INTERNATIONAL JOURNAL OF BUSINESS AND ECONOMICS INSIGHTS

Volume: 16; **Issue:** 8
Pages: 01– 36
Accepted: 05 August 2026
Published: 30 August 2026

Design and Validation of a Human-in-the-Loop AI Governance Framework for Data Quality and Enterprise Decision-Making

Tonmoy Barua[1];

[1]. *Data Analytics & Business Operations Researcher;*
Master of Science in Management Information Systems; Lamar University, Beaumont, Texas, USA;
Email: barua_tnm@yahoo.com

Doi: 10.63125/6m8hep19
This work was peer-reviewed under the editorial responsibility of the IJEI, 2026

Abstract

Enterprises increasingly let models decide on the records their pipelines deliver, and the quality of those records has become a governance problem rather than a data-engineering detail. The EU AI Act asks for data governance and for human oversight that is meaningful; the practical question is where, how much and in what form human involvement pays. This article designs a human-in-the-loop governance framework for data quality in automated decision-making and evaluates it on a transparent simulation testbed. The framework combines validation rules and automated re-fetch, a learned defect detector trained on injected defects, batch monitors with source attribution, a harm score that routes records to capacity-limited human review by evidence of defect, decision sensitivity and stakes, a reviewer model with automation bias and noisy overrides, and feedback loops that escalate alerts and reviewer confirmations into root-cause fixes, within a decision-rights matrix. The testbed replays 1,500 order-approval decisions per day for 180 days under eight governance configurations with identical records, defects and source incidents, in five replications. Defects affected 3–9% of records and raised expected decision cost by 0.8–4.6 million per year, yet only about 5% of defective records changed a decision. Rule-based review queues, the common practice, recovered 9% of the defect-induced cost; harm-based routing without feedback recovered none on average because reviews and overrides cost more than they saved; the full framework recovered 28% (4–46%), with 42% when the routing threshold was set in hindsight, mainly by detecting source incidents within about a day and averting half of their duration. An oracle that routes only decision-changing defects recovered 93% with six reviews a day, which locates the binding constraint in targeting rather than reviewer capacity. The five-year net present value of the framework was 2.5 million on average but negative in two of five replications, and depended most on reviewer override behaviour, time to fix, defect intensity and decision volume. Results describe the simulated stream only; the testbed, assumptions and code are provided for replication with real pipelines.

Keywords

Data Quality, AI Governance, Human-In-The-Loop, Human Oversight, Data Governance, Automation Bias, Machine Learning Monitoring, Enterprise Decision-Making, Design Science, Simulation

Introduction

Most enterprise decisions that are now automated — credit limits and order approvals, claims triage, replenishment, fraud holds, pricing — are taken by a model on a record assembled from several source systems minutes before the decision. The model's accuracy on a held-out test set says nothing about what happens when a field arrives empty because a feed failed overnight, when a currency conversion multiplies an amount by a hundred, when a customer master still carries last year's payment history, or when a join attaches another customer's attributes to the order. These are data-quality defects, and the literature on them is old and consistent: poor data quality is pervasive, costly and under-managed (Redman, 1998; Wang & Strong, 1996; Haug et al., 2011), and in machine-learning systems the data dependencies are the hardest part to control (Sculley et al., 2015; Breck et al., 2017; Sambasivan et al., 2021). What is new is the governance context. The EU AI Act requires that high-risk systems be built on data subject to data-governance practices and that they be designed for effective human oversight (European Parliament and Council, 2024); the NIST AI Risk Management Framework and ISO/IEC 42001 ask organisations to map, measure and manage such risks through accountable roles (NIST, 2023; ISO, 2023). "Human in the loop" has become the standard answer to both demands, and it is an answer that the human-factors literature has questioned for forty years (Bainbridge, 1983; Parasuraman & Riley, 1997; Skitka et al., 1999).

Three gaps motivate this article. First, data quality and human oversight are usually treated as separate topics: data-quality research measures accuracy, completeness and timeliness of data at rest (Pipino et al., 2002; Batini et al., 2009; Ehrlinger & Wöß, 2022), while the human–AI decision-making literature studies how people use model outputs on clean inputs (Lai et al., 2023; Bansal et al., 2021; Green & Chen, 2019). The governance question sits between them: given a stream of decisions on imperfect records, which records should a human see, what can the human realistically do, and how should what the human finds flow back to the data source? Second, the value of human involvement is asserted rather than measured. Reviews cost analyst time and delay decisions; reviewers rubber-stamp model outputs under automation bias and add noise when they override (Parasuraman & Manzey, 2010; Goddard et al., 2012; Kahneman et al., 2021); and a review queue driven by validation rules can fill with cheap-to-fix records while the costly ones pass. Whether a governance design recovers more than it costs is an empirical question that needs a cost model and a counterfactual. Third, regulation describes the properties of oversight — that it be meaningful, that humans can intervene, that limitations be understood (Koulu, 2020; Wagner, 2019; Crootof et al., 2023) — without specifying a design, and organisations are left to build review processes whose effect no one has quantified.

This article addresses the three gaps with a design-science study (Hevner et al., 2004; Peffers et al., 2007; Gregor & Hevner, 2013). The artefact is a human-in-the-loop governance framework for data quality in automated enterprise decision-making. Its components are a monitoring layer with validation rules, a learned defect detector and batch monitors; a routing layer that sends records to a capacity-limited human review queue according to a harm score combining evidence of a defect, the sensitivity of the decision to that record and the stakes; a reviewer model that is explicit about automation bias, workload and overrides; feedback loops that turn alerts and reviewer confirmations into root-cause investigations and fixes at the source; and a decision-rights matrix that assigns each governance decision to a role. The evaluation is a simulation: a synthetic stream of order-approval decisions with a known default probability, into which defects of six types are injected both at the record level and as source-level incidents, replayed under eight governance configurations with common random numbers so that every configuration faces identical records, defects and incidents. Because the generator's truth is known, the evaluation can measure what a governance process cannot measure in practice: the expected cost of each decision, the regret caused by each defect, the share of defect-induced cost that a configuration recovers, and the distance to an oracle that routes exactly the records whose decision a defect has changed.

The results are not uniformly favourable to human involvement, which is the point of measuring. Four findings organise the article. Defects that are easy to see (missing values) are the most frequent and among the least harmful per record, while plausible-looking corruptions (scale errors, mis-joined values) change one decision in five or ten and are nearly invisible to rules and weakly visible to a learned detector; only about 5% of defective records change a decision at all. Untargeted human review

therefore costs more than it saves: full review of every record is catastrophic in cost terms, and harm-based routing without feedback loops recovers nothing on average once review, delay and override costs are counted. The value of the framework comes mainly from its feedback loops: batch alerts and reviewer confirmations detected most source incidents within about a day and averted half of their duration, which is where the framework recovered 28% of the defect-induced cost on average and 42% with a well-set threshold. And the binding constraint is targeting, not capacity: an oracle that routes only decision-changing defects recovers 93% with six reviews a day, while adding reviewers to an imperfectly targeted queue adds cost and nothing else.

The article makes four contributions. It derives ten requirements for data-quality governance in automated decision-making from the data-quality, data-governance, machine-learning-operations, human–AI decision-making and AI-governance literatures and from the relevant standards, and designs a framework that meets them. It builds a transparent testbed with ground truth for defects, incidents and decision costs, and a cost model that includes review, delay, investigation and programme costs. It evaluates eight configurations with replication, sweeps and a value model, and reports where the framework pays, where it does not, and why. And it maps the framework to the EU AI Act, the NIST AI RMF, ISO/IEC 42001, ISO 8000-61, ISO/IEC 25012 and ISO/IEC 38505-1, and proposes an implementation roadmap.

Four research questions structure the study. RQ1: what does the literature require of data-quality governance for automated decisions, and what framework meets those requirements? RQ2: how do data-quality defects of different types affect automated decisions, and how well can rules, learned detectors and a harm score find the defects that matter? RQ3: how do governance configurations that differ in routing, reviewer involvement and feedback compare in decision cost, review effort, defect recovery and incident response, and how sensitive are the comparisons to capacity, reviewer behaviour, thresholds and defect intensity? RQ4: what is the economic value of the framework, and what does it imply for regulatory compliance and implementation?

The limitation to state at the outset is that all data are synthetic. The decision stream does not represent any company, the defect mixes, reviewer parameters and costs are assumptions listed in full, and the effect sizes depend on them. The compensating advantages are that the ground truth is known, that every configuration is compared on identical inputs, and that every assumption can be changed and the study re-run. The simulation script and the complete results file accompany the article.

Background and Related Work

Data quality as a decision problem

The data-quality literature established early that quality is fitness for use: a record is of high quality if it serves the task of its consumer, and the dimensions that matter — accuracy, completeness, consistency, timeliness, believability — are those the consumer values (Wang & Strong, 1996; Pipino et al., 2002). Methodologies for assessing and improving quality followed (Batini et al., 2009; Cai & Zhu, 2015; Ehrlinger & Wöß, 2022), as did standards: ISO/IEC 25012 defines a data-quality model with inherent and system-dependent characteristics (ISO, 2008), and ISO 8000-61 a process reference model for planning, controlling, assuring and improving data quality (ISO, 2016). The cost of poor quality was estimated early as a share of revenue and has been documented in surveys of organisations (Redman, 1998; Haug et al., 2011), and the supply-chain and analytics literatures have argued that data quality is the limiting factor for decision support more often than the analytics themselves (Hazen et al., 2014; Waller & Fawcett, 2013; Schoenherr & Speier-Pero, 2015).

For automated decisions, fitness for use has a precise meaning: a defect matters if it changes the decision, and its cost is the difference between the expected cost of the decision taken on the defective record and that of the decision that would have been taken on the clean one. This framing follows cost-sensitive learning, in which the decision threshold is set by the asymmetric costs of errors (Elkan, 2001), and it implies that defects differ by orders of magnitude in their harm: a missing value that a model imputes with a median changes few decisions, while a value multiplied by a hundred can change many. It also implies that the harm of a defect depends on how close the record's probability was to the decision threshold, which is the sensitivity argument used in selective classification and learning to defer (Chow, 1970; Geifman & El-Yaniv, 2017; Madras et al., 2018). The data-cleaning literature has produced detection methods for many defect types — rule and constraint violations, outliers,

duplicates, mis-joins — and has shown that no single method covers them (Rahm & Do, 2000; Chu et al., 2016; Abedjan et al., 2016); a recent empirical study of how six defect types affect six model families found effects that vary strongly by type and by model (Mohammed et al., 2025). Label and data errors in standard benchmarks are pervasive (Northcutt et al., 2021), and the data-centric AI movement has argued that data quality deserves the attention that modelling has received (Zha et al., 2023; Whang et al., 2023; Sambasivan et al., 2021).

The enterprise context adds two features that the data-cleaning literature, which mostly studies data at rest, does not emphasise. Records are assembled at decision time from several systems with different owners, refresh cycles and failure modes, so that defects arrive both as random record errors and as correlated incidents when a feed, a mapping or a job fails (Hazen et al., 2014; Cai & Zhu, 2015; Shankar et al., 2024). And the consumer of the record is a model whose response to a defect is neither transparent nor uniform: tree ensembles saturate on extreme values, imputation hides missingness, and a plausible but wrong value passes every check. The cost of a defect is therefore a joint property of the defect, the model and the decision, which is why this study measures it as regret on the decision rather than as an error rate on the data.

Monitoring data and models in production

The machine-learning-operations literature treats data dependencies as the main source of technical debt in production systems (Sculley et al., 2015) and has codified what to test: a rubric for production readiness covers data, model, infrastructure and monitoring, with tests for feature schemas, distributions and dependencies (Breck et al., 2017). Data validation for machine-learning pipelines at scale checks each batch against an inferred schema and alerts on drift, with the design lesson that alerts must be actionable by the engineer who receives them (Breck et al., 2019; Polyzotis et al., 2017). Unit tests for data, declared as constraints on completeness, uniqueness and distributions, extend this to general pipelines (Schelter et al., 2018). Dataset shift has a formal literature (Quiñonero-Candela et al., 2009), and an empirical comparison of shift detectors found that simple two-sample tests on low-dimensional representations detect most shifts, while the harder problem is characterising what shifted (Rabanser et al., 2019). Concept drift, in which the relation between inputs and outcome changes, is a related but distinct problem with its own detection and adaptation methods (Gama et al., 2014; Lu et al., 2019; Widmer & Kubat, 1996; Žliobaitė et al., 2016). Interview and survey studies of practitioners report that monitoring is often improvised, that alerts are ignored when they are not attributed to a cause, and that the organisational loop from alert to fix is where deployments fail (Paleyes et al., 2023; Shankar et al., 2024). Anomaly detection offers record-level scores, with isolation forests a common default (Chandola et al., 2009; Liu et al., 2008); whether such scores find the defects that matter for a decision is an empirical question that this study tests.

Humans in the loop: oversight, automation bias and complementarity

The human-factors literature on automation predates machine learning. The ironies of automation are that the operator is left with the tasks the designer could not automate, which are the hardest, and is expected to monitor a system that rarely fails, which people do badly (Bainbridge, 1983). Models of function allocation distinguish types and levels of automation across information acquisition, analysis, decision selection and action implementation (Parasuraman et al., 2000), and the use, misuse, disuse and abuse of automation have been documented across domains (Parasuraman & Riley, 1997). Automation bias — the tendency to accept automated recommendations and to miss errors that one would have caught unaided — has been shown experimentally and reviewed systematically (Skitka et al., 1999; Goddard et al., 2012; Parasuraman & Manzey, 2010), and complacency rises with workload and with the reliability of the automation. Trust in automation is calibrated by experience and design (Lee & See, 2004; Hoff & Bashir, 2015), and the path from assistance to autonomy has been argued to run through the same human-performance problems (Endsley, 2017).

The recent human–AI decision-making literature has refined these findings for model-assisted decisions. People show both algorithm aversion after seeing a model err and algorithm appreciation in other conditions (Dietvorst et al., 2015; Logg et al., 2019). Explanations and confidence displays shift reliance but do not reliably produce complementary performance, in which the human–AI team beats both the human and the AI alone (Bansal et al., 2021; Zhang et al., 2020; Wang & Yin, 2021; Lai & Tan, 2019); cognitive forcing functions reduce over-reliance at some cost in satisfaction (Buçinca et al., 2021),

and proxy tasks mislead about real decision performance (Buçinca et al., 2020). A review of empirical studies finds complementary performance to be the exception (Lai et al., 2023), and over-reliance has been reviewed as a design problem with partial mitigations (Passi & Vorvoreanu, 2022; Schemmer et al., 2023; Vasconcelos et al., 2023). Field experiments with generative tools show productivity gains that depend on the task and on whether the tool is right (Dell'Acqua et al., 2023; Noy & Zhang, 2023). Human judgement is noisy as well as biased (Kahneman et al., 2021; Tversky & Kahneman, 1974; Kahneman, 2011), and studies of human overrides find that they help in some conditions and hurt in others (Vodrahalli et al., 2022; Green & Chen, 2019; Mozannar et al., 2024). On the modelling side, learning to defer and selective prediction formalise when a model should hand a case to a human (Madras et al., 2018; Mozannar & Sontag, 2020; Raghu et al., 2019; Kamath et al., 2020), and humans-in-the-loop are argued to be most valuable where the model's information differs from theirs, for instance when the record the model saw is wrong (De-Arteaga et al., 2020). Hybrid intelligence, mixed-initiative interfaces and human-centred AI provide design vocabularies for these systems (Dellermann et al., 2019; Horvitz, 1999; Amershi et al., 2019; Shneiderman, 2020; Shneiderman, 2022; Raisch & Krakowski, 2021; Shrestha et al., 2019), and a survey of human-in-the-loop machine learning catalogues where humans enter the pipeline, from labelling to review (Wu et al., 2022; Settles, 2012).

Explanations are part of this literature but play a limited role in the present design. Explanations of a model's output help reviewers when they reveal that an input is implausible, and interpretable models make such checks easier (Rudin, 2019; Lipton, 2018; Doshi-Velez & Kim, 2017; Miller, 2019); question-driven and human-centred explainability frameworks list the questions a reviewer asks, of which "is the input right?" is one (Liao et al., 2020; Ehsan et al., 2021; Kulesza et al., 2015). The framework uses the decision model's probability for sensitivity and a separate detector for evidence of a defect, rather than relying on reviewers to infer defects from explanations, because the evidence that explanations improve reliance is mixed (Bansal et al., 2021; Wang & Yin, 2021) and because a dedicated detector can be evaluated and retrained on confirmed defects.

Two lessons from this literature shape the framework. Human involvement is valuable where the human has information or access the model lacks — the ability to see that a record is wrong and to fetch the right one, and to notice that many records from one source are wrong — and it is costly where the human merely second-guesses a calibrated model on a correct record. And the design must assume realistic reviewers: capacity-limited, biased towards the automated recommendation, affected by workload, and noisy when they override.

Data governance and AI governance

Data governance assigns decision rights and accountabilities for data: who decides on definitions, quality standards, access and lifecycle (Khatri & Brown, 2010). Contingency studies found that the design of governance depends on the organisation and that one size does not fit all (Weber et al., 2009); case studies describe how roles of data owner and data steward are organised in practice (Otto, 2011), and the governance of big data has been framed around value, risk and cost (Tallon, 2013). A structured review proposed a conceptual framework with scope, antecedents, mechanisms and consequences (Abraham et al., 2019), and the DAMA body of knowledge and ISO/IEC 38505-1 provide practitioner and standards vocabularies for roles and for the governing body's accountability for data (DAMA International, 2017; ISO, 2017). Data governance for public-sector and trustworthy AI has been argued to require the same mechanisms extended to algorithms (Janssen et al., 2020).

AI governance has developed in parallel. Principles have converged across dozens of guidelines on transparency, fairness, non-maleficence, responsibility and privacy (Jobin et al., 2019; Floridi et al., 2018), with a persistent gap between principles and practice (Morley et al., 2020). Instruments for closing it include auditing frameworks (Raji et al., 2020; Mökander & Floridi, 2021; Koshiyama et al., 2024), documentation of models and datasets (Mitchell et al., 2019; Gebru et al., 2021), and management-system standards and risk frameworks (ISO, 2023; NIST, 2023; OECD, 2019). Information-systems scholarship has defined organisational AI governance and its layers (Mäntymäki et al., 2022; Schneider et al., 2023; Papagiannidis et al., 2023) and framed the management of AI as a problem of autonomy, learning and inscrutability (Berente et al., 2021; Benbya et al., 2021). Regulation now makes some of this mandatory. The EU AI Act requires, for high-risk systems, data governance over training, validation and testing data (Article 10) and human oversight that lets natural persons understand,

monitor and intervene (Article 14) (European Parliament and Council, 2024). Legal scholarship has warned that human control can become nominal — a human who cannot in practice change the outcome but absorbs liability — and has asked what meaningful oversight requires (Koulu, 2020; Wagner, 2019; Crootof et al., 2023; Enarsson et al., 2022); the debate over a right to explanation raised similar questions about the gap between legal text and technical possibility (Wachter et al., 2017; Goodman & Flaxman, 2017; Selbst & Barocas, 2018). Model risk management in banking supervision offers a mature analogue: validation, ongoing monitoring and governance of models in use (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011; Office of the Comptroller of the Currency, 2021), and enterprise risk management supplies the generic vocabulary (ISO, 2018). The framework in this article takes its roles from data governance, its oversight requirements from the AI Act and the legal critique, and its controls from model risk management and machine-learning operations.

Business value of analytics and the cost of human effort

The analytics-capability literature links data quality, analytics and decision performance to firm outcomes, with data quality and organisational routines as enablers (Mikalef et al., 2019; Dubey et al., 2020; Fosso Wamba et al., 2017; Gunasekaran et al., 2017). Practitioner accounts emphasise that most of the value of AI in enterprises comes from many small automated decisions rather than from a few spectacular ones (Davenport & Ronanki, 2018; Agrawal et al., 2018), and the productivity literature has documented the lag between investment in intangibles such as data and governance and measurable returns (Brynjolfsson et al., 2021). Simulation-based evaluation of decision policies has been used in operations to compare prescriptive methods on identical demand draws (Elmachtoub & Grigas, 2022; Bertsimas & Kallus, 2020; Ban & Rudin, 2019) and to model disruption and recovery (Ivanov, 2017; Ivanov & Dolgui, 2021); judgemental adjustment of forecasts has been evaluated with the same question this article asks of reviewers — when does the human improve on the model? (Fildes et al., 2009). The present study adopts the simulation approach and extends it with a cost model in which human effort, delay and investigation are priced, so that governance configurations can be compared on total cost rather than on data-quality metrics alone.

Requirements and the framework

Table 1 distils the literature and standards into ten requirements. R1 states that quality must be judged by its effect on the decision. R2 and R3 require continuous monitoring and the detection of silent defects. R4 and R5 require selective human involvement designed for realistic reviewers. R6 requires feedback to the source. R7 and R8 require clear decision rights and oversight that can change outcomes. R9 requires alignment with regulation and standards, and R10 measurable value.

Table 1: Requirements for data-quality governance in automated decision-making, derived from the literature and standards

Requirement	What the framework must provide	Grounding in the literature and standards
R1 Fitness for use, not abstract accuracy	Data quality is judged by its effect on the decision; defects that do not change decisions are low priority	Wang & Strong (1996); Pipino et al. (2002); Redman (1998); Haug et al. (2011); ISO/IEC 25012 (ISO, 2008)
R2 Continuous monitoring of data in production	Validation rules, drift and anomaly monitors, and alerting on pipelines feeding models	Breck et al. (2019); Polyzotis et al. (2017); Schelter et al. (2018); Rabanser et al. (2019); Shankar et al. (2024)
R3 Detection of silent, plausible-looking defects	Learned detectors beyond range and null checks; cross-field consistency	Abedjan et al. (2016); Chu et al. (2016); Mohammed et al. (2025); Northcutt et al. (2021)
R4 Selective human involvement	Route to humans only where the expected harm justifies the cost; capacity-aware triage	Chow (1970); Madras et al. (2018); Mozannar & Sontag (2020); Raghu et al. (2019); De-Arteaga et al. (2020)
R5 Realistic reviewer behaviour	Design for automation bias, workload effects and imperfect overrides rather than assuming perfect oversight	Parasuraman & Manzey (2010); Skitka et al. (1999); Goddard et al. (2012); Green & Chen (2019); Bansal et al. (2021); Buçinca et al. (2021)
R6 Feedback to the source	Reviewer findings and batch alerts must reach data owners and close incidents, not only fix records	Redman (1998); Otto (2011); Khatri & Brown (2010); Sculley et al. (2015); Breck et al. (2017)
R7 Clear decision rights	Named owners, stewards, model owners and reviewers with accountability for each governance decision	Khatri & Brown (2010); Weber et al. (2009); Abraham et al. (2019); DAMA International (2017); ISO/IEC 38505-1 (ISO, 2017)
R8 Meaningful rather than nominal oversight	Human control that can actually change outcomes; avoid oversight that is only a liability sink	Koulu (2020); Wagner (2019); Crootof et al. (2023); Enarsson et al. (2022); Bainbridge (1983)
R9 Regulatory alignment	Data governance (Art. 10) and human oversight (Art. 14) under the EU AI Act; risk management functions in NIST AI RMF and ISO/IEC 42001	European Parliament and Council (2024); NIST (2023); ISO/IEC 42001 (ISO, 2023); Mäntymäki et al. (2022); Schneider et al. (2023)
R10 Measurable value	KPIs that connect data quality to decision cost, review cost and time to detect; a value model	Haug et al. (2011); Mikalef et al. (2019); Davenport & Ronanki (2018); Brynjolfsson et al. (2021)

Figure 1 shows the framework that meets them, as a layered architecture with four lanes and three feedback loops. Records are assembled from source systems under data contracts. The monitoring layer runs validation rules, a learned defect detector trained on injected defects, batch monitors (population-stability drift, null share, duplicate share and the detector's alarm rate) and the decision model, and computes a harm score as the product of evidence of a defect, the decision's sensitivity to the record and the stakes. The routing layer first re-fetches flagged records automatically — repairing the record-level defects that a second request fixes — and then routes the most harmful records to a prioritised, capacity-limited review queue with a service level, together with a random audit sample. Reviewers detect and correct defects, override decisions and confirm defects and their source; the design treats their automation bias, workload and override noise as parameters, not as absent. The decision gate records whether a decision was automatic, corrected or human. Three feedback loops close the design: confirmed defects retrain the detector and rules; alerts and confirmation spikes go to the data owner for root-cause investigation and a fix at the source; and governance records re-tune the threshold, capacity and audit share. Figure 2 assigns the seven governance decisions that the framework requires — rules and thresholds, routing threshold and audit share, capacity and service level, record correction and override, incident declaration, fix at source, and retraining and release — to five roles in a RACI matrix, and lists the evidence each decision needs. Table 2 maps the components to design principles, to the requirements and to their operationalisation in the testbed.

Figure 1: Human-in-the-loop AI governance framework for data quality: layered architecture with monitoring, routing and review, decision and feedback lanes, three feedback loops and governance roles

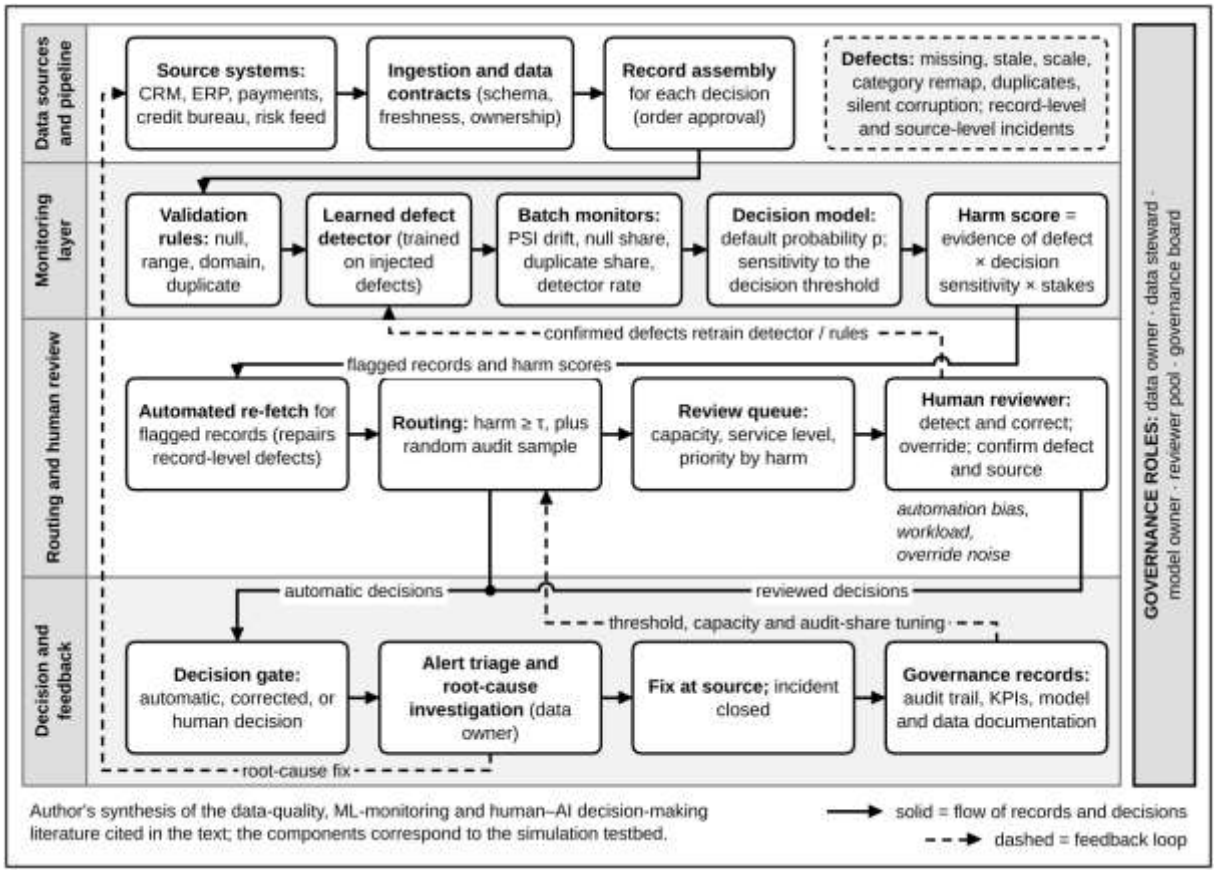


Figure 2: Decision-rights matrix: governance roles × framework decisions (RACI), with the evidence each decision needs

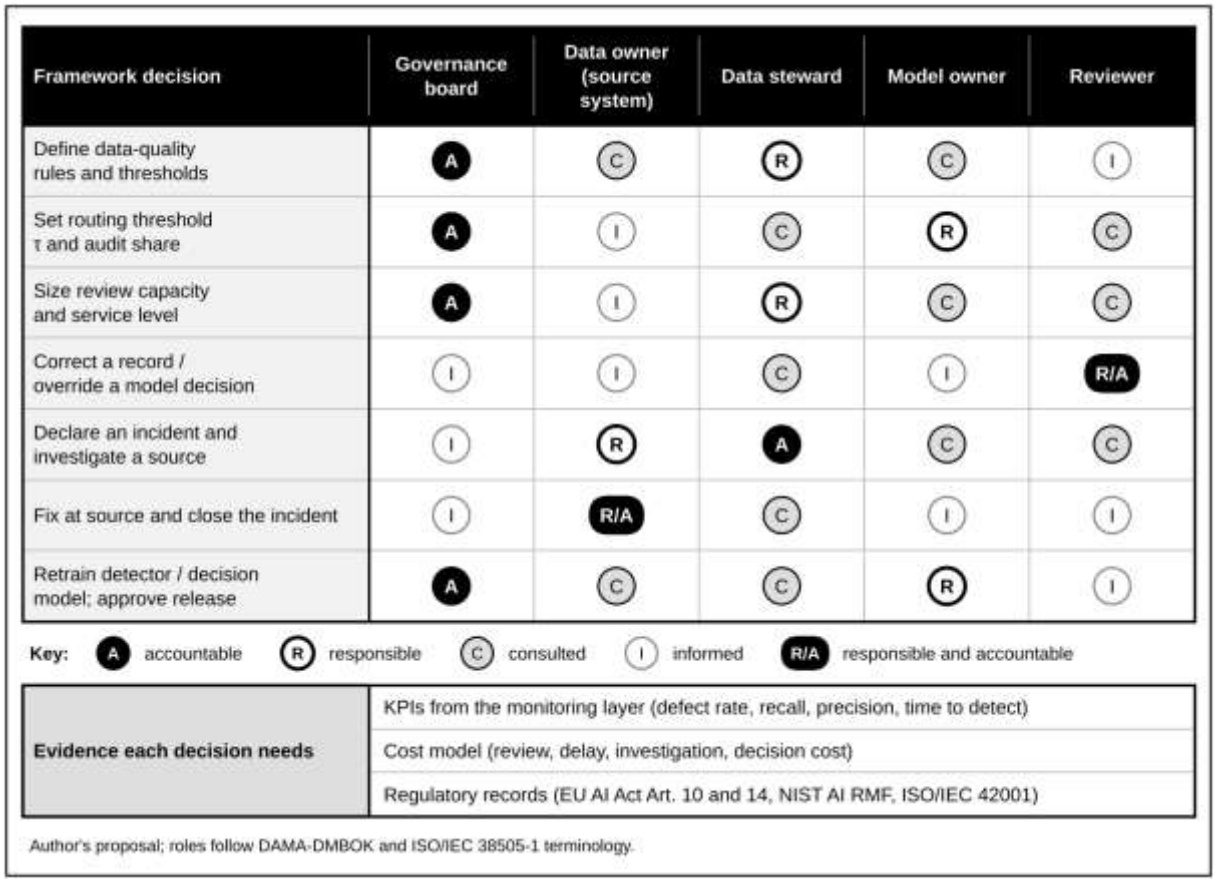


Table 2: Framework components, design principles, requirements addressed and operationalisation in the simulation testbed

Component (Figure 1)	Design principle	Requirements	Operationalisation in the testbed
Data sources and contracts	Every attribute has a named source and owner; contracts specify schema, freshness and ownership	R6, R7	Five sources (CRM, ERP, payments, bureau, risk feed); incidents are source-level events with an owner
Validation rules	Cheap deterministic checks run on every record: null, range, domain, duplicate	R2	Rules on eight numeric ranges, two categorical domains, an exact-duplicate check catching 60% of duplicates
Automated re-fetch	Flagged records are re-requested before a human sees them; automation handles the easy half	R4	Re-fetch repairs record-level defects with probability 0.5 after a 0.25-day delay; incident defects persist
Learned defect detector	A classifier trained on injected defects of the known types scores every record for plausible-looking defects	R3	LightGBM on raw values plus null indicators, trained on the fit window with 30% injected defects
Batch monitors	Distribution drift, null share, duplicate share and detector rate per batch trigger alerts with source attribution	R2, R6	PSI > 0.10 or null share > 2% on any field, duplicate share > 2%, detector rate 3 points above its clean baseline
Harm score and routing	Route by expected harm = evidence of defect × decision sensitivity × stakes, with a threshold τ and a capacity cap; add a random audit sample	R1, R4	Evidence = 0.5·flag + 0.5·detector; sensitivity = closeness of p to the decision threshold; stakes = order value / mean; τ tuned on validation; 0.5% audit
Review queue	Prioritised, capacity-limited, with a service level; no unbounded backlogs	R4, R5	Capacity 60 records/day, priority by harm, release after 2 days
Human reviewer model	Reviewers detect and correct with type-specific probability; automation bias, workload and overrides reduce their value	R5, R8	Detection 0.5–0.95 by type × (1 − bias 0.3) × workload factor; 10% noisy overrides on undetected records
Decision gate	Automatic, corrected or human decision recorded with its provenance	R8, R9	Cost-sensitive approve/reject at p* = 0.167; decision cost evaluated against the true default probability
Alert triage and root-cause fix	Alerts and reviewer-confirmation spikes trigger an investigation of the source; incidents are fixed with a lag	R6	Investigation cost 500 (1,500 for an all-source search); fix lag 2 days; cool-down 3 days per source
Feedback to detectors and thresholds	Confirmed defects retrain detectors and rules; KPIs re-tune τ, capacity and audit share	R6, R10	Represented by the confirmation-escalation rule and by the sweeps; detector retraining not simulated
Decision rights and records	RACI matrix (Figure 2); audit trail, KPIs and documentation for regulators	R7, R9, R10	KPIs: decision, review, delay and investigation cost; recall, precision, correction rate; time to detect; regret

The design choices that distinguish the framework from a conventional review queue are three. Routing by harm rather than by rule flags (R1, R4) sends to humans the records whose defect, if present, would

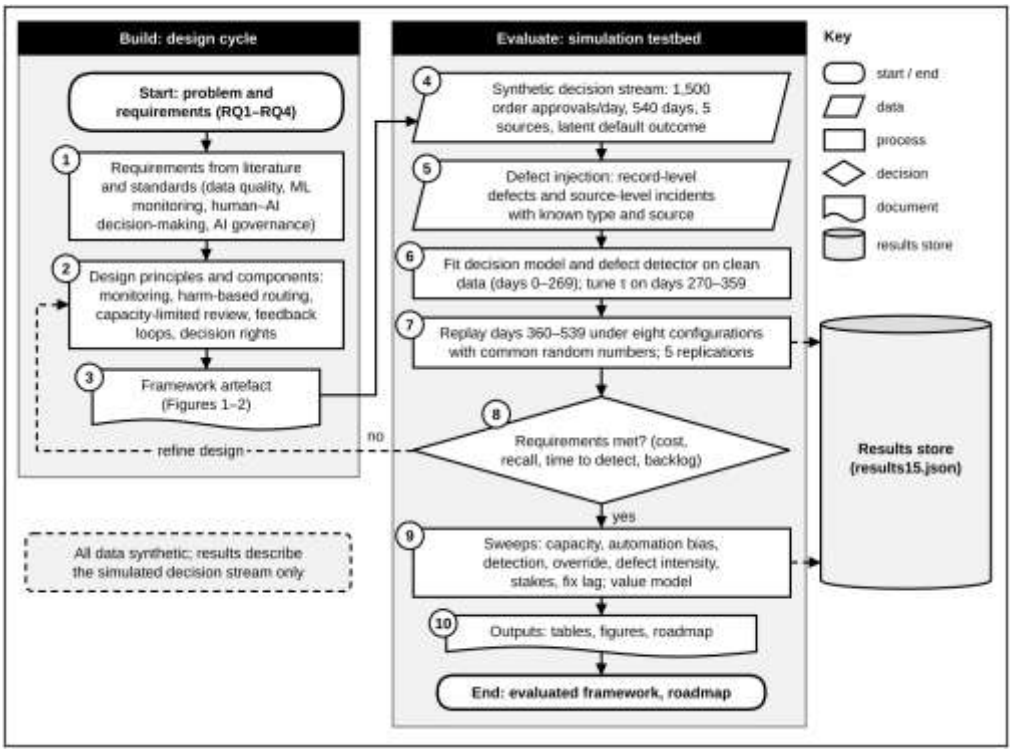
change a costly decision, rather than the records that are easiest to flag. Capacity-aware triage with a service level (R4, R5) prevents the unbounded backlogs that incidents create in rule-driven queues. And the escalation of reviewer confirmations to the data owner (R6) makes the reviewers a sensor for incidents as well as a repair mechanism for records. Whether these choices pay, under realistic reviewer behaviour and costs, is what the evaluation tests.

Methods

Study design

The study follows the design-science research process: problem identification, requirements, design of the artefact, demonstration and evaluation, and communication (Peffers et al., 2007; Hevner et al., 2004; March & Smith, 1995). The evaluation is an artificial, summative evaluation in the sense of the FEDS framework, chosen because a real deployment would neither reveal the ground truth about defects nor permit identical inputs under alternative governance designs (Venable et al., 2016; Sonnenberg & vom Brocke, 2012); the contribution is positioned as a nascent design theory with an evaluated artefact (Gregor & Hevner, 2013; Baskerville et al., 2018). Figure 3 shows the design as a flowchart with a build column (requirements, design principles, artefact) and an evaluate column (synthetic decision stream, defect injection, model and detector fitting, replay under eight configurations, sweeps and value model), with a refinement loop from the evaluation back to the design principles. The simulation follows the conventions of simulation studies — replication, common random numbers, separate verification of each component and an explicit list of assumptions (Law, 2015; Sargent, 2013) — and of chronological evaluation, with tuning on a validation window that precedes the test window (Bergmeir & Benítez, 2012; Kaufman et al., 2012). Five decision streams were generated with seeds 4040 to 4044; all configuration comparisons are reported across the five, and the first is used for the record-level and daily illustrations. The pipeline is implemented in Python with NumPy, pandas, SciPy, scikit-learn and LightGBM (Harris et al., 2020; Virtanen et al., 2020; Pedregosa et al., 2011; Ke et al., 2017). All data are synthetic.

Figure 3: Design-science study design as a flowchart: build column (requirements, design principles, artefact) and evaluate column (synthetic stream, defect injection, fitting and tuning, replay under eight configurations, sweeps and value model)



The synthetic decision stream

Table 3 lists the generator. Each day, 1,500 B2B order-approval decisions arrive for 540 days: days 0–269 are the fit window, in which one third of the records is used to fit the decision model and the defect detector on clean data; days 270–359 are the validation window, on which the routing threshold is tuned; and days 360–539 are the test window, 270,000 decisions on which the configurations are compared. Each record carries eight numeric attributes and two categorical ones drawn from five source systems: tenure, segment and channel from a CRM; order value and days since the last order from an ERP; payment score, average payment delay and disputes from a payments system; credit utilisation from a credit bureau; and a region-risk index from a risk feed. The attributes are correlated through a latent customer-quality factor, so that a value taken from another record creates cross-field inconsistencies that a detector could in principle learn. The outcome is payment default, a logistic function of the attributes only, with prevalence 13.5%; the decision is to approve or reject the order, with a loss of 60% of the order value if a defaulter is approved and 12% of the order value if a good order is rejected, so that the optimal rule rejects when the default probability exceeds $p^* = 0.167$ (Elkan, 2001; Thomas et al., 2017; Siddiqi, 2017). The decision model is a gradient-boosting classifier fitted on clean fit-window data with median imputation, with a hold-out AUC of 0.86–0.87 against 0.874 for the generator's own probability (Ke et al., 2017; Friedman, 2001; Fawcett, 2006). Decision costs are evaluated in expectation given the generator's true probability, a conditional Monte Carlo estimator that removes outcome noise, which is roughly fifty times larger than the effect of defects on a 180-day window; realised-outcome costs are kept as a check (Law, 2015).

Table 3: Generator of the synthetic decision stream: attributes, sources, outcome model, decision costs and evaluation

Element	Specification	Value or distribution
Volume and horizon	Order-approval decisions per day; days; windows	1,500 per day; 540 days: fit 0–269 (one third of records), validation 270–359, test 360–539 (270,000 decisions)
Customer quality factor	Latent q correlating attributes so that silent corruption creates cross-field inconsistencies	q ~ N(0, 1); enters pay score (+9q), average delay (−2.5q), disputes (−0.5q in the log rate), credit utilisation (−0.08q), tenure and recency (±0.1–0.15q)
Attributes (source)	Eight numeric and two categorical	order value (ERP): log-normal, median 8,000 × e^{0.3 ·segment} × seasonal factor; tenure months (CRM): gamma; pay score (payments): N(70 − 5 ·segment + 9q, 11) clipped 0–100; credit utilisation (bureau): Beta(2, 3) + 0.05 ·segment − 0.08q; days since last order (ERP): gamma; average delay (payments): gamma − 2 + 1.5 ·segment − 2.5q, clipped 0–90; disputes in 12 months (payments): Poisson; region risk (risk feed): N(0.4 + 0.1 ·channel, 0.2) clipped 0–1; segment (4 levels) and channel (3 levels) (CRM)
Outcome	Payment default, a logistic function of the attributes only	logit = −3.75 + 1.8 × (−0.025 ·(pay − 70) + 1.6 ·(util − 0.45) + 0.05 ·(delay − 5) + 0.35 ·disputes + 1.2 ·(region − 0.45) − 0.012 ·(tenure − 36) + 0.15 ·log(value/8,000) + 0.25 ·[segment 3] + interaction); prevalence 13.5%
Decision	Cost-sensitive approve or reject	Loss 60% of order value if a defaulter is approved; 12% of value if a good order is rejected; reject if $p > p^* = 0.12/0.72 = 0.167$
Decision model	Gradient boosting fitted on clean fit-window data	LightGBM, 300 trees, 31 leaves, learning rate 0.05; median imputation; hold-out AUC 0.86–0.87; AUC of the true probability 0.874
Cost evaluation	Expected cost given the generator's true probability (conditional Monte Carlo)	Removes outcome noise, which is 50 times larger than the defect effects; realised-outcome costs kept as a check
Replications	Independent streams	Seeds 4040–4044; all configurations replay identical records, defects and incidents (common random numbers)

Defects and incidents

Defects alter what the decision model sees, never the true outcome. Table 4 defines six types, with the mechanism, the mix parameters, their share of defects in the test windows, and their detectability. Missing values replace a value with null; stale values replace it with an earlier snapshot correlated 0.5 with the current value; scale errors multiply by 100 or 0.01; category remaps shift a categorical level; silent corruption takes the value from another record, as a mis-join would; and duplicates process a record twice. Defects arrive at two levels. Record-level defects affect 1.5% of records at random, one field each, with a mix that favours missing values. Source-level incidents start with probability 0.6% per source and day, last 1 + geometric(0.25) days (mean five), affect 10–50% of the records of one field of that source, and have a kind drawn from the same taxonomy; 4–7 incidents overlapped each 180-day test window. The resulting defect rate was 3.4–9.0% of records (mean 6.3%), of which 58–85% came from incidents. No duplicate incident fell in a test window, so the duplicate mechanism is implemented but untested in these runs. The taxonomy follows the data-cleaning literature (Rahm & Do, 2000; Abedjan et al., 2016; Chu et al., 2016) and the defect types used in a recent empirical study of their

effects on models (Mohammed et al., 2025); the two-level structure follows practitioner accounts that distinguish random record errors from pipeline incidents (Breck et al., 2019; Shankar et al., 2024; Paleyes et al., 2023).

Table 4: Defect taxonomy: mechanism, mix parameters, share of defects in the test windows, rule recall, detector AUC and reviewer detection probability

Defect type	Mechanism in the generator	Mix parameters	Share of defects in test windows, mean (range)	Rule recall	Detector AUC	Reviewer detection probability
Missing value	Value replaced by null	50% of numeric and 30% of categorical record-level defects; 28% of incidents	43% (20–77%)	1.00 (1.00–1.00)	1.00 (1.00–1.00)	0.95
Stale value	Earlier snapshot: mean + 0.5 (value − mean) + noise, kept in range	15% of numeric record-level defects; 22% of incidents	12% (2–28%)	0.01 (0.00–0.02)	0.59 (0.54–0.63)	0.50
Scale error	Value × 100 or × 0.01 (unit or decimal error)	20% of numeric record-level defects; 17% of incidents	8% (2–26%)	0.46 (0.43–0.47)	0.94 (0.91–0.95)	0.90
Category remap	Categorical level shifted to another level	70% of categorical record-level defects; 16% of incidents (numeric fields remapped to silent corruption)	8% (2–27%)	0.00 (0.00–0.01)	0.62 (0.61–0.62)	0.70
Silent corruption (mis-join)	Value taken from another record (mis-join)	15% of numeric record-level defects; 17% of incidents	29% (3–72%)	0.01 (0.00–0.01)	0.53 (0.52–0.54)	0.60
Duplicate record	Record processed twice; extra handling cost 10% of value	12% of ERP incidents only; none fell in a test window	0%	–	–	0.95

The monitoring layer

Table 5 lists the signals. Validation rules flag nulls, values outside eight plausible ranges, invalid categorical levels and exact duplicates (catching 60% of duplicates). A learned defect detector is a gradient-boosting classifier trained on the clean fit data into which 30% of records had defects of the known types injected, with raw values and null indicators as features; it scores every incoming record and its share of records scoring above 0.5 is tracked per batch against a clean baseline of 1.5% (Mohammed et al., 2025; Northcutt et al., 2021). An isolation forest on the same features is computed as an unsupervised benchmark (Liu et al., 2008; Chandola et al., 2009). Batch monitors compute, per field and day, a population-stability index against reference deciles of the fit window (with reference shares, so that discrete fields are handled correctly), the null share and the duplicate share (Quiñonero-Candela et al., 2009; Rabanser et al., 2019; Siddiqi, 2017; Schelter et al., 2018). Alerts fire when the PSI exceeds 0.10, the null share 2%, the duplicate share 2%, or the detector rate its baseline by three points. Decision sensitivity is $1 - |p - p^*| / p^*$, clipped to [0, 1], from the decision model's probability on the record as received, following the reject-option logic of selective classification (Chow, 1970; Geifman & El-Yaniv, 2017). The harm score is (0.5 ·flag + 0.5 ·detector) × (0.3 + 0.7 ·sensitivity) × (order value / mean order value): evidence that the record is wrong, weighted by how much the decision depends on it and by what is at stake.

The multiplicative form of the harm score is a design choice with alternatives. An additive score would route high-stakes records with no evidence of a defect, which wastes reviews; a score without the sensitivity term would route defective records whose decision is far from the threshold, whose correction changes nothing; and a score with only the rule flag would ignore the plausible-looking defects that the detector partly sees. The weights 0.5 and 0.5 on flag and detector, and 0.3 and 0.7 on the floor and the sensitivity term, were set by the author before the test runs and were not tuned; the threshold τ absorbs their scale. The score is reported as a ranking instrument, with its precision and regret capture as functions of the share of records routed, so that an organisation can replace it with a learned routing policy trained on its own regret estimates.

Table 5: Monitoring layer: signals, computation, thresholds and sources

Signal	Computation	Threshold or use	Source
Validation flag	Null, out-of-range (eight numeric fields), out-of-domain (two categorical), exact-duplicate check	Flag → re-fetch; enters harm score with weight 0.5	Rahm & Do (2000); Breck et al. (2019); ISO 8000-61 (ISO, 2016)
Learned defect score	LightGBM trained on the clean fit data with 30% injected defects of the known types; features: raw values and null indicators	Enters harm score with weight 0.5; batch detector rate = share of records with score > 0.5	Mohammed et al. (2025); Abedjan et al. (2016); Northcutt et al. (2021)
Isolation-forest anomaly score (benchmark)	Unsupervised score on imputed values, normalised between the 90th and 99.5th percentiles of clean data	Reported only; AUC ≈ 0.50 for these defects	Liu et al. (2008); Chandola et al. (2009)
Decision sensitivity	$1 - \lvert p - p^* \rvert / p^*$, clipped to [0, 1], from the decision model's probability on the record as received	Multiplier 0.3 + 0.7 × sensitivity in the harm score	Chow (1970); Geifman & El-Yaniv (2017); Elkan (2001)
Stakes	Order value divided by the mean order value in the fit window	Multiplier in the harm score	Elkan (2001); cost-sensitive decisions
Harm score	(0.5 ·flag + 0.5 ·detector) × (0.3 + 0.7 ·sensitivity) × stakes	Route if ≥ τ; priority in the queue; $\tau \in \{0.05 \ldots 0.7\}$ tuned on validation	Author's design (R1, R4)
Population stability index	Per field, against reference deciles (unique edges and reference shares) of the fit window; categorical PSI on level shares	Alert if > 0.10	Quiñonero-Candela et al. (2009); Rabanser et al. (2019); Siddiqi (2017)
Null share and duplicate share	Per field share of nulls in the day's batch; share of records flagged as duplicates	Alert if null share > 2% or duplicate share > 2%	Breck et al. (2019); Schelter et al. (2018)
Detector-rate alert	Share of records with defect score > 0.5 compared with the clean baseline (1.5%)	Alert if 3 points above baseline; triggers an all-source investigation	Author's design
Reviewer-confirmation escalation	Confirmed defects per source per day against the trailing 14-day mean	Escalate if ≥ 4 and ≥ 3 × trailing mean (feedback configuration)	Author's design (R6)

Governance configurations

Eight configurations were compared on identical records, defects and incidents (Table 6). NONE lets the model decide on data as received. RULES adds the automated base layer that all governed configurations share: validation rules; a re-fetch of flagged records after a quarter-day delay that repairs record-level defects with probability 0.5 but cannot repair incident defects, because the source is still broken; and removal of flagged duplicates. RULE-BASED REVIEW, representing common practice, routes the records still flagged after re-fetch to human review, prioritised by order value up to the daily capacity, plus a 0.5% random audit sample. HARM-BASED REVIEW routes records whose harm score exceeds a threshold τ, prioritised by harm up to the free capacity, plus the audit sample; τ was tuned per replication on the validation window over the grid {0.05, 0.08, 0.12, 0.18, 0.25, 0.35, 0.5, 0.7} by total cost. HARM-BASED REVIEW + FEEDBACK adds the feedback loops: batch alerts trigger an investigation of the attributed source (or of all sources on a detector-rate alert), reviewer confirmations of defects from one source that reach at least four in a day and three times the trailing fourteen-day mean trigger an escalation, each investigation costs 500 (1,500 for an all-source search) with a three-day cool-down per source, and an active incident found by an investigation is fixed two days later. A HINDSIGHT variant of this configuration uses the τ that minimises test-window cost, to show the cost of threshold mis-tuning. FULL REVIEW routes every record (capacity 1,500). ORACLE routes exactly the defective records whose decision differs from the clean-data decision, largest regret first, up to the standard capacity, with perfect detection and no overrides: an upper bound on record-level targeting. CLEAN DATA takes every decision on the true record: the floor. The difference between NONE and CLEAN DATA is the defect-induced cost, and the share of it that a configuration recovers is the primary outcome.

Table 6: Governance configurations: automated layer, human review, feedback and role in the study

Configuration	Automated layer	Human review	Feedback and fixes	Role in the study
No governance	None: the model decides on data as received	None	None	Baseline
Rules + re-fetch	Validation rules; flagged records re-fetched after 0.25 day (record-level defects repaired with probability 0.5); flagged duplicates removed	None	None	Automation-only baseline
Rule-based review	As above	Records still flagged after re-fetch, prioritised by order value up to the daily capacity, plus a 0.5% audit sample	None	Current practice: review queue driven by rules
Harm-based review	As above	Records with harm ≥ τ, prioritised by harm up to the free capacity, plus the audit sample	None	Framework without feedback loops
Harm-based review + feedback	As above	As above	Batch alerts and reviewer-confirmation escalations trigger investigations (cost 500; 1,500 for an all-source search); an active incident is fixed after 2 days	Full framework
Harm-based + feedback, hindsight τ	As above	As above, with τ chosen on the test window	As above	Sensitivity: cost of threshold mis-tuning
Full review	As above	Every record reviewed (capacity 1,500); workload factor applies	None	Upper bound on human effort
Oracle routing	As above	Exactly the defective records whose decision differs from the clean decision, largest regret first, up to capacity; perfect detection; no overrides	None	Upper bound on record-level targeting
Clean data	Decisions on the true records	None	None	Floor: cost without defects

Reviewer, queue and cost model

Table 7 lists the parameters. The review queue has a capacity of 60 records per day, priority by harm (or by value for the rule-based queue), and a service level that decides automatically any record waiting more than two days; routing is capacity-aware, so that the queue does not flood during incidents, and the sweep over capacity shows what the cap costs. Reviewers detect a defect with a type-specific probability — 0.95 for missing values and duplicates, 0.90 for scale errors, 0.70 for category remaps, 0.60 for silent corruption, 0.50 for stale values — reduced by automation bias (30% of scrutiny lost) and by a workload factor of 0.85 when the day's reviews reach 90% of capacity (Parasuraman & Manzey, 2010; Skitka et al., 1999; Goddard et al., 2012; Buçinca et al., 2021). A detected defect is corrected, which here means that the decision is taken on the true record; a detected duplicate is removed. When no defect is detected, the reviewer accepts the model's decision in 90% of cases and in 10% overrides it

with a judgement modelled as the model's logit plus noise of standard deviation 0.8, so that overrides add noise on average, in line with the evidence that humans rarely improve on a calibrated model from the same information (Green & Chen, 2019; Kahneman et al., 2021; Vodrahalli et al., 2022; Lai et al., 2023). Costs are 25 per review, 500 per investigation, 0.1% of order value per day of delay, and 10% of order value for a duplicate processed twice. The regret of a decision is its expected cost minus that of the clean-data decision; it is zero for a defect that does not change the decision, and can be negative for an individual record because the model is imperfect. Key performance indicators are total cost and its components (decision, review, delay, investigation), the share of defect-induced cost recovered, reviews per day, review precision (share of reviewed records that were defective), defect recall and correction rate, regret on defective records, incidents detected, time to detect and incident-days averted.

Table 7: Reviewer, queue and cost parameters with baseline values, sweeps and rationale

Parameter	Baseline value	Sweep	Rationale and sources
Review capacity	60 records per day	20–250	A small reviewer pool for a 1,500-decision stream; capacity-aware triage (R4)
Service level	Records waiting more than 2 days are decided automatically	–	Bounded delay; a backlog that is never cleared is a known failure mode of review queues
Audit sample	0.5% of records at random	–	Monitoring the defect rate outside the routed population; Raji et al. (2020)
Reviewer detection probability	Missing 0.95, duplicate 0.95, scale 0.90, category 0.70, silent 0.60, stale 0.50	× 0.6–1.2	Visible defects are easy, plausible values hard; author's assumption
Automation bias	Scrutiny reduced by 30%	0–70%	Parasuraman & Manzey (2010); Skitka et al. (1999); Goddard et al. (2012); Passi & Vorvoreanu (2022)
Workload effect	Detection × 0.85 when the day's reviews reach 90% of capacity	–	Buçinca et al. (2021); Parasuraman et al. (2000)
Override	10% of reviews without a detected defect replace the model's probability by a noisy judgement (SD 0.8 on the logit)	0–30%	Humans add noise as well as insight: Green & Chen (2019); Kahneman et al. (2021); Vodrahalli et al. (2022)
Review cost	25 per record	–	About ten minutes of analyst time
Investigation cost	500 per source; 1,500 for an all-source search; 3-day cool-down per source	–	Data-owner time; author's assumption
Delay cost	0.1% of order value per day of waiting; re-fetch 0.25 day	0–0.3%	Lost sales and customer dissatisfaction; stylised
Fix lag	2 days from investigation to fix	1–7	Depends on the data owner's responsiveness
Programme cost	400,000 set-up; 250,000 per year	× 0.5–4	Monitoring platform, stewards' tooling, model operations

Sweeps, value model and statistical analysis

For every replication, one parameter at a time was varied with the others at baseline: review capacity (20–250), routing threshold (the full grid, on the test window), automation bias (0–0.7), reviewer detection (× 0.6–1.2), override probability (0–0.3), fix lag (1–7 days), delay cost (0–0.3% per day), defect intensity (× 0.5–3, which moves the defect rate from 3.6% to 16.1%) and stakes (× 0.5–2). Sweep results are reported as means over the five replications with their ranges, because the feedback configuration's response to a single parameter is discontinuous — whether an incident is caught on its first or second day changes the outcome — and single-replication curves are noisy. A five-year net-present-value model converts the annualised difference in total cost between a baseline configuration and the framework into savings, against a set-up cost of 400,000 and 250,000 per year for the monitoring platform, stewards' tooling and model operations, at an 8% discount rate, with sensitivities to decision volume, defect intensity, stakes and programme cost and a break-even volume. Differences in total cost between the framework and each other configuration were analysed as paired observations across the five replications with paired t-tests and ranges; with five replications the tests are descriptive (Nadeau

& Bengio, 2003; Demšar, 2006). Verification checks were that the clean-data floor equals the cost without defects, that the oracle approaches the floor, that cost components add to totals, that common random numbers give identical defect counts across configurations until a fix changes them, and that expected and realised costs agree in sign and order of magnitude.

Results

The decision stream, its defects and their cost (RQ2)

Table 8 summarises the five replications. The decision model reached an AUC of 0.870 on clean test records and 0.869 on records as received, against 0.874 for the generator's own probability: at the aggregate level, defects affecting 3–9% of records are invisible in the model's discrimination, which is one reason why accuracy monitoring does not substitute for data-quality monitoring. The expected decision cost was 771 per record on clean data and 776 on data as received, against 759 for a perfect-information decision and 1,844 for approving every order. The gap of five per record is small in relative terms — 0.6% of decision cost — but at 1,500 decisions a day it is a defect-induced cost of 0.8–4.6 million per year (mean 2.7 million), and it is concentrated: Figure 4 shows, for the first replication, that defects arrive as a low daily baseline of record-level errors punctuated by source incidents that raise the count from about twenty to several hundred records a day, that the batch monitors react to incidents of missing values within the day while a stale-value incident on order values passed below the PSI threshold, and that harm scores separate defective from clean records but only weakly separate the defects that change a decision from those that do not. Of 16,877 defective records in that replication, 764 (4.5%) changed the decision, and almost all of them had a clean-data probability near the decision threshold (Figure 4d).

Table 8: Diagnostics per replication: defect rate, incident share, model and detector performance, decision cost and defect-induced cost per year

Replication (seed)	Defect rate	Share from incidents	Incidents / incident-days in test window	AUC of the decision model: clean / as received	Defect-detection AUC: learned detector / isolation forest	Harm-score AUC: any defect / decision-changing	Expected decision cost per record: clean / as received	Defect-induced cost per year (M)	Tuned τ: harm-based / + feedback
1 (4040)	6.3%	77%	4 / 21	0.871 / 0.870	0.92 / 0.50	0.88 / 0.90	771 / 774	1.68	0.25 / 0.25
2 (4041)	9.0%	85%	4 / 41	0.871 / 0.869	0.64 / 0.50	0.62 / 0.70	771 / 778	3.73	0.7 / 0.5
3 (4042)	7.9%	83%	7 / 45	0.870 / 0.869	0.84 / 0.50	0.80 / 0.82	772 / 776	2.55	0.5 / 0.7
4 (4043)	4.6%	69%	4 / 22	0.870 / 0.868	0.79 / 0.53	0.74 / 0.85	771 / 780	4.58	0.5 / 0.5
5 (4044)	3.4%	58%	4 / 17	0.871 / 0.870	0.68 / 0.51	0.64 / 0.81	770 / 771	0.78	0.5 / 0.7
Mean	6.3%	74%	4.6 / 29	0.870 / 0.869	0.77 / 0.51	0.74 / 0.82	771 / 776	2.66	–

Figure 4: Decision stream of replication 1: (a) defects per day and source incidents, (b) batch monitors, (c) harm-score distributions for clean, defective and decision-changing records, (d) regret of defective records against the clean-data probability

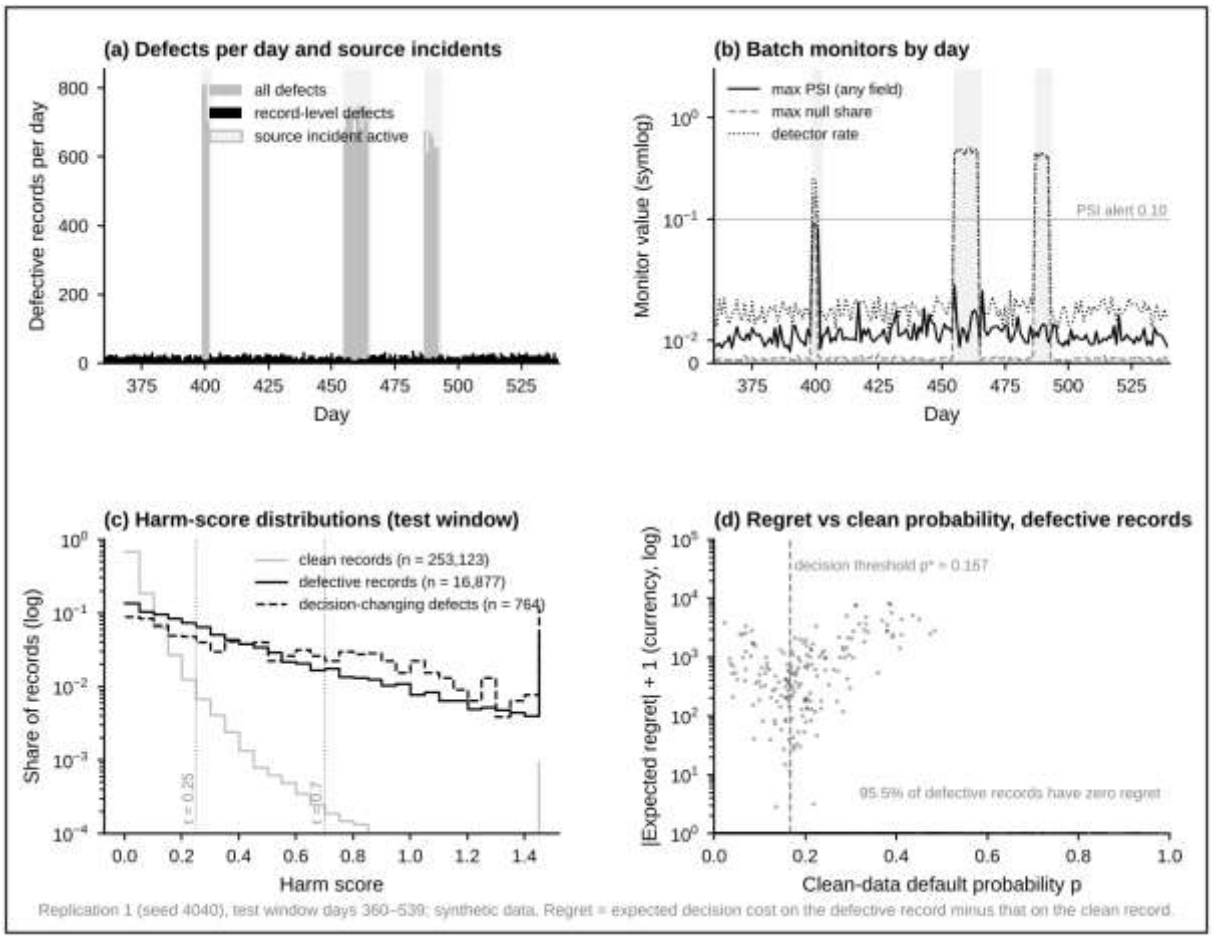


Figure 5 and Table 9 decompose the harm by defect type. Missing values were the most common defect (43% of defects on average, 20–77% across replications) and changed 4.2% of decisions at an expected regret of 39 per record, because median imputation returns the model to a typical probability; category remaps changed 3.1% at 12 per record. Scale errors changed 21% of decisions at 274 per record, silent corruption 9.1% at 108, and stale values 5.6% at 33. Weighted by frequency, silent corruption (33%), missing values (28%) and scale errors (26%) accounted for most of the defect-induced regret, with wide variation across replications depending on which incidents occurred. Detectability runs the other way. Validation rules caught every missing value and 46% of scale errors (those that left the plausible range) and essentially none of the other types, for an overall rule recall of 47% (23–79%). The learned detector separated missing values perfectly and scale errors well (AUC 0.94), but category remaps (0.62), stale values (0.59) and silent corruption (0.53) only slightly better than chance, because a plausible value drawn from another record is inconsistent with the rest of the record only through the weak cross-field correlations of the generator; the isolation forest, at an AUC of 0.50–0.53, did not detect these defects at all. The defects that are hardest to see are, per record, the most harmful.

Figure 5: Defect impact by type over five replications: (a) decision flip rate, (b) expected regret per defective record, (c) rule recall, (d) AUC of the learned detector and the isolation forest

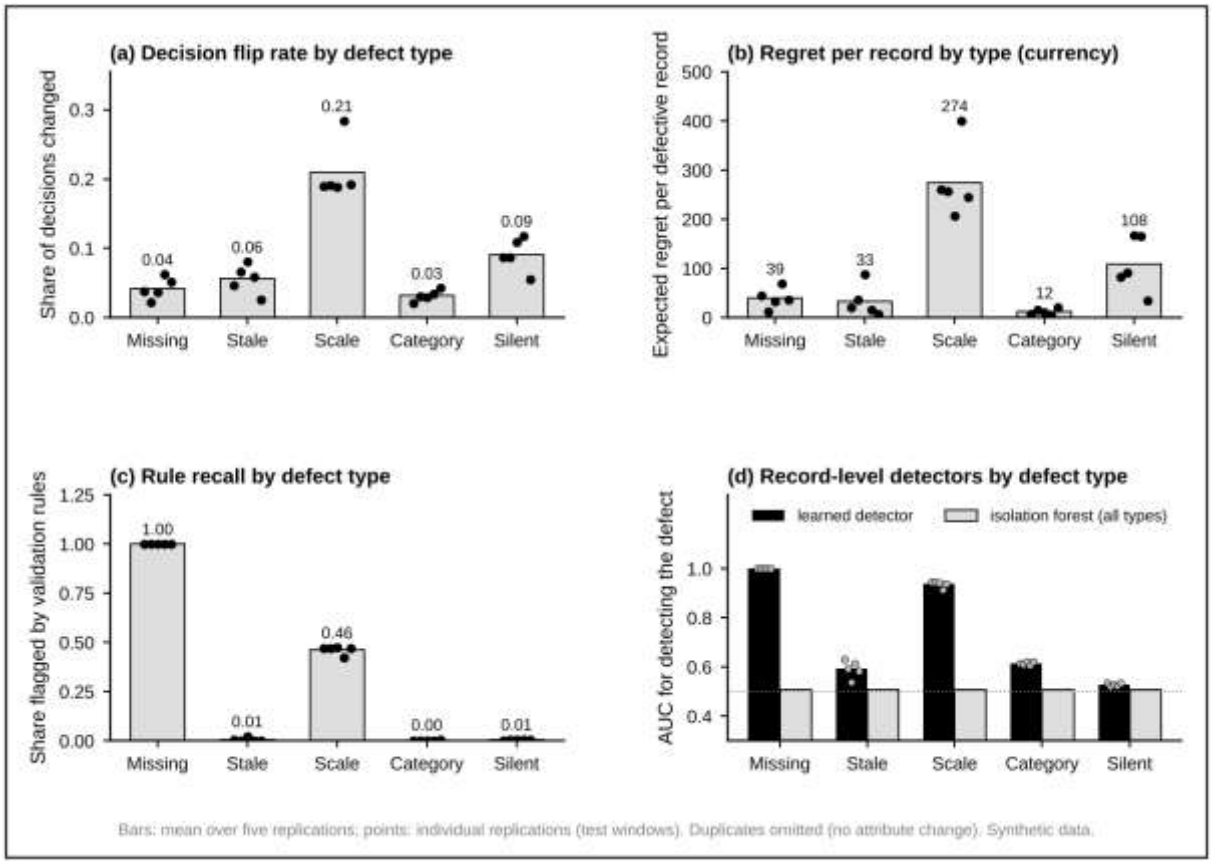


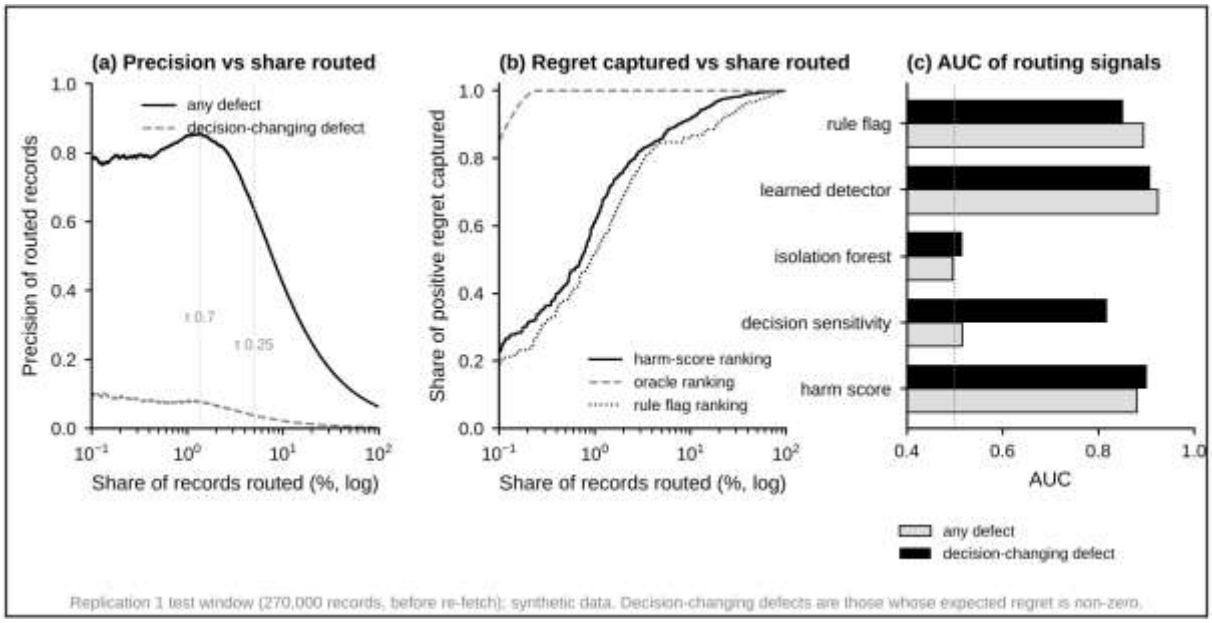
Table 9: Decision flip rate, expected regret, share of defect-induced regret, rule recall and detector AUC by defect type (means and ranges over five replications)

Defect type	Share of decisions changed, mean (range)	Expected regret per defective record, mean (range)	Share of defect-induced regret	Rule recall	Learned-detector AUC
Missing value	0.042 (0.023–0.062)	39 (13–69)	28% (4–71%)	1.00 (1.00–1.00)	1.00 (1.00–1.00)
Stale value	0.056 (0.026–0.081)	33 (7–88)	10% (0–41%)	0.01 (0.00–0.02)	0.59 (0.54–0.63)
Scale error	0.209 (0.189–0.284)	274 (208–400)	26% (8–57%)	0.46 (0.43–0.47)	0.94 (0.91–0.95)
Category remap	0.031 (0.020–0.043)	12 (6–20)	3% (0–13%)	0.00 (0.00–0.01)	0.62 (0.61–0.62)
Silent corruption (mis-join)	0.091 (0.055–0.118)	108 (35–168)	33% (4–87%)	0.01 (0.00–0.01)	0.53 (0.52–0.54)
Clean records (reference)	0.000 (0.000–0.000)	0	–	0 (no false flags)	–
All defective records	0.074 (0.045–0.136)	81 (42–181)	100%	0.47 (0.23–0.79)	0.77 (0.64–0.92)

Figure 6 examines targeting in the first replication. Ranked by harm score, the top 1% of records were 85% defective and captured 61% of the positive regret; the top 5% captured 86%. A ranking by rule flags and order value captured 52% and 85% at the same shares, so the harm score's advantage over a conventional queue is concentrated in the first percent. The oracle ranking captured all regret with 0.25% of records, because only 667 records had positive regret. Precision for decision-changing defects, however, never exceeded 8% at any routing share: the harm score finds defective records well (AUC 0.88 for any defect) and finds the records whose decision a defect changed moderately well (AUC 0.90), but the latter are so rare that even a good ranking routes a dozen records that need no correction for every one that does. Decision sensitivity alone had an AUC of 0.82 for decision-changing defects and 0.52 for defects in general, confirming its role in the score: among defective records, those with

sensitivity above 0.5 carried a mean absolute regret of 186 against 15 for the rest. The implication for the configuration comparison is that record-level review will spend most of its effort on records whose correction does not change the decision.

Figure 6: Targeting in replication 1: (a) precision of routed records against share routed, (b) share of positive regret captured by harm-score, rule-flag and oracle rankings, (c) AUC of routing signals for any defect and for decision-changing defects



Governance configurations compared (RQ3)

Table 10 and Figure 7 give the main comparison, Table 11 the paired differences. Annualised, the stream cost 424.8 million per year without governance and 422.2 million on clean data. Rules with automated re-fetch recovered 8% of the defect-induced cost (3–23%), at a delay cost but no review cost; the re-fetch repaired about a thousand record-level defects per test window but could not touch incident defects. Rule-based review, the common practice, recovered 9% (2–17%) with 17 reviews a day at a precision of 0.58 and a correction rate of 6.9% of defects; it recovered less than its precision suggests because the defects that rules flag — missing values — are the cheap ones, because reviewers under automation bias corrected only about 70% of what they saw, and because the 10% of reviews that ended in an override added decision cost. Harm-based review without feedback recovered nothing on average (−12%, range −41% to +21%): with the threshold tuned on the validation window it routed 14–42 records a day at a precision of 0.30, corrected 4% of defects, and spent more on reviews, delay and overrides than it saved. Its regret on defective records was lower than the rule-based queue's (2.1 against 2.2 million per year), so the targeting did find costlier defects, but not enough of them to pay for the reviews.

Table 10: Governance configurations: total cost, saving, share of defect-induced cost recovered, review load, precision, correction rate, regret and incident response (means and ranges over five replications)

Configuration	Total cost per year (M), mean (range)	Saving vs no governance per year (M)	Defect-induced cost recovered	Reviews per day	Review precision	Correction rate (share of defects)	Regret on defective records per year (M)	Incidents detected; time to detect; incident-days averted
No governance	424.8 (422.4–426.9)	0.00 (0.00–0.00)	0% (0–0%)	0.0 (0.0–0.0)	–	–	2.66 (0.78–4.58)	–
Rules + re-fetch	424.7 (422.2–426.8)	0.12 (0.07–0.18)	8% (3–23%)	0.0 (0.0–0.0)	–	–	2.47 (0.59–4.39)	–
Rule-based review	424.6 (422.3–426.3)	0.22 (0.07–0.64)	9% (2–17%)	17.2 (14.1–21.6)	0.58 (0.46–0.68)	0.069 (0.041–0.080)	2.19 (0.47–3.71)	–
Harm-based review	424.8 (422.7–425.9)	−0.01 (−0.43 to 0.98)	−12% (−41 to 21%)	22.2 (14.1–42.0)	0.30 (0.16–0.49)	0.040 (0.019–0.059)	2.11 (0.62–3.46)	–
Harm-based review + feedback	423.8 (422.4–424.8)	0.97 (0.03–2.05)	28% (4–46%)	19.7 (12.3–40.7)	0.19 (0.13–0.27)	0.040 (0.020–0.079)	1.21 (0.51–1.99)	3.6 of 4.6; 1.1 d; 14 of 29 d
Harm-based + feedback, hindsight τ	423.5 (422.4–424.2)	1.33 (0.03–2.73)	42% (4–60%)	22.5 (12.3–39.4)	0.19 (0.13–0.24)	0.050 (0.020–0.063)	0.85 (0.51–1.30)	4.0 of 4.6; 0.5 d; 17 of 29 d
Full review	441.7 (440.2–442.6)	−16.84 (−17.83 to −15.74)	−949% (−2277 to −344%)	1500.0 (1500.0–1500.0)	0.06 (0.03–0.09)	0.417 (0.351–0.477)	1.52 (0.46–2.35)	–
Oracle routing	422.3 (421.7–422.8)	2.47 (0.74–4.11)	93% (90–96%)	5.7 (2.4–9.5)	1.00 (1.00–1.00)	0.061 (0.041–0.094)	0.07 (0.00–0.36)	–
Clean data (floor)	422.2 (421.6–422.5)	2.66 (0.78–4.58)	100% (100–100%)	0.0 (0.0–0.0)	–	–	0.00 (0.00–0.00)	–

Figure 7: Governance configurations across five replications: (a) annual saving against no governance, (b) share of defect-induced cost recovered, (c) cost decomposition above the clean floor, (d) review load and precision

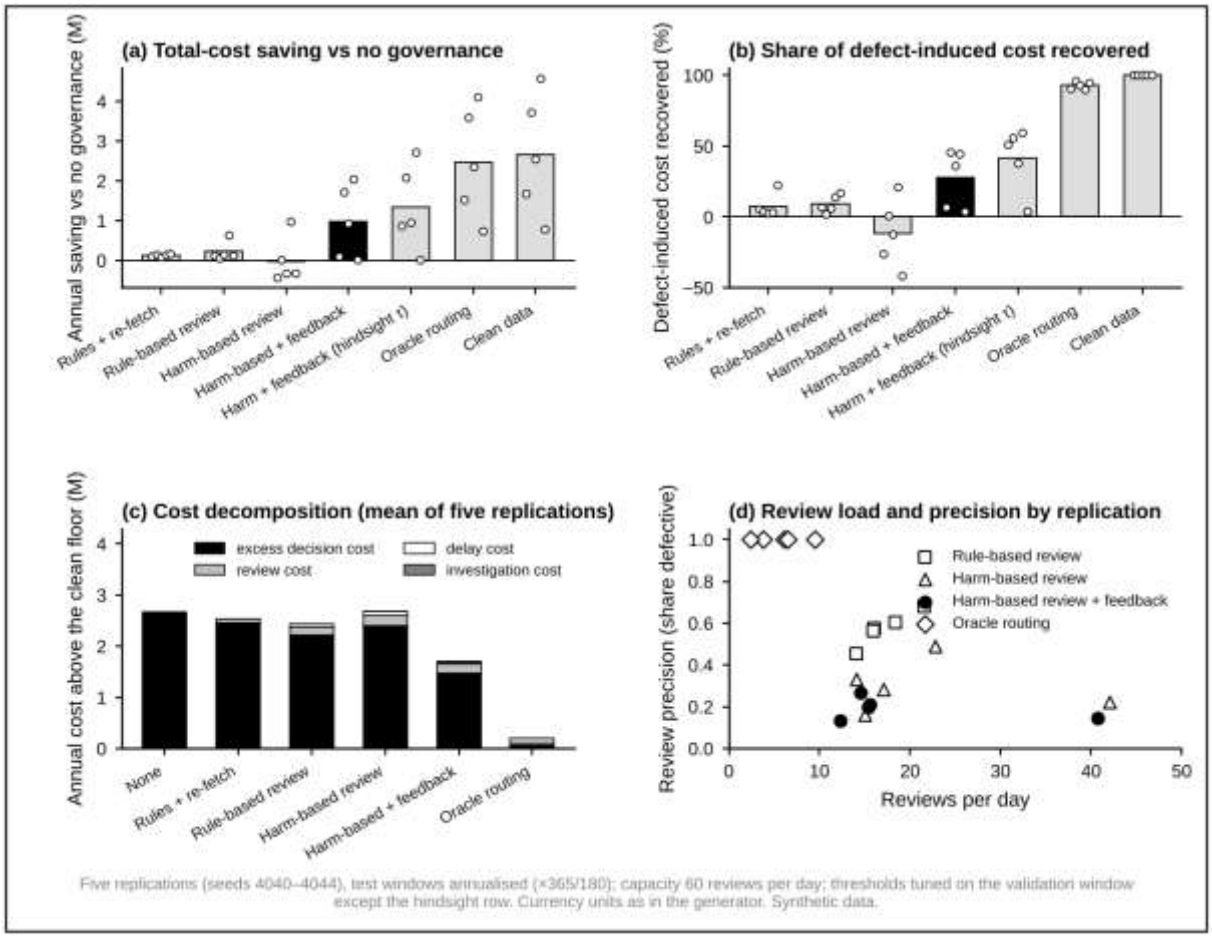


Table 11: Paired comparisons of total cost against harm-based review with feedback, and share of defect-induced cost recovered by replication

Comparator	Cost of reference minus comparator, per year (M)	Difference (% of comparator), mean (range)	Paired t-test (5 replications; descriptive)	Defect-induced cost recovered by the comparator, replications 1–5
No governance	−0.97	−0.23% (−0.48 to −0.01)	t = −2.37, p = 0.077	0%; 0%; 0%; 0%; 0%
Rules + re-fetch	−0.84	−0.20% (−0.45 to +0.04)	t = −2.06, p = 0.108	6%; 4%; 3%; 3%; 23%
Rule-based review	−0.74	−0.17% (−0.39 to +0.03)	t = −2.08, p = 0.106	7%; 2%; 6%; 14%; 17%
Harm-based review	−0.98	−0.23% (−0.40 to −0.08)	t = −4.06, p = 0.015	−26%; 1%; −13%; 21%; −41%
Harm-based review, hindsight τ	−0.75	−0.18% (−0.40 to −0.02)	t = −2.52, p = 0.065	3%; 1%; 7%; 21%; −20%
Harm-based + feedback, hindsight τ	+0.37	+0.09% (+0.00 to +0.18)	t = 2.33, p = 0.080	51%; 56%; 38%; 60%; 4%
Full review	−17.80	−4.03% (−4.11 to −3.86)	t = −84.91, p = 0.000	−1007%; −440%; −678%; −344%; −2277%
Oracle routing	+1.50	+0.36% (+0.17 to +0.49)	t = 6.42, p = 0.003	91%; 96%; 93%; 90%; 95%
Clean data (floor)	+1.70	+0.40% (+0.18 to +0.60)	t = 5.80, p = 0.004	100%; 100%; 100%; 100%; 100%
Harm-based review + feedback (reference)	0	0	–	7%; 46%; 36%; 45%; 4%

The full framework — harm-based review with feedback loops — recovered 28% of the defect-induced cost (4–46%), a saving of 0.97 million per year on average (0.03–2.05 million) and 0.23% of total cost (paired t = −2.37, p = 0.08 against no governance; p = 0.11 against the rule-based queue). The difference from harm-based review without feedback, which shares its routing, was consistent across replications (−0.23%, p = 0.015) and came almost entirely from incidents: batch alerts and reviewer-confirmation escalations detected 3.6 of 4.6 incidents per test window after a mean of 1.1 days (0 to 2.3), and the fixes averted 14 of 29 incident-days (49%). Figure 8 shows the mechanism in the first replication: three incidents of missing values were detected on their first day and ended two days later instead of running for two, ten and six days, so that the daily defect count fell back to its baseline while the rule-based queue was still saturated by flagged records; the stale-value incident on order values, which no field-level monitor flagged, was caught on its first day by the all-source search that the detector-rate alert triggered, because a missing-value incident in another source began on the same day and raised the detector rate. Regret on defective records fell from 2.66 to 1.21 million per year. The hindsight variant, with the threshold set on the test window, recovered 42% (4–60%) and saved 1.33 million per year, which measures the cost of threshold mis-tuning: in two replications the validation-tuned threshold was too low and filled the queue with low-precision candidates, and in the fifth replication the incidents were short and few, so that no configuration recovered more than 23%.

Across replications, incidents of missing values were detected on their first day by the null-share monitor whenever the affected share exceeded 2%, which every incident did; scale and category incidents triggered the PSI monitor; incidents of stale and silently corrupted values, whose marginal distributions barely move, were found by the detector-rate alert in some cases and by the escalation of reviewer confirmations in others, or not at all, which is why 1 of 4.6 incidents per window on average went undetected and why the mean time to detect was 2.3 days in the replication whose incidents were of those kinds. Investigations were imprecise: in the first replication 4 of 9 found an active incident, because alerts persisted for the length of an incident and re-triggered investigations after the cool-down; the cost was small (9,000 per window) relative to the review cost, but the all-source search triggered by the detector rate cost three times a targeted one and was the only route to incidents without a signature in any single field.

Figure 8: Daily dynamics in replication 1: (a) defects per day without and with root-cause fixes, (b) records routed to review under three configurations against capacity, (c) batch alerts and active incidents

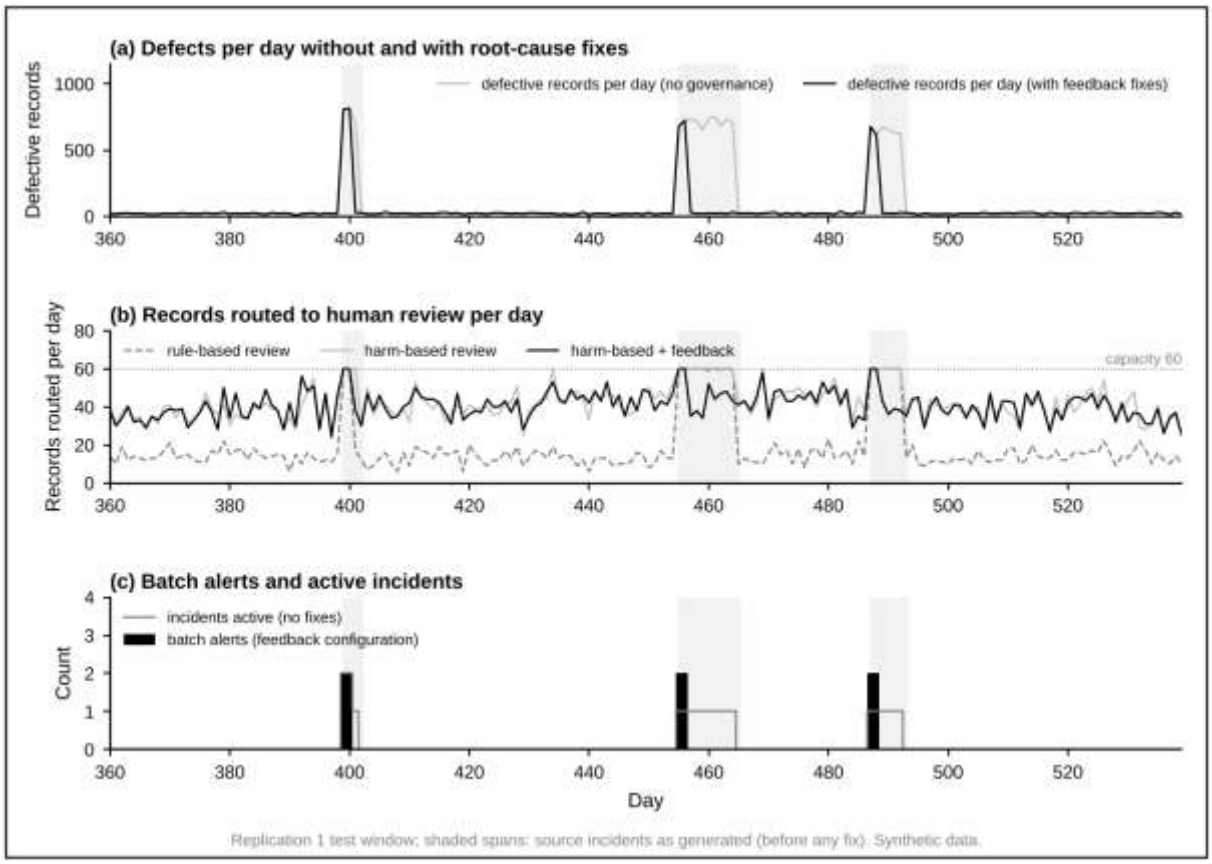


Figure 7c decomposes the annual cost above the clean floor. Without governance it is entirely excess decision cost, 2.7 million. Rules reduce the excess slightly and add a small delay cost. The rule-based queue adds 0.16 million of review cost and recovers 0.4 million of decision cost. Harm-based review without feedback adds 0.2 million of review cost and, because of overrides on correct records, leaves the excess decision cost almost unchanged. The framework adds the same review cost, a few thousand of investigation cost, and removes 1.2 million of excess decision cost. The oracle removes almost all of it for 0.05 million of reviews. Two checks on the accounting are worth recording. Expected and realised decision costs agreed in sign and order of magnitude — the realised regret per defect was 36–158 against 42–181 expected — but the realised totals carried outcome noise of about 1.2 million per window, fifty times the defect effect, which is why the expected-cost estimator was used. And the validation-window sweeps showed the same shape as the test-window sweeps, a loss at low thresholds and a plateau above 0.25, with a flatter plateau that let a lower threshold win in two replications.

The bounds frame these results. Full review of every record recovered −949% of the defect-induced cost: it found 93% of defects and corrected 42% of them, but 6.75 million of review cost per test window and 26,000 overrides cost ten times what the defects did. The oracle recovered 93% (90–96%) with 5.7 reviews a day, routing only records whose decision a defect had changed. The distance between the oracle and the framework is not reviewer capacity, which the oracle barely uses, but the ability to identify the few hundred records per window whose correction changes a decision — a targeting problem to which the current monitors offer an AUC of about 0.9 and a precision below 10%.

Sensitivity to capacity, reviewer behaviour, thresholds and defect intensity (RQ3)

Figures 9 and 10 and Table 12 report the sweeps as means over the five replications. Review capacity hardly mattered: the rule-based queue recovered 9–12% and the framework 24–29% at any capacity between 20 and 250 records a day, because routing was capacity-aware and the queue rarely needed more than the twenty most harmful candidates; the oracle rose from 85% at a capacity of 20 to 94% at

100, where it saturated. The routing threshold mattered most. At thresholds of 0.18 and below, harm-based review filled its capacity with candidates of low precision and recovered −50% of the defect-induced cost; at 0.35 and above it broke even, and the framework recovered 27–39% at thresholds between 0.35 and 0.7 with 13–23 reviews a day, against −10% to −15% at low thresholds. The validation-tuned thresholds fell on both sides of this range. Automation bias reduced the rule-based queue's recovery from 18% with fully attentive reviewers to 2% with 70% of scrutiny lost, and the framework's from 31% to 11%; reviewer detection skill moved the rule-based queue from 4% to 15% and the framework from 14% to 28%. Reviewer overrides were the most consequential behavioural parameter: with no overrides the framework recovered 44% and the rule-based queue 14%; at the baseline 10% override rate 28% and 9%; at 30%, the framework recovered −3% and full review −1,450%. Every review of a correct record is an opportunity to make a decision worse, and the number of such reviews, not the number of defects corrected, determined whether human involvement paid.

Figure 9: Sweeps over five replications: (a) review capacity, (b) routing threshold with the validation-tuned values marked, (c) automation bias, (d) reviewer detection skill

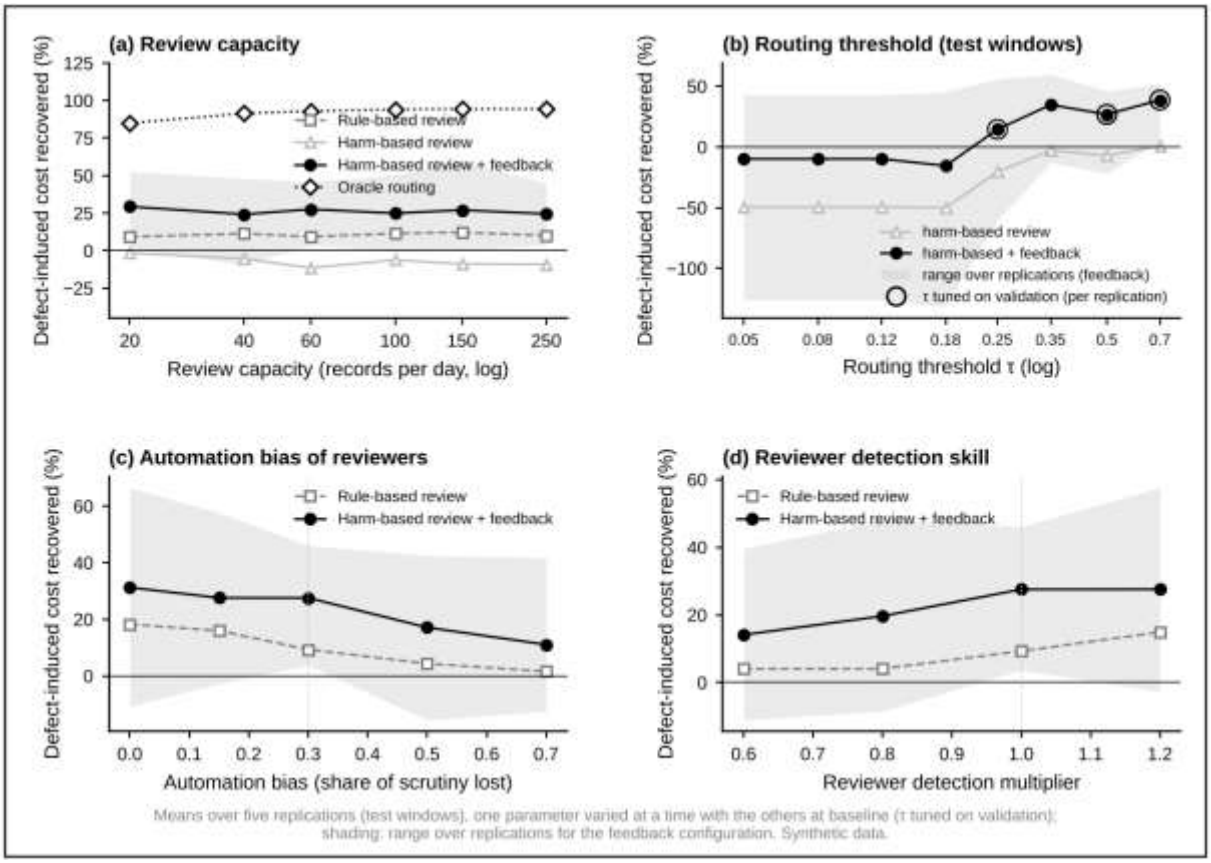


Figure 10: Sweeps over five replications: (a) reviewer overrides, (b) time to fix an incident, (c) cost of holding a record, (d) defect intensity, (e) decision stakes, (f) defect-induced cost and saving per year by defect intensity

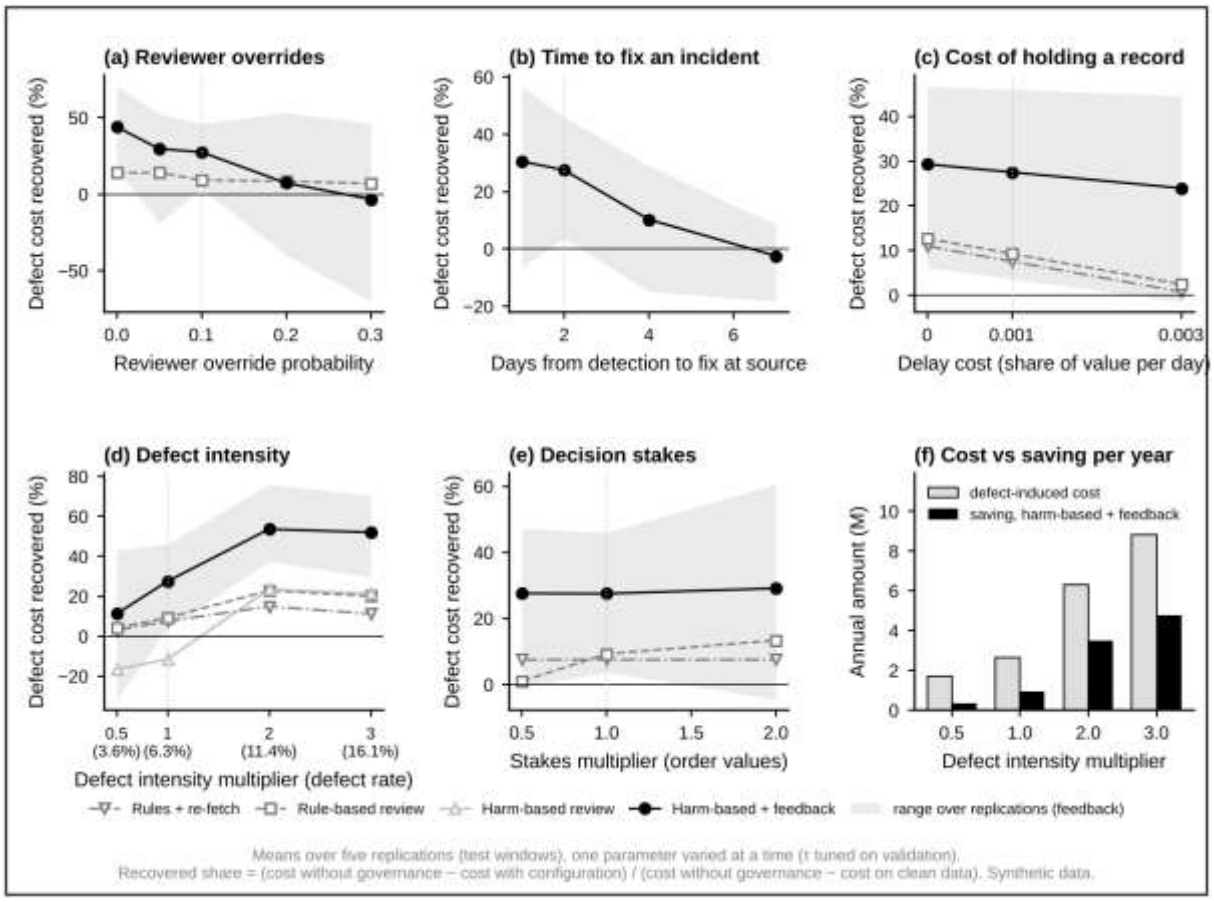


Table 12: Share of defect-induced cost recovered under one-at-a-time parameter sweeps (means over five replications)

Parameter	Values	Recovered share, baselines	Recovered share, framework	Other
Review capacity (records/day)	20 / 40 / 60 / 100 / 150 / 250	Rule-based: 9% / 11% / 9% / 11% / 12% / 10%	Harm + feedback: 29% / 24% / 28% / 25% / 27% / 25%	Oracle: 85% / 91% / 93% / 94% / 94% / 94%
Routing threshold τ (test windows)	0.05 / 0.12 / 0.18 / 0.25 / 0.35 / 0.5 / 0.7	Harm-based: −50% / −50% / −50% / −20% / −3% / −7% / 1%	Harm + feedback: −10% / −10% / −15% / 15% / 35% / 27% / 39%	Reviews/day (feedback): 60 / 60 / 59 / 39 / 23 / 16 / 13
Automation bias	0 / 0.15 / 0.3 / 0.5 / 0.7	Rule-based: 18% / 16% / 9% / 4% / 2%	Harm + feedback: 31% / 28% / 28% / 17% / 11%	Scrutiny lost reduces detection proportionally
Reviewer detection multiplier	0.6 / 0.8 / 1.0 / 1.2	Rule-based: 4% / 4% / 9% / 15%	Harm + feedback: 14% / 20% / 28% / 28%	Training and tooling for reviewers
Override probability	0 / 0.05 / 0.1 / 0.2 / 0.3	Rule-based: 14% / 14% / 9% / 9% / 7%	Harm + feedback: 44% / 30% / 28% / 7% / −3%	Full review: −705% / −829% / −949% / −1204% / −1450%
Fix lag (days)	1 / 2 / 4 / 7	–	Harm + feedback: 31% / 28% / 10% / −2%	Incident-days averted: 18 / 14 / 9 / 4 of 29
Delay cost (share of value per day)	0 / 0.001 / 0.003	Rules: 11% / 8% / 1%; rule-based review: 13% / 9% / 2%	Harm + feedback: 29% / 28% / 24%	Holding records is costly when defects are cheap
Defect intensity (defect rate)	0.5 (3.6%) / 1 (6.3%) / 2 (11.4%) / 3 (16.1%)	Rules: 3% / 8% / 15% / 11%; rule-based review: 4% / 9% / 23% / 20%	Harm + feedback: 11% / 28% / 54% / 52%	Saving per year (M): 0.36 / 0.97 / 3.51 / 4.79
Stakes multiplier	0.5 / 1 / 2	Rules: 8% / 8% / 8%; rule-based review: 1% / 9% / 13%	Harm + feedback: 28% / 28% / 29%	Recovered share stable; absolute saving scales with stakes

The feedback loop's value depended on the time to fix. With a one-day fix lag the framework recovered 31% and averted 18 of 29 incident-days; with two days 28% and 14 days; with four days 10% and 9 days; with seven days −2%, because the incidents were over before the fix arrived and the framework was left with the costs of review. The cost of holding a record mattered for the automated layer: with no delay cost, rules and the rule-based queue recovered 11–13%, and at 0.3% of order value per day 1–2%, because the quarter-day re-fetch of every flagged record during an incident of missing values held thousands of orders. Defect intensity moved the economics in the expected direction. At half the

baseline intensity (3.6% of records defective) the framework recovered 11% and saved 0.36 million per year; at twice (11.4%) 54% and 3.5 million; at three times (16.1%) 52% and 4.8 million, while the rule-based queue reached 20–23%. Stakes scaled the absolute saving without changing the recovered share, as the harm score's stakes term implies.

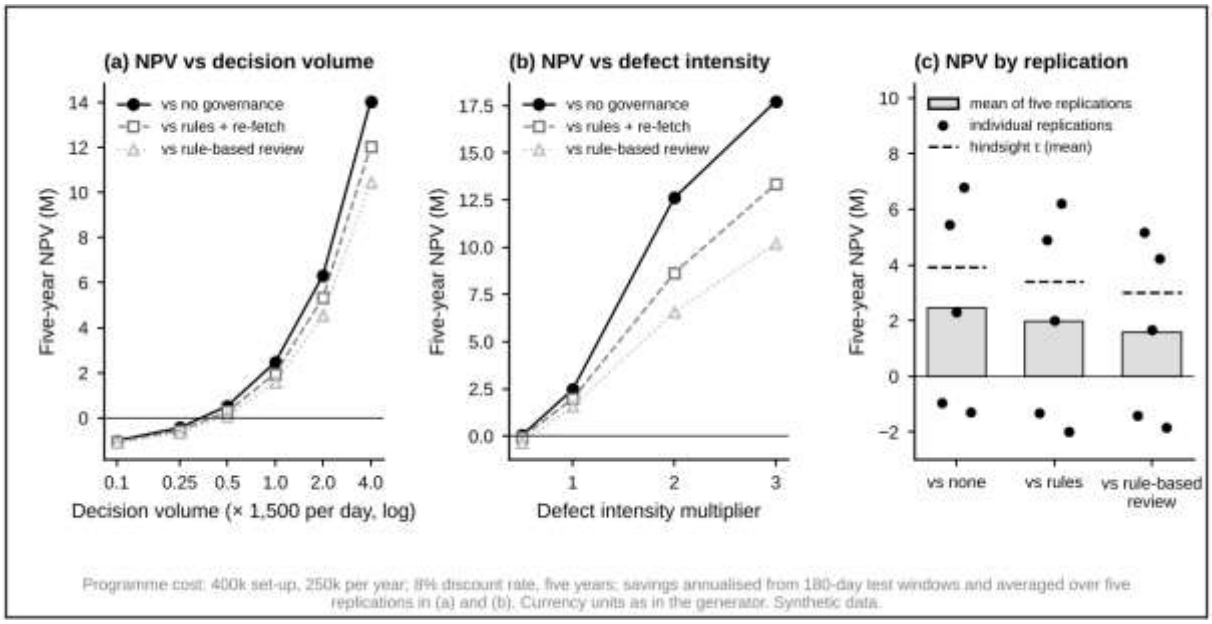
Value (RQ4)

Table 13 and Figure 11 give the value model. Against no governance, the framework saved 0.97 million per year on average, for a five-year net present value of 2.5 million after programme costs of 400,000 up front and 250,000 per year, with payback in about seven months; against the rule-based queue it saved 0.74 million per year for a net present value of 1.6 million. The ranges are wide: the net present value was negative in two of the five replications (−1.3 and −1.8 million against no governance and the rule-based queue respectively) and reached 6.8 million in the best. With the threshold set in hindsight the mean net present value rose to 3.9 million. Harm-based review without feedback had a negative net present value against every baseline. Break-even against no governance required about 540 decisions a day (0.36 of the baseline volume) and against the rule-based queue about 700; at four times the volume the net present value was 14 million, at a quarter of the volume −0.4 million. Doubling defect intensity raised the net present value against no governance to 12.6 million; halving it reduced it to zero. Doubling programme costs left 1.1 million; quadrupling them made the programme unprofitable against every baseline. The value of the framework is therefore insurance-like: it is earned in the windows in which incidents occur and are caught early, it depends on volume and on how dirty the data are, and a configuration without the incident-response loop does not earn it at all.

Table 13: Five-year value of the framework against three baselines, with hindsight and no-feedback variants, break-even volume and sensitivities

Case (1,500 decisions per day; set-up 400,000; 250,000 per year; 8%; five years; mean of five replications)	Saving per year (thousand)	NPV, five years (M), mean (range over replications)	Payback (years)
Harm-based review + feedback (τ tuned on validation) vs no governance	967	2.46 (−1.28 to 6.78)	0.6
Harm-based review + feedback (τ tuned on validation) vs rules + re-fetch	842	1.96 (−1.99 to 6.21)	0.7
Harm-based review + feedback (τ tuned on validation) vs rule-based review	743	1.57 (−1.83 to 5.18)	0.8
Harm-based + feedback, hindsight τ vs no governance	1,333	3.92 (−1.28 to 9.52)	0.4
Harm-based + feedback, hindsight τ vs rules + re-fetch	1,208	3.42 (−1.99 to 8.95)	0.4
Harm-based + feedback, hindsight τ vs rule-based review	1,108	3.03 (−1.83 to 6.96)	0.5
Harm-based review without feedback vs no governance	−14	−1.46 (−3.13 to 2.52)	none
Harm-based review without feedback vs rules + re-fetch	−139	−1.95 (−3.51 to 1.95)	none
Harm-based review without feedback vs rule-based review	−239	−2.35 (−3.61 to −0.04)	none
Break-even decision volume vs no governance / rules / rule-based review (× 1,500 per day)	–	0.36 / 0.42 / 0.47	–
Sensitivity: decision volume × 0.25 / 0.5 / 1 / 2 / 4 (vs no governance)	–	−0.43 / 0.53 / 2.46 / 6.32 / 14.05	–
Sensitivity: defect intensity × 0.5 / 1 / 2 / 3 (vs no governance; vs rule-based review)	–	0.03 / 2.46 / 12.62 / 17.74; −0.36 / 1.57 / 6.56 / 10.21	–
Sensitivity: stakes × 0.5 / 1 / 2 (vs no governance)	–	0.48 / 2.46 / 7.21	–
Sensitivity: programme cost × 0.5 / 1 / 2 / 4 (vs no governance; vs rule-based review)	–	3.16 / 2.46 / 1.07 / −1.73; 2.27 / 1.57 / 0.17 / −2.63	–

Figure 11: Value model: (a) net present value against decision volume, (b) against defect intensity, (c) net present value by replication against three baselines, with the hindsight-threshold mean



Discussion

What the evaluation says about the design

The evaluation supports three of the framework's design choices and qualifies a fourth. The feedback loops — batch alerts with source attribution, escalation of reviewer confirmations, and a fix at the source with a short lag — were the source of most of the recovered cost, which is consistent with the data-quality literature's long-standing argument that quality is improved at the source rather than by cleaning downstream (Redman, 1998; Otto, 2011; ISO, 2016) and with the finding from machine-learning operations that the organisational loop from alert to fix is where monitoring succeeds or fails (Breck et al., 2019; Shankar et al., 2024; Paleyes et al., 2023). The reviewers' second role, as a sensor for incidents that the batch monitors miss, is built into the design through the confirmation escalation; in these runs the escalations fired alongside the batch alerts rather than instead of them, so their separate contribution could not be isolated. Capacity-aware triage with a service level kept the review queue stable through incidents that would have flooded a rule-driven queue, so that the sweep over capacity showed almost no dependence on the number of reviewers. And the harm score's sensitivity term did what cost-sensitive and selective-classification theory predicts (Elkan, 2001; Chow, 1970; Madras et al., 2018): regret concentrated among records near the decision threshold, and the score found decision-changing defects with an AUC of about 0.9.

The qualification concerns record-level review itself. Only about 5% of defective records changed a decision, so even a well-ranked queue spent most of its reviews on records whose correction did not matter, and every such review carried the risk of a harmful override. The result that harm-based review without feedback recovered nothing on average is therefore not a failure of the score but a property of the problem: when the defects that matter are rare and the reviewer is a noisy second opinion on the rest, the economics of review are governed by precision for decision-changing defects, which no signal in this study raised above 10%. This is the data-quality counterpart of the finding that human–AI teams rarely achieve complementary performance (Lai et al., 2023; Bansal et al., 2021; Green & Chen, 2019), and of the argument that humans add value where their information differs from the model's (De-Arteaga et al., 2020; Raghu et al., 2019). The framework's reviewers added value when they could see something the model could not — a wrong record, a broken source — and subtracted it when they second-guessed a correct decision. Designs that restrict overrides to cases with a detected defect, that make the override a deliberate act with a cognitive forcing function (Buçinca et al., 2021), or that route to reviewers only when the model's own uncertainty is high (Hüllermeier & Waegeman, 2021; Kendall & Gal, 2017; Lakshminarayanan et al., 2017) follow directly.

The threshold result deserves emphasis because it is a governance result rather than a modelling one. A threshold tuned on a 90-day validation window was too low in two of five replications and cost the framework a third of its value; the hindsight threshold recovered 42% against 28%. The validation window contained different incidents from the test window, and a threshold that was right for one incident mix was wrong for another. The decision-rights matrix assigns the threshold to the model owner with the board accountable; the evaluation adds that the threshold must be re-tuned on a rolling basis and bounded by the free capacity, that the audit sample provides the data for re-tuning, and that the KPIs that drive it — review precision and regret recovered — must be in the governance records. This is the operational content of continual improvement in ISO/IEC 42001 and of the Manage function in the NIST framework (ISO, 2023; NIST, 2023).

Design principles as a nascent design theory

The evaluation allows the framework to be restated as five design principles, in the form that design-science research uses to communicate a nascent theory (Gregor & Hevner, 2013; Baskerville et al., 2018). P1, route by harm: send to humans the records whose correction would change a costly decision, scored as evidence of a defect times decision sensitivity times stakes, not the records that are easiest to flag; the evaluation shows that the sensitivity term carries most of the discrimination for decision-changing defects. P2, automate the easy half: re-fetch, impute and remove before a human looks, because the defects that rules catch are the ones that automation can also repair, and because every review of a correct record risks a harmful override. P3, cap and prioritise: a review queue must be sized to its capacity, ordered by harm and bounded by a service level, so that incidents do not convert into backlogs and delay costs. P4, close the loop to the source: alerts, escalations, named owners and short fix lags recover more than record-level correction, and the reviewers' confirmations are a sensor for incidents the monitors miss. P5, measure regret, not defects: the governance KPIs that matter are the expected decision cost recovered, the precision of review for decision-changing defects and the time from incident to fix, and the routing threshold, capacity and audit share should be re-tuned against them. The principles are supported in the simulated setting and are offered as testable propositions for naturalistic evaluation; P1 and P5 in particular depend on being able to estimate regret, which requires either a calibrated model and a sample of re-fetched records or a reviewer-labelled sample of decisions. The principles transfer to decision types other than order approval to the extent that three conditions hold: the decision is taken by a calibrated model with asymmetric error costs, so that sensitivity and stakes are defined; records are assembled from sources whose incidents can be detected at the batch level and fixed by an owner; and reviewers can obtain a better record than the model saw. Claims triage, credit limits, replenishment and fraud holds meet these conditions. Decisions in which the human's information is not a better record but a different judgement — hiring, medical triage — fall outside the framework's scope, and the literature on complementary performance applies to them directly (Lai et al., 2023; De-Arteaga et al., 2020).

Implications for meaningful oversight and regulation

Table 14 maps the framework to the instruments that organisations must satisfy. For Article 10 of the EU AI Act, the monitoring layer and the audit trail of corrections are the data-governance practices; the evaluation adds the warning that such practices find visible defects reliably and plausible-looking ones poorly, so that documentation should state which defect types the controls cover (European Parliament and Council, 2024; Mohseni et al., 2021). For Article 14, the framework's answer to the question of what meaningful oversight is (Koulu, 2020; Wagner, 2019; Crootof et al., 2023) is operational: oversight is meaningful when it is targeted at the decisions a human can improve, when the human has the means to improve them (the true record, the source owner), and when its effect is measured. Oversight of every decision — the full-review configuration — is not more meaningful; it is more expensive and, through overrides, more harmful. Oversight that routes by rule flags is meaningful for the defects that rules catch and blind to the rest. The decision-rights matrix (Figure 2) supplies the accountability structure that the NIST framework's Govern function and ISO/IEC 38505-1 require (NIST, 2023; ISO, 2017; DAMA International, 2017; Khatri & Brown, 2010), and the KPIs supply the measurement that Article 14's requirement to understand the system's limitations presupposes. Model risk management offers the template for institutionalising this: data-quality controls validated, monitored and governed like the models they feed (Board of Governors of the Federal Reserve System & Office of the

Comptroller of the Currency, 2011; Office of the Comptroller of the Currency, 2021).

Table 14: Mapping of the framework to the EU AI Act, NIST AI RMF, ISO/IEC 42001, ISO 8000-61, ISO/IEC 25012, ISO/IEC 38505-1, OECD principles, DAMA-DMBOK and model risk management

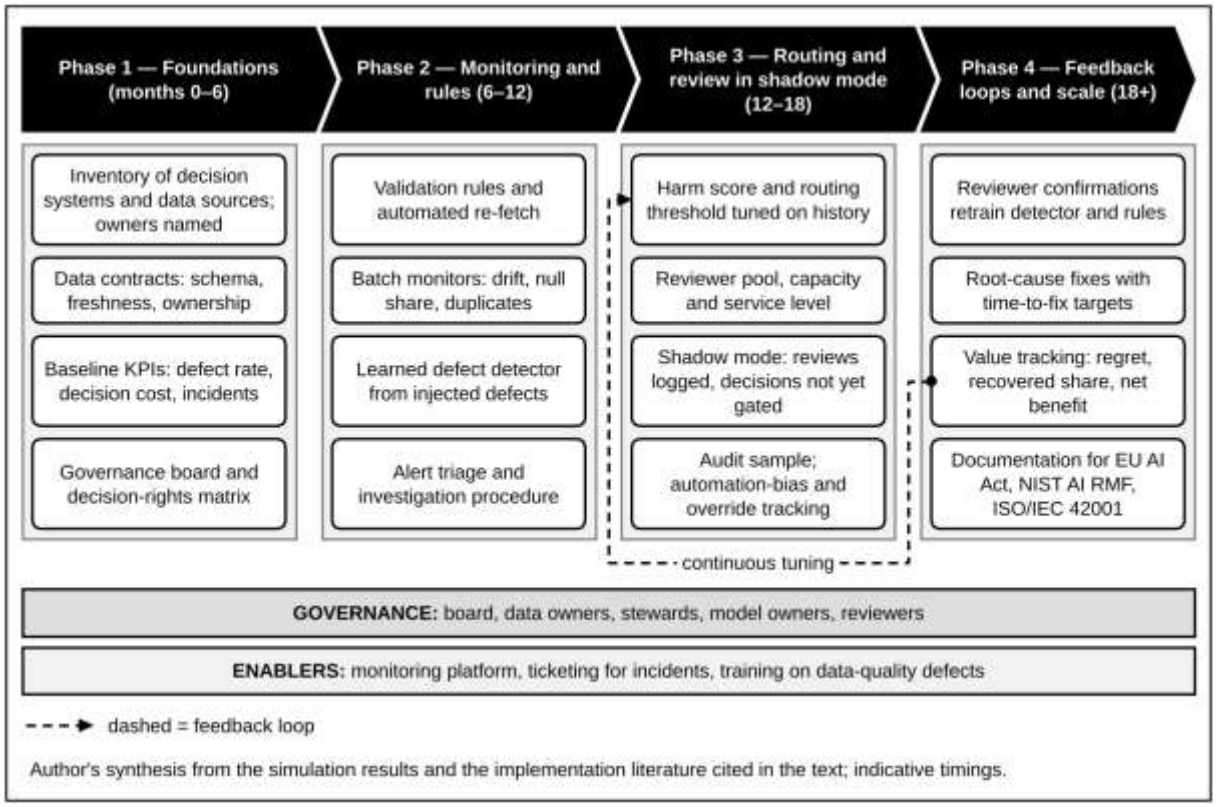
Instrument	What it asks	Framework component that answers it	Evidence this study provides
EU AI Act, Art. 10 (data and data governance)	Training, validation and testing data subject to data-governance practices; examination for possible biases; data-quality criteria	Data contracts, validation rules, learned detector, batch monitors; audit trail of corrections	Monitoring detects visible defects reliably, plausible-looking ones poorly (Table 9)
EU AI Act, Art. 14 (human oversight)	High-risk systems designed so that natural persons can oversee them, understand limitations, decide not to use the output, and intervene	Harm-based routing, decision gate, override logging, reviewer training on automation bias	Oversight is only valuable when targeted; untargeted review costs more than it saves (Tables 10–12)
NIST AI RMF (Govern, Map, Measure, Manage)	Accountable roles, risk mapping, measurement of trustworthiness characteristics, and management of incidents	RACI matrix (Figure 2); KPIs; incident investigation and fix; threshold re-tuning	Time to detect 0.4–2.3 days and 49% of incident-days averted on average with feedback loops (Table 10)
ISO/IEC 42001 (AI management system)	Plan–do–check–act management of AI, including data quality, impact assessment and continual improvement	Phased roadmap (Figure 12); feedback loops; value tracking	Programme NPV positive on average but negative in two of five replications at baseline (Table 13)
ISO 8000-61 (data-quality management process)	Process reference model: planning, control, assurance and improvement of data quality	Rules and re-fetch (control), monitoring (assurance), root-cause fixes (improvement)	Root-cause fixes recover more than record-level correction (Table 10)
ISO/IEC 25012 (data-quality model)	Characteristics such as accuracy, completeness, consistency, currentness	Defect taxonomy maps to characteristics: missing (completeness), stale (currentness), scale and silent (accuracy), category (consistency)	Harm per defect differs by two orders of magnitude across characteristics (Table 9)
ISO/IEC 38505-1 (governance of data)	Governing body accountability for data: value, risk and constraints	Governance board owns thresholds, capacity and release decisions	Decision-rights matrix (Figure 2); author's proposal
OECD AI Principles (2019)	Robustness, security and safety; accountability; transparency	Monitoring, decision records, documented roles	Qualitative mapping only
DAMA-DMBOK (2017)	Data governance, data quality and stewardship functions	Stewards triage alerts; owners fix sources; terminology of roles	Qualitative mapping only
Model risk management (SR 11-7; OCC, 2021)	Validation, ongoing monitoring, and governance of models in use	Decision-model monitoring, detector retraining, release approval	Analogy: data-quality controls as part of model risk management

Implications for practice and an implementation roadmap

Four practical implications follow. First, measure the defect-induced cost before buying reviewers: the comparison of decisions on records as received with decisions on re-fetched or corrected records, on a sample, gives the regret per defect by type and the share of defects that change decisions, which determine whether review can pay at all. Second, invest in the loop to the source before investing in the queue: incident detection with attribution, a named owner per source, a time-to-fix target and a ticketing path recovered more in this study than any amount of record-level review, and the fix-lag sweep shows that a fix that takes a week is worth nothing. Third, design the queue for realistic reviewers: cap it, prioritise it by harm, set a service level, train reviewers on the defect types that matter, track overrides and their outcomes, and expect automation bias. Fourth, treat the routing threshold, capacity and audit share as governance parameters with owners, evidence and a re-tuning cadence, not as configuration.

Figure 12 arranges these into a roadmap. Phase 1 establishes foundations: an inventory of decision systems and sources with named owners, data contracts, baseline KPIs and the governance board with its decision-rights matrix. Phase 2 builds monitoring and rules: validation rules with automated re-fetch, batch monitors, a learned detector trained from injected defects, and an alert-triage procedure. Phase 3 introduces routing and review in shadow mode, with the harm score and threshold tuned on history, a reviewer pool with capacity and service level, reviews logged but decisions not yet gated, and tracking of audits, automation bias and overrides. Phase 4 closes the loops: reviewer confirmations retrain detectors and rules, root-cause fixes carry time-to-fix targets, value is tracked as regret recovered and net benefit, and documentation is assembled for the AI Act, the NIST framework and ISO/IEC 42001. The timings are indicative; the sequence — source loop before queue, shadow mode before gating — is the recommendation.

Figure 12: Implementation roadmap for human-in-the-loop data-quality governance in four phases with governance and enabler bands



Limitations and further research

Table 15 lists the assumptions and their likely effects. The decision stream has one decision type and attributes that are nearly independent apart from a latent quality factor, which makes plausible-looking defects harder to detect than they might be in richer data and understates what a learned detector could achieve; a detector trained on an organisation's own defect history, with the reviewer confirmations the framework collects, is the obvious next step and was not simulated. The defect taxonomy has six types with fixed mixes, incidents affect one field of one source, and no duplicate incident fell in a test window. The reviewer model is stylised — type-specific detection, a single automation-bias parameter, a workload factor and a 10% noisy override — and the override assumption drives the cost of untargeted review; the sweeps show how much, and reviewer experiments or override logs from a real queue would replace the assumption with evidence (Green & Chen, 2019; Lai et al., 2023; Vodrahalli et al., 2022). The decision model is not retrained, so model drift is absent. Costs are stylised and the value model is a five-year NPV with four sensitivities. Five replications with 180-day windows and 4–7 incidents each give wide ranges, and the paired tests are descriptive. The evaluation is artificial in the FEDS sense (Venable et al., 2016); a naturalistic evaluation — the framework in shadow mode on a real

decision stream, with re-fetched records as partial ground truth — is the next step in the design cycle, and the testbed is offered as the instrument for specifying what to measure.

Table 15: Assumptions, limitations, likely effects on results, and verification or tests with real data

Component	Assumption or limitation	Likely effect on results	Verification done / how to test with real data
Decision stream	One decision type (order approval); attributes nearly independent apart from a latent quality factor	Plausible-looking defects are harder to detect than in richer data; understates detector value	Generator checked for prevalence, AUC and cost magnitudes; replace with historical decisions and labelled defects
Defects	Six types with fixed mixes; incidents affect one field of one source; no duplicate incident fell in a test window	Duplicate mechanism untested; correlated multi-field incidents absent	Incident log of a real pipeline; defect audits (Haug et al., 2011)
Decision model	Fitted once on clean data; not retrained	Drift-related model risk not represented	Monthly AUC on new data; retraining policy
Cost evaluation	Expected cost given the generator's true probability (conditional Monte Carlo); regret relative to the clean decision	Removes outcome noise; regret can be negative for a record because the model is imperfect	Realised-outcome costs kept as a check (regret per defect 36–158 realised vs 42–181 expected)
Reviewer model	Type-specific detection, automation bias, workload and 10% noisy overrides; perfect correction once a defect is detected	Override assumption drives the cost of untargeted review; sweeps 0–30%	Reviewer experiments (Green & Chen, 2019; Lai et al., 2023); override logs
Queue	FIFO with daily capacity and 2-day service level; capacity-aware triage	Removes unbounded backlogs; a rule-based queue without a cap would do worse in incidents	Queue logs; service-level reports
Threshold tuning	τ tuned once on a 90-day validation window	Mis-tuning cost visible: hindsight τ recovers 42% vs 28%	Rolling re-tuning on longer histories
Feedback loops	Batch alerts and confirmation escalations find incidents; fixes after 2 days; detector retraining not simulated	Incident value depends on fix lag (Table 12); retraining value not quantified	Time-to-fix statistics from incident tickets
Replication	Five streams, 180-day test windows, 4–7 incidents per window	Wide ranges; paired tests descriptive	Longer windows; more replications
Value model	Stylised review, delay, investigation and programme costs	Drive NPV; sensitivities reported	Actual cost accounting
Verification	Unit checks: clean-data floor equals the no-defect cost; oracle recovers 93%; accounting identities; common random numbers; realised vs expected costs	–	Sargent (2013); Law (2015)

Three research directions follow. The first is detection of plausible-looking defects, where cross-field consistency models, record-linkage signals and detectors trained on confirmed defects could raise the precision for decision-changing defects that limited every configuration here. The second is learning to defer for data quality: a routing policy trained to send a record to a human only when the expected improvement from correction exceeds the review cost, extending the learning-to-defer literature (Mozannar & Sontag, 2020; Madras et al., 2018) from predictions to data. The third is the empirical study of reviewers in data-quality roles — their detection rates by defect type, their override behaviour and their value as incident sensors — which the human–AI decision-making literature has not yet covered.

Conclusion

This article designed a human-in-the-loop AI governance framework for data quality in automated enterprise decision-making and evaluated it on a transparent simulation testbed in which the ground truth about defects, incidents and decision costs is known. The framework's components — validation

rules with automated re-fetch, a learned defect detector, batch monitors with source attribution, harm-based routing to a capacity-limited review queue, an explicit reviewer model, feedback loops to the source and a decision-rights matrix — meet ten requirements derived from the data-quality, machine-learning-operations, human–AI decision-making and AI-governance literatures and from the relevant standards.

On the simulated stream, defects affecting 3–9% of records cost 0.8–4.6 million per year in expected decision cost, but only about 5% of defective records changed a decision, and the defects hardest to detect were the most harmful per record. Rule-based review queues recovered 9% of the defect-induced cost, harm-based routing without feedback recovered nothing on average, and the full framework recovered 28% — 42% with a well-set threshold — almost entirely through the detection of source incidents within about a day and fixes that averted half of their duration. Full review was ruinous, and an oracle that routed only decision-changing defects recovered 93% with six reviews a day, which places the binding constraint in targeting rather than in capacity. The framework's five-year net present value was positive on average and negative in two of five replications, and depended most on reviewer overrides, time to fix, defect intensity and decision volume.

For organisations building oversight to satisfy the EU AI Act and comparable frameworks, the evaluation offers a concrete reading of "meaningful": oversight that is targeted at decisions a human can improve, equipped with the means to improve them, connected to the source of the data, and measured. Human involvement in data-quality governance pays when humans do what models cannot — see that a record is wrong, fetch the right one, and notice that a source is broken — and costs when they second-guess correct decisions on correct records. The testbed, assumptions and code are provided so that the design can be re-evaluated against an organisation's own defects, reviewers and costs.

REFERENCES

[1]. Abedjan, Z., Chu, X., Deng, D., Fernandez, R. C., Ilyas, I. F., Ouzzani, M., Papotti, P., Stonebraker, M., & Tang, N. (2016). Detecting data errors: Where are we and what needs to be done? *Proceedings of the VLDB Endowment*, *9*(12), 993–1004. https://doi.org/10.14778/2994509.2994518

[2]. Abraham, R., Schneider, J., & vom Brocke, J. (2019). Data governance: A conceptual framework, structured review, and research agenda. *International Journal of Information Management*, *49*, 424–438. https://doi.org/10.1016/j.ijinfomgt.2019.07.008

[3]. Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction machines: The simple economics of artificial intelligence*. Harvard Business Review Press. https://openlibrary.org/isbn/9781633695672

[4]. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300233

[5]. Bainbridge, L. (1983). Ironies of automation. *Automatica*, *19*(6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8

[6]. Ban, G.-Y., & Rudin, C. (2019). The big data newsvendor: Practical insights from machine learning. *Operations Research*, *67*(1), 90–108. https://doi.org/10.1287/opre.2018.1757

[7]. Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445717

[8]. Baskerville, R., Baiyere, A., Gregor, S., Hevner, A., & Rossi, M. (2018). Design science research contributions: Finding a balance between artifact and theory. *Journal of the Association for Information Systems*, *19*(5), 358–376. https://doi.org/10.17705/1jais.00495

[9]. Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), Article 16. https://doi.org/10.1145/1541880.1541883

[10]. Benbya, H., Pachidi, S., & Jarvenpaa, S. L. (2021). Special issue editorial: Artificial intelligence in organizations: Implications for information systems research. *Journal of the Association for Information Systems*, *22*(2), 281–303. https://doi.org/10.17705/1jais.00662

[11]. Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, *45*(3), 1433–1450. https://doi.org/10.25300/MISQ/2021/16274

[12]. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences, 191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

[13]. Bertsimas, D., & Kallus, N. (2020). From predictive to prescriptive analytics. *Management Science*, *66*(3), 1025–1044. https://doi.org/10.1287/mnsc.2018.3253

[14]. Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). *Supervisory guidance on model risk management* (SR Letter 11-7). Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

[15]. Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 1123–1132). IEEE. https://doi.org/10.1109/BigData.2017.8258038

[16]. Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data validation for machine learning. In *Proceedings of Machine Learning and Systems 1 (MLSys 2019)* (pp. 334–347). MLSys. https://proceedings.mlsys.org/paper_files/paper/2019/hash/928f1160e52192e3e0017fb63ab65391-Abstract.html

[17]. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. *American Economic Journal: Macroeconomics*, *13*(1), 333–372. https://doi.org/10.1257/mac.20180386

[18]. Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces (IUI '20)* (pp. 454–464). ACM. https://doi.org/10.1145/3377325.3377498

[19]. Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), Article 188. https://doi.org/10.1145/3449287

[20]. Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, *14*, Article 2. https://doi.org/10.5334/dsj-2015-002

[21]. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, *41*(3), Article 15. https://doi.org/10.1145/1541880.1541882

[22]. Chow, C. K. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, *16*(1), 41–46. https://doi.org/10.1109/TIT.1970.1054406

[23]. Chu, X., Ilyas, I. F., Krishnan, S., & Wang, J. (2016). Data cleaning: Overview and emerging challenges. In *Proceedings of the 2016 International Conference on Management of Data (SIGMOD 2016)* (pp. 2201–2206). ACM. https://doi.org/10.1145/2882903.2912574

[24]. Crootof, R., Kaminski, M. E., & Price, W. N., II. (2023). Humans in the loop. *Vanderbilt Law Review*, *76*(2), 429–510. https://scholarship.law.vanderbilt.edu/vlr/vol76/iss2/2/

[25]. DAMA International. (2017). *DAMA-DMBOK: Data management body of knowledge* (2nd ed.). Technics Publications. https://openlibrary.org/isbn/9781634622349

[26]. Davenport, T. H., & Ronanki, R. (2018). Artificial intelligence for the real world. *Harvard Business Review*, *96*(1). https://hbr.org/2018/01/artificial-intelligence-for-the-real-world

[27]. De-Arteaga, M., Fogliato, R., & Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–12). ACM. https://doi.org/10.1145/3313831.3376638

[28]. Dell'Acqua, F., McFowland, E., III, Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality*. SSRN. https://doi.org/10.2139/ssrn.4573321

[29]. Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

[30]. Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30. https://jmlr.org/papers/v7/demsar06a.html

[31]. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

[32]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning (arXiv:1702.08608). arXiv. https://arxiv.org/abs/1702.08608

[33]. Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., & Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, *226*, Article 107599. https://doi.org/10.1016/j.ijpe.2019.107599

[34]. Ehrlinger, L., & Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, *5*, Article 850611. https://doi.org/10.3389/fdata.2022.850611

[35]. Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)* (pp. 1–19). ACM. https://doi.org/10.1145/3411764.3445188

[36]. Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence* (pp. 973–978). Morgan Kaufmann. https://cseweb.ucsd.edu/~elkan/rescale.pdf

[37]. Elmachtoub, A. N., & Grigas, P. (2022). Smart "predict, then optimize". *Management Science*, *68*(1), 9–26. https://doi.org/10.1287/mnsc.2020.3922

[38]. Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. *Information & Communications Technology Law*, *31*(1), 123–153. https://doi.org/10.1080/13600834.2021.1958860

[39]. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

[40]. European Parliament and Council of the European Union. (2024). *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)* (OJ L, 2024/1689). Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[41]. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[42]. Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting*, *25*(1), 3–23. https://doi.org/10.1016/j.ijforecast.2008.11.010

[43]. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

[44]. Fosso Wamba, S., Gunasekaran, A., Akter, S., Ren, S. J.-F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research*, *70*, 356–365. https://doi.org/10.1016/j.jbusres.2016.08.009

[45]. Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[46]. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44. https://doi.org/10.1145/2523813

[47]. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé, H., III, & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

[48]. Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates. https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html

[49]. Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, *19*(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

[50]. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

[51]. Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–24. https://doi.org/10.1145/3359152

[52]. Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Quarterly*, *37*(2), 337–355. https://doi.org/10.25300/MISQ/2013/37.2.01

[53]. Gunasekaran, A., Papadopoulos, T., Dubey, R., Fosso Wamba, S., Childe, S. J., Hazen, B., & Akter, S. (2017). Big data and predictive analytics for supply chain and organizational performance. *Journal of Business Research*, *70*, 308–317. https://doi.org/10.1016/j.jbusres.2016.08.004

[54]. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[55]. Haug, A., Zachariassen, F., & van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management*, *4*(2), 168–193. https://doi.org/10.3926/jiem.2011.v4n2.p168-193

[56]. Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, *154*, 72–80. https://doi.org/10.1016/j.ijpe.2014.04.018

[57]. Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105. https://doi.org/10.2307/25148625

[58]. Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

[59]. Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)* (pp. 159–166). ACM. https://doi.org/10.1145/302979.303030

[60]. Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

[61]. International Organization for Standardization. (2008). *ISO/IEC 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model*. ISO. https://www.iso.org/standard/35736.html

[62]. International Organization for Standardization. (2016). *ISO 8000-61:2016 Data quality — Part 61: Data quality management: Process reference model*. ISO. https://www.iso.org/standard/63086.html

[63]. International Organization for Standardization. (2017). *ISO/IEC 38505-1:2017 Information technology — Governance of IT — Governance of data — Part 1: Application of ISO/IEC 38500 to the governance of data*. ISO. https://www.iso.org/standard/56639.html

[64]. International Organization for Standardization. (2018). *ISO 31000:2018 Risk management — Guidelines*. ISO. https://www.iso.org/standard/65694.html

[65]. International Organization for Standardization. (2023). *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. ISO. https://www.iso.org/standard/81230.html

[66]. Ivanov, D., & Dolgui, A. (2021). A digital supply chain twin for managing the disruption risks and resilience in the era of Industry 4.0. *Production Planning & Control*, *32*(9), 775–788. https://doi.org/10.1080/09537287.2020.1768450

[67]. Ivanov, D. (2017). Simulation-based ripple effect modelling in the supply chain. *International Journal of Production Research*, *55*(7), 2083–2101. https://doi.org/10.1080/00207543.2016.1275873

[68]. Janssen, M., Brous, P., Estévez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. *Government Information Quarterly*, *37*(3), Article 101493. https://doi.org/10.1016/j.giq.2020.101493

[69]. 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

[70]. Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Little, Brown Spark. https://openlibrary.org/isbn/9780316451406

[71]. Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux. https://openlibrary.org/isbn/9780374275631

[72]. Kamath, A., Jia, R., & Liang, P. (2020). Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5684–5696). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.503

[73]. Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data, 6*(4), Article 15. https://doi.org/10.1145/2382577.2382579

[74]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*, 3146–3154. https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree

[75]. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates. https://proceedings.neurips.cc/paper_files/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[76]. Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148–152. https://doi.org/10.1145/1629175.1629210

[77]. Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, Ł., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S., ... Chatterjee, S. (2024). Towards algorithm auditing: Managing legal, ethical and technological risks of AI, ML and associated algorithms. *Royal Society Open Science*, *11*(5), Article 230859. https://doi.org/10.1098/rsos.230859

[78]. Koulu, R. (2020). Human control over automation: EU policy and AI ethics. *European Journal of Legal Studies*, *12*(1), 9–46. https://doi.org/10.2924/EJLS.2019.019

[79]. Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)* (pp. 126–137). ACM. https://doi.org/10.1145/2678025.2701399

[80]. Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)* (pp. 29–38). ACM. https://doi.org/10.1145/3287560.3287590

[81]. Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., & Tan, C. (2023). Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)* (pp. 1369–1385). ACM. https://doi.org/10.1145/3593013.3594087

[82]. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*. Curran Associates. https://proceedings.neurips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html

[83]. Law, A. M. (2015). *Simulation modeling and analysis* (5th ed.). McGraw-Hill Education. https://www.mheducation.com/

[84]. Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

[85]. Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)* (pp. 1–15). ACM. https://doi.org/10.1145/3313831.3376590

[86]. Lipton, Z. C. (2018). The mythos of model interpretability. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

[87]. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

[88]. Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

[89]. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., & Zhang, G. (2019). Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering, 31*(12), 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857

[90]. Madras, D., Pitassi, T., & Zemel, R. (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*. Curran Associates. https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html

[91]. 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

[92]. March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

[93]. Mikalef, P., Boura, M., Lekakos, G., & Krogstie, J. (2019). Big data analytics and firm performance: Findings from a mixed-method approach. *Journal of Business Research*, *98*, 261–276. https://doi.org/10.1016/j.jbusres.2019.01.044

[94]. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[95]. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

[96]. Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2025). The effects of data quality on machine learning performance on tabular data. *Information Systems*, *132*, Article 102549. https://doi.org/10.1016/j.is.2025.102549

[97]. Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*, *11*(3-4), Article 24. https://doi.org/10.1145/3387166

[98]. Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds and Machines*, *31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

[99]. Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics*, *26*(4), 2141–2168. https://doi.org/10.1007/s11948-019-00165-5

[100]. Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (PMLR 119)* (pp. 7076–7087). PMLR. https://proceedings.mlr.press/v119/mozannar20b.html

[101]. Mozannar, H., Bansal, G., Fourney, A., & Horvitz, E. (2024). Reading between the lines: Modeling user behavior and costs in AI-assisted programming. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)* (pp. 1–16). ACM. https://doi.org/10.1145/3613904.3641936

[102]. Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. *Machine Learning*, *52*(3), 239–281. https://doi.org/10.1023/A:1024068626366

[103]. National Institute of Standards and Technology. (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

[104]. Northcutt, C. G., Athalye, A., & Mueller, J. (2021). *Pervasive label errors in test sets destabilize machine learning benchmarks* (arXiv:2103.14749). arXiv. https://doi.org/10.48550/arXiv.2103.14749

[105]. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

[106]. OECD. (2019). *Recommendation of the Council on Artificial Intelligence* (OECD/LEGAL/0449). OECD. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

[107]. Office of the Comptroller of the Currency. (2021). *Comptroller's handbook: Model risk management* (Version 1.0). Office of the Comptroller of the Currency. https://www.occ.gov/publications-and-resources/publications/comptrollers-handbook/files/model-risk-management/index-model-risk-management.html

[108]. Otto, B. (2011). Organizing data governance: Findings from the telecommunications industry and consequences for large service providers. *Communications of the Association for Information Systems*, *29*, Article 3. https://doi.org/10.17705/1CAIS.02903

[109]. Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2023). Challenges in deploying machine learning: A survey of case studies. *ACM Computing Surveys*, *55*(6), Article 114. https://doi.org/10.1145/3533378

[110]. Papagiannidis, E., Enholm, I. M., Dremel, C., Mikalef, P., & Krogstie, J. (2023). Toward AI governance: Identifying best practices and potential barriers and outcomes. *Information Systems Frontiers*, *25*(1), 123–141. https://doi.org/10.1007/s10796-022-10251-y

[111]. Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

[112]. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

[113]. Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *30*(3), 286–297. https://doi.org/10.1109/3468.844354

[114]. Passi, S., & Vorvoreanu, M. (2022). *Overreliance on AI: Literature review* (MSR-TR-2022-12). Microsoft Research. https://www.microsoft.com/en-us/research/publication/overreliance-on-ai-literature-review/

[115]. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-

learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

[116]. Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

[117]. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. https://doi.org/10.1145/505248.506010

[118]. Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2017). Data management challenges in production machine learning. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD 2017)* (pp. 1723–1726). ACM. https://doi.org/10.1145/3035918.3054782

[119]. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). *Dataset shift in machine learning*. MIT Press. https://doi.org/10.7551/mitpress/9780262170055.001.0001

[120]. Rabanser, S., Günnemann, S., & Lipton, Z. C. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems 32*. Curran Associates. https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html

[121]. Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., & Mullainathan, S. (2019). *The algorithmic automation problem: Prediction, triage, and human effort* (arXiv:1903.12220). arXiv. https://arxiv.org/abs/1903.12220

[122]. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, *23*(4), 3–13. https://sites.computer.org/debull/A00DEC-CD.pdf

[123]. Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

[124]. 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 (FAT* '20)* (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

[125]. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82. https://doi.org/10.1145/269012.269025

[126]. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[127]. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 39, 1–15). ACM. https://doi.org/10.1145/3411764.3445518

[128]. Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24. https://doi.org/10.1057/jos.2012.20

[129]. Schelter, S., Lange, D., Schmidt, P., Celikel, M., Biessmann, F., & Grafberger, A. (2018). Automating large-scale data quality verification. *Proceedings of the VLDB Endowment*, *11*(12), 1781–1794. https://doi.org/10.14778/3229863.3229867

[130]. Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)* (pp. 410–422). ACM. https://doi.org/10.1145/3581641.3584066

[131]. Schneider, J., Abraham, R., Meske, C., & vom Brocke, J. (2023). Artificial intelligence governance for businesses. *Information Systems Management*, *40*(3), 229–249. https://doi.org/10.1080/10580530.2022.2085825

[132]. Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics*, *36*(1), 120–132. https://doi.org/10.1111/jbl.12082

[133]. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems, 28*, 2503–2511. https://proceedings.neurips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

[134]. Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review*, *87*(3), 1085–1139. https://ir.lawnet.fordham.edu/flr/vol87/iss3/11/

[135]. Settles, B. (2012). *Active learning*. Morgan & Claypool. https://doi.org/10.2200/S00429ED1V01Y201207AIM018

[136]. Shankar, S., Garcia, R., Hellerstein, J. M., & Parameswaran, A. G. (2024). "We have no idea how models will behave in production until production": How engineers operationalize machine learning. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), Article 206. https://doi.org/10.1145/3653697

[137]. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

[138]. Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press. https://doi.org/10.1093/oso/9780192845290.001.0001

[139]. Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. (2019). Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

[140]. Siddiqi, N. (2017). *Intelligent credit scoring: Building and implementing better credit risk scorecards* (2nd ed.). Wiley. https://doi.org/10.1002/9781119282396

[141]. Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

[142]. Sonnenberg, C., & vom Brocke, J. (2012). Evaluations in the science of the artificial – Reconsidering the build-evaluate pattern in design science research. In *Design Science Research in Information Systems. Advances in Theory and Practice (DESRIST 2012), Lecture Notes in Computer Science, Vol. 7286* (pp. 381–397). Springer. https://doi.org/10.1007/978-3-642-29863-9_28

[143]. Tallon, P. P. (2013). Corporate governance of big data: Perspectives on value, risk, and cost. *Computer*, *46*(6), 32–38. https://doi.org/10.1109/MC.2013.155

[144]. Thomas, L. C., Crook, J., & Edelman, D. (2017). *Credit scoring and its applications* (2nd ed.). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611974560

[145]. Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[146]. Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW1), 1–38. https://doi.org/10.1145/3579605

[147]. Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. *European Journal of Information Systems*, *25*(1), 77–89. https://doi.org/10.1057/ejis.2014.36

[148]. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[149]. Vodrahalli, K., Daneshjou, R., Gerstenberg, T., & Zou, J. (2022). Do humans trust advice more if it comes from AI? An analysis of human-AI interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)* (pp. 763–777). ACM. https://doi.org/10.1145/3514094.3534150

[150]. Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

[151]. Wagner, B. (2019). Liable, but not in control? Ensuring meaningful human agency in automated decision-making systems. *Policy & Internet*, *11*(1), 104–122. https://doi.org/10.1002/poi3.198

[152]. Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: A revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84. https://doi.org/10.1111/jbl.12010

[153]. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, *12*(4), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[154]. Wang, X., & Yin, M. (2021). Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)* (pp. 318–328). ACM. https://doi.org/10.1145/3397481.3450650

[155]. Weber, K., Otto, B., & Österle, H. (2009). One size does not fit all—A contingency approach to data governance. *Journal of Data and Information Quality*, *1*(1), Article 4. https://doi.org/10.1145/1515693.1515696

[156]. Whang, S. E., Roh, Y., Song, H., & Lee, J.-G. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *The VLDB Journal*, *32*(4), 791–813. https://doi.org/10.1007/s00778-022-00775-9

[157]. Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23*(1), 69–101. https://doi.org/10.1007/BF00116900

[158]. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, *135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

[159]. Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). *Data-centric artificial intelligence: A survey* (arXiv:2303.10158). arXiv. https://doi.org/10.48550/arXiv.2303.10158

[160]. Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)* (pp. 295–305). ACM. https://doi.org/10.1145/3351095.3372852

[161]. Žliobaitė, I., Pechenizkiy, M., & Gama, J. (2016). An overview of concept drift applications. In N. Japkowicz & J. Stefanowski (Eds.), *Big data analysis: New algorithms for a new society* (pp. 91–114). Springer. https://doi.org/10.1007/978-3-319-26989-4_4