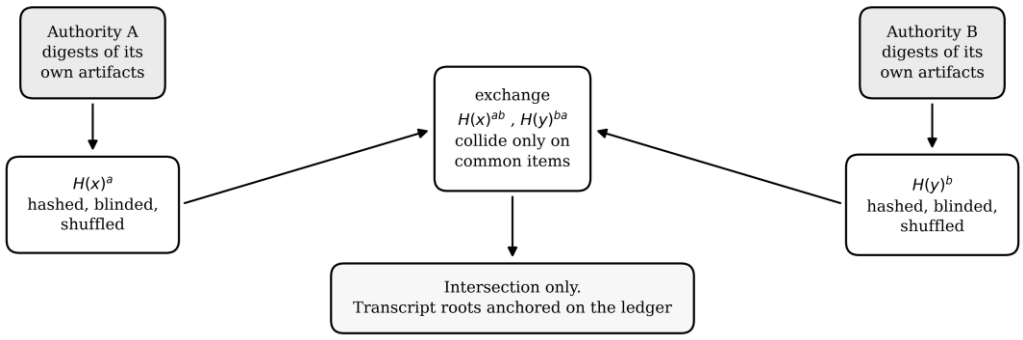


Fig. 1. The exchange. Each authority blinds its own digests with a single-use exponent, both sides re-blind what they receive, and only elements held by both collide. The transcript roots and the result digest are anchored.

Fig. 1 shows the exchange. Each side maps its digests onto curve points and raises them to a secret exponent drawn fresh for this query:

A sends { H(x)^a : x in X }, shuffled (1)

B sends { H(y)^b : y in Y }, shuffled (2)

Each side then raises what it received to its own exponent, producing H(x)^ab and H(y)^ba. Because the group is commutative, an element present in both sets yields the same point on both sides, and an element present in only one yields a point the other party cannot

invert. Matching the two doubly blinded sets gives the intersection, and the shuffle before transmission means position discloses nothing about input order.

Hashing to the curve is by try-and-increment on the compressed encoding, which is not constant time. That is acceptable because the values hashed are the party's own digests, not secrets held from itself; the quantity that must not leak is the exponent, and scalar multiplication is handled by libsecp256k1.

2.4 What the Transcript Commits

The exchange is summarised by four Merkle roots, one over each of the two blinded sets and each of the two doubly blinded sets, together with a digest of the agreed intersection. These are placed in a fixed field vector and each field is committed separately:

$$c = H(0x02 \,\|\, s \,\|\, n \,\|\, enc(v)),\ \ s \leftarrow \{0,1\}^{128} \tag{3}$$

The Merkle root of the field commitments is the exchange digest, and it is that root, not the exchange, which reaches the ledger. A party wanting to establish later what was asked releases the salts for the fields it chooses to open; the rest stay opaque. Because the wire values are already blinded with an exponent that is discarded after the query, opening the transcript roots discloses no artifact digest even to a party that later obtains them.

2.5 Corpus and Scenarios

The corpus is a real medium. A 3.5 GB FAT32 volume was populated with files from 280 distribution packages, 30 percent were deleted with mtools, and the volume was refilled to capacity with further real files, so that clusters were genuinely reused. Artifacts were then taken from it with libtsk and foremost, giving 62,973 artifacts and 57,080 distinct digests. Two scenarios partition those digests, and in both the true intersection arises from duplication that is really present in the corpus rather than being planted.

- Two-device. Authority A holds the digests of the original population, 32,988 of them; authority B holds those of the material written later, 24,154. The true overlap is 64 digests, which is the realistic case: two unrelated devices sharing a handful of common files.
- Even-odd. The artifacts are split by index, giving 29,310 against 29,377 with a true overlap of 1,607. This is the high-overlap operating point, and it tests whether cost depends on how much the sets have in common.

2.6 Experimental Setup

Measurements were taken on two allocated cores under CPython 3.11 with libsecp256k1 through coincurve, single-threaded. A scalar multiplication costs 49 microseconds on this hardware, which is the unit of account for everything that follows. The ledger is a deployed three-node etcd 3.4.30 Raft cluster with four endorsing processes holding Ed25519 keys and answering over TCP sockets.

3. Results

3.1 Correctness and Cost

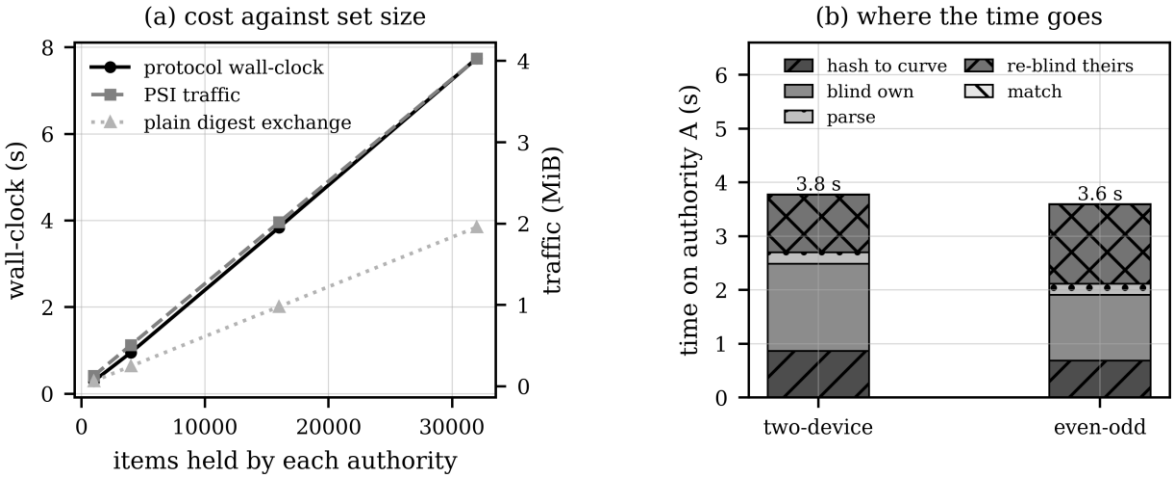


Fig. 2. (a) Wall-clock and traffic against set size, with the plain digest exchange for comparison. (b) Where authority A's time goes in each scenario.

Both scenarios returned the intersection exactly: 64 of 64 and 1,607 of 1,607, with no false positives and no false negatives. Correctness here is not a statistical claim, since the protocol is exact and the ground truth is set equality on digests; the experiment confirms that the implementation realises the protocol. Table 2 gives the costs. The two-device query took 8.4 s and moved 3.60 MiB; the even-odd query, with twenty-five times the overlap, took 7.4 s and moved 3.69 MiB. Cost tracks set size and is indifferent to how much the sets share, which is the privacy property showing up in the performance numbers.

Table 2. Both scenarios, executed on digests of real recovered artifacts

Scenario	A items	B items	True overlap	Found	Wall-clock	Traffic
Two-device	32,988	24,154	64	64 exact	8.4 s	3.60 MiB
Even-odd	29,310	29,377	1,607	1,607 exact	7.4 s	3.69 MiB

Fig. 2(b) breaks down where authority A's time goes in the two-device scenario: 0.87 s hashing its own digests onto the curve, 1.62 s blinding them, 0.21 s parsing what arrived, 1.07 s re-blinding it and 6 ms matching. Two thirds is scalar multiplication, as expected, and the matching step that people intuitively expect to dominate is noise.

3.2 Against the Plain Digest Exchange

The comparison that matters operationally is with what agencies do now. Sending raw digests costs 32 bytes per item once; the protocol costs 66 bytes per item, being a 33-byte compressed point in each of two rounds. The privacy therefore costs 2.06 times the

traffic in both scenarios and about 0.24 ms of computation per item held. On a 30,000-item holding that is under four megabytes and under ten seconds, against a plain exchange that would disclose 30,000 artifacts to obtain 64 answers. Put that way the trade is not close.

3.3 Scaling

Table 3 repeats the protocol on random subsets. Wall-clock rises from 0.29 s at 1,000 items per side to 7.74 s at 32,000, and traffic from 0.13 MiB to 4.03 MiB, both linear in set size to within measurement noise, and every run returned the exact intersection. Linearity is the expected behaviour of the construction and the measurement is a check that no accidental quadratic step crept into the implementation, which is a common way for a matching phase to go wrong.

Table 3. Scaling on random subsets. Every run returned the exact intersection

Items per side	Wall-clock	Traffic	Bytes per item	Plain digest exchange
1,000	0.29 s	0.13 MiB	66	0.06 MiB
4,000	0.94 s	0.50 MiB	66	0.24 MiB
16,000	3.83 s	2.01 MiB	66	0.98 MiB
32,000	7.74 s	4.03 MiB	66	1.95 MiB

3.4 Exponent Reuse

A working implementation can be correct and still leak, and this is the way we expect it to happen in practice. If a party keeps its blinding exponent between queries, its blinded set is a deterministic function of its holdings, so two transcripts are directly comparable. We measured it: with a fresh exponent per query, two runs over the same 3,000 items share none of their wire values, an overlap of 0.000; with the exponent reused, the overlap is 1.000 and every unchanged artifact is recognisable across the two runs. The intersections agree in both cases, so nothing in the output would alert an operator. A counterparty that keeps old transcripts thereby obtains a persistent fingerprint of the other's holdings and can watch it change, which is precisely the surveillance the protocol was adopted to prevent. Deriving the exponent freshly per query costs one random scalar.

3.5 Anchoring the Exchange

Committing the transcript and anchoring it on the Raft cluster costs 632 bytes of ledger state and 13.4 ms in the two-device scenario, against 3.60 MiB and 8.4 s for the exchange it attests: four ten-thousandths of the traffic. The disclosable record, releasing 12 of its 18 fields, is 2,166 bytes and verifies in 0.96 ms including three real Ed25519 signature checks. Table 4 collects these. The practical reading is that there is no cost argument against attesting every exchange.

Table 4. Cost of attesting the exchange, and the effect of reusing the blinding exponent

Quantity	Two-device	Even-odd	Note
Ledger state per exchange	632 B	632 B	constant
Anchor latency	13.4 ms	6.5 ms	real Raft commit
Disclosable record	2,166 B	2,166 B	12 of 18 fields
Record verification	0.96 ms	0.82 ms	3 Ed25519 checks
Wire overlap, fresh exponent	0.000	0.000	runs unlinkable
Wire overlap, reused exponent	1.000	1.000	runs linkable

4. Discussion

4.1 Limitations

Three. The security model is semi-honest: a party that inflates its input set with guessed digests learns whether those guesses are held, and this construction does not prevent it, though the anchored transcript does record the set size that was submitted. The cluster runs on loopback, so the anchoring latency is LAN-class, while the byte counts do not depend on topology. And the corpus, while genuinely recovered from a real medium by real tools, is not a seized device; the overlap between our two scenarios arises from duplicate files in a software distribution rather than from the behaviour of a suspect, so the overlap sizes should be read as operating points and not as estimates of what real cases share.

4.2 What the Anchor Is For

It is worth being precise about the division of work. The cryptography protects the inputs during the exchange. The anchor protects the account of the exchange afterwards, and only that: it establishes that a query with these transcript roots and this result digest completed no later than its epoch was anchored. It says nothing about whether the inputs were honestly assembled. Presenting it as more would repeat the error the custody literature has already made once.

4.3 Fitting Mutual Legal Assistance

Assistance under the Budapest Convention and its second additional protocol [12] is slow partly because the request itself discloses. An authority must often explain what it is looking for in order to ask, and that explanation reaches a foreign administration. A private query inverts the order: the parties establish that there is something to talk about, and only then does the formal process begin, now with a narrow and well-founded scope. The anchored transcript gives the later formal step something to point at.

4.4 Data Protection

Nothing identifiable is written to the ledger. Ledger state consists of epoch roots, sizes, linkage and signatures, and every field commitment is salted with 128 bits of entropy, so the anchored record is not personal data recoverable by anyone, which is the property that makes erasure obligations under the GDPR [13] satisfiable off-chain. The blinded transcripts themselves are not retained by the protocol; if a party keeps them, the exponent that produced them is gone.

5. Conclusion

Two agencies can find out what they hold in common without either handing over its holdings, and the cost of doing so on real forensic artifacts is about twice the traffic of the disclosure-everything method practised today and a quarter of a millisecond per item. Both of our scenarios returned the intersection exactly. The part that the cryptographic literature leaves out, and that a case file needs, is the record: anchoring the transcript costs 632 bytes and 13 ms, four ten-thousandths of the exchange it attests, and leaves a claim about what was asked and answered that neither party can later revise. The failure mode to watch is not the mathematics but the housekeeping. An implementation that reuses its blinding exponent produces identical transcripts across queries and turns a private question into a standing one, and nothing in the protocol's output reveals that this has happened.

Author Contributions

Conceptualization – B.Zh; Methodology – N.K. and G.J.; Software – B.Zh; Validation – B.Zh; Writing - original draft – B.Zh; Writing - review and editing N.K. and G.J.; Supervision – B.Zh.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1]. Meadows, C. (1986). A more efficient cryptographic matchmaking protocol for use in the absence of a continuously available third party. In *Proceedings of the IEEE Symposium on Security and Privacy* (pp. 134–137). IEEE.

[2]. Huberman, B. A., Franklin, M., & Hogg, T. (1999). Enhancing privacy and trust in electronic communities. In *Proceedings of the 1st ACM Conference on Electronic Commerce* (pp. 78–86). ACM.

[3]. De Cristofaro, E., & Tsudik, G. (2010). Practical private set intersection protocols with linear complexity. In *Financial cryptography and data security* (LNCS, Vol. 6052, pp. 143–159). Springer.

[4]. Pinkas, B., Schneider, T., & Zohner, M. (2014). Faster private set intersection based on OT extension. In *Proceedings of the 23rd USENIX Security Symposium* (pp. 797–812). USENIX Association.

[5]. National Institute of Standards and Technology. (2024). *National Software Reference Library and the Reference Data Set*.

[6]. Kolesnikov, V., Kumaresan, R., Rosulek, M., & Trieu, N. (2016). Efficient batched oblivious PRF with applications to private set intersection. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security* (pp. 818–829). ACM.

[7]. Lone, M. M., & Mir, R. N. (2019). Forensic-chain: Blockchain based digital forensics chain of custody with PoC in Hyperledger Composer. *Digital Investigation, 28*, 44–55.

[8]. Bonomi, S., Casini, M., & Ciccotelli, C. (2018). *B-CoC: A blockchain-based chain of custody for evidences management in digital forensics* (arXiv:1807.10359). arXiv.

[9]. Dasaklis, T. K., Casino, F., & Patsakis, C. (2021). SoK: Blockchain solutions for forensics. In *Technology development for security practitioners* (pp. 21–40). Springer.

[10]. Merkle, R. C. (1988). A digital signature based on a conventional encryption function. In *Advances in Cryptology - CRYPTO 1987* (LNCS, Vol. 293, pp. 369–378). Springer.

[11]. Laurie, B., Langley, A., & Kasper, E. (2013). *Certificate transparency* (RFC 6962). Internet Engineering Task Force.

[12]. Council of Europe. (2001). *Convention on Cybercrime* (ETS No. 185), and Second Additional Protocol (CETS No. 224, 2022). Council of Europe.

[13]. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). (2016). *Official Journal of the European Union, L 119*, 1–88.

[14]. Ongaro, D., & Ousterhout, J. (2014). In search of an understandable consensus algorithm. In *Proceedings of the USENIX Annual Technical Conference* (pp. 305–319). USENIX Association.

[15]. Faz-Hernandez, A., Scott, S., Sullivan, N., Wahby, R. S., & Wood, C. A. (2023). *Hashing to elliptic curves* (RFC 9380). Internet Engineering Task Force.

[16]. Carrier, B. (2005). *File system forensic analysis*. Addison-Wesley.

[17]. Casey, E. (2011). *Digital evidence and computer crime: Forensic science, computers, and the Internet* (3rd ed.). Academic Press.

[18]. International Organization for Standardization. (2012). *Information technology - Security techniques - Guidelines for identification, collection, acquisition and preservation of digital evidence* (ISO/IEC Standard No. 27037:2012).

Yerassyl Aikenov*
Master's student, School of Cybersecurity
255374@astanait.edu.kz, ORCID: https://orcid.org/0009-0003-9287-7183
Astana IT University, Kazakhstan

ON-CHAIN AND OFF-CHAIN DATA PLACEMENT IN HYBRID BLOCKCHAIN HEALTH SYSTEMS

Abstract: Which clinical data belongs on the ledger? Hybrid blockchain designs answer class by class, quietly, and the answer usually hardens into habit before anyone writes down why. We formalise it. Placement becomes constrained optimisation over storage tiers — each carrying its own verifiability, erasability, exposure, read latency and ledger footprint — and over clinical data classes carrying regulatory and service-level duties. The problem turns out to be a multiple-choice knapsack instance, hence NP-hard; compliance then prunes it so hard that the hardness stops mattering. Kazakhstan supplies the setting, and its statutory duties supply the constraints. Nine FHIR-aligned classes, six tiers, one worked instance. Regulatory and latency limits alone shrink 10,077,696 candidate assignments to 48; a ledger budget takes it to 12; and every uniform placement policy in common use lands outside that region entirely. The optimiser then buys per-item commitments for seven classes and batched commitments for the two carrying the heaviest instance counts — volume drives that split, not sensitivity. Re-running the model ten thousand times with every estimated parameter perturbed by up to twenty-five per cent left the pruning and the infeasibility of uniform policies intact in every draw, while the batching assignment held in 76.1 per cent of draws. Parameters are illustrative and tabulated in full, and every number reported here comes out of the model rather than off a deployment.

Keywords: *hybrid blockchain; electronic health records; data placement; multiple-choice knapsack; right to erasure; design space analysis; HL7 FHIR*

Размещение Данных в Реестре и Вне Его в Гибридных Медицинских Блокчейн-Системах

Аннотация: Какие клинические данные должны попадать в реестр? Гибридные блокчейн-архитектуры отвечают на этот вопрос класс за классом, негласно, и ответ обычно закрепляется как привычка раньше, чем кто-либо записывает его обоснование. Мы формализуем этот выбор. Размещение становится задачей оптимизации с ограничениями над уровнями хранения, каждый из которых характеризуется своей проверяемостью, возможностью удаления, степенью раскрытия, задержкой чтения и объёмом в реестре, и над классами клинических данных, несущими регуляторные обязательства и требования к уровню обслуживания. Задача оказывается экземпляром задачи о многовариантном рюкзаке и потому NP-трудна; однако требования соответствия отсекают пространство

решений настолько сильно, что эта трудность перестаёт иметь значение. Контекст задаёт Казахстан, а его законодательные требования задают ограничения. Рассмотрены девять классов данных, согласованных с FHIR, шесть уровней хранения и один рабочий пример. Одни только регуляторные ограничения и ограничения по задержке сокращают 10 077 696 возможных назначений до 48; бюджет реестра оставляет 12; и каждая распространённая единообразная политика размещения оказывается целиком за пределами этой области. Оптимизатор затем выбирает поэлементные обязательства для семи классов и пакетные — для двух классов с наибольшим числом экземпляров: это разделение определяет объём данных, а не их чувствительность. Десять тысяч повторных прогонов модели при отклонении каждого оценочного параметра до двадцати пяти процентов сохранили отсечение и недопустимость единообразных политик во всех прогонах, а распределение пакетирования — в 76,1 процента прогонов. Параметры носят иллюстративный характер и приведены полностью, а все числа получены из модели, а не из реального развёртывания.

Ключевые слова: *гибридный блокчейн; электронные медицинские записи; размещение данных; задача о многовариантном рюкзаке; право на удаление; анализ пространства проектных решений; HL7 FHIR*

Медициналық Гибридті Блокчейнде Деректерді Тізілімде және Одан Тыс Орналастыру

Аңдатпа: Қандай клиникалық деректер тізілімге түсуі керек? Гибридті блокчейн архитектуралары бұл сұраққа әр класс бойынша, үнсіз жауап береді, ал бұл жауап әдетте оның негіздемесі жазылмай тұрып-ақ әдетке айналып кетеді. Біз осы таңдауды формалдаймыз. Деректерді орналастыру сақтау деңгейлері бойынша шектеулері бар оңтайландыру есебіне айналады: әр деңгейдің өз тексерілу дәрежесі, жою мүмкіндігі, ашылу дәрежесі, оқу кідірісі және тізілімдегі көлемі бар, ал клиникалық деректер кластары нормативтік міндеттер мен қызмет көрсету деңгейіне қойылатын талаптарды көтереді. Есеп көп нұсқалы рюкзак есебінің дербес жағдайы болып шығады, демек NP-күрделі; алайда сәйкестік талаптары шешімдер кеңістігін соншалықты қатты қысқартатындықтан, бұл күрделіліктің іс жүзінде маңызы қалмайды. Зерттеудің контексін Қазақстан анықтайды, ал оның заңнамалық міндеттері шектеулерді құрайды. FHIR стандартымен үйлестірілген тоғыз деректер класы, алты сақтау деңгейі және бір жұмыс мысалы қарастырылды. Тек нормативтік және кідіріс шектеулерінің өзі 10 077 696 ықтимал тағайындауды 48-ге дейін қысқартады; тізілім бюджеті оларды 12-ге дейін азайтады; ал кең таралған бірыңғай орналастыру саясаттарының әрқайсысы осы аймақтан толықтай тыс қалады. Оңтайландырушы жеті класс үшін жеке міндеттемелерді, ал даналар саны ең көп екі класс үшін топтамалық міндеттемелерді таңдайды: бұл бөліністі деректердің

сезімталдығы емес, олардың көлемі анықтайды. Әрбір бағаланған параметр жиырма бес пайызға дейін өзгертілген он мың қайталама есептеу шешімдер кеңістігінің қысқаруын және бірыңғай саясаттардың жарамсыздығын барлық жағдайда сақтады, ал топтамалау бөлінісі есептеулердің 76,1 пайызында сақталды. Параметрлер иллюстрациялық сипатта және толық келтірілген, ал барлық сандар нақты ендіруден емес, модельден алынған.

Түйін сөздер: *гибридті блокчейн; электрондық медициналық жазбалар; деректерді орналастыру; көп нұсқалы рюкзак есебі; деректерді жою құқығы; жобалау кеңістігін талдау; HL7 FHIR*

1. Introduction (Literature Review)

Placement decides the rest. Once an architect has settled which clinical data classes go on a ledger, which go off it behind a commitment, and which never touch a ledger at all, the compliance posture, the storage bill and the audit story are largely determined — and reversing the choice later means migrating live patient records, which nobody does willingly.

The literature treats this decision unevenly. Payloads off-chain, hashes on-chain — proposals state it, then hurry on to consensus tuning or key management, as though the split were self-evident. It is not. Ask what each option actually proves. A hash on a permissioned ledger, a Merkle commitment anchored in public, and a batched commitment covering sixty-four records differ in what they demonstrate, what they leak, what they cost, and whether anyone can ever undo them. Health data is exactly where those differences bite: statutory retention pulling one way, erasure duties the other [1], [2].

Kazakhstan gives this work its concrete setting. Clinical documentation across the national system already runs electronically; the Code on Public Health defines personal medical data and the aggregator role that handles it [3]; a WHO-supported programme has been reshaping statutory forms toward a digital-friendly information model [4]. Walk a ledger-based service into that environment and it inherits the whole obligation set before a single design choice gets made.

This work converts those obligations into constraints and the architectural preferences into an objective, then asks what remains. The contribution is fourfold. A tier model separates the properties an architect actually trades off, rather than the platform labels usually compared. Regulatory duties, audit duties and service-level targets become hard constraints over that model. The resulting placement problem turns out to be a multiple-choice knapsack instance, which locates it precisely in a well-studied complexity class. And an instantiation on nine FHIR-aligned classes reports a design-space contraction of five orders of magnitude, together with an optimal placement that no uniform policy reproduces.

Three strands bear on the problem, and they have largely developed apart from one another.

1.1 Deciding Whether and Where to Use a Ledger

Wüst and Gervais handed the field its decision flowchart: do you need a blockchain at all, or will a managed database do [5]? Eberhardt and Tai catalogued the off-chaining patterns whose trade-offs we formalise here [6]. Xu and colleagues built a taxonomy meant to carry architectural reasoning [7]. Useful work, all of it. It classifies. Nobody optimises, and nobody carries regulatory duty into the constraint set.

1.2 Ledger Architectures for Health Records

MedRec put permission management on a public ledger [8]. Ancile added proxy re-encryption and interoperable access control [9]; BBDS and MedBlock went after cloud-resident sharing [10], [11]; Healthcare Data Gateways moved the patient to the centre of risk control [12], and later work carried that pattern into mobile-cloud settings [13]. Permissioned substrates followed — Fabric as the usual base [14], PREHEALTH bolting on anonymous credentials [15], latency behaviour by now reasonably well charted [16]. Off-chain custody hardened into practice through IPFS-backed designs [17], [18], [19], [20], [21], and reviews map what resulted [22], [23], [24]. Every one of these systems fixes a placement. Not one derives it.

1.3 Mutability and Erasure

Ateniese and colleagues opened the redactable-blockchain line with chameleon hashing [25]; Derler et al. sharpened rewriting down to attribute-level policies [26]; Deuber et al. threw out trapdoor custody in favour of on-chain voting [27]. Politou's group documents the legal pressure driving all of it [28], [2]. We do not treat redactability as the answer. It is one tier among six, carrying its own price and its own weakened verifiability.

The remainder of the paper is organized as follows. Section 2 sets out the system model, the problem formulation, the solution method and the experimental setup. Section 3 reports the results, Section 4 discusses their implications and the threats to validity, and Section 5 concludes.

2. Methodology

2.1 System Model

Let I index clinical data classes and T index storage tiers. A placement is a map $x : I \to T$ assigning every class to exactly one tier.

2.1.1 Data Classes

Class i carries eight parameters. Payload size s_i, instances per patient-year n_i, reads per instance f_i. Then the duties: a required verifiability level θ_i, a required erasability level e_i, and a ledger-presence flag p_i marking classes whose existence has to be independently attestable. Finally a read-latency ceiling λ_i and a sensitivity weight σ_i, the latter standing in for how much damage a disclosure would do.

Two deserve comment. Erasability and ledger presence fight each other by construction — consent grants and access log entries have to survive for audit, so they carry

$p_i = 1$ and $e_i = 0$, while ordinary clinical records carry the reverse. Uniform placement dies right there. The model puts that tension on the table instead of burying it.

2.1.2 Storage Tiers

Tier t carries a ledger cost rate κ_t, a verifiability level ver_t, an erasability level era_t, a presence flag $pres_t$, an exposure coefficient ρ_t, a read latency lat_t, and a per-instance ledger footprint. Table 2 lists the six tiers used here. They span the practical range: full payload on a public chain, full payload on a permissioned chain, redactable payload via chameleon hashing [25], [26], off-chain custody with a per-item commitment, off-chain custody with commitments batched into a Merkle tree of sixty-four [29], and off-chain custody with no ledger record whatsoever.

Batching is where the interesting economics live. A batched commitment costs one sixty-fourth of the ledger bytes but proves membership at coarser granularity, so an auditor learns that a record belonged to a batch rather than that it stood alone. The model encodes that difference as ver = 0.75, against 0.95 for per-item commitments.

2.1.3 Cost Components

Three costs and one quality loss feed the objective. Storage combines ledger and off-chain bytes at their respective rates. Read latency piles up as $n_i\, f_i\, lat_t$, disclosure exposure as $n_i\, \sigma_i\, \rho_t$. Verification granularity loss accumulates as $n_i\, \sigma_i\, (1 - ver_t)$, which docks the cheap tiers for weakening what an auditor can establish. Each component gets normalised by its maximum over $I \times T$. That puts the four weights on one scale and lets a reader treat them as priorities rather than as unit conversions.

2.2 Problem Formulation

2.2.1 Hard Constraints

Compliance does not belong in an objective function. A regulator does not accept a slightly cheaper design that is slightly less lawful, so duties enter as feasibility constraints, and cost operates only inside the region they leave.

C1 verifiability. $ver_t \geq \theta_i$. Integrity of the class must be demonstrable to an external auditor at the level the class demands.

C2 erasability. $era_t \geq e_i$. Classes subject to withdrawal of consent must sit where the data can be rendered irrecoverable [1], [2], through destruction of the key material that protects it [30].

C3 ledger presence. $pres_t = 1$ whenever $p_i = 1$. Audit-bearing classes must leave a trace no operator can silently drop.

C4 latency. $lat_t \leq \lambda_i$. Interactive classes cannot be parked behind public-chain confirmation.

C5 ledger budget. $\Sigma_i\, L(i, x(i)) \leq B$, where L is the annual on-ledger footprint of class i under its tier and B is the budget per patient-year. This is the only constraint that couples classes to one another, and it is what makes the problem combinatorial rather than separable.

Residency slips in implicitly. Every tier from the fourth onward keeps payloads inside the national perimeter, so a jurisdiction that forbids foreign replication simply deletes the first tier from every admissible set.

2.2.2 Optimisation Problem

Writing $x_{it} \in \{0,1\}$ for the indicator that class i lands on tier t, the problem becomes

$$\min \Sigma_i \Sigma_t x_{it} (\alpha c^s_{it} + \beta c^l_{it} + \gamma r_{it} + \delta d_{it}) \quad (1)$$

subject to $\Sigma_t x_{it} = 1$ for every i, to $x_{it} = 0$ whenever tier t violates any of C1 through C4 for class i, and to $\Sigma_i \Sigma_t x_{it} L_{it} \leq B$.

The structure is exact, and it deserves a name. Classes partition into disjoint groups; exactly one option gets picked from each; one linear capacity constraint ties the picks together. That is the multiple-choice knapsack problem in minimisation form [31], equivalent by negation to the maximisation form people have studied since the late nineteen-seventies. MCKP generalises 0–1 knapsack and inherits its NP-hardness [32], [33].

Hardness is not the practical obstacle here, though. Pruning is so severe that the surviving space is trivially enumerable, and the useful observation is which constraint does the pruning.

2.3 Solution Method

C1 through C4 apply class by class and hand back admissible sets $A_i \subseteq T$. Enumerate $\Pi_i |A_i|$ candidates exactly, drop whatever breaches the budget. At the scale reported below, that runs in milliseconds.

Real deployments will not stay this small. Several hundred classes across multiple facility types, separate budgets per consortium channel — that instance leaves enumeration far behind. Standard MCKP machinery covers the regime: prune dominated options, order what survives within each class by incremental efficiency, solve the linear relaxation in O(n) by partitioning, then close the integrality gap with branch and bound along the lines of Pisinger's minimal algorithm [34]. The relaxation throws in a bound as well, telling an architect the most a given budget increase could possibly buy. In a procurement conversation that bound is often worth more than the optimum itself.

2.4 Experimental Setup

The instantiation uses nine data classes aligned to FHIR resource types [35], [36] and the six tiers of Section 2.1.2. Parameters appear in Tables 1 and 2. They are illustrative: sizes and instance counts reflect plausible ambulatory workloads, and sensitivity weights encode ordinary clinical intuition about relative disclosure damage. Tables 1 and 2 state them in full, so a reader who disagrees can substitute their own and rerun the model.

Weights are $\alpha = 0.25$ for storage, $\beta = 0.15$ for latency, $\gamma = 0.35$ for exposure and $\delta = 0.25$ for granularity loss. The ledger budget is $B = 8$ KB per patient-year, chosen to bind: the

minimum feasible budget under C1 through C4 is 5.22 KB, so the constraint has real force without making the instance infeasible.

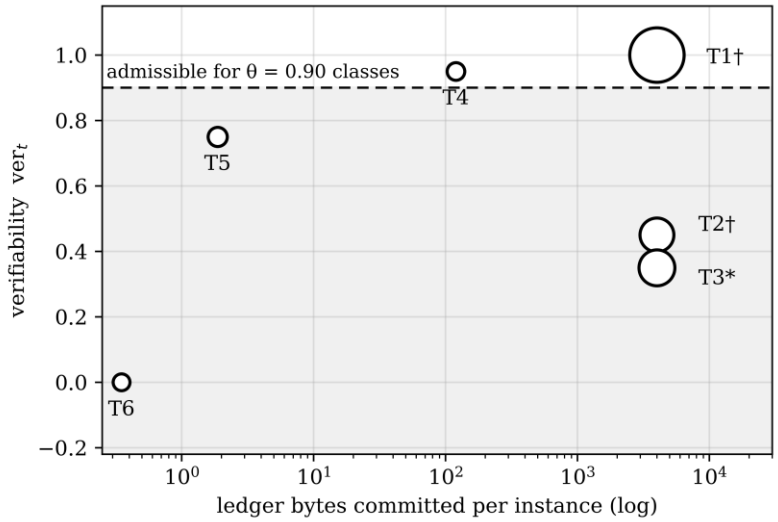


Fig. 1. The six tiers in the footprint–verifiability plane. Marker area scales with exposure ρ_t. † = not erasable, * = redactable only.

Fig. 1 places the six tiers in the plane that matters most for this decision, ledger footprint against verifiability, with marker area scaling to exposure. The θ = 0.90 line cuts through it and eliminates most of the plane for the strictest classes.

Table 1. Data classes (illustrative parameters, per patient-year)

Class	**Size**	**n**	**θ_i**	**e_i**	**λ_i (ms)**	**σ_i**
Patient demographics	2 KB	1	0.00	1.0	400	0.60
Encounter metadata	4 KB	6	0.40	1.0	400	0.40
Observation	1 KB	120	0.75	1.0	250	0.50
DiagnosticReport	30 KB	8	0.90	1.0	500	0.80
ImagingStudy	25 MB	1.5	0.90	1.0	2000	0.90
MedicationRequest	2 KB	14	0.75	1.0	250	0.70
Dispense record *	1 KB	10	0.90	0.0	600	0.55
Consent grant *	0.3 KB	20	0.90	0.0	200	0.15
Access log entry *	0.2 KB	180	0.75	0.0	600	0.20

* audit-bearing: $p_i = 1$, so a ledger trace is mandatory.

Table 2. Storage tiers

Tier	Ledger B / inst.	ver	era	ρ	lat
T1 public, payload	s_i	1.00	0.00	1.00	2500
T2 permissioned, payload	s_i	0.45	0.00	0.35	300
T3 redactable payload	s_i	0.35	0.85	0.40	420
T4 off-chain, per-item	120	0.95	1.00	0.05	180
T5 off-chain, batched/64	1.875	0.75	1.00	0.07	150
T6 off-chain only	0	0.00	1.00	0.04	120

3. Results

3.1 Contribution of the Compliance Envelope

Six tiers, nine classes, $6^9 = 10{,}077{,}696$ unconstrained placements. Verifiability alone cuts that to 10,368. Erasability takes it to 288. Ledger presence removes nothing further — the classes demanding it already sit on tiers that supply it. The latency ceiling leaves 48. The budget leaves 12. Fig. 2 shows the funnel.

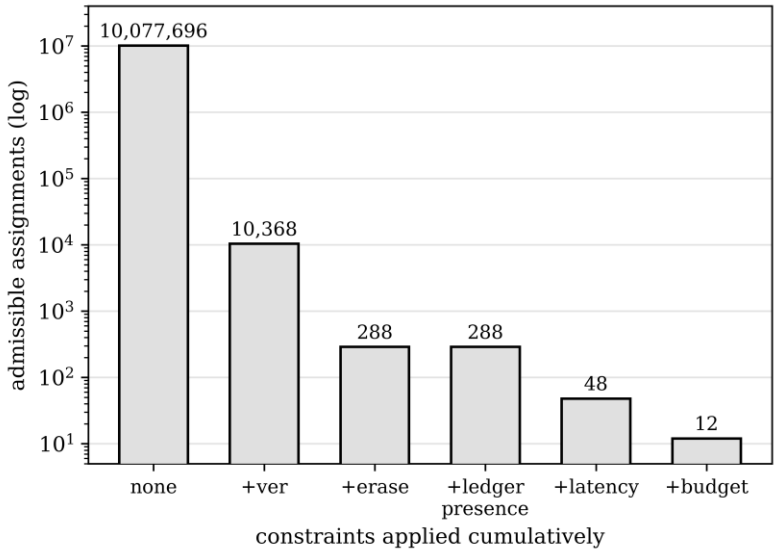


Fig. 2. Compliance constraints remove five orders of magnitude from the design space before any cost term is evaluated.

Five orders of magnitude vanish before anyone consults a single cost coefficient. That is probably the most transferable thing in this paper, and it flips the usual emphasis: architects tune weights and argue storage economics while the compliance envelope has already decided nearly everything. Optimise without writing that envelope down and you are searching a region you never characterised.

3.2 Infeasibility of Uniform Policies

Table 3 evaluates the four uniform policies that recur in the literature against the optimum. Not one of them is feasible. Placing payloads on a permissioned ledger consumes 38,866 KB per patient-year, roughly 4,858 times the budget, and the ImagingStudy class alone accounts for most of it. The budget is not its only problem: a permissioned payload tier cannot clear a 0.75 verifiability floor, cannot erase, and cannot answer inside 250 ms, so C1, C2 and C4 fall with C5. Committing every class per-item needs 42.25 KB, over budget by a factor of five. Batching every class is cheap at 0.66 KB but fails C1 for the four classes demanding $\theta = 0.90$. Pure off-chain storage fails both C1 and C3.

Table 3. Policies compared at B = 8 KB per patient-year

Policy	Ledger KB	J	Gran. loss	Status
All payload, permissioned	38 866	0.584	68.8	C1, C2, C4, C5 fail
All per-item commitment	42.25	0.082	6.25	C5 fails
All batched commitment	0.66	0.063	31.3	C1 fails
All off-chain only	0.00	0.059	125.1	C1, C3 fail
Model optimum	7.64	0.172	25.5	feasible

The unconstrained objective values of the infeasible policies are lower than the optimum's, which is exactly as it should be. Feasibility costs something. A design that ignores C1 through C5 will always look cheaper on paper, and that is precisely how ledger pilots reach procurement before failing legal review.

3.3 Volume Rather Than Sensitivity Governs Batching

The optimum assigns per-item commitments to seven classes and batched commitments to two: Observation and Access log entry. Those are the two classes with the highest instance counts, 120 and 180 per patient-year. They are not the two least sensitive classes — Consent grant carries the lowest sensitivity in Table 1 at $\sigma = 0.15$, yet it receives a per-item commitment, because $\theta = 0.90$ forbids anything coarser.

Granularity is therefore rationed by volume subject to a verifiability floor, not by how much anyone would mind the data leaking. That result is not obvious from the outside, and it contradicts the intuition that cheap proofs should be spent on unimportant data.

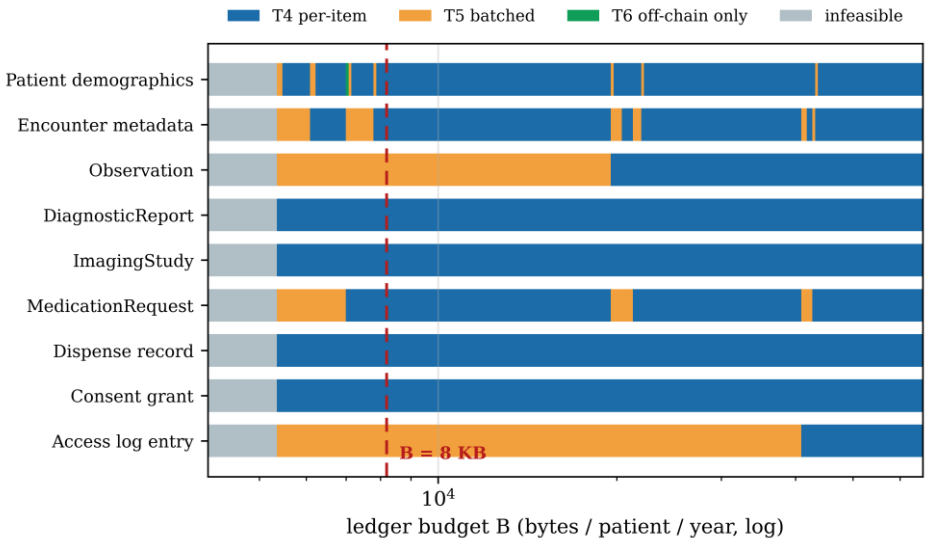


Fig. 3. Optimal placement as the ledger budget grows. Per-item commitments are bought first for low-volume classes; high-volume classes migrate last.

Fig. 3 traces how the placement evolves as B grows from 4 KB to 64 KB. The optimiser buys per-item commitments first for low-volume classes, and the high-volume classes migrate last, at budgets several times the minimum. An architect reading this chart can see directly what an extra kilobyte of ledger budget buys and where it stops mattering.

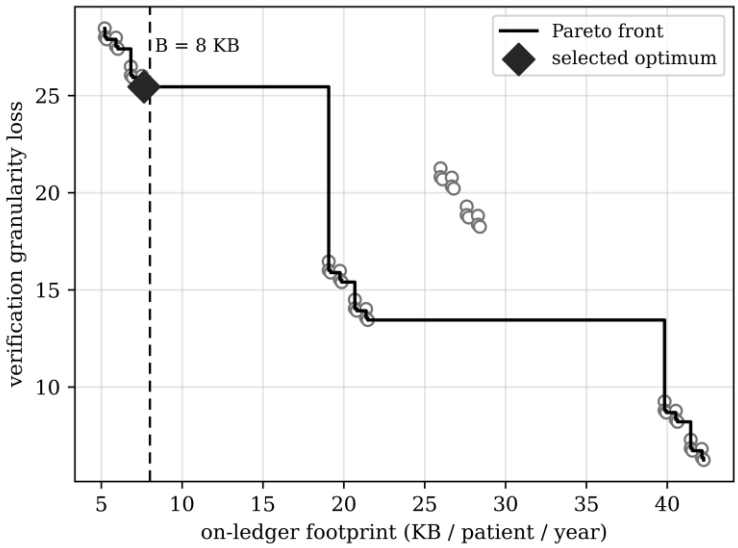


Fig. 4. All 48 admissible assignments. The budget line selects one point on the front; relaxing it buys granularity at a decreasing rate.

Fig. 4 shows all 48 admissible assignments in the footprint–granularity plane together with the Pareto front. The front is short and steep near the minimum feasible budget, then flattens: past roughly 42 KB per patient-year, further ledger spending buys almost no additional verification granularity, because every class that can use a per-item commitment already has one.

3.4 Sensitivity to Parameter Error

Fair question, and the honest answer needs numbers rather than a promise. Tables 1 and 2 hold estimates. An estimate that drifts twenty-five per cent, in a health system nobody has instrumented yet, is ordinary rather than alarming. So the model was rerun ten thousand times with every payload size, instance count and sensitivity weight drawn independently from ±25 per cent of its tabulated value. θ, e, p and λ stayed fixed throughout, because those are duties and service targets rather than measurements — a jurisdiction does not jitter its own retention law.

Two of the three headline results never broke. Pruning stayed above four orders of magnitude in every draw. All four uniform policies stayed infeasible in every draw. Neither result touches the numbers at all: both turn on which tier clears a verifiability floor and which does not, and moving a payload size does not move a floor.

The third result is softer, and saying so costs nothing. Batching landed on the highest-volume classes in 76.1 per cent of draws. In the remaining quarter the instance counts had shuffled far enough that a mid-volume class captured the batched commitment instead. Volume rations granularity robustly; which classes end up batched does not survive the same

abuse. An architect substituting their own counts should expect the assignment to move even where the principle holds.

A reference implementation accompanies this paper. It reads Tables 1 and 2 and recomputes the funnel, Table 3, the optimum and the budget sweep from scratch, so every figure reported above can be checked against the parameters that produced it rather than taken on trust.

4. Discussion and Future Trends

4.1 Implications for Practice

Three implications follow for anyone building a hybrid ledger service in a regulated health system.

First, publish the envelope before the architecture. If constraints eliminate 99.9995 percent of the design space, then a paper reporting a placement without reporting its constraints has withheld the part that did the reasoning. Future system descriptions should tabulate θ, e, p and λ as routinely as they already tabulate throughput.

Second, treat batching as a budgeted resource. The literature usually sells batched commitments as a scalability trick. Here they behave like currency — spent where volume runs highest, withheld where an auditor needs to isolate one record — and the allocation shifts predictably as the budget moves.

Third, be careful comparing objective values across feasibility boundaries. Every infeasible policy in Table 3 scores better than the optimum. Any benchmark that ranks architectures on cost or latency alone, without checking the envelope first, will reliably recommend designs that cannot be deployed.

4.2 Threats to Validity

Parameters are illustrative. Nothing here claims that the specific numbers in Tables 1 and 2 describe any particular health system. The qualitative results follow from the ordering of the parameters rather than their exact values, and Section 3.4 puts numbers on how far that holds: severe pruning and the infeasibility of every uniform policy survived all ten thousand perturbed draws, while the batching assignment survived roughly three in four. A system whose ordering differs will reach different conclusions, and the model provides the instrument for finding that out.

The cost model is linear and additive, which understates interactions. Storing a class off-chain changes the operational burden of key management for every other class sharing a key hierarchy [37], [38], and attribute-based access control shifts the exposure coefficients in ways a single ρ_t cannot capture [39]. Exposure itself collapses to a scalar here; a serious risk treatment would use a distribution.

The model also assumes placement is static per class. Real records change status — a study under active care is not a study under statutory retention twelve years later — and a lifecycle-aware version of this model, in which classes migrate between tiers on defined

events, is the obvious next step. That version is not a straightforward extension: migration costs and backup semantics interact awkwardly with erasure duties [40].

Finally, no deployed system has yet tested the model. Nothing here rests on measured latencies, throughput or storage figures from running infrastructure, and the reader should treat every number here as a consequence of stated assumptions.

5. Conclusion

Placement in hybrid blockchain health systems is a constrained optimisation problem, and naming it as one changes what an architect can say about their design. This paper gave the model, located it as a multiple-choice knapsack instance, and instantiated it on nine FHIR-aligned classes. The compliance envelope removed five orders of magnitude from the design space; all four uniform policies in common use turned out infeasible; and the optimum rationed proof granularity by data volume under a verifiability floor. Perturbing every estimated parameter by a quarter left the first two findings untouched and the third standing in three draws out of four. The next step is the lifecycle-aware formulation, together with a Fabric testbed that will replace the illustrative latency and storage parameters with measured ones.

Acknowledgments

Not applicable.

Author Contributions

Yerassyl Aikenov: Conceptualization, Methodology, Software, Formal analysis, Visualization, Writing – original draft, Writing – review & editing.

Conflicts of Interest

The author declares no conflicts of interest.

Data Availability Statement

All parameter values used by the model are given in full in Tables 1 and 2. The reference implementation that reproduces the reported results is available at https://github.com/yerassylaikenov/JSDT_article2_Aikenov.git.

List of Supplementary Materials

Supplementary File S1: reference implementation (source code) that reproduces the constraint funnel, Table 3, the optimum and the budget sweep shown in Figures 1–4.

References

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation). (2016). *Official Journal of the European Union, L119*, 1–88.

[2] Politou, E., Alepis, E., & Patsakis, C. (2018). Forgetting personal data and revoking consent under the GDPR: Challenges and proposed solutions. *Journal of Cybersecurity, 4*(1), tyy001. https://doi.org/10.1093/cybsec/tyy001

[3] Code of the Republic of Kazakhstan No. 360-VI "On Public Health and the Healthcare System". (2020, July 7). https://adilet.zan.kz

[4] Nurgaliyeva, Z., Spatayev, Y., Syla, S., & Yessenbayev, B. (2024). Paving the way to establishing the digital-friendly health and care information model in Kazakhstan. *International Journal of Medical Informatics, 192*, 105610. https://doi.org/10.1016/j.ijmedinf.2024.105610

[5] Wüst, K., & Gervais, A. (2018). Do you need a blockchain? In *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)* (pp. 45–54). IEEE. https://doi.org/10.1109/CVCBT.2018.00011

[6] Eberhardt, J., & Tai, S. (2017). On or off the blockchain? Insights on off-chaining computation and data. In *Service-Oriented and Cloud Computing (ESOCC 2017)* (Lecture Notes in Computer Science, Vol. 10465, pp. 3–15). Springer.

[7] Xu, X., Weber, I., Staples, M., Zhu, L., Bosch, J., Bass, L., Pautasso, C., & Rimba, P. (2017). A taxonomy of blockchain-based systems for architecture design. In *2017 IEEE International Conference on Software Architecture (ICSA)* (pp. 243–252). IEEE. https://doi.org/10.1109/ICSA.2017.33

[8] Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016). MedRec: Using blockchain for medical data access and permission management. In *2016 2nd International Conference on Open and Big Data (OBD)* (pp. 25–30). IEEE. https://doi.org/10.1109/OBD.2016.11

[9] Dagher, G. G., Mohler, J., Milojkovic, M., & Marella, P. B. (2018). Ancile: Privacy-preserving framework for access control and interoperability of electronic health records using blockchain technology. *Sustainable Cities and Society, 39*, 283–297. https://doi.org/10.1016/j.scs.2018.02.014

[10] Xia, Q., Sifah, E. B., Smahi, A., Amofa, S., & Zhang, X. (2017). BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. *Information, 8*(2), 44. https://doi.org/10.3390/info8020044

[11] Fan, K., Wang, S., Ren, Y., Li, H., & Yang, Y. (2018). MedBlock: Efficient and secure medical data sharing via blockchain. *Journal of Medical Systems, 42*(8), 136. https://doi.org/10.1007/s10916-018-0993-7

[12] Yue, X., Wang, H., Jin, D., Li, M., & Jiang, W. (2016). Healthcare data gateways: Found healthcare intelligence on blockchain with novel privacy risk control. *Journal of Medical Systems, 40*(10), 218. https://doi.org/10.1007/s10916-016-0574-6

[13] Nguyen, D. C., Pathirana, P. N., Ding, M., & Seneviratne, A. (2019). Blockchain for secure EHRs sharing of mobile cloud based e-health systems. *IEEE Access, 7*, 66792–66806. https://doi.org/10.1109/ACCESS.2019.2917555

[14] Androulaki, E., Barger, A., Bortnikov, V., Cachin, C., Christidis, K., De Caro, A., Enyeart, D., Ferris, C., Laventman, G., Manevich, Y., Muralidharan, S., Murthy, C., Nguyen, B., Sethi, M., Singh, G., Smith, K., Sorniotti, A., Stathakopoulou, C., Vukolić, M., … Yellick, J. (2018). Hyperledger Fabric: A distributed operating system for permissioned blockchains. In *Proceedings of the Thirteenth EuroSys Conference* (Article 30). ACM. https://doi.org/10.1145/3190508.3190538

[15] Stamatellis, C., Papadopoulos, P., Pitropakis, N., Katsikas, S., & Buchanan, W. J. (2020). A privacy-preserving healthcare framework using Hyperledger Fabric. *Sensors, 20*(22), 6587. https://doi.org/10.3390/s20226587

[16] Xu, X., Sun, G., Luo, L., Cao, H., Yu, H., & Vasilakos, A. V. (2021). Latency performance modeling and analysis for Hyperledger Fabric blockchain network. *Information Processing & Management, 58*(1), 102436. https://doi.org/10.1016/j.ipm.2020.102436

[17] Jayabalan, J., & Jeyanthi, N. (2022). Scalable blockchain model using off-chain IPFS storage for healthcare data security and privacy. *Journal of Parallel and Distributed Computing, 164*, 152–167. https://doi.org/10.1016/j.jpdc.2022.03.009

[18] Zhan, W., Chen, C.-L., Weng, W., Tsaur, W.-J., Lim, Z.-Y., & Deng, Y.-Y. (2022). Incentive EMR sharing system based on consortium blockchain and IPFS. *Healthcare, 10*(10), 1840. https://doi.org/10.3390/healthcare10101840

[19] Benet, J. (2014). *IPFS—Content addressed, versioned, P2P file system* (arXiv:1407.3561). arXiv.

[20] Steichen, M., Fiz, B., Norvill, R., Shbair, W., & State, R. (2018). Blockchain-based, decentralized access control for IPFS. In *2018 IEEE International Conference on Blockchain* (pp. 1499–1506). IEEE.

[21] Kumar, R., & Tripathi, R. (2021). Towards design and implementation of security and privacy framework for Internet of Medical Things (IoMT) by leveraging blockchain and IPFS technology. *The Journal of Supercomputing, 77*(8), 7916–7955.

[22] Hasselgren, A., Kralevska, K., Gligoroski, D., Pedersen, S. A., & Faxvaag, A. (2020). Blockchain in healthcare and health sciences—A scoping review. *International Journal of Medical Informatics, 134*, 104040. https://doi.org/10.1016/j.ijmedinf.2019.104040

[23] Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019). Blockchain technology in healthcare: A systematic review. *Healthcare, 7*(2), 56. https://doi.org/10.3390/healthcare7020056

[24] Lopez, L. J. R., Millan Mayorga, D., Martinez Poveda, L. H., Amaya, A. F. C., & Rojas Reales, W. (2024). Hybrid architectures used in the protection of large healthcare records based on cloud and blockchain integration: A review. *Computers, 13*(6), 152. https://doi.org/10.3390/computers13060152

[25] Ateniese, G., Magri, B., Venturi, D., & Andrade, E. (2017). Redactable blockchain—or—rewriting history in Bitcoin and friends. In *2017 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 111–126). IEEE.

[26] Derler, D., Samelin, K., Slamanig, D., & Striecks, C. (2019). Fine-grained and controlled rewriting in blockchains: Chameleon-hashing gone attribute-based. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*. Internet Society.

[27] Deuber, D., Magri, B., & Thyagarajan, S. A. K. (2019). Redactable blockchain in the permissionless setting. In *2019 IEEE Symposium on Security and Privacy (SP)* (pp. 124–138). IEEE. https://doi.org/10.1109/SP.2019.00039

[28] Politou, E., Casino, F., Alepis, E., & Patsakis, C. (2021). Blockchain mutability: Challenges and proposed solutions. *IEEE Transactions on Emerging Topics in Computing, 9*(4), 1972–1986.

[29] Merkle, R. C. (1988). A digital signature based on a conventional encryption function. In *Advances in Cryptology—CRYPTO '87* (Lecture Notes in Computer Science, Vol. 293, pp. 369–378). Springer.

[30] Kissel, R., Regenscheid, A., Scholl, M., & Stine, K. (2014). *Guidelines for media sanitization* (NIST Special Publication 800-88 Rev. 1). National Institute of Standards and Technology.

[31] Sinha, P., & Zoltners, A. A. (1979). The multiple-choice knapsack problem. *Operations Research, 27*(3), 503–515. https://doi.org/10.1287/opre.27.3.503

[32] Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. W. H. Freeman.

[33] Martello, S., & Toth, P. (1990). *Knapsack problems: Algorithms and computer implementations*. Wiley.

[34] Pisinger, D. (1995). A minimal algorithm for the multiple-choice knapsack problem. *European Journal of Operational Research, 83*(2), 394–410.

[35] HL7 International. (2023). *HL7 FHIR Release 5*. https://hl7.org/fhir/R5/

[36] Zhang, P., White, J., Schmidt, D. C., Lenz, G., & Rosenbloom, S. T. (2018). FHIRChain: Applying blockchain to securely and scalably share clinical data. *Computational and Structural Biotechnology Journal, 16*, 267–278. https://doi.org/10.1016/j.csbj.2018.07.004

[37] Dworkin, M. (2007). *Recommendation for block cipher modes of operation: Galois/Counter Mode (GCM) and GMAC* (NIST Special Publication 800-38D). National Institute of Standards and Technology.

[38] Shamir, A. (1979). How to share a secret. *Communications of the ACM, 22*(11), 612–613.

[39] Goyal, V., Pandey, O., Sahai, A., & Waters, B. (2006). Attribute-based encryption for fine-grained access control of encrypted data. In *Proceedings of the 13th ACM Conference on Computer and Communications Security (CCS)* (pp. 89–98). ACM.

[40] Politou, E., Michota, A., Alepis, E., Pocs, M., & Patsakis, C. (2018). Backups and the right to be forgotten in the GDPR: An uneasy relationship. *Computer Law & Security Review, 34*(6), 1247–1257. https://doi.org/10.1016/j.clsr.2018.08.006

Miraim Atanayeva*
Chief Senior Director, Project Monitoring and Management Center
Digital Government Support Center, Kazakhstan
m.atanayeva@govtec.kz; https://orcid.org/0000-0002-3067-356X
Ulzhan Ospanova
Senior Director, Project Monitoring and Management Center,
Digital Government Support Center, Kazakhstan
u.ospanova8@gmail.com, u.ospanova@govtec.kz; https://orcid.org/0000-0003-0427-7931
Zhanna Isabekova
Analyst at the Project Monitoring and Management Center
Digital Government Support Center, Kazakhstan
zh.issabekova@govtec.kz

AN INTEGRATED PROJECT MANAGEMENT MATURITY MODEL FOR DIGITAL QAZAQSTAN INITIATIVES: EMBEDDING CYBERSECURITY, AI GOVERNANCE AND FORWARD-LOOKING CYBER-RISK REQUIREMENTS INTO ST RK ISO 21502-2022

Abstract: The Nationwide Strategy "Digital Qazaqstan" until 2029, approved by Presidential Decree No. 1311 of 9 June 2026, and its Action Plan, approved by Government Resolution No. 760 of 26 August 2026, commit central and local state bodies and national companies to 136 measures, a large share of which deploy artificial intelligence (AI) in citizen-facing, business-facing and intra-governmental services. Project offices that are now translating these measures into implementation plans must simultaneously satisfy the project-management guidance of ST RK ISO 21502-2022, the information-security requirements of ST RK ISO/IEC 27001-2023 and the Unified Requirements on ICT and information security, and the emerging AI-management requirements of ST RK ISO/IEC 42001-2025, ISO/IEC 42005:2025, ISO/IEC 42006:2025 and the Law of the Republic of Kazakhstan "On Artificial Intelligence". Existing project-management maturity models are domain-neutral and contain no cybersecurity or AI-governance criteria, whereas cybersecurity and AI maturity models ignore the project life cycle. This article proposes an integrated project management maturity model (IPM3-DQ) that extends a ST RK ISO 21502-2022 practice-based baseline with three cross-cutting domains: cybersecurity (S), AI governance (A) and forward-looking cyber risk (F). The model uses a uniform 0–5 anchored scale, AHP-derived weights differentiated by project class, and a non-compensatory gating rule under which the maturity level of a project involving critically important digital objects or AI systems cannot exceed its security or AI-governance level by more than one level. A worked example of a G2C AI-assistant project shows that the conventional arithmetic-mean approach classifies the project as "Managed" (3.67), whereas the integrated model places it at "Defined" (2.67) with AI governance as the binding constraint. A validation protocol based on expert panels, Kendall's W, Cronbach's alpha and intraclass correlation is specified. The model gives project offices of state bodies a single instrument for self-assessment, audit

and semi-annual reporting under the Action Plan, and provides the regulator with a maturity-based lens for the national cybersecurity standardisation programme.

Keywords: *project management maturity; ST RK ISO 21502-2022; Digital Qazaqstan; cybersecurity; ISO/IEC 27001; ISO/IEC 42001; AI management system; AI impact assessment; cyber-risk forecasting; public sector; Kazakhstan.*

Интегрированная модель зрелости проектного управления для инициатив Digital Qazaqstan: встраивание требований кибербезопасности, управления искусственным интеллектом и прогнозной оценки киберрисков в СТ РК ISO 21502-2022

Аннотация: Общенациональная стратегия «Digital Qazaqstan» до 2029 года, утверждённая Указом Президента Республики Казахстан от 9 июня 2026 года № 1311, и План действий по её реализации, утверждённый постановлением Правительства Республики Казахстан от 26 августа 2026 года № 760, возлагают на центральные и местные государственные органы и национальные компании исполнение 136 мероприятий, значительная часть которых предусматривает внедрение искусственного интеллекта (ИИ) в сервисы для граждан, бизнеса и внутригосударственного взаимодействия. Проектные офисы, которые в настоящее время переводят эти мероприятия в планы реализации, должны одновременно соблюдать требования руководства по управлению проектами СТ РК ISO 21502-2022, требования информационной безопасности СТ РК ISO/IEC 27001-2023 и Единых требований в области ИКТ и обеспечения информационной безопасности, а также формирующиеся требования к управлению ИИ, установленные СТ РК ISO/IEC 42001-2025, ISO/IEC 42005:2025, ISO/IEC 42006:2025 и Законом Республики Казахстан «Об искусственном интеллекте». Существующие модели зрелости проектного управления нейтральны к предметной области и не содержат критериев кибербезопасности и управления ИИ, тогда как модели зрелости кибербезопасности и ИИ не учитывают жизненный цикл проекта. В статье предложена интегрированная модель зрелости проектного управления (IPM3-DQ), расширяющая базовый уровень, построенный на практиках СТ РК ISO 21502-2022, тремя сквозными доменами: кибербезопасность (S), управление ИИ (A) и прогнозная оценка киберрисков (F). Модель использует единую шкалу 0–5 с закреплёнными дескрипторами, весовые коэффициенты, полученные методом анализа иерархий (AHP) и дифференцированные по классам проектов, а также некомпенсаторное правило ограничения, согласно которому уровень зрелости проекта, затрагивающего критически важные цифровые объекты или системы ИИ, не может превышать уровень его кибербезопасности или управления ИИ более чем на один уровень. Рабочий пример проекта ИИ-ассистента G2C показывает, что традиционный подход на основе среднего арифметического относит проект к уровню «Управляемый» (3,67), тогда как интегрированная модель определяет его как

«Регламентированный» (2,67), при этом ограничивающим фактором выступает управление ИИ. Описан протокол валидации на основе экспертных панелей, коэффициента конкордации Кендалла (W), альфы Кронбаха и внутриклассовой корреляции. Модель предоставляет проектным офисам государственных органов единый инструмент самооценки, аудита и полугодовой отчётности в рамках Плана действий, а регулятору — основанный на зрелости взгляд на национальную программу стандартизации в сфере кибербезопасности.

Ключевые слова: *зрелость проектного управления; СТ РК ISO 21502-2022; Digital Qazaqstan; кибербезопасность; ISO/IEC 27001; ISO/IEC 42001; система менеджмента искусственного интеллекта; оценка воздействия ИИ; прогнозирование киберрисков; государственный сектор; Казахстан.*

Digital Qazaqstan бастамалары үшін жобалық басқарудың интеграцияланған кемелдік моделі: киберқауіпсіздік, жасанды интеллектті басқару және кибертәуекелдерді болжамды бағалау талаптарын ҚР СТ ISO 21502-2022 стандартына енгізу

Аңдатпа: Қазақстан Республикасы Президентінің 2026 жылғы 9 маусымдағы № 1311 Жарлығымен бекітілген 2029 жылға дейінгі «Digital Qazaqstan» жалпыұлттық стратегиясы және оны іске асыру жөніндегі Қазақстан Республикасы Үкіметінің 2026 жылғы 26 тамыздағы № 760 қаулысымен бекітілген Іс-шаралар жоспары орталық және жергілікті мемлекеттік органдар мен ұлттық компанияларға 136 іс-шараны орындауды жүктейді, олардың едәуір бөлігі азаматтарға, бизнеске және мемлекетішілік өзара іс-қимылға арналған сервистерге жасанды интеллектті (ЖИ) енгізуді көздейді. Қазіргі уақытта осы іс-шараларды іске асыру жоспарларына айналдырып жатқан жобалық офистер бір мезгілде ҚР СТ ISO 21502-2022 жобаларды басқару жөніндегі нұсқаулықтың талаптарын, ҚР СТ ISO/IEC 27001-2023 және Ақпараттық-коммуникациялық технологиялар және ақпараттық қауіпсіздікті қамтамасыз ету саласындағы бірыңғай талаптардың ақпараттық қауіпсіздік талаптарын, сондай-ақ ҚР СТ ISO/IEC 42001-2025, ISO/IEC 42005:2025, ISO/IEC 42006:2025 стандарттарымен және «Жасанды интеллект туралы» Қазақстан Республикасының Заңымен белгіленген, қалыптасып келе жатқан ЖИ-ді басқару талаптарын сақтауға тиіс. Жобалық басқарудың қолданыстағы кемелдік модельдері пәндік салаға бейтарап және киберқауіпсіздік пен ЖИ-ді басқару критерийлерін қамтымайды, ал киберқауіпсіздік пен ЖИ кемелдік модельдері жобаның өмірлік циклін ескермейді. Мақалада ҚР СТ ISO 21502-2022 практикаларына негізделген базалық деңгейді үш өтпелі доменмен — киберқауіпсіздік (S), ЖИ-ді басқару (A) және кибертәуекелдерді болжамды бағалау (F) — кеңейтетін жобалық басқарудың интеграцияланған кемелдік моделі (IPM3-DQ) ұсынылған. Модельде бекітілген дескрипторлары бар бірыңғай 0–5 шкаласы, иерархияларды талдау әдісімен (AHP)

алынған және жоба сыныптары бойынша сараланған салмақ коэффициенттері, сондай-ақ аса маңызды цифрлық объектілерді немесе ЖИ жүйелерін қамтитын жобаның кемелдік деңгейі оның киберқауіпсіздік немесе ЖИ-ді басқару деңгейінен бір деңгейден артық аса алмайтын өтемсіз шектеу ережесі қолданылады. G2C ЖИ-ассистенті жобасының жұмыс мысалы дәстүрлі орташа арифметикалық тәсіл жобаны «Басқарылатын» деңгейге (3,67) жатқызатынын, ал интеграцияланған модель оны ЖИ-ді басқару шектеуші фактор болып табылатын «Регламенттелген» деңгей (2,67) ретінде анықтайтынын көрсетеді. Сараптамалық панельдерге, Кендаллдың конкордация коэффициентіне (W), Кронбах альфасына және сыныпішілік корреляцияға негізделген валидация хаттамасы сипатталған. Модель мемлекеттік органдардың жобалық офистеріне Іс-шаралар жоспары шеңберінде өзін-өзі бағалаудың, аудиттің және жартыжылдық есептіліктің бірыңғай құралын, ал реттеушіге киберқауіпсіздік саласындағы ұлттық стандарттау бағдарламасына кемелдікке негізделген көзқарасты ұсынады.

Түйін сөздер: *жобалық басқарудың кемелдігі; ҚР СТ ISO 21502-2022; Digital Qazaqstan; киберқауіпсіздік; ISO/IEC 27001; ISO/IEC 42001; жасанды интеллект менеджменті жүйесі; ЖИ әсерін бағалау; кибертәуекелдерді болжау; мемлекеттік сектор; Қазақстан.*

1. Introduction

Public administration in Kazakhstan has moved decisively towards project-based delivery of strategic initiatives. The Model Regulation of Project Management of State Bodies, approved by Order No. 2 of the Minister of National Economy of 13 January 2022 and subsequently amended, designates ST RK ISO 21502-2022 as the reference guidance for project management in state bodies [4, 5]. The institutional importance of this shift was confirmed by a recent panel study of sixteen regions of Kazakhstan over 2005–2024, which found that the development of government project offices is positively and significantly associated with regional economic outcomes, while digitalisation on its own shows no robust causal effect unless it is matched by organisational capacity [36]. Technology, in other words, does not substitute for management maturity.

This finding acquires immediate practical weight with the adoption of the Nationwide Strategy for Large-Scale Digitalisation and Total Implementation of Artificial Intelligence Technologies "Digital Qazaqstan" until 2029 (Presidential Decree No. 1311 of 9 June 2026) [1] and of its Action Plan (Government Resolution No. 760 of 26 August 2026) [2]. The Action Plan comprises 136 measures grouped into three directions and twenty priority programmes ("icebreakers"). Responsible bodies must report on implementation twice a year, by 5 July and 20 December, to the Ministry of Artificial Intelligence and Digital Development, which consolidates the information for the Government Office. A large share of the measures are AI-intensive: a universal AI assistant eGov GPT (measure 26), AI assistants for civil servants in G2G, G2B and G2C interactions (measures 127–129), an AI

agent for the e-Otinish petition system (measure 136), AI-based functional analysis of state functions (measures 130–131), and sectoral targets such as a share of clinical decisions supported by AI of at least 80 % by 2028. In parallel, Icebreaker 19 "Cybersecurity" sets a target protection level of at least 90 % for critically important digital objects and tasks the Ministry, the National Security Committee, the Ministry of Trade and Integration and the Digital Government Support Centre with developing national cybersecurity standards through 2029 (measure 117), while Icebreaker 20 "State as a Platform" mandates the migration of state information systems to the QazTech platform (measure 125).

Project offices of state bodies and national companies are therefore drafting implementation plans for projects that are, at the same time, digital-transformation projects, information-security projects and—increasingly—AI projects. The guidance they are instructed to follow, ST RK ISO 21502-2022, is deliberately domain-neutral: it describes integrated and subject-specific management practices but says nothing about threat modelling, security attestation, AI impact assessment or supply-chain risk of machine-learning models. The relevant requirements exist, but in separate instruments: ST RK ISO/IEC 27001-2023 and ISO/IEC 27005:2022 for information security [8, 9]; the Unified Requirements on ICT and information security approved by Government Resolution No. 832 of 20 December 2016 [7]; ST RK ISO/IEC 42001-2025 for AI management systems [10], with national adoption of ISO/IEC 42005 (AI system impact assessment) and ISO/IEC 42006 (requirements for AI management system auditors) planned for 2026 and 2027 respectively [11, 12]; and the Law "On Artificial Intelligence" No. 230-VIII of 17 November 2025, which classifies AI systems by degree of impact and allocates responsibility across the life cycle [3]. No instrument tells a project office how mature its project management must be, as a whole, for these requirements to be met reliably.

The academic literature mirrors this fragmentation. Project management maturity models (PMMMs) such as OPM3, P3M3, CMMI-based models and the Kerzner PMMM have been compared extensively [13] but do not include cybersecurity or AI-governance dimensions; the single published attempt to add cybersecurity extensions to a PMMM for critical-infrastructure projects [21] predates ISO 21502:2020, NIST CSF 2.0 and ISO/IEC 42001. Cybersecurity maturity models [22, 23] and AI maturity models for the public sector [24] are organisational rather than project-oriented. Methods for forecasting cyber threats [27, 28] have not been embedded in project planning. For Kazakhstan, the only baseline measurement of project management maturity dates from 2015 [37], and recent work on the digital maturity of project offices of state bodies [38] contains no security or AI criteria. A further, methodological weakness is shared by most maturity instruments: scores are aggregated by simple arithmetic means with equal weights, without a test of expert agreement or scale reliability, so that a project with excellent scheduling and absent security governance can obtain a respectable "high" rating.

The objective of this article is to propose and specify an integrated project management maturity model for projects implemented under the Digital Qazaqstan Action

Plan—hereafter IPM3-DQ—that (i) retains ST RK ISO 21502-2022 as the normative baseline mandated for state bodies; (ii) embeds cybersecurity requirements derived from ST RK ISO/IEC 27001-2023, ISO/IEC 27005:2022 and the Unified Requirements as a cross-cutting domain; (iii) embeds AI-governance requirements derived from ISO/IEC 42001, 42005, 42006 and the Law on AI as a second cross-cutting domain; (iv) introduces a forward-looking cyber-risk domain that rewards the use of threat forecasting and probabilistic risk quantification in project planning; and (v) replaces the arithmetic-mean index with a weighted, non-compensatory aggregation whose parameters can be derived and validated with established expert-elicitation and psychometric procedures. The contribution is a design artefact: the model, its indicator system, its aggregation logic, a worked example, and a validation protocol.

The remainder of the article is organised as follows. Section 2 reviews related work and identifies the research gap. Section 3 summarises the regulatory and standards baseline of Kazakhstan on which the model is built. Section 4 specifies the model. Section 5 presents a worked example for an AI-assistant project under Icebreaker 20 and the validation protocol. Section 6 discusses implications, positioning and limitations, and Section 7 concludes.

2. Background and related work

2.1. Project management maturity models and the public sector

Maturity models describe an organisation's capability as a staged progression, typically across five levels inherited from the Capability Maturity Model. Nikolaenko and Sidorov [13] compared nine PMMMs (CMMI, OPM3, IPMA Delta, P3M3, SPICE, PM², Kerzner PMMM, ProMMM and PMMMsm) and concluded that the most robust models are those anchored in an international body of knowledge; this supports anchoring a national model in ISO 21502 rather than in a proprietary framework. Early cross-industry evidence showed median maturity of only level 2 in most assessed components [14], and later work in the public sector of Pakistan established, with PLS-SEM on 425 respondents, a significant link between project management maturity and the reputation of public organisations [15]. Žurga's Total Project Management Maturity Model for Slovenian public administration [16] remains one of very few PMMMs designed for government, and systematic reviews report a persistent scarcity of empirical studies on PMMM adoption in the public sector. Recent extensions add new dimensions to PMMMs—most visibly sustainability—which provides a methodological precedent for extending a PMMM with cybersecurity and AI governance. Risk-specific maturity assessment for IT projects has been operationalised by Nikolaenko and Sidorov [17], and fuzzy-AHP has been used to prioritise security-relevant practices in software development [18]; both demonstrate that weighting and consistency checking are feasible within maturity instruments.

2.2. Cybersecurity in project management and cybersecurity maturity models

A 2025 scoping review of 64 studies [19] found that cybersecurity is widely acknowledged as a critical element of project risk management, yet most project managers lack the competence to operationalise it and organisational culture is the dominant moderating factor. Al-Janabi et al. [20] proposed a cyber-resilient IT project management framework that aligns the cyber-risk management process with the stages of the IT project life cycle, from business case and charter to closure; it is the closest process-level precedent for the mapping proposed here. Payette et al. [21] derived six cybersecurity themes from NIST, C2M2 and CERT-RMM and proposed extensions to the project sub-model of P3M3 for critical-infrastructure projects. On the cybersecurity side, Büyüközkan and Güler [22] conducted a systematic review and bibliometric analysis of cybersecurity maturity models and proposed a model with five dimensions, fifteen factors and five levels, noting that earlier models rarely offer an end-to-end organisational assessment and are seldom validated. NIST CSF 2.0 [23] introduces the Govern function and four implementation tiers that are structurally comparable to maturity levels. The harmonised structure shared by ISO/IEC 27001:2022 and ISO/IEC 42001:2023 (clauses 4–10, Plan-Do-Check-Act) facilitates integrated management systems and, by extension, integrated assessment.

2.3. AI management standards and AI governance maturity

ISO/IEC 42001:2023 is the first certifiable management-system standard for AI; its Annex A lists controls grouped by control objectives and Annex C enumerates AI-specific risk sources [10]. ISO/IEC 42005:2025 provides guidance on AI system impact assessment for individuals, groups and society and is referenced by 42001 for high-impact systems [11]; ISO/IEC 42006:2025 specifies requirements for bodies auditing and certifying AI management systems, including auditor competence and audit-time calculation [12]. NIST AI RMF 1.0 structures AI risk management around the Govern, Map, Measure and Manage functions [25], and the NIST taxonomy of adversarial machine learning distinguishes evasion, poisoning, privacy and misuse attacks on predictive and generative AI [30]. For the public sector specifically, Dreyling et al. [24] developed an AI maturity model through a design-science process with iterative expert refinement, drawing on Estonian experience, and Vatamanu and Tofan [26] showed, using the technology–organisation–environment framework, that AI in public administration requires strong governance structures and comprehensive security measures. The European AI Act [43] and the Kazakh Law on AI [3] both adopt impact-based classification of AI systems. Empirical studies of ISO/IEC 42001 implementation remain scarce, and we found no study of an integrated 27001 + 42001 system in a government setting.

2.4. Forward-looking cyber-risk assessment

Project risk management in ISO 21502 is retrospective in practice: risks are registered when identified and re-assessed on request. The cybersecurity literature offers established means of anticipation. Wang, Neil and Fenton [27] implemented and extended the FAIR model with Bayesian networks, obtaining results consistent with Monte-Carlo benchmarks and enabling expected-loss estimates that can inform contingency reserves.

Landauer et al. [28] reviewed time-series methods in security analytics from intrusion detection to long-horizon attack prediction. Cheimonidis and Rantos [29] systematised dynamic risk assessment models, including fuzzy Bayesian networks, which justify re-assessment at project stage gates rather than at project start only. For AI projects, Cinà et al. [31] surveyed training-data poisoning and defences, Wang et al. [32] formalised the three-tier large-language-model supply chain, and practitioner taxonomies—the OWASP Top 10 for LLM Applications [33] and MITRE ATLAS [34]—provide threat catalogues suitable for scenario analysis. Annual threat landscapes such as ENISA's [35] supply empirical base rates. None of these sources has been integrated into a maturity instrument for project offices.

2.5. Kazakhstan context and research gap

Kazakhstan ranks 24th of 193 countries in the 2024 UN E-Government Development Index and first among landlocked countries [39], and its eGov portal offered more than 1,200 services to over 15 million registered users by the end of 2023. Narbaev's 2015 assessment using the project sub-model of P3M3 reported an average maturity of about 2.4 out of 5 across Kazakh organisations [37]. Descriptive studies of project offices in ministries and regions identify systemic shortcomings, and Taizhanova et al. [38] examined the digital maturity of project offices of state bodies, but neither strand includes security or AI criteria. Country-level indices (EGDI, OECD Digital Government Index [40]) do not measure project-management or security maturity at organisational level.

Five gaps follow. First, PMMMs, cybersecurity maturity models and AI-governance maturity models evolve in parallel and have not been integrated. Second, no maturity model is built on ISO 21502, although Kazakhstan's Model Regulation makes it the reference standard. Third, most maturity instruments are psychometrically weak and rely on additive indices without reliability testing. Fourth, forecasting of cyber threats is absent from project maturity assessment. Fifth, no instrument connects project maturity to the Kazakh regulatory architecture created by the Law on AI, the Unified Requirements and the Digital Qazaqstan Action Plan. IPM3-DQ addresses all five.

3. Regulatory and standards baseline

Table 1 lists the instruments on which IPM3-DQ is built and the role each plays in the model. The selection follows a simple rule: the baseline must be the instrument that state bodies are already obliged to apply, and cross-cutting domains must be derived from instruments that are either adopted as national standards or are planned for adoption within the Strategy's horizon.

Table 1. Normative instruments underlying IPM3-DQ

Instrument	Status in Kazakhstan	Role in IPM3-DQ
Decree No. 1311 of 9 June 2026 (Strategy "Digital Qazaqstan" until 2029) [1]; Resolution No. 760 of 26 August 2026 (Action Plan, 136 measures, 20 icebreakers) [2]	In force; semi-annual reporting	Defines the project population, horizon (2029) and reporting cycle; Icebreakers 17, 19, 20 supply targets for domains S, A, F
Model Regulation of Project Management of State Bodies (Order MNE No. 2 of 13.01.2022, as amended) [4]	In force	Mandates ST RK ISO 21502-2022 and 21503 for state bodies; justifies the baseline
ST RK ISO 21502-2022 (ISO 21502:2020, IDT) [5, 6]	National standard, 2022	Baseline: ten practice groups P1–P10
ST RK ISO/IEC 27001-2023 (ISO/IEC 27001:2022, IDT) [8]; ISO/IEC 27005:2022 [9]	National standard, 2023; 27005 adopted as ST RK	Domain S: ISMS scope, risk assessment, Annex A controls
Unified Requirements on ICT and information security (Resolution No. 832 of 20.12.2016, as amended) [7]	In force	Domain S: mandatory security requirements for state information systems and critically important digital objects; attestation
Law "On Artificial Intelligence" No. 230-VIII of 17.11.2025 [3]	In force since 18.01.2026	Domain A: impact-based classification, life-cycle responsibility, national AI platform
ST RK ISO/IEC 42001-2025 (ISO/IEC 42001:2023) [10]	National standard, 2025	Domain A: AI management system roles, risk and impact processes, Annex A controls
ISO/IEC 42005:2025 [11]; ISO/IEC 42006:2025 [12]	National adoption planned 2026 and 2027	Domain A: impact assessment at initiation; audit readiness at closure
NIST CSF 2.0 [23]; NIST AI RMF 1.0 [25]; NIST AI 100-2 [30]; OWASP LLM Top	International references (non-binding)	Domain F: threat catalogues, base rates and forecasting inputs

Instrument	Status in Kazakhstan	Role in IPM3-DQ
10 [33]; MITRE ATLAS [34]; ENISA Threat Landscape [35]		

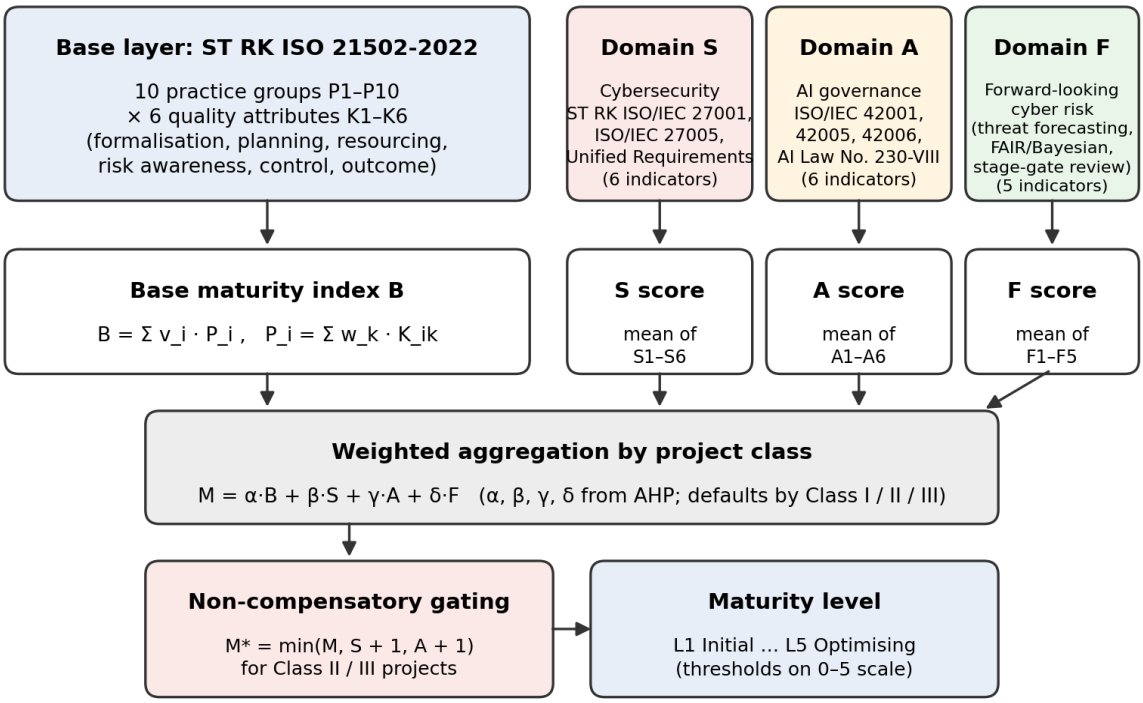
Note. Numbers of measures and icebreakers are cited from the text of Resolution No. 760 as published on adilet.zan.kz.

4. The IPM3-DQ model

4.1. Design approach and requirements

The model was developed as a design-science artefact. From the gaps identified in Section 2 and from the obligations of project offices under the Action Plan, five design requirements were derived: R1—the baseline must be ST RK ISO 21502-2022 without modification of its practice structure, so that the model is auditable against the standard state bodies already apply; R2—cybersecurity and AI governance must be assessed as cross-cutting domains along the project life cycle, not as additional practice groups, because they condition every practice; R3—the scale must be uniform and anchored, so that scores are comparable across practices, domains and assessors; R4—aggregation must be weighted and non-compensatory for projects whose failure has security consequences, so that strength in scheduling cannot mask weakness in security; R5—the instrument must be validatable with standard expert-elicitation and reliability procedures, and its weights must be re-derivable when standards change (for example upon national adoption of ISO/IEC 42005 and 42006). Figure 1 shows the resulting architecture.

Figure 1. Architecture of the IPM3-DQ model



4.2. Scope, assessment units and project classes

The object of assessment is the project management system of an organisation implementing measures of the Action Plan; the subject is the degree to which the practices of ST RK ISO 21502-2022 and the cross-cutting requirements are implemented. Assessment can be performed at three units: the organisation (central or local executive body, national company), its project office, or an individual project. Because the consequences of immaturity differ by project type, three project classes are distinguished and determine both the weights and the gating rule (Section 4.6):

- **Class I —** ordinary digital projects that do not process personal data at scale, do not involve critically important digital objects and do not deploy AI systems (e.g., measure 72, product identification system);
- **Class II —** projects involving state information systems, critically important digital objects or large-scale personal-data processing under the Unified Requirements (e.g., measure 125, migration of state IT systems to QazTech; measure 120, personal-data access control);
- **Class III —** projects deploying AI systems within the meaning of the Law on AI, with the sub-classification by degree of impact prescribed by that Law (e.g., measures 26, 127–129, 136; AI-supported clinical decision systems under Icebreaker 2).

4.3. Baseline layer: practice groups and quality attributes

The baseline layer preserves the practice structure of ST RK ISO 21502-2022. Ten practice groups P1–P10 are assessed: pre-project activities and initiation (P1); planning (P2); stakeholder engagement and communication (P3); resource and team management (P4); risk and issue management (P5); quality management (P6); schedule management (P7); cost management (P8); monitoring, control and change control (P9); and closure, reporting and lessons learned (P10). Each practice group is scored against six quality attributes K1–K6 that describe how well the practice is institutionalised rather than what the practice is—an important correction to earlier formulations in which "risk management" appeared simultaneously as an assessment criterion and as an assessed practice. The attributes are: K1 formalisation (documented procedure, templates, roles); K2 planning quality (completeness and currency of plans and baselines); K3 resourcing and competence (assigned, trained personnel and tooling); K4 risk awareness (practice-specific risks identified and owned); K5 measurement and control (metrics, reporting cadence, corrective action); K6 outcome (evidence that the practice produces its intended result). Appendix A gives the indicators and evidence sources for each practice group.

4.4. Uniform anchored scale

All practices, attributes and domain indicators are scored on the same 0–5 scale with anchored descriptors (Table 2). Using a single scale removes the inconsistency, common in practitioner instruments, between a 1–5 criterion scale, a 0–5 indicator scale and an interpretation table that begins at 0; it also makes the lowest level (absence of the practice) explicitly scorable.

Table 2. Anchored scoring scale

Score	Anchor	Generic descriptor
0	Absent	No evidence that the practice or control exists
1	Initial	Performed ad hoc by individuals; no documentation; results unpredictable
2	Forming	Partially documented; applied inconsistently across projects; roles informal
3	Defined	Documented, approved and applied in all projects; roles and templates defined
4	Managed	Applied systematically; measured with indicators; deviations trigger corrective action
5	Optimising	Continuously improved on the basis of measurements, audits, lessons learned and external benchmarks

4.5. Cross-cutting domains S, A and F

The three cross-cutting domains are assessed with indicators placed along the project life cycle, so that each indicator can be verified at the corresponding stage gate of the Model

Regulation. Table 3 specifies the indicators, their normative source and the evidence an assessor should request. Domain S (cybersecurity) translates ST RK ISO/IEC 27001-2023, ISO/IEC 27005:2022 and the Unified Requirements into six project-level indicators. Domain A (AI governance) translates ISO/IEC 42001, 42005 and 42006 and the Law on AI into six indicators and applies only to Class III projects. Domain F (forward-looking cyber risk) rewards practices that anticipate the threat landscape over the Strategy's horizon rather than registering threats already materialised; its five indicators are deliberately method-agnostic, accepting qualitative scenario analysis at lower scores and probabilistic quantification at higher scores.

Table 3. Indicators of the cross-cutting domains

Code	Indicator (life-cycle stage)	Normative source	Evidence; anchors 0 / 3 / 5
S1	Security classification and threat model at initiation	Unified Requirements; ISO/IEC 27005 cl. 7	Object classified (state IS / critically important digital object); threat model approved. 0 – none; 3 – classification only; 5 – approved threat model linked to charter
S2	Security requirements and ISMS scope in the project plan	ST RK ISO/IEC 27001 cl. 4, 6; Annex A	Security requirements traced to Annex A controls; ISMS scope covers project deliverables. 0 – none; 3 – generic checklist; 5 – traceable, risk-based requirements
S3	Secure procurement and supply-chain requirements	ISO/IEC 27001 A.5.19–5.23; Unified Requirements	Security clauses in tender documentation; supplier assessment. 0 – none; 3 – standard clauses; 5 – supplier risk assessment and acceptance tests
S4	Security testing and attestation before commissioning	Unified Requirements (compliance testing); ISO/IEC 27001 A.8.29	Penetration/compliance test reports; attestation certificate. 0 – none; 3 – testing without remediation tracking; 5 – attestation passed, residual risks accepted formally
S5	Incident-response and monitoring handover	ISO/IEC 27001 A.5.24–5.28; Action Plan measure 119	Connection to monitoring; incident playbooks; roles after go-live. 0 – none; 3 – playbooks exist; 5 – tested through exercises

Code	Indicator (life-cycle stage)	Normative source	Evidence; anchors 0 / 3 / 5
S6	Personal-data protection and access control	Law on Personal Data; Action Plan measure 120	Connection to "Personal-data access control" service; DPIA-type assessment. 0 – none; 3 – partial; 5 – connected and audited
A1	AI impact assessment and legal classification at initiation	ISO/IEC 42005; Law on AI No. 230-VIII	Impact assessment report; classification by degree of impact. 0 – none; 3 – classification without assessment; 5 – assessment approved, mitigations in charter
A2	AI management roles, policy and data governance	ISO/IEC 42001 cl. 5–6, Annex A; NIST AI RMF Govern/Map	Named AI owner; AI policy; training-data provenance and quality controls. 0 – none; 3 – roles only; 5 – documented data lineage, poisoning controls [30, 31]
A3	AI supply-chain and model provenance management	ISO/IEC 42001 A.10; LLM supply-chain [32]	Model/vendor register; licence and update terms; evaluation of third-party models. 0 – none; 3 – register; 5 – contractual controls and re-evaluation on update
A4	Adversarial and safety testing before deployment	NIST AI 100-2 [30]; OWASP LLM Top 10 [33]; MITRE ATLAS [34]	Red-team reports against the relevant threat catalogue. 0 – none; 3 – functional tests only; 5 – adversarial testing with remediation
A5	Human oversight, monitoring and incident handling in operation	ISO/IEC 42001 cl. 8–10; Law on AI	Oversight procedure; model performance and drift monitoring; AI incident process. 0 – none; 3 – oversight without monitoring; 5 – monitored with escalation
A6	Audit readiness of the AI management system	ISO/IEC 42006; ISO/IEC 42001 cl. 9	Internal audit records; readiness for third-party audit. 0 – none; 3 – internal review; 5 – audited by a competent body

Code	Indicator (life-cycle stage)	Normative source	Evidence; anchors 0 / 3 / 5
F1	Use of external threat intelligence in the risk register	ENISA Threat Landscape [35]; national monitoring (measure 119)	Risk register cites current threat landscape; refreshed each reporting period. 0 – none; 3 – cited once; 5 – refreshed semi-annually
F2	Probabilistic quantification of cyber risk	FAIR / Bayesian networks [27]	Expected-loss estimate for top cyber risks. 0 – none; 3 – qualitative likelihood–impact; 5 – quantified distribution
F3	Horizon scanning and scheduled re-assessment at stage gates	Dynamic risk assessment [29]; time-series prediction [28]	Re-assessment minutes at each gate; horizon to 2029. 0 – none; 3 – re-assessed on request; 5 – scheduled at every gate with trend data
F4	Scenario planning for emerging AI-specific threats	NIST AI 100-2; MITRE ATLAS	Documented scenarios (prompt injection, data poisoning, model theft) with triggers. 0 – none; 3 – generic; 5 – project-specific with indicators
F5	Budgeted cyber-risk reserve	ISO 21502 cl. 7.5–7.6; FAIR	Contingency reserve explicitly linked to cyber-risk estimates. 0 – none; 3 – lump-sum reserve; 5 – reserve derived from F2

4.6. Aggregation and interpretation

Scores are aggregated in four steps. First, the score of each practice group is the weighted mean of its six attributes:

$$P_i = \Sigma_k\, w_k \cdot K_{ik}, \quad \Sigma_k\, w_k = 1, \quad k = 1\ldots6 \tag{1}$$

Second, the baseline maturity index is the weighted mean of the ten practice groups:

$$B = \Sigma_i\, v_i \cdot P_i, \quad \Sigma_i\, v_i = 1, \quad i = 1\ldots10 \tag{2}$$

Third, each cross-cutting domain score (S, A, F) is the mean of its indicators, or a weighted mean where the expert panel has assigned indicator weights. Fourth, the integrated maturity index is a weighted combination whose weights depend on the project class:

$$M = \alpha \cdot B + \beta \cdot S + \gamma \cdot A + \delta \cdot F, \quad \alpha + \beta + \gamma + \delta = 1 \tag{3}$$

The weights w, v and (α, β, γ, δ) are derived by the Analytic Hierarchy Process [42] from pairwise comparisons made by an expert panel, with the consistency ratio of each comparison matrix required to be below 0.10. Default values proposed for use until a panel-

derived set is available are given in Table 4; they encode the policy judgement that security and AI governance should carry at least half of the weight for Class III projects.

Table 4. Default domain weights by project class

Project class	α (baseline B)	β (security S)	γ (AI governance A)	δ (forward-looking F)	Gating rule
Class I	0.70	0.20	—	0.10	none
Class II	0.50	0.30	—	0.20	M* = min(M, S + 1)
Class III	0.40	0.25	0.25	0.10	M* = min(M, S + 1, A + 1)

For Class II and Class III projects a non-compensatory gating rule is then applied: the final index M* cannot exceed the security score, or the AI-governance score, by more than one scale point:

$$M^* = \min(M, S + 1, A + 1) \quad (4)$$

The rationale is that a project whose security governance is "initial" cannot be "managed" as a whole, whatever its scheduling discipline; the one-point tolerance avoids the opposite error of reducing the whole assessment to its weakest indicator. Finally, M* is mapped to one of five maturity levels (Table 5). A single interpretation table is used for practices, domains and the integrated index, eliminating the duplicated and inconsistent level names found in earlier instruments.

Table 5. Interpretation of scores

Range of M* (or any score)	Level	Interpretation
0.00 – 1.49	L1 Initial	Practices largely absent or ad hoc; project outcomes unpredictable
1.50 – 2.49	L2 Forming	System being established; partial, inconsistent application
2.50 – 3.49	L3 Defined	Processes documented and applied; limited measurement
3.50 – 4.49	L4 Managed	Processes systematically applied and measured; corrective action routine
4.50 – 5.00	L5 Optimising	Continuous improvement; alignment with best international practice

4.7. Assessment procedure

The procedure follows five steps that map onto the stage gates of the Model Regulation: (1) scoping—select the assessment unit, determine the project class and the

applicable domains; (2) evidence collection—gather the documents listed in Table 3 and Appendix A, complemented by interviews; (3) scoring—at least two assessors score independently, after which disagreements are reconciled and inter-rater agreement is recorded; (4) computation—apply equations (1)–(4) with the current weight set; (5) reporting—produce the practice ranking, the domain profile and the level, together with prioritised recommendations, and feed the result into the semi-annual report under Resolution No. 760. Re-assessment is scheduled at each stage gate for Class II and III projects, which simultaneously satisfies indicator F3.

5. Worked example and validation protocol

5.1. Worked example: a G2C AI-assistant project

To illustrate the computation and the effect of the gating rule, consider a hypothetical Class III project of a central state body implementing measure 129 of the Action Plan (an AI assistant for civil servants in G2C interactions) on the QazTech platform. The scores below are illustrative values chosen to represent a project office with sound conventional project management and an immature AI-governance practice; they are not empirical data. Table 6 gives the attribute scores for the ten practice groups and Table 7 the domain indicator scores.

Table 6. Illustrative attribute scores of the baseline layer

Practice group	K1	K2	K3	K4	K5	K6	P_i
P1 Pre-project and initiation	4	4	4	4	4	4	4.00
P2 Planning	4	4	4	3	4	4	3.83
P3 Stakeholders and communication	4	4	3	3	4	4	3.67
P4 Resources and team	4	3	3	3	4	4	3.50
P5 Risk and issues	3	3	3	3	3	4	3.17
P6 Quality	4	4	3	3	4	4	3.67
P7 Schedule	4	4	4	4	4	4	4.00
P8 Cost	4	4	4	3	4	4	3.83
P9 Monitoring, control and change	4	3	3	3	4	4	3.50
P10 Closure and lessons learned	4	3	3	3	4	4	3.50
Baseline index B (equal weights)							3.67

Table 7. Illustrative domain indicator scores

Domain	Indicator scores	Domain score
S — cybersecurity	S1 = 3, S2 = 3, S3 = 2, S4 = 4, S5 = 3, S6 = 4	3.17
A — AI governance	A1 = 1, A2 = 2, A3 = 2, A4 = 1, A5 = 3, A6 = 1	1.67
F — forward-looking cyber risk	F1 = 2, F2 = 0, F3 = 1, F4 = 1, F5 = 1	1.00

Under the conventional approach—the arithmetic mean of practice scores with no security or AI component—the project obtains 3.67 and is classified as L4 "Managed". Under IPM3-DQ with the Class III default weights, equation (3) gives $M = 0.40 \times 3.67 + 0.25 \times 3.17 + 0.25 \times 1.67 + 0.10 \times 1.00 = 2.78$, and the gating rule (4) gives $M^* = \min(2.78, 4.17, 2.67) = 2.67$, i.e., L3 "Defined", with AI governance as the binding constraint. The domain profile makes the diagnosis explicit: the project has not performed an AI impact assessment (A1 = 1), has not tested the assistant against adversarial threat catalogues (A4 = 1) and has no audit trail (A6 = 1); the forward-looking domain shows that cyber risk is neither quantified nor budgeted (F2 = 0, F5 = 1). The recommendations that follow are concrete and standard-referenced: conduct an ISO/IEC 42005 impact assessment and classify the system under the Law on AI before the next stage gate; commission red-team testing against the OWASP LLM Top 10; and derive a cyber-risk reserve from a FAIR-type estimate. Had the same scores been treated as a Class I project, M would equal 3.30 (L3) without gating—showing that the class assignment, not only the scores, drives the result, which is the intended behaviour.

5.2. Validation protocol

Design requirement R5 demands that the instrument be validated rather than merely proposed. The following protocol is specified and is being implemented with project offices participating in the Action Plan; its results will be reported separately. (a) Content validity: a panel of 15–30 experts drawn from project offices of central state bodies, local executive bodies and national companies, from the Digital Government Support Centre and from information-security specialists, reviews the indicator set in two Delphi rounds; indicators are retained when at least 75 % of the panel rates them relevant. (b) Weights: the panel performs AHP pairwise comparisons for attributes, practice groups and domains separately for each project class; matrices with consistency ratio above 0.10 are returned for revision; group weights are obtained by geometric aggregation. (c) Expert agreement: Kendall's coefficient of concordance W is computed for the rankings of practices and domains, with $W \geq 0.5$ and $p < 0.05$ as the acceptance threshold. (d) Reliability: Cronbach's alpha is computed for the attribute scales within each practice group and for each domain (threshold

0.70). (e) Inter-rater agreement: the intraclass correlation coefficient ICC(2,k) is computed from independent scoring of the same projects by two or more assessors (threshold 0.75). (f) Pilot application: the instrument is applied to 3–5 projects of different classes drawn from Icebreakers 19 and 20, and the resulting profiles are compared with the outcomes of security attestation and with the semi-annual reports, providing a first test of criterion validity. The protocol follows the measurement-theory recommendations that recent psychometric evaluations of public-sector digital-maturity frameworks have shown to be routinely neglected.

6. Discussion

6.1. Positioning relative to existing models

Table 8 positions IPM3-DQ against the models reviewed in Section 2. Its distinctive features are the combination of a project-life-cycle baseline with security and AI domains, the forward-looking component and the non-compensatory aggregation. OPM3 and P3M3 provide richer process libraries but no security or AI content; C2M2 and NIST CSF 2.0 tiers describe organisational cyber capability without a project life cycle; the public-sector AI maturity model of Dreyling et al. [24] covers AI governance dimensions but not project delivery; the Secure-by-Design extensions of Payette et al. [21] are the closest precedent but predate the current standards, lack AI content and were not validated.

Table 8. Comparison of IPM3-DQ with related models

Model	Life-cycle baseline	Security domain	AI domain	Forward-looking	Gating	Public sector
OPM3 / P3M3 / Kerzner PMMM [13]	Yes (PMBOK / PRINCE2)	No	No	No	No	Generic
Secure-by-Design PMMM extensions [21]	Yes (P3M3)	Yes (NIST, C2M2)	No	No	No	Critical infra-structure
Cybersecurity maturity models [22]; NIST CSF 2.0 tiers [23]	No	Yes	Partial	Partial	No	Generic

Model	Life-cycle baseline	Security domain	AI domain	Forward-looking	Gating	Public sector
AI maturity model for the public sector [24]	No	Partial	Yes	No	No	Yes
Digital maturity of project offices of state bodies [38]	Partial	No	No	No	No	Yes (Kazakhstan)
IPM3-DQ (this article)	Yes (ST RK ISO 21502-2022)	Yes (27001/27005, Unified Requirements)	Yes (42001/42005/42006, Law on AI)	Yes	Yes (gating)	Yes (Kazakhstan)

6.2. Implications for project offices and the regulator

For project offices, the model converts a diffuse obligation—"take cybersecurity and AI requirements into account"—into eighteen verifiable indicators tied to the stage gates they already operate, and it produces a ranked list of practices to improve. Because the indicator evidence coincides with the documentation that attestation and audit bodies request, the assessment can be reused for compliance purposes. For the Ministry of Artificial Intelligence and Digital Development, the integrated index offers a candidate component of the digital development index of state bodies foreseen by measure 103 and a maturity-based lens for sequencing the national cybersecurity standardisation programme under measure 117: domains with persistently low scores across many bodies indicate where national standards, templates or training are most needed. For the Digital Government Support Centre, which is both a co-executor of measure 117 and the publisher of this journal, the model provides a basis for guidance material to project offices. The sectoral measures that already call for maturity and risk KPIs (measure 50 in the energy sector; measure 62 in agriculture; the data-governance maturity targets of Icebreaker 17) could adopt the same scale and aggregation logic, improving comparability across sectors.

6.3. Limitations

The study has four limitations. First, the model is presented with a worked example and a validation protocol, not with completed validation; the default weights in Table 4 are policy-informed proposals that the AHP procedure will replace. Second, the one-point

tolerance in the gating rule is a design choice whose sensitivity should be examined in the pilot. Third, the AI-governance domain depends on standards whose national adoption is still in progress (ISO/IEC 42005 and 42006), so indicators A1 and A6 will need to be re-anchored to the final texts of the national standards; the model's modular structure is intended to absorb such revisions. Fourth, self-assessment by project offices is exposed to optimism bias, which the two-assessor requirement and the ICC threshold mitigate but do not eliminate.

7. Conclusion

The Digital Qazaqstan Action Plan places 136 digital and AI measures in the hands of project offices that are obliged to follow ST RK ISO 21502-2022 but have no instrument for judging whether their project management is mature enough to meet the cybersecurity and AI-governance requirements those measures entail. This article has proposed IPM3-DQ, an integrated project management maturity model that keeps ST RK ISO 21502-2022 as its baseline and adds three cross-cutting domains—cybersecurity derived from ST RK ISO/IEC 27001-2023, ISO/IEC 27005 and the Unified Requirements; AI governance derived from ISO/IEC 42001, 42005, 42006 and the Law on AI; and forward-looking cyber risk—scored on one anchored scale, aggregated with class-specific weights and subjected to a non-compensatory gating rule. The worked example shows that the model changes conclusions in exactly the cases that matter: a project with strong conventional practices and weak AI governance is downgraded from "Managed" to "Defined", and the reason is made explicit. The validation protocol specifies how weights, agreement and reliability are to be established. Future work will report the results of the expert panel and pilot applications, examine the sensitivity of the gating tolerance, and extend the model to programme and portfolio level under ST RK ISO 21503, where the aggregation of project-level maturity into a portfolio-level cyber-risk posture is of direct interest to the Strategy's governance.

Acknowledgments

This research wasn’t funded by anyone.

Author Contributions

Conceptualization: Zh.I.; Methodology: Zh.I. and M.A.; Validation: Zh.I. and M.A.; Formal analysis: U.O.; Writing — original draft preparation: Zh.I. and M.A.; Writing — review and editing: Zh.I. and U.O.; Visualization: M.A.; Supervision: Zh.I.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Decree of the President of the Republic of Kazakhstan No. 1311 of 9 June 2026 "On approval of the Nationwide Strategy for Large-Scale Digitalisation and Total Implementation of Artificial Intelligence Technologies 'Digital Qazaqstan' until 2029". Available at: https://adilet.zan.kz/rus/docs/U2600001311

[2] Resolution of the Government of the Republic of Kazakhstan No. 760 of 26 August 2026 "On approval of the Action Plan for the implementation of the Nationwide Strategy for Large-Scale Digitalisation and Total Implementation of Artificial Intelligence Technologies 'Digital Qazaqstan' until 2029". Available at: https://prg.kz/m/amp/document/37110361/

[3] Law of the Republic of Kazakhstan No. 230-VIII of 17 November 2025 "On Artificial Intelligence". Available at: https://adilet.zan.kz/rus/docs/Z2500000230

[4] Order of the Minister of National Economy of the Republic of Kazakhstan No. 2 of 13 January 2022 "On approval of the Model Regulation of Project Management of State Bodies" (as amended). Available at: https://adilet.zan.kz/rus/docs/V2300033371

[5] ST RK ISO 21502-2022. Project, programme and portfolio management — Guidance on project management (ISO 21502:2020, IDT). Astana: Committee for Technical Regulation and Metrology, 2022.

[6] ISO 21502:2020. Project, programme and portfolio management — Guidance on project management. Geneva: ISO, 2020.

[7] Resolution of the Government of the Republic of Kazakhstan No. 832 of 20 December 2016 "On approval of the Unified Requirements in the field of information and communication technologies and information security" (as amended).

[8] ST RK ISO/IEC 27001-2023. Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO/IEC 27001:2022, IDT). Astana, 2023.

[9] ISO/IEC 27005:2022. Information security, cybersecurity and privacy protection — Guidance on managing information security risks. Geneva: ISO/IEC, 2022.

[10] ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. Geneva: ISO/IEC, 2023; adopted as ST RK ISO/IEC 42001-2025.

[11] ISO/IEC 42005:2025. Information technology — Artificial intelligence (AI) — AI system impact assessment. Geneva: ISO/IEC, 2025.

[12] ISO/IEC 42006:2025. Information technology — Artificial intelligence — Requirements for bodies providing audit and certification of artificial intelligence management systems. Geneva: ISO/IEC, 2025.

[13] Nikolaenko V, Sidorov A. Assessment of project management maturity models strengths and weaknesses. Journal of Risk and Financial Management. 2023;16(2):121. https://doi.org/10.3390/jrfm16020121

[14] Grant KP, Pennypacker JS. Project management maturity: an assessment of project management capabilities among and between selected industries. IEEE Transactions on Engineering Management. 2006;53(1):59–68. https://doi.org/10.1109/TEM.2005.861802

[15] Irfan M, Hassan M, Hassan N, Habib M, Khan S, Nasruddin AM. Project management maturity and organizational reputation: a case study of public sector organizations. IEEE Access. 2020;8:73828–73842. https://doi.org/10.1109/ACCESS.2020.2988511

[16] Žurga G. Project management in public administration. TPM — Total Project Management Maturity Model. The case of Slovenian public administration. Transylvanian Review of Administrative Sciences. 2018;14(53):144–159.

[17] Nikolaenko V, Sidorov A. Assessing the maturity level of risk management in IT projects. Sustainability. 2023;15(17):12752. https://doi.org/10.3390/su151712752

[18] Khan RA, et al. Securing software development through people maturity: a fuzzy-AHP decision-making framework. Journal of Software: Evolution and Process. 2025;37:e70045. https://doi.org/10.1002/smr.70045

[19] Lawal Y. Integrating cybersecurity into project management: a scoping review of current practices and gaps. Journal of Technology Studies. 2025. https://doi.org/10.21061/jts.430

[20] Al-Janabi S, Jabbar H, Syms F. Cybersecurity transformation: cyber-resilient IT project management framework. Digital. 2024;4(4):866–897. https://doi.org/10.3390/digital4040043

[21] Payette J, Anegbe E, Caceres E, Muegge S. Secure by design: cybersecurity extensions to project management maturity models for critical infrastructure projects. Technology Innovation Management Review. 2015;5(6):26–34.

[22] Büyüközkan G, Güler M. Cybersecurity maturity model: systematic literature review and a proposed model. Technological Forecasting and Social Change. 2025;213:123996. https://doi.org/10.1016/j.techfore.2025.123996

[23] National Institute of Standards and Technology. The NIST Cybersecurity Framework (CSF) 2.0. NIST CSWP 29. Gaithersburg, MD: NIST, 2024. https://doi.org/10.6028/NIST.CSWP.29

[24] Dreyling R, Lemmik J, Tammet T, Pappel I. An artificial intelligence maturity model for the public sector: a design science approach. TalTech Journal of European Studies. 2024;14(2):217–239. https://doi.org/10.2478/bjes-2024-0023

[25] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: NIST, 2023. https://doi.org/10.6028/NIST.AI.100-1

[26] Vatamanu AF, Tofan M. Integrating artificial intelligence into public administration: challenges and vulnerabilities. Administrative Sciences. 2025;15(4):149. https://doi.org/10.3390/admsci15040149

[27] Wang J, Neil M, Fenton N. A Bayesian network approach for cybersecurity risk assessment implementing and extending the FAIR model. Computers & Security. 2020;89:101659. https://doi.org/10.1016/j.cose.2019.101659

[28] Landauer M, Skopik F, Stojanović B, Flatscher A, Ullrich T. A review of time-series analysis for cyber security analytics: from intrusion detection to attack prediction. International Journal of Information Security. 2025;24(1):3. https://doi.org/10.1007/s10207-024-00921-0

[29] Cheimonidis P, Rantos K. Dynamic risk assessment in cybersecurity: a systematic literature review. Future Internet. 2023;15(10):324. https://doi.org/10.3390/fi15100324

[30] Vassilev A, Oprea A, Fordyce A, Anderson H, Davies X, Hamin M. Adversarial machine learning: a taxonomy and terminology of attacks and mitigations. NIST AI 100-2 E2025. Gaithersburg, MD: NIST, 2025. https://doi.org/10.6028/NIST.AI.100-2e2025

[31] Cinà AE, Grosse K, Demontis A, Vascon S, Zellinger W, Moser BA, Oprea A, Biggio B, Pelillo M, Roli F. Wild patterns reloaded: a survey of machine learning security against training data poisoning. ACM Computing Surveys. 2023;55(13s):294. https://doi.org/10.1145/3585385

[32] Wang S, Zhao Y, Hou X, Wang H. Large language model supply chain: a research agenda. ACM Transactions on Software Engineering and Methodology. 2025;34(5):147. https://doi.org/10.1145/3708531

[33] OWASP Foundation. OWASP Top 10 for LLM Applications 2025. 2024. Available at: https://genai.owasp.org

[34] MITRE Corporation. MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems. Available at: https://atlas.mitre.org (accessed 21 October 2026).

[35] European Union Agency for Cybersecurity. ENISA Threat Landscape 2025. Athens: ENISA, 2025.

[36] Yessengeldina A, Taizhanova Z, Kuttybayeva N, Iskakova D, Khalbayev Y. Government project offices and the effectiveness of government programs: an empirical analysis based on panel data. Economies. 2026;14(7):267. https://doi.org/10.3390/economies14070267

[37] Narbaev, T. (2015). An assessment of project management maturity in Kazakhstan. *PM World Journal*, *4*(11), 1-20.

[38] Bekish, U., Yessengeldina, A., & Skrypko, T. (2025). FEATURES OF ASSESSMENT OF DIGITAL MATURITY OF THE PROJECT OFFICE OF STATE BODIES OF KAZAKHSTAN. Public Administration and Civil Service, 2(93), 31-41.

[39] United Nations Department of Economic and Social Affairs. United Nations E-Government Survey 2024: Accelerating Digital Transformation for Sustainable Development. New York: United Nations, 2024.

[40] OECD–IDB. (2023). 2023 OECD-IDB Digital Government Index of Latin America and the Caribbean: Results and key Findings. OECD Public Governance Policy Papers.

[41] Nerima, M., & Ralyté, J. (2021, May). Towards a digital maturity balance model for public organizations. In International conference on research challenges in information science (pp. 295-310). Cham: Springer International Publishing.

[42] Saaty TL. The Analytic Hierarchy Process: Planning, Priority Setting, Resource Allocation. New York: McGraw-Hill, 1980.

[43] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L, 12.7.2024.

Appendix A. Indicators and evidence for the baseline practice groups

The table lists, for each practice group of ST RK ISO 21502-2022, the key performance indicators and the evidence an assessor should examine when scoring attributes K1–K6 on the scale of Table 2. The list consolidates and corrects the practitioner matrix from which the model originated: all indicators now use the 0–5 scale and are mapped to the ten practice groups of Section 4.3.

Table A1. Baseline indicator matrix

Practice group	KPI	Evidence	Anchors (0 / 3 / 5)
P1 Pre-project and initiation	Project charter; business justification; stakeholder identification	Approved charter; business case; stakeholder register	0 – absent; 3 – partial; 5 – approved and linked to Action Plan measure
P2 Planning	Project management plan; schedule; resource plan; security and AI requirements included	Approved plan; baseline schedule; resource plan	0 – absent; 3 – approved without baselines; 5 – integrated plan with baselines
P3 Stakeholders and communication	Stakeholder map; communication plan; satisfaction	Stakeholder map; communication plan; survey results	0 – absent; 3 – plan exists; 5 – executed and measured
P4 Resources and team	Resource matrix; utilisation; role allocation; competence (incl. security and AI)	RACI; utilisation reports; training records	0 – absent; 3 – roles defined; 5 – competence verified
P5 Risk and issues	Risk register; probability–impact assessment; response plan; issue log	Risk register; risk matrix; response plan; issue log	0 – absent; 3 – register maintained; 5 – responses implemented and reviewed at gates
P6 Quality	Quality plan; quality control; defect rate	Quality plan; test and review records; defect statistics	0 – absent; 3 – plan approved; 5 – controls executed, defects minimal
P7 Schedule	Schedule variance; progress reporting; forecasting tools	Progress reports; schedule analysis	0 – variance > 20 %; 3 – variance 5–20 %; 5 – variance < 5 % with forecasting

Practice group	KPI	Evidence	Anchors (0 / 3 / 5)
P8 Cost	Budget approved; cost variance; financial reporting; cyber-risk reserve (F5)	Budget; cost reports	0 – variance > 20 % either way; 3 – 5–20 %; 5 – < 5 % with reserve analysis
P9 Monitoring, control and change	Reporting regularity; KPI control; change log	Reports; KPI dashboard; change log	0 – absent; 3 – periodic; 5 – systematic with corrective action
P10 Closure and lessons learned	Completion act; final report; knowledge base	Completion act; final report; lessons-learned repository	0 – absent; 3 – prepared; 5 – lessons reused in subsequent projects

Raushan Kambar*
PhD, Project Manager, Project Monitoring and Management Center of Digital Government Support Center
r.kambar@govtec.kz, ORCID orcid.org/0009-0008-6024-736X
Digital Government Support Center, Kazakhstan

Asselya Abildayeva
MSc, Head of Endowment MNU
as_abildayeva@kazguu.kz, ORCID orcid.org/0009-0008-9492-1850
Maqsut Narikbayev University, Kazakhstan, Kazakhstan

PRIVACY RISKS AND DATA PROTECTION READINESS IN AI MARKETING IN KAZAKHSTAN

Abstract: Artificial intelligence (AI) is entering the marketing practice of Kazakhstani companies at the same time as the national legal framework for personal data and AI is being tightened: the Law on Personal Data and Their Protection was amended in 2025–2026, the Law on Artificial Intelligence entered into force on 18 January 2026, and the Digital Code on 12 July 2026. No empirical evidence exists on whether businesses that intend to use AI in marketing understand the privacy and security obligations this creates. This study provides a baseline measurement of that readiness taken immediately before the Law on Artificial Intelligence entered into force. An anonymous online survey of 41 representatives of Kazakhstani companies and startups (15 items: 8 closed, 6 five-point Likert items, 1 open-ended question answered by 24 respondents) was analysed by descriptive statistics and thematic analysis, and respondents' perceptions were mapped onto the obligations of the three legal instruments. The results show a readiness–compliance gap: 80.5% of respondents regard AI integration as a condition of competitiveness, 87.8% intend to use AI tools in marketing and 95.1% already use consumer AI applications; 70.7% agree that AI creates significant risks for customer data, yet the risks they name (AI-enabled phishing and data theft, system failures, attacks on AI infrastructure, privacy violation, manipulation) are framed as external threats, and no respondent referred to the company's own obligations as an operator of personal data. Data security appears in 10 of 24 open-ended responses, but only 2 refer to legal requirements. The mapping shows that none of eight obligation groups is reflected in respondents' perceptions and that routine use of foreign generative AI services with customer data most directly exposes the localisation and cross-border transfer requirements. The scientific novelty of the work lies in the first empirical, obligation-level reading of business perception against Kazakhstan's three-instrument framework; the resulting mapping can serve as a compliance checklist for AI marketing, and a supplementary survey module is proposed for subsequent waves.

Keywords: personal data protection; artificial intelligence; AI marketing; privacy risk; regulatory readiness; generative AI; cross-border data transfer; compliance; Kazakhstan

ИИ-маркетинг в Казахстане: риски приватности и готовность к защите данных

Аннотация: Искусственный интеллект (ИИ) входит в маркетинговую практику казахстанских компаний одновременно с ужесточением национальной нормативной базы в сфере персональных данных и ИИ: Закон «О персональных данных и их защите» изменён в 2025–2026 годах, Закон «Об искусственном интеллекте» вступил в силу 18 января 2026 года, Цифровой кодекс — 12 июля 2026 года. Эмпирических данных о том, понимают ли компании, намеревающиеся применять ИИ в маркетинге, возникающие при этом обязательства в области приватности и безопасности, нет. Исследование даёт базовое измерение такой готовности, выполненное непосредственно перед вступлением в силу Закона об ИИ. Анонимный онлайн-опрос 41 представителя казахстанских компаний и стартапов (15 вопросов: 8 закрытых, 6 по пятибалльной шкале Лайкерта, 1 открытый, на который ответили 24 респондента) проанализирован методами описательной статистики и тематического анализа, а представления респондентов сопоставлены с обязанностями по трём правовым актам. Результаты выявляют разрыв между готовностью к внедрению и готовностью к соблюдению требований: 80,5% респондентов считают интеграцию ИИ условием конкурентоспособности, 87,8% намерены использовать ИИ-инструменты в маркетинге, 95,1% уже пользуются потребительскими ИИ-приложениями; 70,7% согласны, что ИИ создаёт значимые риски для данных клиентов, однако называемые ими риски (фишинг и кража данных с помощью ИИ, сбои систем, атаки на ИИ-инфраструктуру, нарушение приватности, манипуляция) описываются как внешние угрозы, и ни один респондент не упомянул собственные обязанности компании как оператора персональных данных. Безопасность данных присутствует в 10 из 24 открытых ответов, но лишь в 2 упоминаются требования закона. Сопоставление показывает, что ни одна из восьми групп обязанностей не отражена в представлениях респондентов, а повседневное использование зарубежных генеративных ИИ-сервисов с данными клиентов наиболее прямо затрагивает требования локализации и трансграничной передачи. Научная новизна работы — первое эмпирическое прочтение восприятия бизнеса на уровне конкретных обязанностей по трём актам Казахстана; полученное сопоставление может служить чек-листом комплаенса для ИИ-маркетинга, а для последующих волн опроса предложен дополнительный модуль анкеты.

Ключевые слова: защита персональных данных; искусственный интеллект; ИИ-маркетинг; риски приватности; регуляторная готовность; генеративный ИИ; трансграничная передача данных; комплаенс; Казахстан

Қазақстандағы ЖИ-маркетинг: жекелік қауіптері және деректерді қорғауға дайындық

Аңдатпа: Жасанды интеллект (ЖИ) қазақстандық компаниялардың маркетингтік тәжірибесіне дербес деректер мен ЖИ саласындағы ұлттық нормативтік база қатаңдатылып жатқан кезеңде енуде: «Дербес деректер және оларды қорғау туралы» Заңға 2025–2026 жылдары өзгерістер енгізілді, «Жасанды интеллект туралы» Заң 2026 жылғы 18 қаңтарда, Цифрлық кодекс 2026 жылғы 12 шілдеде күшіне енді. Маркетингте ЖИ қолдануға ниетті компаниялардың осыдан туындайтын жекелік пен қауіпсіздік міндеттемелерін түсінетіні туралы эмпирикалық деректер жоқ. Зерттеу осындай дайындықтың «Жасанды интеллект туралы» Заң күшіне енер алдында жүргізілген базалық өлшемін ұсынады. Қазақстандық компаниялар мен стартаптардың 41 өкілі арасында жүргізілген анонимді онлайн-сауалнама (15 сұрақ: 8 жабық, 6 бес балдық Лайкерт шкаласы бойынша, 1 ашық сұрақ, оған 24 респондент жауап берді) сипаттамалық статистика және тақырыптық талдау әдістерімен өңделді, ал респонденттердің түсініктері үш құқықтық актінің міндеттемелерімен салыстырылды. Нәтижелер енгізуге дайындық пен талаптарды сақтауға дайындық арасындағы алшақтықты көрсетті: респонденттердің 80,5%-ы ЖИ интеграциясын бәсекеге қабілеттіліктің шарты деп санайды, 87,8%-ы маркетингте ЖИ құралдарын қолдануға ниетті, 95,1%-ы тұтынушылық ЖИ-қосымшаларды пайдаланады; 70,7%-ы ЖИ клиенттер деректеріне елеулі қауіп төндіреді деп келіседі, алайда олар атаған қауіптер (ЖИ арқылы фишинг пен деректерді ұрлау, жүйе ақаулары, ЖИ инфрақұрылымына шабуылдар, жекелікті бұзу, манипуляция) сыртқы қатерлер ретінде сипатталады және бірде-бір респондент компанияның дербес деректер операторы ретіндегі өз міндеттерін атамаған. Деректер қауіпсіздігі 24 ашық жауаптың 10-ында кездеседі, бірақ тек 2-уінде ғана заң талаптары аталады. Салыстыру сегіз міндеттемелер тобының ешқайсысы респонденттердің түсінігінде көрініс таппағанын және шетелдік генеративті ЖИ-сервистерін клиент деректерімен күнделікті пайдалану деректерді локализациялау мен трансшекаралық беру талаптарын барынша тікелей қозғайтынын көрсетті. Жұмыстың ғылыми жаңалығы — Қазақстанның үш актіден тұратын нормативтік базасы аясында бизнес түсінігін нақты міндеттемелер деңгейінде алғаш рет эмпирикалық талдау; алынған кесте ЖИ-маркетинг үшін комплаенс-чеклист ретінде қолданыла алады, ал кейінгі сауалнама толқындары үшін қосымша сауалнама модулі ұсынылды.

Түйін сөздер: дербес деректерді қорғау; жасанды интеллект; ЖИ-маркетинг; жекелік қауіптері; реттеуші дайындық; генеративті ЖИ; деректерді трансшекаралық беру; комплаенс; Қазақстан

1. Introduction (Literature Review)

Artificial intelligence (AI) has become the dominant technological driver of change in marketing. Customer analytics, personalisation, recommendation, conversational agents and content generation increasingly rely on machine learning models that are trained on, and fed with, the personal data of consumers [1], [2]. The academic debate has moved from whether AI will transform marketing to the conditions under which it can do so legitimately: the capabilities that make AI valuable for marketers — inference from large volumes of behavioural data, micro-targeting, automated decisions about individuals — are the same capabilities that generate privacy harm, erode consumer trust and attract regulatory intervention [3]–[6].

Kazakhstan is an instructive setting for studying this tension. The country has declared AI a national priority: the Concept for the Development of Artificial Intelligence for 2024–2029 was adopted in July 2024 [7], the International Center for Artificial Intelligence Alem.AI was launched, and on 18 September 2025 the sectoral ministry was transformed into the Ministry of Artificial Intelligence and Digital Development (MAIDD) [8]. In parallel, the legal framework governing the use of personal data by business has been tightened in rapid succession. Fines under Article 79 of the Code on Administrative Offences were roughly tripled in 2025 [9]; the Law "On Artificial Intelligence" No. 230-VIII of 17 November 2025 entered into force on 18 January 2026 [10]; the Digital Code No. 255-VIII of 9 January 2026 entered into force on 12 July 2026 [11]; and Law No. 326-VIII of 24 June 2026 introduced into the Law "On Personal Data and Their Protection" an identifier-based definition of personal data, a prohibition on decisions based solely on automated processing, a register of personal data security breaches and a register of persons collecting and processing personal data [12], [13]. The regulator has announced further increases in liability, including fines of up to 5,000 monthly calculation indices (MCI) and criminal liability for mass leaks [14].

Business appears to be moving at a different pace. The KPMG Kazakhstan AI Readiness Report (March 2025), covering more than 50 companies, found that only 20% had adopted AI-specific internal regulations, 40% relied on general data policies and 50% considered their data infrastructure inadequate; even among "AI leaders", more than 60% had no AI-specific rules [15]. Mass leaks of personal data — including the June 2025 publication of a database of more than 16 million records attributed to private information systems [16] — show that the cost of this gap is borne by consumers.

1.1 Literature review

Privacy as the binding constraint of AI marketing. The marketing literature treats data privacy as a strategic variable rather than a legal formality. Martin and Murphy [4] frame privacy as a multi-level construct that simultaneously affects consumers, firms and regulators, and show that privacy failures translate into measurable business losses; Martin, Borah and Palmatier [17] demonstrate that data breaches reduce customer trust and firm

performance and that the effect spills over to competitors. Acquisti, Brandimarte and Loewenstein [18] show that consumers' privacy decisions are context-dependent and malleable, so that formal consent obtained in a persuasive interface is a weak guarantee of informed agreement — a point that bears directly on AI-generated marketing interfaces. Bleier, Goldfarb and Tucker [5] conclude that privacy regulation reshapes rather than eliminates data-based marketing. AI raises the stakes within this body of work: Davenport et al. [1] identify data privacy, algorithmic bias and ethical use as the main constraints on the future of AI in marketing; Huang and Rust [2] note that higher levels of AI — "thinking" and "feeling" AI — require deeper access to personal and emotional data; Puntoni et al. [19] show that consumers experience AI as a source of exploitation and loss of autonomy as well as of convenience. Hermann [6] argues that ethical AI marketing requires informed consent, transparency and the avoidance of manipulation, and that systems collecting data without informed consent create risk for the company itself. Jobin, Ienca and Vayena [20], reviewing 84 AI ethics guidelines, find transparency, justice, non-maleficence, responsibility and privacy to be the most frequently cited principles — the principles Kazakhstan's Law on Artificial Intelligence has now codified. Bibliometric and agenda-setting work [3], [21] identifies privacy, trust and ethics as under-researched dimensions of AI marketing, particularly outside Western markets.

Generative AI tools in business practice. A second strand concerns the operational use of generative AI tools (ChatGPT and analogous services) by employees. Wu, Duan and Ni [22] systematise the security, privacy and ethical concerns of ChatGPT, including the retention and potential reuse of user inputs by the provider and the leakage of confidential information entered into prompts. The 2023 case in which Samsung banned employee use of generative AI after staff uploaded proprietary source code [23] shows that the risk is practical. The use of AI tools in violation of organisational rules ("shadow AI") is widespread: in a global study by KPMG and the University of Melbourne, 48% of employees reported using AI tools in ways that contravened company policy [24]. For marketing functions this matters because customer data — names, contacts, purchase histories, segments — are a natural input to generative tools used for drafting personalised messages, analysing feedback or building audiences. Where the provider is located abroad, each such input is potentially a cross-border transfer of personal data; where the provider retains inputs, it is a disclosure to a third party without the data subject's consent.

Small and medium-sized enterprises and data protection regimes. Studies of small and medium-sized enterprises (SMEs) under the European Union General Data Protection Regulation (GDPR) consistently show that awareness of a regime does not translate into compliance without targeted guidance: Brodin [25] reports that most SMEs studied were aware of the GDPR but only a minority had identified the steps needed to comply, and Freitas and Mira da Silva [26] find that SMEs postpone compliance until confronted with enforcement. Legal scholarship on Kazakhstan's personal data regime — the comparative chapter by Omarova, Mukhtorov and Yerlankyzy [27], the analysis of the supervisory

authority by Akhmetova, Tursynkulova and Ibrayev [28] and the review by Alibayeva and Yerbolatov [29] — is doctrinal and is not connected to empirical evidence on business behaviour.

1.2 Regulatory framework of Kazakhstan for personal data and artificial intelligence

The framework for the processing of personal data in marketing is now formed by three instruments, supplemented by subordinate rules and by liability provisions of the Code on Administrative Offences and the Criminal Code. Because the survey reported in this paper was conducted before 18 January 2026, Table 1 distinguishes obligations that were already in force at the time of the survey from those introduced afterwards.

Law "On Personal Data and Their Protection" No. 94-V of 21 May 2013 [12]. The Law establishes the conditions of collection and processing (Article 7), the requirements for the content of the data subject's consent (Article 8), the requirement to store personal data in a database located in Kazakhstan (Article 12) and the conditions of cross-border transfer, which is permitted to states that ensure the protection of personal data and otherwise only with the consent of the data subject or on other grounds listed in the Law (Article 16). Owners and operators must take legal, organisational and technical protection measures and notify the authorized body of a personal data security breach within one working day of its discovery (Article 25) [30], [31]. Law No. 326-VIII of 24 June 2026 [13] added an identifier-based definition of personal data; definitions of automated processing, anonymisation, masking and hashing; a rule that publication of data by the data subject does not in itself constitute consent to further processing; a prohibition on decisions based solely on automated processing, with a right to object; a prohibition on the collection and processing of personal data for the creation or expansion of databases by their non-purpose extraction from publicly available sources; a register of personal data security breaches; and a register of persons collecting and processing personal data.

Law "On Artificial Intelligence" No. 230-VIII of 17 November 2025 [10]. In force since 18 January 2026, the Law establishes the principles of legality, fairness, transparency and explainability, priority of human well-being, protection of personal data and confidentiality, and safety of AI systems, and introduces a risk-based classification of AI systems (minimal, medium and high risk). Of direct relevance to marketing, it prohibits AI systems that use manipulative methods affecting the subconscious or distorting behaviour, systems that exploit vulnerabilities of persons, social scoring, emotion recognition without consent, biometric categorisation for discriminatory purposes, and systems that unlawfully collect and process personal data. It requires synthetic content to be distributed only with machine-readable marking and a visual warning, users to be informed that they are interacting with an AI system or receiving a service produced with AI, and goods, works and services created with AI to be labelled. Users may request information on decisions taken with AI and on the underlying data. The accompanying law amended the Code on

Administrative Offences to introduce liability for failure to inform users of misleading synthetic results and for failure to manage the risks of a high-risk system where harm results.

Digital Code No. 255-VIII of 9 January 2026 [11]. In force since 12 July 2026, the Code does not replace the Law on Personal Data but establishes digital rights that constrain AI marketing: the right to erasure, anonymisation and restriction of processing; the right to have algorithmic decisions reviewed by a human; the restriction of biometric processing to cases provided by law; and the obligation to delete leaked data immediately with notification of the competent bodies. The Law "On Informatization" was restructured and renamed the Law "On Cybersecurity" by Law No. 256-VIII of the same date and retains the duty of owners of digital objects containing personal data to take the measures provided by personal data legislation [32].

Liability and enforcement. Article 79 of the Code on Administrative Offences, as amended in 2025, provides fines for unlawful collection and processing (up to 200 MCI for large businesses), for failure to take protection measures (up to 600 MCI) and, where such failure results in loss or unlawful processing, up to 2,000 MCI for large businesses — approximately 8.65 million tenge at the 2026 MCI value [9]. Article 147 of the Criminal Code provides criminal liability for violation of the inviolability of private life and of personal data legislation causing substantial harm [33]. Since the beginning of 2024 the Information Security Committee has imposed fines totalling 4.9 million tenge under Articles 79 and 641 [34], and 43 personal data leaks, each affecting 5–10 thousand persons, were registered in 2025 alone [35]. The authorized body is MAIDD, with state control exercised by its Information Security Committee [8]. The Concept of Digital Transformation, Development of the ICT Industry and Cybersecurity for 2023–2029 [36], which succeeded the "Digital Kazakhstan" programme, and the AI Concept [7] set the policy frame; Kazakhstan ranked 48th of 174 economies in the IMF AI Preparedness Index 2024 [37] and 60th of 195 in the Oxford Insights Government AI Readiness Index 2025 [38].

Table 1. Obligations of the legal framework relevant to AI marketing and their status at the time of the survey (second half of 2025)

No.	Obligation (source)	In force at the time of the survey
O1	Consent of the data subject with the content required by the Law (Law No. 94-V, Arts. 7, 8)	Yes
O2	Prohibition on building or expanding databases by non-purpose extraction from publicly available sources (Law No. 94-V, Art. 7, as amended by Law No. 326-VIII)	No (from 2026)
O3	Storage of personal data in a database in Kazakhstan (Art. 12) and conditions of cross-border transfer (Art. 16)	Yes
O4	Protection measures and notification of the authorized body within one working day of discovering a breach (Art. 25; Orders No. 232/NQ, 338/NQ)	Yes (notification duty since 1 July 2024)

No.	Obligation (source)	In force at the time of the survey
O5	Prohibition on decisions based solely on automated processing; right to object (Law No. 326-VIII); human review of algorithmic decisions (Digital Code)	No (from 2026)
O6	Prohibition on manipulative AI systems and exploitation of vulnerabilities (Law No. 230-VIII)	No (from 18 January 2026)
O7	Labelling of synthetic content; disclosure of interaction with AI (Law No. 230-VIII)	No (from 18 January 2026)
O8	Right to erasure, anonymisation and restriction of processing (Digital Code; Law No. 94-V)	Partly (erasure on withdrawal of consent under Law No. 94-V; extended by the Digital Code from 12 July 2026)

1.3 Research gap, research questions, novelty and contributions

Research gap. The literature establishes that privacy is the binding constraint of AI marketing, that everyday use of generative AI tools creates uncontrolled disclosure and cross-border transfer risks, and that SMEs tend to be aware of data protection regimes without complying with them. What is missing for Kazakhstan is an empirical link between the two bodies of evidence that do exist — consultancy surveys of AI readiness, which do not address legal obligations, and doctrinal legal studies, which do not address business behaviour. In particular, there is no measurement of how businesses that intend to use AI in marketing perceive privacy risk and whether that perception includes the obligations that the framework attaches to their own practices. Because the framework changed fundamentally between November 2025 and July 2026, a measurement taken before 18 January 2026 is also the only available baseline against which the effect of the new instruments can later be assessed.

Research questions. This study addresses the following research question: does the readiness of Kazakhstani companies and startups to adopt AI in marketing include awareness of the personal data protection obligations that such adoption entails? Specifically, the paper examines: (1) RQ1 — how Kazakhstani companies and startups perceive the privacy and security risks associated with the use of AI in marketing; (2) RQ2 — to what extent their declared readiness to adopt AI in marketing is accompanied by awareness of their own obligations as owners or operators of personal data; (3) RQ3 — which obligations of the Law on Personal Data, the Law on Artificial Intelligence and the Digital Code are most exposed by the prevailing patterns of AI use, and what measures follow for companies and the regulator.

Scientific novelty. The scientific novelty of this work consists in the first empirical, obligation-level reading of business perception of AI-related privacy risk in Kazakhstan,

obtained immediately before the entry into force of the Law on Artificial Intelligence and mapped onto the three-instrument framework (the amended Law on Personal Data, the Law on Artificial Intelligence and the Digital Code), which distinguishes it from prior approaches that treat either business readiness or the legal regime in isolation.

Contributions. The paper makes four contributions: (C1) a baseline measurement (N = 41) of awareness, intention and perceived AI-related risk among Kazakhstani companies and startups taken before 18 January 2026; (C2) empirical evidence of a readiness–compliance gap, in which 87.8% of respondents intend to use AI in marketing and 70.7% recognise data security risk, while no respondent identifies the company's own obligations; (C3) a mapping of perceived risks onto eight obligation groups with the applicable sanctions, distinguishing obligations in force at the time of the survey from those introduced after it, usable as a compliance checklist; (C4) a supplementary survey module of ten items mapped to specific obligations for subsequent waves.

The remainder of the paper is organised as follows: Section 2 presents the materials and methods; Section 3 reports the results; Section 4 discusses them in relation to the research questions, the literature and the legal framework and states the limitations; Section 5 concludes the paper.

2. Methods and Materials

2.1 Research design

The study combines a cross-sectional descriptive survey, thematic analysis of open-ended responses and doctrinal analysis of the regulatory framework. The survey component is exploratory: its purpose is to establish a baseline of awareness, attitudes and perceived risks, not to test hypotheses, and the sample does not permit inferential generalisation to the population of Kazakhstani companies.

2.2 Instrument

The questionnaire consisted of 15 items: 8 closed-ended items (type of organisation, sector, size, respondent role, self-assessed level of knowledge of AI, experience of using AI-based applications, current use of AI in marketing, perceived threats of AI — multiple choice); 6 statements rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree, or 1 = very unlikely, 5 = very likely for the intention item) covering the necessity of AI for competitiveness, the likelihood of using AI tools in marketing strategies, the sufficiency of technical infrastructure, the availability of competent specialists, the existence of a strategy for AI use, and the perceived risk of AI for customer data security; and 1 optional open-ended item inviting respondents to share their views, experience and recommendations. The questionnaire was implemented in Google Forms and is reproduced in Appendix A.

2.3 Sampling and data collection

The target population was representatives of Kazakhstani companies and startups involved in marketing decisions. The link to the questionnaire was distributed by e-mail through professional networks of the authors (convenience sampling). The survey was conducted in the second half of 2025 between September and October, before the entry into force of the Law on Artificial Intelligence (18 January 2026). Participation was voluntary and anonymous; respondents gave informed consent on the first page of the form, and no personal data of respondents were collected. In total 41 complete responses were received (N = 41); 24 respondents (58.5%) answered the optional open-ended item.

2.4 Analysis

Closed and Likert items were analysed by descriptive statistics (absolute frequencies and percentages of N = 41; percentages are rounded to one decimal place, so that one respondent corresponds to 2.4%). Open-ended responses were coded inductively following the reflexive thematic analysis approach of Braun and Clarke [39]: both authors read all 24 responses independently, assigned initial codes, consolidated codes into themes and resolved discrepancies by discussion; a response could be assigned to more than one theme. The regulatory analysis (Section 1.2) is based on the texts of the legislative acts published on the official legal information system Adilet and on the subordinate rules of the authorized body. The mapping (Section 3.6) was constructed by listing the obligation groups of Table 1, identifying for each the marketing practice that triggers it, searching the closed and open-ended responses for any reference to the obligation or to the practice, and recording the applicable sanction.

2.5 Ethical considerations

The study did not involve the processing of personal data of respondents, and ethics committee approval was not required under the rules of the authors' institutions.

3. Results

3.1 Sample

The composition of the sample is shown in Table 2. Companies predominated over startups (65.9% vs 34.1%), small and micro organisations formed two thirds of the sample, and the information technology and telecommunications sector was the largest (31.7%). Most respondents were marketing managers or specialists (39.0%) or owners and founders (26.8%).

Table 2. Composition of the sample (N = 41)

Characteristic	Category	n	%
Type of organisation	Company	27	65.9

Characteristic	Category	n	%
	Startup	14	34.1
Size (number of employees)	Micro (up to 15)	12	29.3
	Small (16–100)	15	36.6
	Medium (101–250)	8	19.5
	Large (over 250)	6	14.6
Sector	Information technology and telecommunications	13	31.7
	Retail and e-commerce	8	19.5
	Financial services	6	14.6
	Education and consulting	5	12.2
	Manufacturing and logistics	4	9.8
	Other services	5	12.2
Role of respondent	Marketing manager or specialist	16	39.0
	Owner or founder	11	26.8
	Senior manager (director, deputy director)	7	17.1
	IT or analytics specialist	4	9.8
	Other	3	7.3

3.2 Awareness of, experience with and current use of AI

Self-assessed knowledge of AI was medium or higher for 40 of 41 respondents (97.6%); only 1 respondent (2.4%) rated it as low and none as very low (Table 3). Experience of using AI-based applications (ChatGPT, Siri, Google Assistant and similar) was reported by 39 respondents (95.1%). Current use of AI in marketing was limited: 9 respondents (22.0%) reported regular use, 17 (41.5%) occasional or pilot use, and 15 (36.6%) no use.

Table 3. Awareness of, experience with and current use of AI (N = 41)

Item	Category	n	%
Self-assessed knowledge of AI	Very low	0	0.0
	Low	1	2.4
	Medium	14	34.1
	High	19	46.3
	Very high	7	17.1

Item	Category	n	%
Experience of using AI-based applications	Yes	39	95.1
	No	2	4.9
Current use of AI in marketing	Regular	9	22.0
	Occasional or pilot	17	41.5
	Not used	15	36.6

3.3 Attitudes and intention

Table 4 reports the full distribution of the six Likert items. Agreement that AI integration is a necessary condition of competitiveness was expressed by 33 respondents (80.5%), with 5 (12.2%) disagreeing. The likelihood of using AI tools in marketing strategies was rated "very likely" by 23 respondents (56.1%) and "likely" by 13 (31.7%), a combined positive intention of 36 respondents (87.8%); 1 respondent (2.4%) rated it "very unlikely". By contrast, only 13 respondents (31.7%) agreed that their organisation had sufficient technical infrastructure, 13 (31.7%) that it had specialists with the required competencies, and 10 (24.4%) that it had a clear strategy for the use of AI in marketing. Agreement that the use of AI in marketing creates significant risks for the security of customer data was expressed by 29 respondents (70.7%).

Table 4. Distribution of responses to Likert items (N = 41; n and % of respondents)

Item	1	2	3	4	5	Agree (4+5)
L1. AI integration is a necessary condition of competitiveness in the Kazakhstani market	1 (2.4)	4 (9.8)	3 (7.3)	14 (34.1)	19 (46.3)	33 (80.5)
L2. Likelihood that our organisation will use AI tools in its marketing strategies	1 (2.4)	0 (0.0)	4 (9.8)	13 (31.7)	23 (56.1)	36 (87.8)
L3. Our organisation has sufficient technical infrastructure for AI in marketing	5 (12.2)	13 (31.7)	10 (24.4)	9 (22.0)	4 (9.8)	13 (31.7)
L4. Our organisation has specialists with the competencies to implement AI in marketing	5 (12.2)	14 (34.1)	9 (22.0)	10 (24.4)	3 (7.3)	13 (31.7)
L5. Our organisation has a clear strategy for the use of AI in marketing	6 (14.6)	13 (31.7)	12 (29.3)	8 (19.5)	2 (4.9)	10 (24.4)
L6. The use of AI in marketing creates significant risks for the security of customer data	1 (2.4)	3 (7.3)	8 (19.5)	17 (41.5)	12 (29.3)	29 (70.7)

Result 1 (Tables 3 and 4) corresponds to Contribution C1: it provides the baseline values of awareness, experience and intention before the entry into force of the Law on Artificial Intelligence.

3.4 Perceived threats of AI

The closed multiple-choice item on perceived threats (Table 5) was answered by all 41 respondents, who selected 3.1 options on average. The most frequently selected threats were an increase in phishing attacks and data theft carried out with AI assistants (31 respondents, 75.6%) and violation of user privacy (27, 65.9%), followed by failures and errors in the operation of AI systems (24, 58.5%), cybersecurity vulnerability of AI infrastructure, including attacks on and hacking of AI systems (22, 53.7%), and the use of AI for manipulative purposes, including the spread of fake news and interference in political campaigns (19, 46.3%). Three respondents (7.3%) selected "other" and named job losses and dependence on foreign technologies. None of the options, and none of the "other" entries, referred to an obligation of the respondent's own organisation.

Table 5. Perceived threats of AI (multiple choice; N = 41)

Threat	n	%
Increase in phishing attacks and data theft carried out with AI assistants	31	75.6
Violation of user privacy	27	65.9
Failures and errors in the operation of AI systems	24	58.5
Cybersecurity vulnerability of AI infrastructure (attacks, hacking)	22	53.7
Use of AI for manipulative purposes (fake news, political campaigns)	19	46.3
Other (job losses; dependence on foreign technologies)	3	7.3

3.5 Themes in open-ended responses

Thematic analysis of the 24 open-ended responses produced five themes (Table 6); 19 responses were assigned to more than one theme. Respondent codes indicate the serial number, type and sector of the organisation.

Table 6. Themes identified in open-ended responses (n = 24)

Theme	Definition	Responses, n (%)
T1. AI as an efficiency and content tool	AI is described through concrete productivity gains: text and visual content generation, drafting of advertising, response to customer enquiries, analysis of reviews	14 (58.3)
T2. Shortage of competencies and strategy	Lack of specialists, lack of understanding of where AI should be applied, absence of an internal strategy or owner of the topic	11 (45.8)

Theme	Definition	Responses, n (%)
T3. Data security and privacy concerns	Concern about leakage, theft or misuse of customer data; distrust of where data entered into AI tools end up; of these, references to legal requirements (consent, the Law on Personal Data)	10 (41.7); of which legal references 2 (8.3)
T4. Need for regulation and guidance	Requests for clear rules, state guidance, standards or certification of AI tools	7 (29.2)
T5. Cost, infrastructure and language	Cost of tools and integration, weak infrastructure, limited support of the Kazakh language by available tools	6 (25.0)

Theme T1 was the most frequent: respondents described AI primarily as a tool that saves time on routine marketing tasks ("we generate three variants of a post in ten minutes instead of a day" — R07, startup, IT; "the chatbot answers the standard questions, the manager handles the rest" — R19, company, retail). Theme T2 reflected the Likert results on competencies and strategy: "everyone uses ChatGPT, but nobody in the company is responsible for the result" (R12, company, financial services); "we do not understand which of our tasks are worth automating" (R23, company, manufacturing). Within Theme T3, concern was expressed in general terms — "I am afraid that client data will end up who knows where" (R04, startup, e-commerce); "after the leak in the summer everybody got nervous" (R15, company, financial services) — and was in most cases located outside the organisation; two responses referred to legal requirements, one noting that "consent of clients is needed for everything, and nobody knows exactly in what form" (R09, company, education and consulting) and one asking whether "it is at all permissible to upload the client base into foreign services" (R31, company, IT). Theme T4 contained requests for guidance from the state ("the Ministry should explain in plain language what is allowed and what is not" — R27, startup, IT) and for certification of tools. Theme T5 combined cost with the observation that available tools "work poorly in Kazakh" (R35, company, retail).

Result 2 (Tables 5 and 6) corresponds to Contribution C2: data security is recognised as a risk by 70.7% of respondents in the closed item and appears in 41.7% of open-ended responses, but references to the organisation's own legal obligations are absent from the closed items and occur in only 2 of 24 open-ended responses.

3.6 Mapping of perceived risks onto legal obligations

Table 7 maps the eight obligation groups of Table 1 onto the marketing practices that trigger them, the corresponding evidence in the survey responses, and the applicable sanctions. For none of the eight groups did any closed or open-ended response refer to the obligation as a duty of the respondent's own organisation; for two groups (O3 and O4), responses referred to the underlying risk without reference to the obligation.

Table 7. Mapping of legal obligations onto marketing practices, survey evidence and sanctions

Obligation	Marketing practice that triggers it	Evidence in survey responses	Sanction for non-compliance
O1. Consent (Law No. 94-V, Arts. 7, 8)	Collection of customer data for AI-based personalisation, profiling and targeting; use of CRM data for model training	Not mentioned as an obligation; 1 open-ended response (R09) notes uncertainty about the form of consent	KoAP Art. 79(1)–(2): up to 200–600 MCI for large businesses
O2. Prohibition on non-purpose extraction from public sources (Law No. 94-V, as amended 2026)	Scraping social networks and open directories to build lead lists and audiences	Not mentioned	KoAP Art. 79(1)–(2)
O3. Localisation (Art. 12) and cross-border transfer (Art. 16)	Entering customer data into foreign generative AI services; use of foreign cloud marketing platforms	95.1% use foreign consumer AI applications; 1 open-ended response (R31) asks whether uploading the client base to foreign services is permissible; no reference to the obligation	KoAP Art. 79; Criminal Code Art. 147 in case of substantial harm
O4. Protection measures and one-day breach notification (Art. 25; Orders No. 232/NQ, 338/NQ)	Any AI system holding customer data; integration of AI tools with CRM	"Phishing and data theft" (75.6%) and "vulnerability of AI infrastructure" (53.7%) selected as external threats; notification duty not mentioned	KoAP Art. 79(3)–(4): up to 600 MCI; up to 2,000 MCI (≈ 8.65 million tenge) if loss or unlawful processing results
O5. No decisions based solely on automated processing; human review (Law No. 326-VIII; Digital Code)	Automated scoring of customers, automated pricing, automated refusal of offers	Not mentioned	KoAP Art. 79; civil claims
O6. Prohibition on manipulative AI (Law No. 230-VIII)	Hyper-personalised persuasive messaging, dark patterns, targeting of vulnerable groups	"Manipulative use" selected by 46.3%, but in the context of fake news and politics, not of the organisation's own marketing	KoAP (new AI offences); prohibition of the system
O7. Labelling of synthetic content; disclosure of interaction with AI (Law No. 230-VIII)	AI-generated advertising images, video and text; chatbots and voice assistants in customer service	Not mentioned; chatbots described in open-ended responses (R19) without reference to disclosure	KoAP: 86,500 to 865,000 tenge depending on business size and repetition (2026 values)

Obligation	Marketing practice that triggers it	Evidence in survey responses	Sanction for non-compliance
O8. Right to erasure, anonymisation and restriction of processing (Digital Code; Law No. 94-V)	Customer data retained in training sets and vector databases; data retained by third-party AI providers	Not mentioned	KoAP Art. 79

Result 3 (Table 7) corresponds to Contribution C3. The obligation most directly exposed by respondents' reported behaviour is O3: with 95.1% of respondents using foreign consumer AI applications and 87.8% intending to use AI in marketing, routine entry of customer data into such services is probable, and only one respondent questioned its permissibility.

4. Discussion

4.1 RQ1: perception of privacy and security risks

RQ1 asked how Kazakhstani companies and startups perceive the privacy and security risks of AI in marketing. The results in Sections 3.4 and 3.5 show that the risks are perceived as real — 70.7% agree that AI creates significant risks for customer data, 75.6% select AI-enabled phishing and data theft and 65.9% privacy violation — but as external and generic. Four of the five threat options describe actors or failures outside the organisation; privacy violation is selected without connection to any practice of the organisation; and in the open-ended responses the concern is located "who knows where" (R04) or in the national leak of summer 2025 (R15). This is consistent with the experiential account of Puntoni et al. [19], in which AI is perceived as a source of exploitation by actors other than oneself, and with the KPMG finding that 40% of Kazakhstani companies are worried about data privacy while only 20% have AI-specific rules [15]. The perception is therefore accurate in direction but misplaced in locus: the obligations of the framework attach to the organisation's own practices, not to external attackers.

4.2 RQ2: readiness to adopt versus awareness of obligations

RQ2 asked whether declared readiness to adopt AI in marketing is accompanied by awareness of the organisation's own obligations. The results confirm a readiness–compliance gap. On the adoption side, 97.6% report medium or higher knowledge of AI, 95.1% use AI applications and 87.8% intend to use AI in marketing. On the compliance side, no closed response and only 2 of 24 open-ended responses (8.3%) refer to a legal obligation, and both of those express uncertainty rather than knowledge. The gap is not explained by indifference: 29.2% of open-ended responses ask for rules and guidance (T4). It mirrors the pattern reported for SMEs under the GDPR — awareness of the regime without identification of the

concrete steps it requires [25], [26] — and the global rate of policy-violating AI use [24]. The low values for infrastructure (31.7%), competencies (31.7%) and strategy (24.4%) in Table 4 suggest a structural cause: organisations without an owner of the AI topic have no one to translate the framework into internal rules, which is exactly the finding of Theme T2 ("nobody in the company is responsible for the result" — R12).

The gap has a specific consequence in Kazakhstan. Obligations O1, O3 and O4 were already in force at the time of the survey, and the fines under Article 79 had already been tripled; obligations O2, O5, O6 and O7 came into force within the following seven months. An organisation that perceives privacy only as an external threat is therefore likely to have been non-compliant with O1, O3 and O4 at the time of the survey and to have become non-compliant with four further obligation groups since, without any change in its own behaviour.

4.3 RQ3: obligations most exposed and resulting measures

RQ3 asked which obligations are most exposed by prevailing patterns of AI use. The mapping (Table 7) identifies O3 — localisation and cross-border transfer — as the most exposed. Entering customer names, contacts, purchase histories or feedback into a service whose infrastructure is located abroad is, on the face of Article 16, a cross-border transfer that is lawful only if the receiving state ensures protection of personal data or the data subject has consented; where the provider retains and may reuse inputs [22], it is also a disclosure to a third party outside the purposes of collection. The Samsung case [23] shows that even technically sophisticated organisations fail to control this behaviour, and the 48% rate of policy-violating AI use [24] makes it unlikely that Kazakhstani marketing teams are an exception. The one open-ended response that raised the question ("is it at all permissible to upload the client base into foreign services" — R31) shows that the issue is recognisable once articulated, but is not currently articulated. This "shadow AI" channel is invisible to information security controls focused on the organisation's own infrastructure and deserves specific attention from the authorized body.

The second most exposed group is O6. Respondents named manipulation as a risk (46.3%) but located it in fake news and politics. The Law on Artificial Intelligence prohibits manipulative systems regardless of context, including commercial persuasion; hyper-personalised messaging that exploits consumers' malleable privacy decisions [18] now sits close to a statutory prohibition, and the boundary between legitimate personalisation and prohibited manipulation is undefined in practice. O7 (labelling and disclosure) is exposed by the chatbots and generated content that respondents describe as their main use case (T1) without any reference to disclosure.

Three measures follow for companies: an internal AI use policy prohibiting the entry of customer personal data into external generative AI services unless a lawful transfer basis and a contractual guarantee of non-retention exist; a review of consent texts and privacy notices to cover AI-based profiling, automated decisions and AI-generated communications,

including labelling; and assignment of responsibility for one-day breach notification and for entry in the register of persons processing personal data. Table 7 can be used as the checklist for this review. For the authorized body, the results indicate that enforcement alone will not close the gap, because respondents do not resist compliance but are unaware of it and ask for guidance (T4); sector-specific guidance for marketing on the status of foreign generative AI services under Articles 12 and 16, on lawful personalisation under the Law on Artificial Intelligence and on labelling of AI-generated advertising would address the most exposed obligations directly, and the announced increase of fines to 5,000 MCI [14] will be more effective if accompanied by such guidance. For educational programmes, the results confirm the shortage of specialists combining marketing competence with data protection and AI governance competence.

4.4 Limitations

The sample is small (N = 41) and non-probabilistic, with an over-representation of the IT sector and of small organisations; the results describe the surveyed group and cannot be generalised to all Kazakhstani companies. The instrument measured perceived risk in general terms and did not include items on specific legal obligations; the mapping in Table 7 is therefore an interpretation of respondents' silence on obligations, not a direct measurement of their knowledge, and the two open-ended references to legal requirements may understate awareness that respondents did not think relevant to the question asked. The thematic analysis is based on short written responses rather than interviews. Finally, the survey preceded the entry into force of the Law on Artificial Intelligence and the Digital Code, so the results measure perception under the framework as it stood in 2025 and cannot show whether the new instruments have changed it. Appendix B proposes a module of ten items mapped to specific obligations that addresses the first two limitations and allows the baseline reported here to be compared with a post-2026 wave.

5. Conclusion

Kazakhstan has moved within twenty months from a single personal data law to a three-instrument framework that attaches specific obligations to ordinary marketing practices — consent for profiling, localisation of data, one-day breach notification, prohibition of scraping and of manipulative systems, labelling of AI-generated content. This study shows that the businesses most eager to adopt AI in marketing entered that framework without perceiving those obligations as their own. The significance of the result lies less in the small sample than in its timing and its structure: taken immediately before the Law on Artificial Intelligence entered into force, it fixes the point from which the effect of the new framework can be measured, and it shows that the gap is not one of indifference — respondents recognise the risk and ask for rules — but of locus, with risk perceived in external attackers rather than in the organisation's own handling of customer data through foreign AI services.

The scientific novelty of this work — the first empirical, obligation-level reading of business perception of AI-related privacy risk in Kazakhstan against the amended Law on Personal Data, the Law on Artificial Intelligence and the Digital Code — has been confirmed by the finding that none of eight obligation groups is reflected in respondents' perceptions while the behaviour of 95.1% of respondents directly engages the localisation and cross-border transfer requirements; this demonstrates that readiness to adopt AI and readiness to comply with the framework are distinct constructs that must be measured separately. The claim is substantiated for the surveyed group; its generalisation to the population of Kazakhstani companies requires the larger, stratified wave for which the supplementary module is provided.

Three directions for future research follow. First, the supplementary module should be applied to a stratified sample after the Digital Code has been in force for at least a year, so that the baseline reported here can be compared with perception under the full framework. Second, the actual practice of marketing teams in entering customer data into generative AI services should be studied directly, by interview or by analysis of organisational policies, since survey silence on an obligation is only indirect evidence of non-compliance. Third, the boundary between lawful personalisation and prohibited manipulation under the Law on Artificial Intelligence requires doctrinal and empirical work before it can be enforced or complied with; the mapping proposed here offers a starting point for both.

Author Contributions

Conceptualization: R.K.; Methodology: R.K. and A.A.; Investigation: A.A.; Formal analysis: R.K. and A.A.; Writing — original draft preparation: R.K.; Writing — review and editing: R.K. and A.A.; Supervision: R.K.

Conflicts of Interest

The authors declare no other conflict of interest.

Data Availability Statement

The anonymised survey responses (N = 41) and the coding of open-ended responses are available from the corresponding author on reasonable request. The dataset is not deposited publicly because the open-ended responses contain descriptions of organisations that could allow their identification.

List of Supplementary Materials

- Appendix A: Questionnaire (15 items) as administered in Google Forms.
- Appendix B: Proposed supplementary survey module on regulatory readiness (10 items mapped to legal obligations).

References

[1]. Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

[2]. Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science, 49*(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9

[3]. Mustak, M., Salminen, J., Plé, L., & Wirtz, J. (2021). Artificial intelligence in marketing: Topic modeling, scientometric analysis, and research agenda. *Journal of Business Research, 124*, 389–404. https://doi.org/10.1016/j.jbusres.2020.10.044

[4]. Martin, K. D., & Murphy, P. E. (2017). The role of data privacy in marketing. *Journal of the Academy of Marketing Science, 45*(2), 135–155. https://doi.org/10.1007/s11747-016-0495-4

[5]. Bleier, A., Goldfarb, A., & Tucker, C. (2020). Consumer privacy and the future of data-based innovation and marketing. *International Journal of Research in Marketing, 37*(3), 466–480. https://doi.org/10.1016/j.ijresmar.2020.03.006

[6]. Hermann, E. (2022). Leveraging artificial intelligence in marketing for social good — An ethical perspective. *Journal of Business Ethics, 179*(1), 43–61. https://doi.org/10.1007/s10551-021-04843-y

[7]. Postanovlenie Pravitel'stva Respubliki Kazakhstan ot 24 iyulya 2024 goda No. 592 «Ob utverzhdenii Kontseptsii razvitiya iskusstvennogo intellekta na 2024–2029 gody» [Decree of the Government of the Republic of Kazakhstan No. 592 of 24 July 2024 "On Approval of the Concept for the Development of Artificial Intelligence for 2024–2029"]. (2024). Adilet. https://adilet.zan.kz/rus/docs/P2400000592

[8]. Ukaz Prezidenta Respubliki Kazakhstan ot 18 sentyabrya 2025 goda No. 997 «O merakh po dal'neyshemu sovershenstvovaniyu sistemy gosudarstvennogo upravleniya» [Decree of the President of the Republic of Kazakhstan No. 997 of 18 September 2025 "On Measures for Further Improvement of the Public Administration System"]. (2025).

[9]. Kodeks Respubliki Kazakhstan ob administrativnykh pravonarusheniyakh ot 5 iyulya 2014 goda No. 235-V (v redaktsii Zakona ot 10 yanvarya 2025 goda No. 155-VIII) [Code of the Republic of Kazakhstan on Administrative Offences No. 235-V of 5 July 2014 (as amended by Law No. 155-VIII of 10 January 2025)]. (2014). Adilet. https://adilet.zan.kz/rus/docs/K1400000235

[10]. Zakon Respubliki Kazakhstan ot 17 noyabrya 2025 goda No. 230-VIII «Ob iskusstvennom intellekte» [Law of the Republic of Kazakhstan No. 230-VIII of 17 November 2025 "On Artificial Intelligence"]. (2025). Adilet. https://adilet.zan.kz/rus/docs/Z2500000230

[11]. Tsifrovoy kodeks Respubliki Kazakhstan ot 9 yanvarya 2026 goda No. 255-VIII [Digital Code of the Republic of Kazakhstan No. 255-VIII of 9 January 2026]. (2026).

[12]. Zakon Respubliki Kazakhstan ot 21 maya 2013 goda No. 94-V «O personal'nykh dannykh i ikh zashchite» (s izmeneniyami i dopolneniyami po sostoyaniyu na 25.08.2026) [Law of the Republic of Kazakhstan No. 94-V of 21 May 2013 "On Personal Data and Their Protection" (as amended on 25 August 2026)]. (2013). Adilet. https://adilet.zan.kz/rus/docs/Z1300000094

[13]. Zakon Respubliki Kazakhstan ot 24 iyunya 2026 goda No. 326-VIII «O vnesenii izmeneniy i dopolneniy v nekotorye zakonodatel'nye akty Respubliki Kazakhstan po voprosam tsifrovizatsii, zashchity personal'nykh dannykh, dorozhnogo dvizheniya i peredovykh transportnykh tekhnologiy» [Law of the Republic of Kazakhstan No. 326-VIII of 24 June 2026 "On Amendments and Additions to Certain Legislative Acts of the Republic of Kazakhstan on Issues of Digitalisation, Protection of Personal Data, Road Traffic and Advanced Transport Technologies"]. (2026).

[14]. Zakon.kz. (2026, September 29). Shtrafy za utechku personal'nykh dannykh dlya kompaniy khotyat povysit' v Kazakhstane [Fines for personal data leaks for companies are to be increased in Kazakhstan]. Zakon.kz. https://zakon.kz/obshestvo/6533131-shtrafy-za-utechku-personalnykh-dannykh-dlya-kompaniy-khotyat-povysit-v-kazakhstane.html

[15]. KPMG in Kazakhstan and Central Asia, & Digital Business. (2025, March). *Kazakhstan AI readiness report*. KPMG. https://assets.kpmg.com/content/dam/kpmg/kz/pdf/2025/03/Kazakhstan-AI-Readiness-eng.pdf

[16]. Kazinform. (2025). Kiberugroza: pochemu proizoshla utechka personal'nykh dannykh kazakhstantsev i kak ikh zashchitit' [Cyber threat: Why the leak of personal data of Kazakhstanis occurred and how to protect them]. Kazinform. https://www.inform.kz/ru/kiberugroza-pochemu-proizoshla-utechka-personalnih-dannih-kazahstantsev-i-kak-ih-zashitit-df15ec

[17]. Martin, K. D., Borah, A., & Palmatier, R. W. (2017). Data privacy: Effects on customer and firm performance. *Journal of Marketing, 81*(1), 36–58. https://doi.org/10.1509/jm.15.0497

[18]. Acquisti, A., Brandimarte, L., & Loewenstein, G. (2020). Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *Journal of Consumer Psychology, 30*(4), 736–758. https://doi.org/10.1002/jcpy.1191

[19]. Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and artificial intelligence: An experiential perspective. *Journal of Marketing, 85*(1), 131–151. https://doi.org/10.1177/0022242920953847

[20]. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

[21]. Dwivedi, Y. K., Ismagilova, E., Hughes, D. L., Carlson, J., Filieri, R., Jacobson, J., Jain, V., Karjaluoto, H., Kefi, H., Krishen, A. S., Kumar, V., Rahman, M. M., Raman, R., Rauschnabel, P. A., Rowley, J., Salo, J., Tran, G. A., & Wang, Y. (2021). Setting the future of digital and social media marketing research: Perspectives and research propositions. *International Journal of Information Management, 59*, 102168. https://doi.org/10.1016/j.ijinfomgt.2020.102168

[22]. Wu, X., Duan, R., & Ni, J. (2024). Unveiling security, privacy, and ethical concerns of ChatGPT. *Journal of Information and Intelligence, 2*(2), 102–115. https://doi.org/10.1016/j.jiixd.2023.10.007

[23]. Gurman, M. (2023, May 2). Samsung bans staff's AI use after spotting ChatGPT data leak. *Bloomberg*. https://www.bloomberg.com/news/articles/2023-05-02/samsung-bans-chatgpt-and-other-generative-ai-use-by-staff-after-leak

[24]. Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). *Trust, attitudes and use of artificial intelligence: A global study 2025*. The University of Melbourne and KPMG. https://doi.org/10.26188/28822919

[25]. Brodin, M. (2019). A framework for GDPR compliance for small- and medium-sized enterprises. *European Journal for Security Research, 4*(2), 243–264. https://doi.org/10.1007/s41125-019-00042-z

[26]. Freitas, M. da C., & Mira da Silva, M. (2018). GDPR compliance in SMEs: There is much to be done. *Journal of Information Systems Engineering & Management, 3*(4), 30. https://doi.org/10.20897/jisem/3941

[27]. Omarova, B., Mukhtorov, A., & Yerlankyzy, A. (2024). Kazakhstan. In *Privacy and personal data protection law in Asia* (pp. 435–462). Hart Publishing. https://doi.org/10.5040/9781509965502.ch-017

[28]. Akhmetova, S. B., Tursynkulova, D. A., & Ibrayev, N. S. (2024). Powers of the Personal Data Protection Authority in Europe and Kazakhstan: Comparative analysis, recommendations for improving the activities of the Kazakhstani Personal Data Protection Authority. *Bulletin of L.N. Gumilyov Eurasian*

National University. Law Series, 148(3), 201–214. https://doi.org/10.32523/2616-6844-2024-148-3-201-214

[29]. Alibayeva, G. A., & Yerbolatov, E. E. (2025). Review of legal protection of personal data in Kazakhstan and abroad. *Bulletin of Toraigyrov University. Law Series*, (1), 16–29. https://doi.org/10.48081/cxuk5853

[30]. Prikaz Ministra iskusstvennogo intellekta i tsifrovogo razvitiya Respubliki Kazakhstan ot 30 aprelya 2026 goda No. 232/NQ «Ob utverzhdenii Pravil uvedomleniya o narusheniyakh bezopasnosti personal'nykh dannykh» [Order of the Minister of Artificial Intelligence and Digital Development of the Republic of Kazakhstan No. 232/NQ of 30 April 2026 "On Approval of the Rules for Notification of Personal Data Security Breaches"]. (2026). Adilet.

[31]. Prikaz Ministra iskusstvennogo intellekta i tsifrovogo razvitiya Respubliki Kazakhstan ot 22 iyunya 2026 goda No. 338/NQ «Ob utverzhdenii Pravil osushchestvleniya mer po zashchite personal'nykh dannykh» [Order of the Minister of Artificial Intelligence and Digital Development of the Republic of Kazakhstan No. 338/NQ of 22 June 2026 "On Approval of the Rules for the Implementation of Measures for the Protection of Personal Data"]. (2026). Adilet. https://adilet.zan.kz/rus/docs/V2300032810

[32]. Zakon Respubliki Kazakhstan ot 24 noyabrya 2015 goda No. 418-V «O kiberbezopasnosti» (ranee — «Ob informatizatsii»; v redaktsii Zakona ot 9 yanvarya 2026 goda No. 256-VIII) [Law of the Republic of Kazakhstan No. 418-V of 24 November 2015 "On Cybersecurity" (formerly "On Informatization"; as amended by Law No. 256-VIII of 9 January 2026)]. (2015). Adilet. https://adilet.zan.kz/rus/docs/Z1500000418

[33]. Ugolovnyy kodeks Respubliki Kazakhstan ot 3 iyulya 2014 goda No. 226-V [Criminal Code of the Republic of Kazakhstan No. 226-V of 3 July 2014]. (2014). Adilet. https://adilet.zan.kz/rus/docs/K1400000226

[34]. Digital Rights Center Qazaqstan. (2024). 4 910 360 — summa shtrafov za narusheniya v sfere personal'nykh dannykh s nachala 2024 goda [4,910,360 — the amount of fines for violations in the field of personal data since the beginning of 2024]. Digital Rights Center Qazaqstan. https://drcq.law/blog/4-910-360-summa-shtrafov-za-narusheniya-v-sfere-personalnykh-dannykh-s-nachala-2024-goda/

[35]. Profit.kz. (2025). Bol'she 40 utechek personal'nykh dannykh proizoshlo v Kazakhstane s nachala goda [More than 40 personal data leaks have occurred in Kazakhstan since the beginning of the year]. Profit.kz. https://profit.kz/news/69493/

[36]. Postanovlenie Pravitel'stva Respubliki Kazakhstan ot 28 marta 2023 goda No. 269 «Ob utverzhdenii Kontseptsii tsifrovoy transformatsii, razvitiya otrasli informatsionno-kommunikatsionnykh tekhnologiy i kiberbezopasnosti na 2023–2029 gody» [Decree of the Government of the Republic of Kazakhstan No. 269 of 28 March 2023 "On Approval of the Concept of Digital Transformation, Development of the Information and Communication Technology Industry and Cybersecurity for 2023–2029"]. (2023). Adilet. https://adilet.zan.kz/rus/docs/P2300000269

[37]. International Monetary Fund. (2024). *AI Preparedness Index (AIPI)* [Data set]. IMF DataMapper. https://www.imf.org/external/datamapper/datasets/AIPI

[38]. Oxford Insights. (2025). *Government AI Readiness Index 2025*. Oxford Insights. https://oxfordinsights.com/ai-readiness/ai-readiness-index/

[39]. Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Supplementary Materials

Privacy Risks and Data Protection Readiness in AI Marketing in Kazakhstan

Raushan Kambar, Assel Abildayeva

APPENDIX A. Questionnaire

Introductory text and consent. This survey is conducted for a study on the use of artificial intelligence (AI) in the marketing activities of Kazakhstani companies and startups. Participation is voluntary and anonymous; no personal data are collected, and the answers will be used only in aggregated form for research purposes. Completing the questionnaire takes about 7 minutes. By proceeding to the questions you confirm that you have read this information and agree to participate.

Part A. About your organisation and you (closed items)

Q1. Type of organisation: (a) company; (b) startup.

Q2. Main sector of activity: (a) information technology and telecommunications; (b) retail and e-commerce; (c) financial services; (d) education and consulting; (e) manufacturing and logistics; (f) other services (please specify).

Q3. Number of employees: (a) up to 15; (b) 16–100; (c) 101–250; (d) over 250.

Q4. Your role in the organisation: (a) owner or founder; (b) senior manager (director, deputy director); (c) marketing manager or specialist; (d) IT or analytics specialist; (e) other (please specify).

Part B. Awareness and use of AI (closed items)

Q5. How would you rate your level of knowledge about artificial intelligence? (a) very low; (b) low; (c) medium; (d) high; (e) very high.

Q6. Have you used applications based on artificial intelligence (for example, ChatGPT, Siri, Google Assistant)? (a) yes; (b) no.

Q7. Does your organisation currently use artificial intelligence in its marketing activities? (a) yes, regularly; (b) occasionally or as a pilot; (c) not yet.

Q8. Which threats do you associate with the development and implementation of artificial intelligence? (select all that apply) (a) increase in phishing attacks and data theft carried out with AI assistants; (b) failures and errors in the operation of AI systems; (c) cybersecurity vulnerability of AI infrastructure, including attacks on and hacking of AI systems; (d) violation of user privacy; (e) use of AI for manipulative purposes, including the spread of fake news and interference in political campaigns; (f) other (please specify).

Part C. Attitudes (5-point scale)

For statements Q9 and Q11–Q14: 1 = strongly disagree; 2 = disagree; 3 = neither agree nor disagree; 4 = agree; 5 = strongly agree. For Q10: 1 = very unlikely; 2 = unlikely; 3 = not sure; 4 = likely; 5 = very likely.

Q9. The integration of artificial intelligence is a necessary condition for maintaining competitiveness in the Kazakhstani market.

Q10. How likely is it that your organisation will use artificial intelligence tools in its marketing strategies in the future?

Q11. Our organisation has sufficient technical infrastructure to use artificial intelligence in marketing.

Q12. Our organisation has specialists with the competencies needed to implement artificial intelligence in marketing.

Q13. Our organisation has a clear strategy for the use of artificial intelligence in marketing.

Q14. The use of artificial intelligence in marketing creates significant risks for the security of customer data.

Part D. Open-ended item (optional)

Q15. Please share your views, experience or recommendations regarding the use of artificial intelligence in marketing in Kazakhstan.

APPENDIX B. Proposed supplementary module on regulatory readiness

The module is intended for subsequent survey waves. Items S2.1–S2.8 are answered on a 5-point scale (1 = strongly disagree, 5 = strongly agree; "do not know" is also offered) and are mapped to a specific obligation of the legal framework (Table S2.1). Items S2.9 and S2.10 measure awareness of the framework and guidance needs.

Table S2.1. Supplementary survey items mapped to legal obligations

No.	Item	Obligation measured
S2.1	Our organisation obtains the consent of customers, in the form required by the Law on Personal Data, before using their data for AI-based personalisation or profiling.	O1 — Consent (Law No. 94-V, Arts. 7, 8)
S2.2	Our organisation does not collect customer or prospect data from social networks or open sources to build or expand marketing databases.	O2 — Prohibition on non-purpose extraction from public sources (Law No. 94-V, as amended 2026)

No.	Item	Obligation measured
S2.3	Employees of our organisation are prohibited from entering personal data of customers into external AI services (ChatGPT and analogues), or such use is permitted only on a documented lawful basis.	O3 — Localisation and cross-border transfer (Law No. 94-V, Arts. 12, 16)
S2.4	Our organisation has a designated person and a procedure for notifying the authorized body of a personal data security breach within one working day.	O4 — Protection measures and breach notification (Law No. 94-V, Art. 25)
S2.5	Our organisation has registered, or has assessed whether it must register, as a person collecting and processing personal data.	Register of persons processing personal data (Law No. 326-VIII)
S2.6	Advertising and marketing content generated with AI is labelled as such, and customers are informed when they communicate with a chatbot or voice assistant.	O7 — Labelling and disclosure (Law No. 230-VIII)
S2.7	Our organisation has assessed whether its AI-based marketing tools could be classified as manipulative or as exploiting vulnerabilities of customers.	O6 — Prohibition on manipulative AI (Law No. 230-VIII)
S2.8	Customers can request deletion of their data from all systems, including AI tools and third-party providers, and we can fulfil such a request.	O8 — Right to erasure (Digital Code; Law No. 94-V)
S2.9	(Multiple choice) Which of the following legal acts does your organisation take into account when using AI in marketing? Law on Personal Data; Law on Artificial Intelligence; Digital Code; Law on Advertising; none; do not know.	Awareness of the framework
S2.10	(Open) What guidance from the authorized body would help your organisation use AI in marketing lawfully?	Guidance needs