Diagnosing resistance, designing adoption: a systematic review of human and organizational factors in artificial-intelligence implementation in engineering organizations

Helena Gutiérrez[1,2], Enrique Gutiérrez[3,4], and Martin Eigner[5]

[1]Department of Mechanical and Process Engineering, RPTU University of Kaiserslautern-Landau, Gottlieb-Daimler-Straße 44, 67663 Kaiserslautern, Germany
[2]Share PLM (Share Enterprises S.L.), C/ Irún 23, 1º, 28008 Madrid, Spain
[3]Departamento de Matemática Aplicada a las Tecnologías de la Información y las Comunicaciones, Escuela Técnica Superior de Ingeniería de Sistemas Informáticos, Universidad Politécnica de Madrid, Calle de Alan Turing s/n, Campus Sur, 28031 Madrid, Spain
[4]MIT linQ, Massachusetts Institute of Technology, Cambridge, MA 02139, United States
[5]Former Head, Chair of Virtual Product Development (VPE), Department of Mechanical and Process Engineering, RPTU University of Kaiserslautern-Landau, Gottlieb-Daimler-Straße 44, 67663 Kaiserslautern, Germany

Corresponding author: Enrique Gutiérrez, enrique.gutierrez.alvarez@upm.es, ORCID 0000-0002-0045-7372

Abstract

Practitioner surveys, self-reported and contested, find that most organizations use artificial intelligence (AI) and few see returns. People-related explanations of the gap are rarely tested with engineers. This systematic review asks which people-side factors shape the adoption, sustained use and value of AI in engineering organizations. Following PRISMA 2020, we searched Scopus and Web of Science, snowballed once, and screened with a human reviewer and a large language model whose sensitivity was checked by blind audits. The 679 included reports (522 primary empirical studies) were synthesized on a framework crossing Maurer's three levels of resistance with Shea and Solomon's work-system levers. Of 274 field studies able to observe organizational change, 176 (64 %) report no change to any work-system lever, and 7 measure realized organizational value. In software engineering, which supplies 37 % of the reports, engineers relied on existing tests and reviews to check AI output, and senior engineers more often rejected or constrained it. Identity threat showed as identity-defining tasks kept from AI and concealed use; controls judged impractical went with unsanctioned use. Change and configuration management (12 studies) and validation and compliance (16), which carry traceability obligations, are the least studied. Outputs include a lifecycle gap map, an intervention matrix and ten hypothesized propositions. For engineering managers, we suggest planning and staffing the checking of AI output, which in software teams consumed much of the time generation saved, and judging roll-outs by logged use and value after the pilot.

Keywords: artificial intelligence adoption; generative AI; resistance to change; work design; professional identity; systematic literature review

1 Introduction

1.1 The adoption–value gap

In McKinsey's 2025 global survey, 88 % of organizations reported regular use of AI in at least one business function, while about 6 % of respondents qualified as high performers (Singla et al.,

2025). A preliminary report of July 2025 found 95 % of organizations getting zero return from generative AI (Challapally et al., 2025), a figure that rests partly on a convenience sample of conference attendees, counts pilots with no measurable impact *yet* as failures and is publicly contested. Both are practitioner figures resting on self-assessment and not peer reviewed; they motivate the question and do not enter the synthesis.

1.2 Organizational explanations of the gap

A special-issue editorial places much of what separates valuable deployments from the rest inside the organization (Benbya et al., 2020). Task-level effects of AI become firm performance only when complementary organizational resources are in place (Enholm et al., 2022), and a validated measure of firm AI capability includes coordination, change capacity and risk proclivity alongside data and technology (Mikalef and Gupta, 2021). Most key readiness factors identified in a review of 52 studies are organizational (Ali and Khan, 2025), although an expert ranking of nineteen success factors gave the technology category the greatest weight (Merhi, 2023). This evidence comes from frequency counts, expert ratings and cross-sectional surveys, none of engineers, and few mechanisms have been tested in the field.

1.3 AI compared with earlier engineering IT

Earlier engineering IT such as computer-aided design (CAD) and product lifecycle management (PLM) fits the task model of Autor et al. (2003), in which computers substitute for routine, rule-based work and complement non-routine problem solving, where engineering judgement lies. AI reaches into that layer: it lowers the cost of prediction and leaves judging what is wanted as the scarce human input (Agrawal et al., 2018). It also keeps learning after go-live, increasingly acts on its own and cannot be inspected by those who supervise it (Berente et al., 2021). Its gains are uneven: resolutions per hour rose by 15 % among 5,172 support agents, mostly the less experienced (Brynjolfsson et al., 2025), while among 758 consultants an assistant improved work inside its competence and cut correct solutions by 19 percentage points on a task just outside it (Dell'Acqua et al., 2026).

1.4 Engineering organizations as a setting

Engineering expertise rests on technical judgement exercised through tacit and relational knowledge acquired slowly in practice (Trevelyan, 2014), and in Abbott's terms a technology that performs parts of diagnosis and inference is a jurisdictional event for a profession (Abbott, 1988). Engineers rely on solvers and simulations with reproducible, traceable results, whereas the reliability of AI output cannot be read from the output itself (Endsley, 2023). Appropriate reliance then requires trust calibrated to actual capability (Lee and See, 2004), and engineers show more ambiguity aversion than the general population (Brown and Utley, 2019). Much engineering work is also regulated, and its designs are valuable intellectual property (IP). Experienced engineering designers raised IP concerns (Edgecomb et al., 2025), and in an automotive advanced-engineering unit, IP, data security and skill atrophy limited generative-AI use (Luke et al., 2026).

The closest precedent is PLM, which organizes engineering work around the product lifecycle. The PLM literature prescribes organizational change alongside the technology (Garetti et al., 2005; Eigner and Stelzer, 2009), and experts ranked management commitment, user involvement and cultural change among its critical success factors (Singh et al., 2020). Since PLM, unlike enterprise resource planning, prescribes little of the work, its use varied with how far employees perceived an advantage (Cantamessa et al., 2012). Engineering Management Journal has reported a similar dependence on organizational conditions for 3D scanning in automotive product development (Pereira and Liker, 2021) and, for model-based systems engineering, correlations between adoption and organizational structure (Henderson and Salado, 2024). Among

21 software engineers, the most frequent displacement reported was a shift from writing code to reviewing machine output (Tavantzis et al., 2026). Software differs from physical engineering in how cheaply output can be tested and changes reversed, and the review therefore treats it as its own stratum (Section 7.2).

1.5 What prior reviews cover

The 22 prior reviews and syntheses in Table 1 were chosen purposively before the database search, at least one per adjacent stream; no formal search identified them, so the comparison is illustrative. Acceptance, trust and aversion reviews pool end users, decision-aid users and laboratory participants (Kelly et al., 2023; Glikson and Woolley, 2020; Burton et al., 2020; Mahmud et al., 2022). Romeo and Lacko (2026) apply the technology–organization–environment (TOE) model to organizations and list resistance among AI's challenges and leadership and training among its enablers, Magara and Phahlane (2026) apply it to the industrial Internet of Things, and AI research in Engineering Management Journal concentrates on use cases, tools and firm capability (Jallow et al., 2023; Harrir and Triqui Sari, 2025; Ashergawi et al., 2026). The three reviews set in engineering domains survey AI methods by design stage (Yüksel et al., 2023) and by life-cycle-engineering subfield (Rahman et al., 2022), and deep generative models for design (Regenwetter et al., 2022). Stornelli et al. (2021) map firm-level barriers to generic adoption stages, the precedent we adapt for lifecycle mapping. Cieslak and Valor (2024) model resistance to digital transformation in three stages, and Parker and Grote (2022) how work design can be shaped under automation. None takes engineers as its population, links diagnoses of resistance to the organizational levers that could answer them, or maps that evidence onto engineering lifecycle stages.

1.6 Review questions and contributions

Five review questions were fixed in the protocol before screening.

RQ-A. What people-side factors, at individual, team and organizational level, are reported to enable or hinder AI adoption and sustained use in engineering organizations, and how are those factors conceptualized and measured?

RQ-B. Through which mechanisms do professional identity and trust, skills and learning, work design and autonomy, and leadership and participation operate, and how do they interact?

RQ-C. How does the evidence distribute across engineering lifecycle stages, AI types, organization types and study designs, and where are the gaps?

RQ-D. Which theories and frameworks are used, and how well do they accommodate AI's distinctive features: autonomous learning, opacity, and agency?

RQ-E. Which organizational interventions are associated with reduced resistance and sustained use?

We synthesized 679 reports, 157 from Scopus and Web of Science and 522 from snowballing; 522 of the 679 are primary empirical studies. Software engineering supplies 37 % of the reports, and the headline results are re-tested in nine subsets. The review contributes a framework matching Maurer's levels of resistance to Shea and Solomon's work-system levers, an evidence gap map by lifecycle stage with an estimate of the studies missed, ten propositions traced to the search arm and stratum of their support, and a research agenda with evidence-labelled guidance for engineering managers.

Table 1: Prior reviews compared with this review (selection in Section 1.5). Values follow each review's own reporting; n.r. not reported; — conceptual work without a corpus; TAM, UTAUT technology acceptance models; HRM human resource management; SCOR Supply Chain Operations Reference model.

Review	Corpus (n)	Population	Focus	Framework	Lifecycle mapping
Kelly et al. (2023)	60	Consumers, patients, students	AI acceptance predictors	TAM, UTAUT	None
Glikson and Woolley (2020)	n.r.	Laboratory and field users	Trust in AI	Own: representation × machine intelligence	None
Burton et al. (2020)	61	Decision-aid users, mostly laboratory	Algorithm aversion	Five inductive themes	None
Mahmud et al. (2022)	80	Decision makers, mostly student and online	Algorithm aversion	Inductive four-theme taxonomy	None
Jussupow et al. (2024)	—	None of its own	Algorithm aversion and appreciation	Four decision configurations	None
Bagočiūnaitė and Endriulaitienė (2026)	29	Employees, managers, HR professionals	Attitudes to AI: definition, measurement	Cognitive–affective	None
Vrontis et al. (2022)	45	Firms and employees, via HRM	AI and robotics in HRM	HRM strategies and activities	HR processes only
Pereira et al. (2023)	60	Employees at three levels	AI and workplace outcomes	Antecedents–phenomenon–outcomes	None
Oldemeyer et al. (2025)	71	Small and medium firms	AI implementation challenges	Inductive	None
Romeo and Lacko (2026)	196	Public and private organizations	Enablers and challenges of AI	TOE	None
Heimberger et al. (2026b)	47	Production firms	35 AI adoption factors	Own internal/external scheme	None
Stornelli et al. (2021)	87	Manufacturing firms	Advanced manufacturing technology barriers, enablers	Inductive	Three adoption-process stages
Magara and Phahlane (2026)	46	Industrial Internet of Things firms	AI applications, drivers, barriers	TOE	None
Haefner et al. (2023)	n.r.	Firms, cross-industry	Implementing and scaling AI	Own phase framework	Firm implementation phases
Hanelt et al. (2021)	279	Firms	Digital transformation	Own multi-dimensional framework	None
Vial (2019)	282	Organizations	Digital transformation	Inductive process model	None
Cieslak and Valor (2024)	63	Employees, generic	Resistance to digital transformation	Own three-stage model	Resistance-episode stages
Parker and Grote (2022)	—	Employees across sectors	Work design under automation	Work design, sociotechnical	None
Harrir and Triqui Sari (2025)	n.r.	None (supply-chain processes)	AI tools in supply chains	SCOR	SCOR processes (tools)
Yüksel et al. (2023)	n.r.	None (AI methods)	AI in engineering design	AI technique × design activity	Design stages (technical)
Regenwetter et al. (2022)	n.r.	None (algorithms)	Deep generative models in design	None	None
Rahman et al. (2022)	42	None (AI techniques)	AI in life cycle engineering	Life cycle engineering subfields	Same subfields (technical)
This review	679	Engineers and engineering organizations	People-side factors in AI adoption	Maurer's levels × work-system levers	Engineering lifecycle stages

2 Conceptual background

2.1 Technology acceptance models and AI

The Technology Acceptance Model explains an individual's use of a technology through perceived usefulness and perceived ease of use (Davis, 1989), and TAM2 adds social influence, including subjective norm under mandatory use (Venkatesh and Davis, 2000). The Unified Theory of Acceptance and Use of Technology (UTAUT) consolidated eight competing models into four determinants that explained about 70 % of the variance in behavioural intention (Venkatesh et al., 2003). At firm level, the technology–organization–environment framework treats implementation as a problem distinct from the adoption decision (Tornatzky and Fleischer, 1990), as diffusion theory does in separating the decision and implementation stages of the innovation-decision process (Rogers, 2003).

What these models predict is behavioural intention, of which usefulness and ease of use explain about half in meta-analyses (King and He, 2006), with coefficients that shift with respondent type and culture (Schepers and Wetzels, 2007). An organization needs sustained use and realized value (Section 2.4). Their treatment of emotion is thin, although Dwivedi et al. (2019) showed that attitude, which UTAUT had dropped, partially mediates the core paths to intention, and Beaudry and Pinsonneault (2010) found that the same emotion can raise or lower use depending on the coping behaviour it triggers. Identity and power are missing, although algorithmic systems, in the terms of Kellogg et al. (2020), restrict, record, rate and replace, and workers resist them individually and collectively. The models also treat the artefact as static, whereas an AI system learns and increasingly acts; in a meta-analysis of 136 studies acceptance was lower for AI in a performative role than in an advisory one (Li et al., 2026). AI literacy names the competencies needed to evaluate and work with such systems (Long and Magerko, 2020).

2.2 Maurer's three levels of resistance

Maurer's three levels of resistance serve as the diagnostic lens (Maurer, 1996, 2010). From the model we take the prescription to identify the level before responding, since answering emotional or relational resistance with more information is, in Maurer's account, a common failure.

Level 1 resistance, "I don't get it", is informational, and readiness for change is created chiefly through a change message that conveys discrepancy and efficacy (Armenakis et al., 1993), although readiness also has an affective and collective component (Rafferty et al., 2013). At Level 2, "I don't like it", resistance is emotional and identity-based. Dispositional resistance is measurable as routine seeking, emotional reaction to imposed change, short-term focus and cognitive rigidity (Oreg, 2003), and Piderit (2000) argued that responses to one change can diverge across cognition, emotion and behaviour into ambivalence. Identity research specifies what is threatened. Threat to IT identity, a technology become part of the self (Carter and Grover, 2015), has been operationalized as one measurable construct (Craig et al., 2019). In a bank that replaced credit decisions with AI, experienced consultants felt their role identity threatened while newcomers felt theirs strengthened (Strich et al., 2021). AI strengthens identity when it complements or creates tasks and threatens it when it replaces defining ones (Selenko et al., 2022), and loss of status is the strongest measured correlate of AI identity threat (Mirbabaie et al., 2022). Level 3, "I don't trust you", is relational and concerns those leading the change. Change agents contribute to resistance by breaking agreements, violating trust and failing to legitimize the change (Ford et al., 2008), and resistance to a new system forms before go-live (Kim and Kankanhalli, 2009). In trust terms, Level 3 is a judgement about the benevolence and integrity of management (Mayer et al., 1995), whereas trust in the system's ability belongs to the professional identity and trust dimension (Section 2.4).

Maurer's model implies escalation, in which unanswered Level 1 concerns harden into Level 2 and dismissed Level 2 concerns become Level 3 distrust. Part of this sequence appears in hospital

implementations of clinical information systems, where the object of physicians' resistance shifted from the system to its significance and then to its advocates (Lapointe and Rivard, 2005).

2.3 The Work Systems Model

In the founding sociotechnical study an efficient longwall mining layout failed because it destroyed the self-regulating work groups on which production depended (Trist and Bamforth, 1951), and Cherns (1976) codified the lesson into design principles. Clegg (2000) made task allocation between humans and machines an explicit design choice, and Baxter and Sommerville (2011) found that these methods rarely reach systems engineering practice, partly because social and organizational questions arrive too late in the lifecycle to change anything.

The intervention lens of this review is the Work Systems Model of Shea and Solomon (2020), stated in peer-reviewed form by Day and Shea (2020). Its premise is that behaviour follows the work environment more reliably than exhortation does, and eight levers describe that environment: organization, workplace design, task, people, rewards, measurement, information distribution and decision allocation. The authors' rule of thumb is that at least four levers must change coherently before people experience a different environment. Technology does not determine work design: the same AI can enrich or impoverish a job depending on managerial choices and participation (Parker and Grote, 2022), and automating the easy parts of a task leaves the human the hardest cases, with skills that decay through disuse (Bainbridge, 1983).

2.4 The analytical framework

Figure 1 combines the two lenses. Its rows are the four people-side dimensions D1 professional identity and trust, D2 skills and learning, D3 work design and autonomy and D4 leadership and participation, whose operational definitions in Table 2 double as extraction categories. One axis records where resistance sits (Maurer's levels), the other what the organization changed (Shea and Solomon levers). A lever is coded only when an organization actually changed it, so recommending training does not count as the people lever. A context layer wraps both axes: engineering lifecycle stage, AI type and autonomy (assistive, advisory, substitutive or agentic), organization type and regulatory regime. Both lenses are practitioner heuristics without empirical validation. We therefore use the levels and levers as sensitizing coding categories, read the framework as an interpretive map, and treat Maurer's escalation as a hypothesis (P5) that allows movement in either direction.

The paper keeps four outcome terms apart, and none of them is intention to use, the dependent variable of most acceptance studies. *Adoption* is the decision, by an organization or an individual, to introduce or take up an AI system. *Use* is actual work with the system, logged or self-reported. *Sustained use* persists beyond the pilot or the first months of roll-out. *Realized value* is a measured organizational benefit (cost, quality, throughput or return on investment); task-level gains and perceived benefits fall outside it.

3 Method

3.1 Design and reporting standards

The review is reported according to PRISMA 2020 (Page et al., 2021) and, for the search, PRISMA-S (Rethlefsen et al., 2021). The process follows Tranfield et al. (2003) and, for software-engineering studies, Kitchenham and Charters (2007), and the synthesis is concept-centric (Webster and Watson, 2002). A versioned protocol fixed the eligibility criteria, screening design and stopping rules, extraction form and synthesis plan before any screening decision. It was not registered before the search, which ran on 4 September 2026; the screening, extraction and synthesis protocol is deposited with its dated amendment log at `https://osf.io/7ys25/?view_`

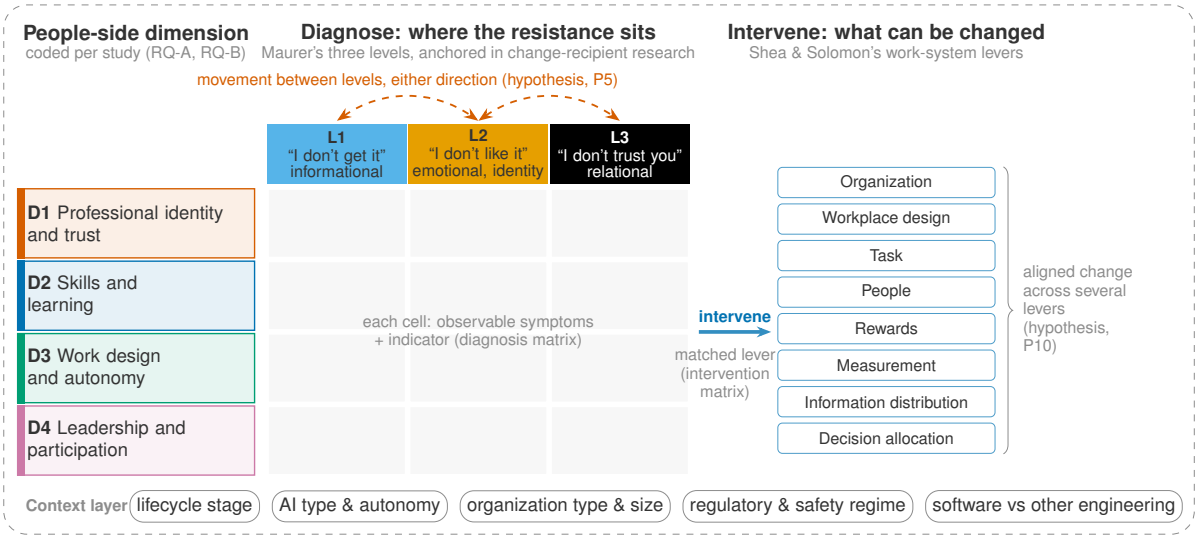


Figure 1: Analytical framework of the review. The four people-side dimensions (D1–D4) form the rows. Maurer's three levels of resistance (L1 informational, L2 emotional and identity, L3 relational) mark where a barrier sits, and Shea and Solomon's eight work-system levers mark what an organization can change. The dashed two-way arrows mark movement between levels in either direction (P5).

Table 2: Operational definitions of the four people-side dimensions used to code the included studies. Verifiability need is a trust construct (D1); outcomes are coded separately (field group H, Appendix D).

Dimension	Operational definition	Example constructs	Typical measures
D1 professional identity and trust	How the AI system bears on the engineer's sense of professional self and expertise, and the willingness to rely on the system and on those introducing it	identity threat or enhancement; perceived devaluation of expertise; trust calibration; algorithm aversion; verifiability need; AI anxiety; technostress; job insecurity	identity-threat scales (Craig et al., 2019); trust and reliance items; technostress creators (Tarafdar et al., 2007); interview codes
D2 skills and learning	The competence engineers hold and the learning conditions they are given to use, check and adapt to AI	AI and data literacy; self-efficacy; training dose and relevance; protected learning time; communities of practice; skill atrophy; upskilling	AI-literacy competencies (Long and Magerko, 2020); self-efficacy scales; training hours; interview codes
D3 work design and autonomy	How tasks, decision rights and control are allocated between engineers, AI and management, and the job content that results	task allocation; level of automation; autonomy; decision rights; procedural fairness; toolchain integration; workload	job-design scales (Hackman and Oldham, 1976); automation-level coding; logged or self-reported use
D4 leadership and participation	How the change is led, framed, communicated and governed, and how far engineers take part in shaping it	vision clarity; transformation versus IT-project framing; leader credibility; participation and voice; governance through engineering review; change fatigue	psychological-safety scale (Edmondson, 1999); leadership-support and participation items; interview codes

Table 3: Eligibility criteria (version 2.0). Each full-text exclusion was assigned the first criterion it failed.

Criterion	Include	Exclude
Population	Engineers, designers, technical specialists or engineering managers; or an engineering, product-development or manufacturing organization or production process as the unit of analysis	Students or any education setting; consumers, patients or the public; online convenience samples with no identified workplace; a non-technical function (e.g., HR, accounting) studied in isolation
Technology	AI introduced into or evaluated in work (machine learning, generative and agentic AI, decision support, coding assistants, AI in CAD/CAE/PLM)	Digitalization, automation or Industry 4.0 with no AI system; AI only as the engineered product
People-side factor	At least one factor measured, analysed or explicitly theorized (e.g., trust, identity, skills, autonomy, leadership, resistance)	Factor merely implied by the application; purely technical algorithm or benchmark papers
Study type	Peer-reviewed empirical study of any design; conceptual work with explicit propositions about an in-scope setting; peer-reviewed review	Editorials, commentaries, consultancy reports, dissertations; macro, legal or policy analysis with no technical anchor
Setting and sector	Engineering, product-development, manufacturing and software-development work, including consumer-goods manufacturing and food production	Healthcare, finance, HR, retail, tourism, primary agriculture, public administration
Language and date	English or German; 2015 to the search date	Other languages; earlier years

`only=0e99091e533343fbbea52413a335ed60` (view-only anonymized link for review).

3.2 Eligibility criteria

A study was eligible when AI had actually been introduced into, or evaluated in, engineering work and at least one people-side factor was measured, analysed or explicitly theorized (Table 3). Healthcare and finance are excluded; strong identity and trust studies from them appear only as anchors in Section 2 or as comparators marked outside the corpus. Because the string accepted *Industry 4.0* as a population term, many records mentioned AI without studying an AI system. The exclusion code for them (`no_ai`), added before any full-text decision, was the most frequent full-text reason in both arms (97 database and 264 snowball reports).

3.3 Information sources and search

Scopus and the Web of Science Core Collection were searched on 4 September 2026 via Universidad Politécnica de Madrid institutional access. The string intersects three blocks (AI technology, engineering population, people-side constructs), limited to articles, conference papers and reviews in English or German from 2015 to the search date (Appendix A). No information specialist peer-reviewed the strings against the PRESS checklist, and we did not search the complementary sources named in the protocol (IEEE Xplore, ACM Digital Library, AIS eLibrary, Design Society). Instead, backward and forward snowballing (Wohlin, 2014) on the 157 database includes identified 5,875 further candidates. Wohlin iterates until no new studies appear, and our single round is a departure from the method.

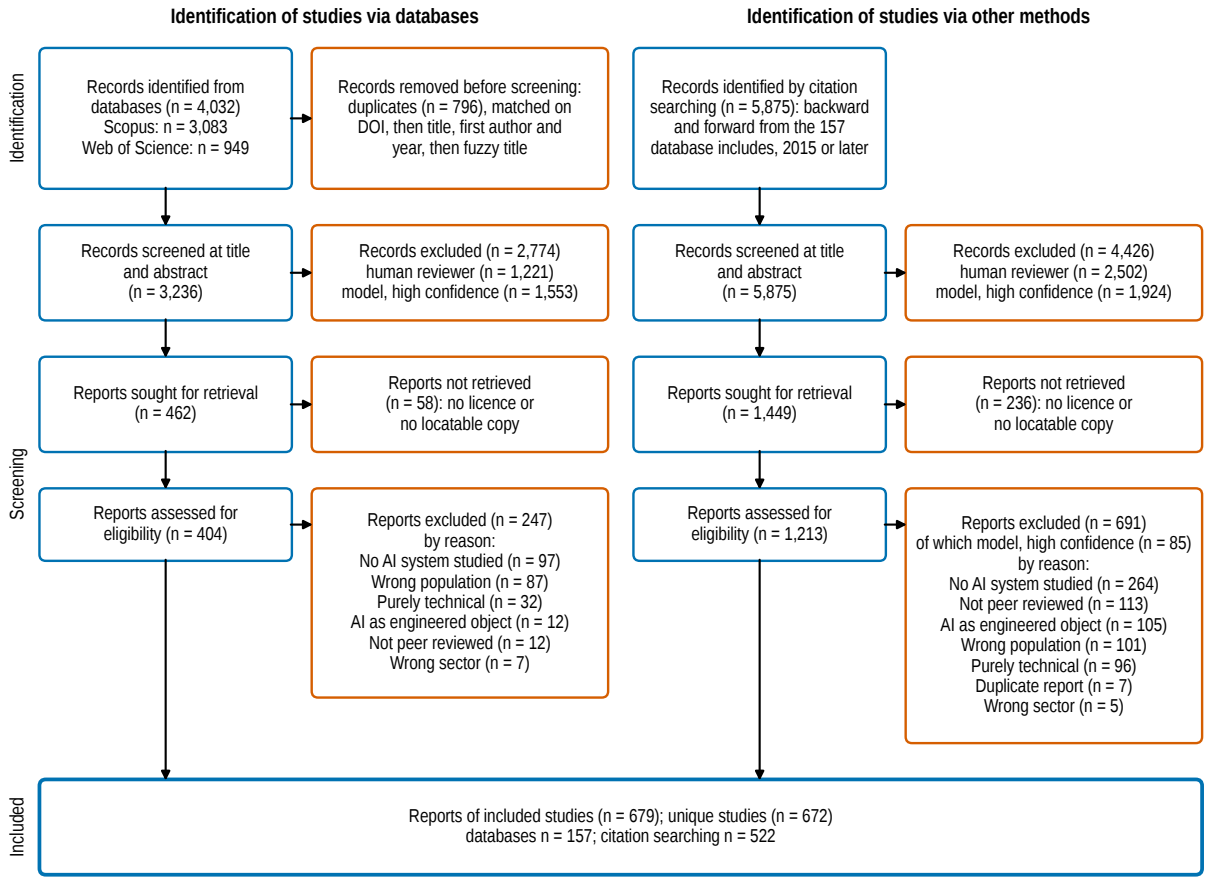


Figure 2: Snowballing on the 157 database includes supplied 522 of the 679 included reports (77 %). Flow of records, database arm left and snowball arm right. Model exclusions are those made at high confidence that stood after the blind audits. Counts are reports: the 679 included reports describe 672 unique studies.

3.4 Study selection

After deduplication (796 duplicates removed), 3,236 database records were screened (Fig. 2). With no second human screener available, the reviewer screened alongside a large language model (Anthropic's `claude-sonnet-5`) that applied the same written criteria, gave a verdict and a confidence per criterion, and was told to prefer inclusion when in doubt. Published validations report high model sensitivity against human reference decisions (Tran et al., 2024; Zhan et al., 2025), in one case comparable to individual human screeners (Matsui et al., 2024), but performance depends on the prompt and on the balance of included to excluded records (Khraisha et al., 2024); we therefore measured sensitivity on this corpus against the reviewer's blind labels. Neither screener saw the other's decisions (Appendix B).

Table 4 and Fig. 3 give every automated stage in order. Round 1 (ungated) and round 2 (confidence-gated, so that a model exclusion stands only at high confidence) failed the 0.95 floor, round 2 in a test confounded by a prompt wrapper that told the model the rule. All ten round-2 misses failed on the population criterion, which the reviewer had also applied differently across the two sessions (Appendix B). Model misses cluster by criterion; human–model agreement is therefore not inter-rater reliability. The reviewer's rulings on five scope questions drawn from these disagreements produced criteria v2.0, which is not independent of the calibration data. In round 3 the reviewer reproduced all 40 hidden repeats, and the gated rule reached 0.916 [0.836, 0.959] against a floor lowered beforehand to 0.90 (Appendix B). After scoring we relaxed the precision condition from 0.85 to 0.80, without which the rule would have failed. No untouched validation set followed, and we treat round 3 as development evidence.

Table 4: Automated screening stages in the order they were run. Round 3 is treated as development evidence, not confirmatory validation: its precision leg was relaxed after scoring, and that change decided the outcome. *Round 2 was confounded by a change to the model's instructions.

Stage	Sample	Rule	Sensitivity [95 % CI]	Floor	Floor set	Outcome
Title/abstract, round 1	blind 200, criteria v1.0	ungated	0.923 [0.797, 0.973]	0.95	before scoring	failed; design stopped
Title/abstract, round 2	fresh blind 200	gated*	0.796 [0.664, 0.885]	0.95	before scoring	failed; confounded
Title/abstract, round 3	400 + 40 repeats, criteria v2.0	gated	0.916 [0.836, 0.959]	0.90; precision leg 0.80	floor before, precision leg after scoring	deployed
Title/abstract audit, databases	250 model exclusions, blind	deployed gate	1.000 [0.936, 1.000]	0.90	before audit	passed
Title/abstract audit, snowball	100 model exclusions, blind	deployed gate	1.000 [0.957, 1.000]	0.90	before audit	passed
Full-text validation	404 database reports, blind	gated	0.994 [0.965, 0.999]	0.90	before scoring	deployed on snowball arm
Full-text audit, snowball	100 model exclusions, blind	deployed gate	0.979 [0.966, 0.987]	0.90	before audit	passed; 13 of 100 kept

In deployment the reviewer screened every record the model had not excluded at high confidence, and blind audits, mixed unmarked into the reviewer's sheets, estimated what the unread exclusions held. At title and abstract the reviewer kept 0 of 250 audited database exclusions and 0 of 100 snowball ones. At full text the reviewer read every retrieved database report; a model screener validated blind on those decisions would have removed only 38 of 404 from human reading. In the snowball arm the reviewer kept 13 of 100 audited high-confidence exclusions, a miss rate far above the validation's; the expected loss is in Table 7. Reports that could not be obtained (58 database, 236 snowball) are counted as not retrieved (Appendix B).

3.5 Data extraction

A coding book (Appendix D) defines each field and its controlled values. The same model drafted each extraction in an isolated session from a pinned rulebook, with a page number and a verbatim quotation behind every substantive value. A stratified 10 % sample (68 studies) was coded blind from the PDF by the reviewer and by a second coder, after a five-study calibration round. Mean Cohen's κ between the two humans ranged from 0.22 to 0.48, fair to moderate. The drafts agreed with the reviewer at least as well as the second coder did in four of six field groups, but not for the Maurer-level and lever fields ($\kappa = 0.16$ against 0.42 between the humans) or for appraisal (0.32 against 0.48). Human codings replace the drafts for 73 studies (the 68 double-coded ones and the five calibration studies, coded once). Every figure and table built from the coding sheet therefore mixes 73 human codings with 606 AI drafts. Claims resting on the framework fields or on the appraisal weights carry that caveat (Appendix B).

3.6 Quality appraisal

Each study was appraised with the Mixed Methods Appraisal Tool (MMAT) (Hong et al., 2018) and a five-item relevance rubric (Appendix E). Its authors discourage a summary score; we derived one nonetheless and combined it with relevance into a high, medium or low weight, under a rule fixed before any weight was used (Table 17). The weight orders the evidence and never excluded a study; Section 4.8 tests each headline result without weighting and in the high- and medium-weight subset. Reviews and conceptual papers were weighted on relevance alone and capped at medium, because no review-specific tool was applied.

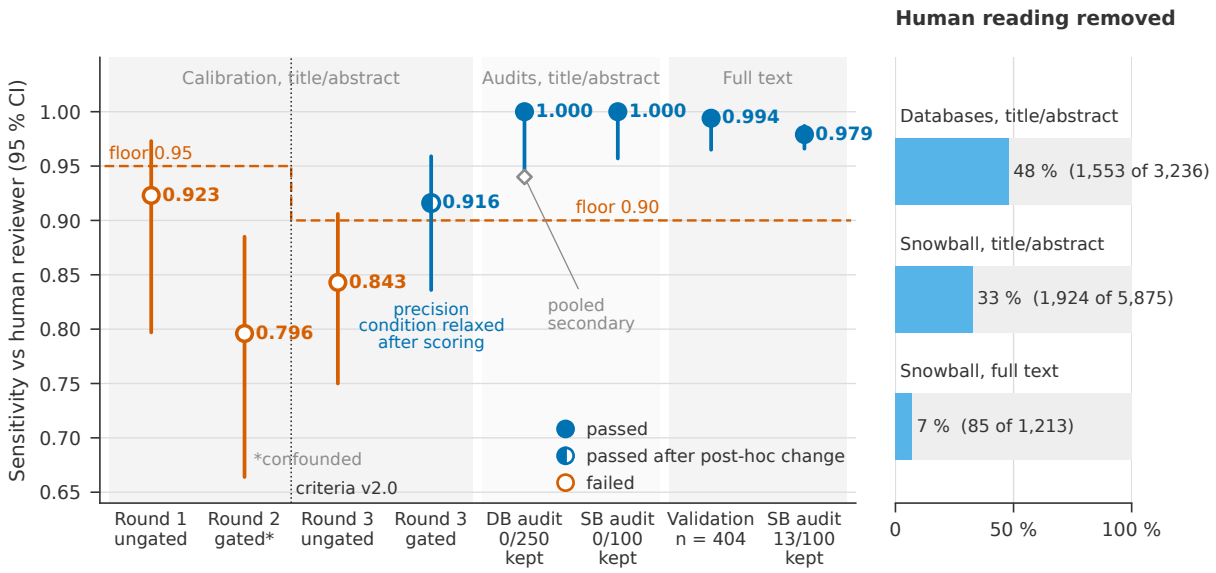


Figure 3: In calibration only the gated rule under the amended criteria cleared the floor, lowered from 0.95 to 0.90, and it passed only after the post-hoc change to the precision condition. Left: sensitivity of each gate against the human reviewer, in time order, with Wilson 95 % intervals; hollow markers failed, the half-filled marker passed only after the post-hoc change, the dashed line is the floor in force, and the diamond is the pooled database-audit estimate. Right: share of records each deployed gate removed from human reading.

3.7 Synthesis

The evidence is synthesized thematically (Thomas and Harden, 2008) within the four dimensions. Framework synthesis (Carroll et al., 2011) then maps each barrier onto a Maurer level and each intervention onto a Shea and Solomon lever, and the propositions are derived from the resulting matrices. An evidence gap map (Saran and White, 2018) crosses lifecycle stage with dimension (Fig. 8). Designs and outcomes are too heterogeneous for meta-analysis.

The unit of screening and counting is the report, which the text calls a study; linked reports of the same study are listed in the coding sheet (Fig. 2). Counts of interventions and value use only the 274 primary empirical field studies of an AI system at pilot stage or beyond, which could have observed an organizational change; reviews and conceptual papers inform interpretation only.

Nine subsets re-test the headline results and each proposition's support (Section 4.8), among them the database arm alone, the human-coded studies and the software stratum, analysed separately so that software does not stand in for engineering at large.

3.8 Deviations from protocol

The amendment log (Supplementary Material) dates every deviation. The four that matter for interpretation are criteria v2.0 with the two threshold changes, access through another institution than planned, one snowballing round in place of the complementary databases, and closure of retrieval before interlibrary loan. Harvest plots and bibliometric maps named in the protocol were not produced. A post-hoc peer-review audit replaced 20 preprints by their version of record and excluded one (Appendix C).

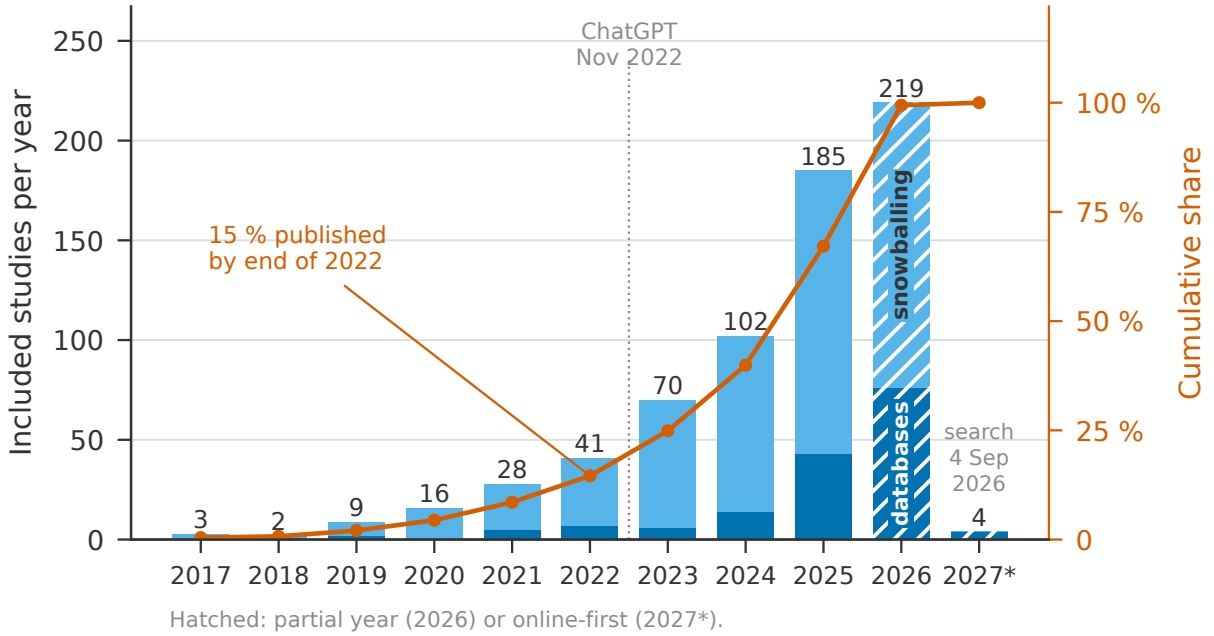


Figure 4: Most of the evidence postdates generative AI, with 85 % of the 679 included studies appearing from 2023 onward. Bars count included studies per publication year, split by arm (database, snowballing), with the cumulative share as a line. The 2026 bar ends at the search date of 4 September 2026; 2027 holds online-first articles.

4 Descriptive results

4.1 Study selection

The database searches identified 4,032 records, of which 3,236 remained after deduplication and 157 were included after full-text screening (Fig. 2). Snowballing added 522, 77 % of the 679 reports in the corpus. Of these, 522 are primary empirical, 110 reviews and 47 conceptual. A study counts once for each value it carries in a field, so the counts are not additive, and coded fields mix human codings with AI drafts (Section 3.5).

4.2 Growth of the field

No included study predates 2017, although the search reached back to 2015, and 404 appeared in 2025 and 2026 alone (Fig. 4). Generative AI for text or code is the subject of 63 % of the studies from 2023 onward, and only 45 studies (7 %) are longitudinal.

4.3 Where the evidence comes from

Software and embedded engineering supplies 37 % of the studies, and most deployed systems studied are pilots (Fig. 5). Context often goes unreported: organization size is missing in 501 studies (74 %), region in 354 and informant role in 297, and only 15 study small or medium-sized firms.

4.4 What kind of AI, and where in the lifecycle

Generative AI for text and code (375 studies) and agentic systems (103) concentrate in software development and testing (Fig. 6). The thinnest stages are change and configuration management (12 studies) and validation and compliance (16), the stages that carry engineering's traceability obligations. No lifecycle stage is reported for 245 studies.

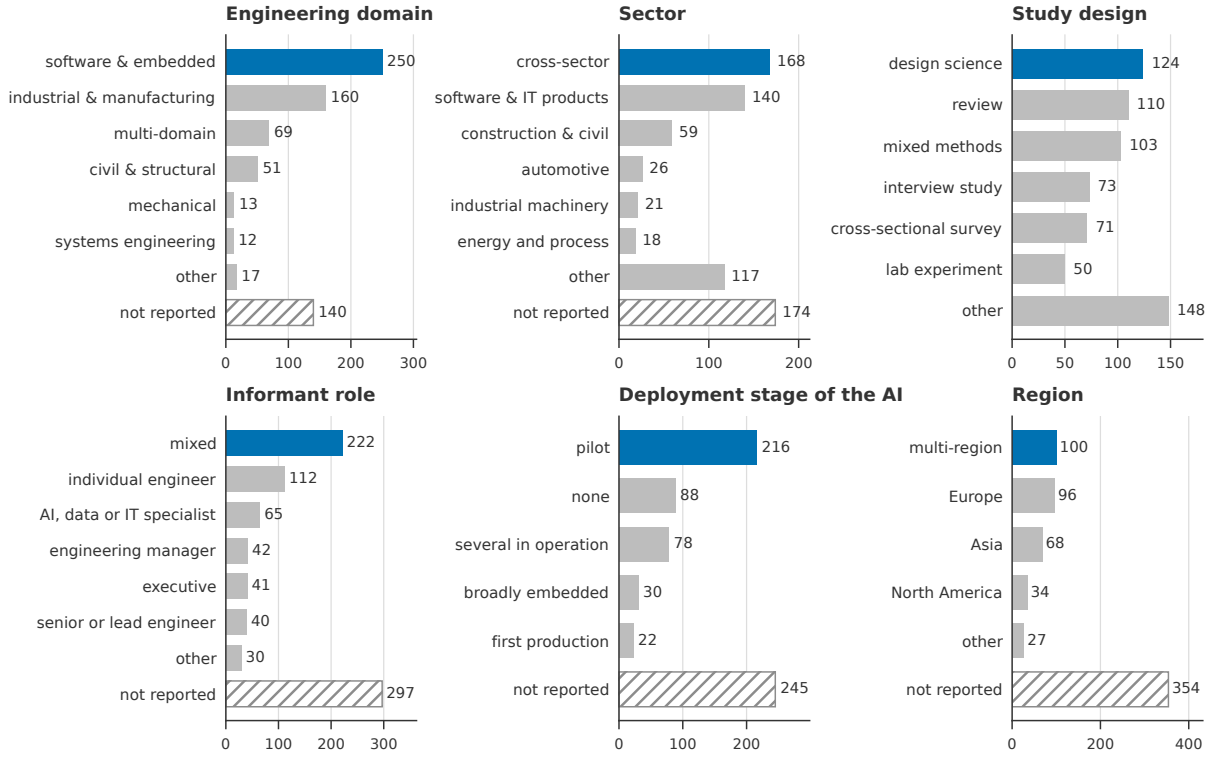


Figure 5: Characteristics of the included studies, one panel per coded field. Each panel counts studies per coded value, and hatched bars mark studies that do not report the field. Sector is the studied organization's industry, domain the engineering work studied.

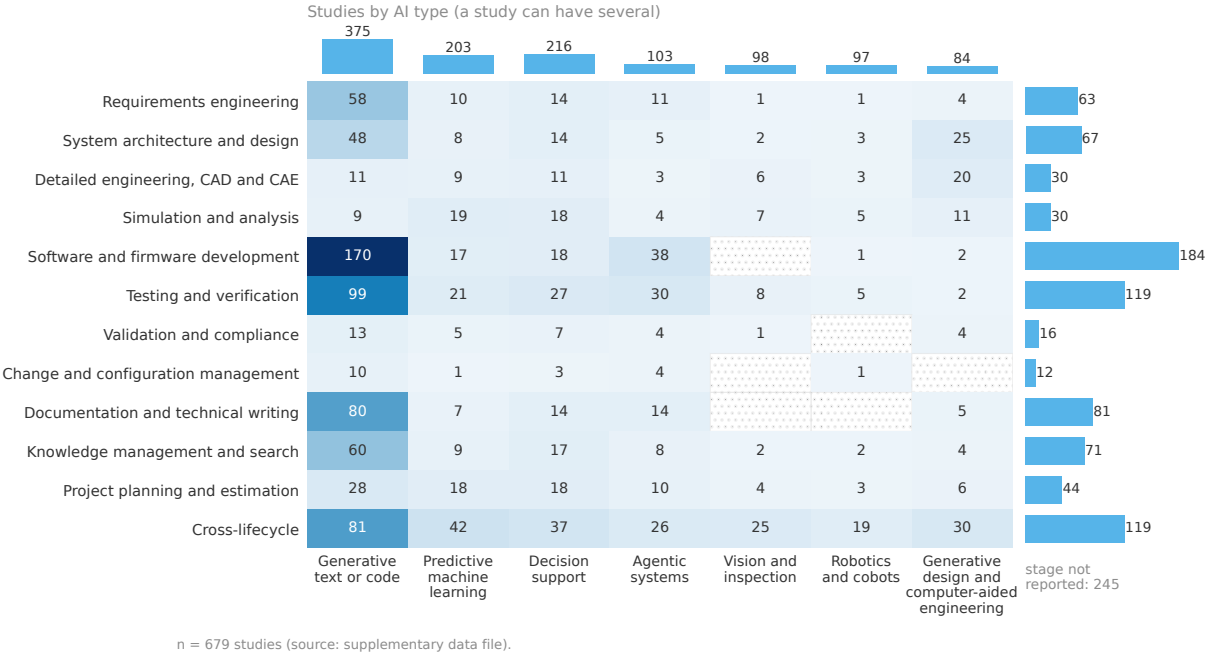


Figure 6: Included studies by AI type and lifecycle stage. Cells count studies coding each AI type at each lifecycle stage, and the marginal bars give row and column totals. Dotted cells are empty.

Table 5: Theories named in the included studies and the role each plays. A study naming several theories counts in each row; the role is coded once per study, and four studies do not report it.

Theory	Studies	Tested	Extended	Sensitizing	Proposed
Sociotechnical systems	52	2	9	34	6
Trust theories	48	5	10	28	4
TAM (incl. TAM2)	33	3	9	18	3
Technology–organization–environment	25	5	9	8	3
UTAUT (incl. UTAUT2)	22	4	11	6	1
Dynamic capabilities	14	3	6	5	0
Diffusion of innovations	13	3	4	6	0
Self-determination theory	11	2	1	7	1
Job characteristics model	11	1	3	7	0
Social or role identity	4	1	0	3	0
Conservation of resources	3	0	1	2	0
Institutional theory	2	1	1	0	0
Change-management models	2	0	1	1	0
Other or study-specific	233	25	64	118	24
No theory stated	355				

4.5 Theories used

No theory is named in 355 studies (52 %). Of the 176 studies that name at least one established theory (Table 5), 24 test one, most using it as a sensitizing lens. Social or role identity theories, central to the professional identity and trust dimension, appear in four studies.

4.6 How adoption and value are measured

No adoption measure is reported in 284 studies, 42 % of the corpus (Fig. 7). Of the 379 studies that report a defined measure, 106 (28 %) log use; the others rely on self-reported use, deployment status or intention alone. We set aside the 17 studies coded under the adoption category *value realized*, because the coding book never defined it and it admits perceived value. The outcome behind the adoption–value gap (Section 1.1) is measured far more rarely. A study measures *realized organizational value* when it measures an outcome of organizational value or return on investment; perceived value and task-level gains in a trial do not count. In the corpus 12 studies do so, and 7 of the 274 that could have observed it (Section 3.7). Some benefit, almost always at task level, is measured in 66 of those 274 (whole corpus in Fig. 7).

4.7 Appraisal distribution

Appraisal weights 134 studies high, 159 medium and 386 low. Reviews, rated on relevance alone and capped at medium, are low in 98 of 110 cases. High weight is twice as common in the snowball arm (117 of 522, 22 %) as in the database arm (17 of 157, 11 %). Software engineering has 92 high-weight studies among its 250 (37 %), the other 429 engineering studies 42 (10 %).

4.8 Sensitivity and completeness

Table 6 re-tests five headline results in nine subsets, without weights except in the high- and medium-weight column. Among studies that could observe a change, 64 % report no lever change, and no subset falls below 59 %. No subset has more than 4 % of them measuring realized organizational value (none of the 29 human-coded ones), and in every subset under half of the studies reporting an adoption measure log use.

The Maurer levels split almost evenly in the full corpus, but the split tilts towards L1 in the database arm and towards L3 among the human-coded studies. The reviewer coded 64 of these

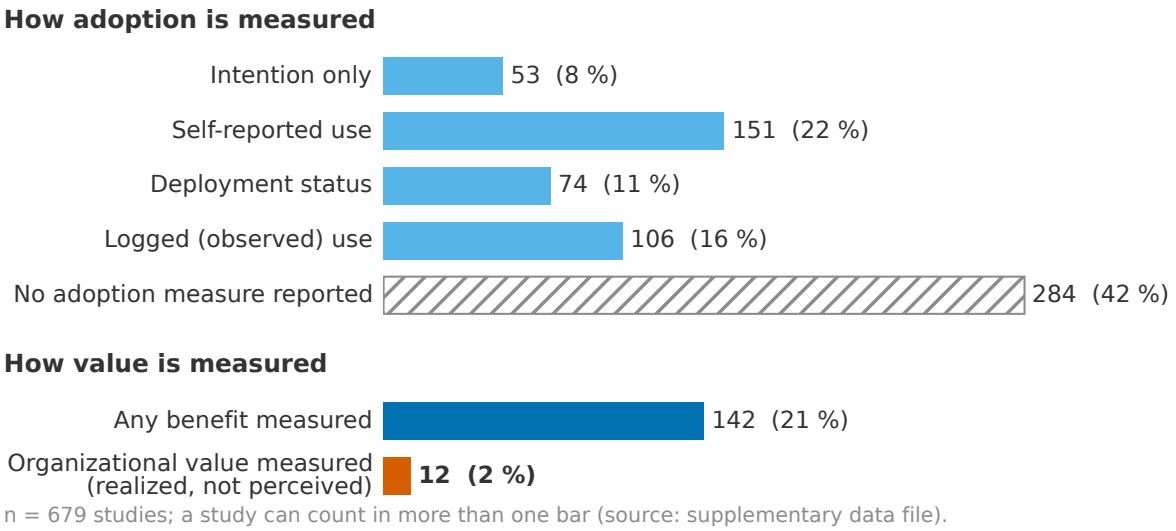


Figure 7: Organizational value is measured in 12 of the 679 studies and in 7 of the 274 that could observe it; the rest stop at use or task-level benefit. The first group counts studies by the adoption measure they report, with a hatched bar for the 284 that report none; a study using several measures counts in each bar, and studies coded only as *value realized*, a category we do not use, are not shown. The second group counts studies that measure any benefit and studies that measure (not perceive) an organizational value or return-on-investment outcome; the two groups come from different coding fields and are not nested.

73 studies (88 %) as not classifiable, where the AI drafts of the same studies had 39 (53 %), with $\kappa = 0.16$ between drafts and reviewer on this field (Section 3.5). Leadership and participation is the least coded dimension in every subset, although the reviewer assigned it more often (41 % of the human-coded studies against 25 % of the corpus). Its thinness may partly be a search artefact, since the executed people-side block lacked leadership, management support, readiness, culture, governance, training, empowerment and autonomy. An indicative re-run in OpenAlex, not a database the review searched, found that adding these terms triples the matching records, and leadership alone adds 41 %.

In the database arm alone, P3, P5, P6 and P10 have no key supporting study (lower block of Table 6). We read this as a sampling result, since only citation searching reached their support and its studies are twice as often high-weight (Section 4.7). By stratum, P6 to P9 have no support outside software and P4 none inside it, and P7 has none among the studies that could observe a change.

About 135 eligible studies (121–151), 17 % of all eligible, are expected to be missing, most among the 236 unretrieved snowball reports (Table 7, which includes the unread model exclusions). If the missing studies resembled the corpus, they would add about three studies each to validation and compliance and to small and medium-sized firms, and two to change and configuration management.

5 Thematic synthesis

5.1 Professional identity, trust and early buy-in

Table 8 profiles the four dimensions and the constructs the synthesis draws on. Software engineering supplies about two in five studies in each of the first three dimensions and more than half of those at high or medium weight, so the patterns below describe developers unless another sample is named.

Developers protect work central to their identity while still using the tool. Among 860 developers, identity-central tasks went with lower openness to AI support, standardized $\beta =$

Table 6: Sensitivity of the headline results to how the corpus is cut. Upper block: each claim re-tested in each subset (✓ holds, × does not), with the value. Lower block: for each proposition, the key supporting studies of Table 11 that remain in the subset; every one of them is a high- or medium-weight study. "Could observe": primary empirical field studies of an AI system at pilot stage or beyond.

Claim	All	Primary	Could observe	High/med.	Database	Snowball	Human-coded	Software	Other eng.
n	679	522	274	293	157	522	73	250	429
Most studies that could observe a lever change report none (>50 %)	✓64 % of 274	✓64 % of 274	✓64 % of 274	✓59 % of 149	✓61 % of 75	✓65 % of 199	✓72 % of 29	✓67 % of 115	✓62 % of 159
Few studies that could observe it measure realized value (<10 %)	✓3 % of 274	✓3 % of 274	✓3 % of 274	✓1 % of 149	✓4 % of 75	✓2 % of 199	✓0 % of 29	✓2 % of 115	✓3 % of 159
Adoption is mostly measured by intention or self-report (logged use <50 %)	✓28 % of 379	✓30 % of 350	✓28 % of 191	✓29 % of 223	✓21 % of 89	✓30 % of 290	✓44 % of 27	✓38 % of 175	✓20 % of 204
Maurer levels are spread evenly (L1/L2/L3 shares within 15 points)	✓36/35/30	✓37/34/29	✓37/35/29	✓34/37/29	× 42/33/25	✓34/35/31	× 30/10/60	✓31/37/32	✓38/33/28
Leadership and participation is the least coded dimension	✓25 %	✓22 %	✓25 %	✓26 %	✓32 %	✓23 %	✓41 %	✓21 %	✓28 %
P1	5	5	4	5	2	3	2	4	1
P2	5	5	3	5	1	4	2	4	1
P3	2	2	1	2	0	2	0	1	1
P4	3	3	3	3	2	1	1	0	3
P5	3	3	2	3	0	3	1	2	1
P6	3	3	2	3	0	3	0	3	0
P7	1	1	0	1	1	0	0	1	0
P8	3	3	2	3	2	1	0	3	0
P9	3	3	1	3	1	2	0	3	0
P10	3	3	3	3	0	3	1	1	2

Table 7: Eligible studies the review is expected to have missed, by source. Expected counts apply the inclusion rate observed where the review did read to the records it did not read; ranges are Wilson 95 % intervals on that rate. The unretrieved snowball reports dominate the shortfall.

Source of possible loss	Records	Rate applied	Expected eligible (range)
Snowball: high-confidence model exclusions not read by a human	85	13 of 100 audited	11 (7–18)
Databases: reports not retrieved	58	157 of 404 assessed	23 (20–25)
Snowballing: reports not retrieved	236	522 of 1213 assessed	102 (95–108)
Total expected missing			135 (121–151)
Included studies	679		
Expected missing as a share of all eligible			17 % (15–18 %)

Table 8: Dimension profile of the 679 included studies. A: studies per dimension (N; a study can code several), from the database arm (DB), at high or medium weight (Wtd), from software engineering (SW; SW wtd among weighted studies), able to observe a lever change (Opp., Section 4.6) and, of those, reporting none. B: studies per construct, and the software count (SW).

	N	DB	Wtd	SW	SW wtd	Opp.	None
A. Dimensions							
D1 Professional identity and trust	482	111	230	187 (39 %)	126 (55 %)	204	126
D2 Skills and learning	442	107	207	165 (37 %)	114 (55 %)	178	102
D3 Work design and autonomy	552	128	260	215 (39 %)	146 (56 %)	235	148
D4 Leadership and participation	171	50	77	52 (30 %)	34 (44 %)	69	21
B. Constructs (dimension)							
Professional identity (D1)	59			25			
Job insecurity or AI anxiety (D1)	105			36			
Verifiability need (D1)	124			71			
Training provision or quality (D2)	117			23			
AI literacy (D2)	107			32			
Informal and social learning (D2)	49			23			
Deskilling and skill atrophy (D2)	86			46			
Task allocation, human and AI (D3)	344			138			
Workflow and toolchain integration (D3)	233			116			
Shadow or unsanctioned use (D3)	21			13			
Sustained use versus abandonment (D3)	55			40			
Participation and co-design (D4)	54			3			
Governance through existing reviews (D4)	63			28			

$-.09$, yet with higher reported use, $\beta = .15$ (Choudhuri et al., 2026). Developers who enjoyed an activity less were more willing to delegate it to AI, most clearly for writing natural-language artefacts ($r = -.32$) (Sergeyuk et al., 2025). A threat account would predict refusal; the pattern sits closer to the augmentation account of Selenko et al. (2022). Workers also erased traces of AI use to signal expertise (Xia et al., 2026), and 14 of 20 architecture, engineering and construction (AEC) design professionals passed AI output to clients unacknowledged (Xu et al., 2026). Only one of the professional-identity studies concerns mechanical, electrical or systems engineers.

Seniority splits reliance, much as it split role responses in the banking case of Strich et al. (2021). In logs of about one million code suggestions at one firm, developers with under six months' tenure accepted slightly more (odds ratio 1.07) and those with three to five years slightly fewer (0.96) (Brown et al., 2024). In a timed programming exam, experts rejected the AI more often, even where it could have corrected their errors (Qian and Wexler, 2024). Outside software, user-experience and industrial designers appraised AI as threat or opportunity according to perceived control, which seniority and policy clarity shaped (Uusitalo et al., 2024), and AEC design professionals mostly reported that senior colleagues resist the tools, though a minority saw seniors as the more receptive (Xu et al., 2026).

We distinguish *verification*, the checking of an AI output against a reference independent of the system (a test suite, a solver, a standard or a physical measurement), from *explanation*, the system's own account of why it produced the output. In design-scenario interviews, eleven software engineers said they would accept machine-translated code after their usual review and testing, without needing to understand the model (Weisz et al., 2021). Yet developers also read an agent's plan as a proxy for its work and passing tests as proof of correctness (Dhanorkar et al., 2026), and in the exam above dependence rose while stated trust fell (Qian and Wexler, 2024). Parasuraman and Manzey (2010) offer an attentional account that fits such divergences: when other tasks compete for attention, people monitor an automated aid less. In visual inspection of electronic products, heatmap explanations raised factory workers' balanced accuracy by 7.7 percentage points over an unexplained quality score, although both groups also had a reference image of a faultless product (Senoner et al., 2024). Whether verification can replace explanation for testable output has not been tested, and nearly all testable cases are software.

Reflexive algorithm aversion finds little support. Programmers weighted advice from generative AI above identical advice attributed to a human expert (Amrollahi et al., 2026), as lay participants did with algorithmic advice (Logg et al., 2019). Studies of job insecurity and AI anxiety mostly report expectations, most often about juniors (Määttä et al., 2026; Li et al., 2024).

Only 29 of the 482 professional identity and trust studies are longitudinal, most measure by self-report, and the construct counts in Table 8 indicate emphasis ($\kappa = 0.35$ between AI and human coders). Maurer-level shares in this and the next two dimensions are unstable (Section 4.8).

5.2 Skills, learning and enablement

None of three German organizations with agile software teams offered practical training in generative AI, and one organization's mandated sessions only explained its guideline (Neumann et al., 2026), while software engineers building copilots learned through online communities (Parnin et al., 2025). Studies of training provision come mostly from outside software (94 of 117, Table 8). People was the lever most often changed, in 60 studies, although most skills studies that could have observed a lever change report none.

Support for learning recurs as an enabler of use. In 27 matched pairs of developers from the same teams, frequent and infrequent users differed mainly in how they perceived the tool and handled its failures; champions, social learning and protected time were reported as supports, and delivery pressure without learning support as pushing developers back to old methods (Miller et al., 2026). Among 1,130 German manufacturers, those with high organizational AI readiness,

a bundle of training, skills and culture, had twice the odds of adopting AI in production (odds ratio 2.11), whereas technological readiness alone was marginal (Heimberger et al., 2026a). No high- or medium-weight study evaluated training against measured use, so there is no convincing evidence yet that training dose alone improves sustained use.

As the corpus uses the term, AI literacy is mostly the judgement of whether an AI output can go forward into review and testing, narrower and more operational than the critical evaluation of AI in generic definitions (Long and Magerko, 2020). AI can also erode the social learning through which such judgement spreads. Hiding AI use reduced colleagues' opportunities to learn from one another's use (Xia et al., 2026), and in a small cybersecurity software firm, interviewees perceived generative AI as displacing face-to-face exchange, especially among younger staff (Liccardo and Cerchione, 2025).

Whether AI deskills engineers is unresolved, because the case for loss rests on expectation. Novices gained most from AI in the corpus, on open-ended questions (Qian and Wexler, 2024), and outside it the gains were largest for less experienced workers in a field study (Brynjolfsson et al., 2025) and for the weakest writers among professionals in an experiment (Noy and Zhang, 2023). Practitioners, however, expect lost understanding among developers (Felder et al., 2026) and fewer on-the-job learning opportunities for junior industrial designers (Süner-Pla-Cerdà et al., 2026). About half of the deskilling studies rely on self-report alone, and only three are longitudinal and three rest on observed data alone. In a three-week field study, developers' belief in their unique skills rose equally with and without Copilot (Butler et al., 2025).

5.3 Work design, autonomy and ways of working

In the studies that report it, checking AI output took much of the time that generating it saved: user-experience and software designers said so for early-stage design work (Mulgund et al., 2026), and among 124 practitioners in requirements engineering, validation burden correlated at $\rho = -0.87$ with reported adoption success (Shah et al., 2026). In an interview study of 27 software professionals, about half sent AI-generated code through peer review before merging (Klemmer et al., 2024), and R&D formulation specialists at a confectionery maker kept physical and sensory testing after adopting an explainable model, though on far fewer prototypes (Amarilli et al., 2025). Self-report cannot show how well this verification route (Section 5.1) works. Scientists were confident they could catch errors in generated code, yet their logs recorded three cases of errors accepted unnoticed (O'Brien, 2025).

Work design and autonomy is the most often coded dimension and task allocation between human and AI its commonest construct (Table 8); 243 of the 344 task-allocation studies report mixed effects. Most studies that could have observed a lever change report none. Coders recorded only what papers report; redesign may be more common than its documentation. No study tested toolchain integration as a route to organizational use. P7 rests on a survey of 184 software engineers in which perceived benefits related to intention to use only through compatibility with existing workflows, with no significant direct path (Russo, 2024).

In Feng et al. (2026), organizational policy largely fixed developers' agency before individual preference mattered. In a three-wave survey of 368 engineering-consulting professionals, perceived autonomy when working with an AI system went with higher trust, directly and through warmer perceptions of the system (Cui et al., 2026). Where compliance-led whitelists and a mandatory two-person review of AI output were judged impractical, developers moved to unsanctioned tools (Neumann et al., 2026). Some practitioners freely shared production code with ChatGPT, although the authors could not tell whether company policy allowed it (Khojah et al., 2024), and the few organizational AI policies software professionals reported were mainly motivated by data-leakage and privacy concerns (Klemmer et al., 2024). Unsanctioned use is coded in only 21 studies (Table 8), and the legal exposure it can create, over intellectual property, contracts and export control, went unexamined.

Adoption itself is weakly measured. Among the 552 work-design studies, it is self-reported

in 130 and logged in 94. Logs can mislead as well, since about half of the 500 repositories with the largest fall to zero in AI-assisted commits had afterwards edited an agent guidance or configuration file (Robbes et al., 2026). Realized organizational value, as defined in Section 4.6, is measured in 6 of the 235 work-design studies that could have observed it, and the rewards lever was changed in three studies, all at low weight.

5.4 Leadership, vision, communication and participation

Leadership recurs in every phase of a process model grounded in 17 Canadian organizations across industries (Taherizadeh and Beaudry, 2023), yet sponsorship was often weak. In only 31.8 % of 44 business units did respondents agree that the executive sponsor of AI in product development had the capability and credibility to lead it (Cooper and Brem, 2024).

Leadership and participation is the thinnest dimension (Table 8), possibly in part because of the search string (Section 4.8). None of its studies is an experiment and only 14 follow a case over months, so what leaders did is known mostly from respondents' accounts. Unlike in the other dimensions, most studies here that could observe a lever change report one, probably because describing what leaders did usually means describing an organizational change.

Participation went with acceptance, and its evidence comes almost entirely from outside software. At an industrial machinery maker, involving all departments from the start was credited with raising acceptance, and weekly joint meetings aimed at a common language (Krzywdzinski and Butollo, 2022). At a confectionery maker, feedback loops between R&D formulation experts and an explainable model accompanied the easing of early fears of displacement (Amarilli et al., 2025). Stouten et al. (2018) count participation among the best-supported elements of practitioner change models, because taking part gives those affected both information about a change and some influence over it (see also Lines, 2004; Oreg et al., 2011). Participation also has costs: a one-day participatory design workshop at a manufacturer overloaded its practitioners and held only three of four planned sessions, which the authors attribute to guidelines that set no order for tackling concerns (Yamamoto et al., 2024).

After three weeks of randomly assigned access to a coding assistant, developers' agreement with "I like using AI coding tools" rose from 2.72 to 3.61 in a pre-post comparison, while underpowered telemetry showed no significant change in activity (Butler et al., 2025). In one interview study, six of twenty developers described a subtle loss of agency in their employers' top-down push to use AI (Feng et al., 2026). Perhaps exposure changes what people believe about a tool while pressure to use it erodes their sense of control; engineering studies have not tested this.

Review arrangements drew mixed responses in software. One practitioner reported that AI-generated code went through the same review process as any other code (Klemmer et al., 2024), and designers and developers accepted AI framed as a tool under human control within existing governance, sometimes by adding review steps (Vishwarupe et al., 2026). A mandated two-person approval of every AI output was judged unworkable by the teams subject to it (Neumann et al., 2026), and adding legal and security reviewers to design reviews raised coordination costs (Mulgund et al., 2026). No study compares review arrangements directly, and P8 remains untested.

5.5 Cross-cutting themes

Software engineering differs from the rest of the corpus in the stage of adoption it observes. Of its 250 studies, 79 (32 %) examine AI beyond the pilot and 51 (20 %) a pilot, whereas in all other engineering the proportions reverse, with 51 of 429 studies (12 %) beyond the pilot and 165 (38 %) in pilots. Most evidence on sustained use comes from software developers.

Agentic AI is coded in 103 studies, 79 of them published from 2025 and 52 in software. In one firm's deployment of an agent, practitioners approved 82 % of agent-generated plans, yet about 8 % of the issues handled ended in a merged pull request (Takerngsaksiri et al., 2025).

Experience or seniority recurs as a moderator, often proposed without a test: with reliance and control in professional identity and trust and work design and autonomy (Sections 5.1 and 5.3), and with who gains in skills and learning (Section 5.2).

Of the 679 studies, 53 concern regulated settings, 68 partly regulated and 13 unregulated ones; the other 545 do not report regulation. Among ten requirements engineers in regulated industries, nine reported AI-generated design artefacts that lacked transparency, and three of the ten had been recruited as former users who had abandoned the tools (Ullah Shah et al., 2025).

Of the 155 pilot studies that could have observed a lever change, 38 (25 %) report one and 26 (17 %) code leadership, compared with 60 (50 %) and 43 (36 %) of the 119 such studies of systems beyond the pilot. We cannot tell from these counts whether lever change carries AI beyond the pilot or scaling brings the change, because studies of scaled systems have more organizational change to describe, and a change a pilot study leaves unreported may still have occurred. One software company took more than six months to move from pilot to licence distribution, and none of the seven companies studied could quantify a benefit (Kemell et al., 2025).

6 An integrative framework and testable propositions

Figure 1 organizes the results of Section 5. Its Maurer-level and lever codes are the least reliable extraction fields (Section 3.5), and their counts indicate emphasis only.

6.1 Mapping the evidence onto the engineering lifecycle

The evidence gap map (Fig. 8) has no empty cell, but 216 of the 434 studies with a stage (50 %) sit in its software and firmware development or testing and verification rows, which hold the best-weighted cells in every dimension. Change and configuration management has 12 studies, 10 in software and one systems-engineering pilot with four users (Gräßler et al., 2022). Validation and compliance has 16, six of high or medium weight, and detailed engineering with CAD and CAE has nine in leadership and participation.

6.2 A diagnosis matrix

Table 9 crosses the four dimensions with Maurer's three levels, naming in each cell an observable symptom and an indicator that could detect it. Level shares do not survive re-coding (Section 4.8), and we offer the matrix as a checklist for managers.

Two reading rules separate cells that the coding book left overlapping. Trust in the AI system is a professional identity and trust construct. A barrier sits at L3 only when the distrust is directed at the people who introduce the system, whether management, a change agent or a vendor. A construct spanning dimensions is placed by the cause the study reports. Unsanctioned use around controls judged impractical thus belongs to work design and autonomy, and concealment for fear of peers' judgement to professional identity and trust. Because the extraction rulebook did not state these rules, some drafted L3 codes may record distrust of the tool. Validated instruments are rare, and no included study used the identity-threat or psychological-safety scales that the table proposes.

6.3 An intervention matrix

Table 10 lists the practices observed for each lever. Of the 139 studies that describe a lever change (Fig. 9), 41 are reviews, conceptual, laboratory or pre-pilot studies outside the 274 that could observe one. The matrix counts practices wherever described; the opportunity statistic (Section 4.8) concerns field studies. Task redesign, training and information distribution carry the most evidence, and no high- or medium-weight study changed rewards. Only two high-weight studies compared a lever change against a control group: randomized access to a coding

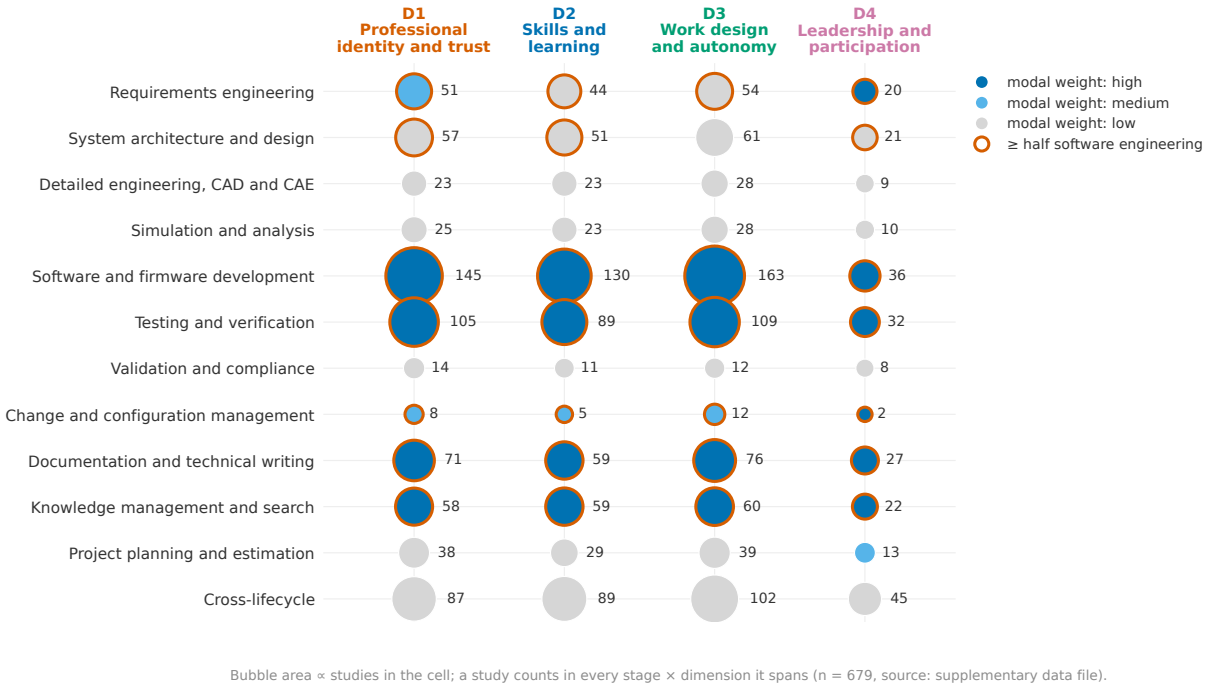


Figure 8: Evidence gap map: lifecycle stages (rows) by people-side dimension (columns). Bubble area gives the number of studies and fill the modal appraisal weight; an outline marks cells in which at least half of the studies are from software engineering. Studies count in every stage and dimension they span; the 245 with no reported stage are omitted.

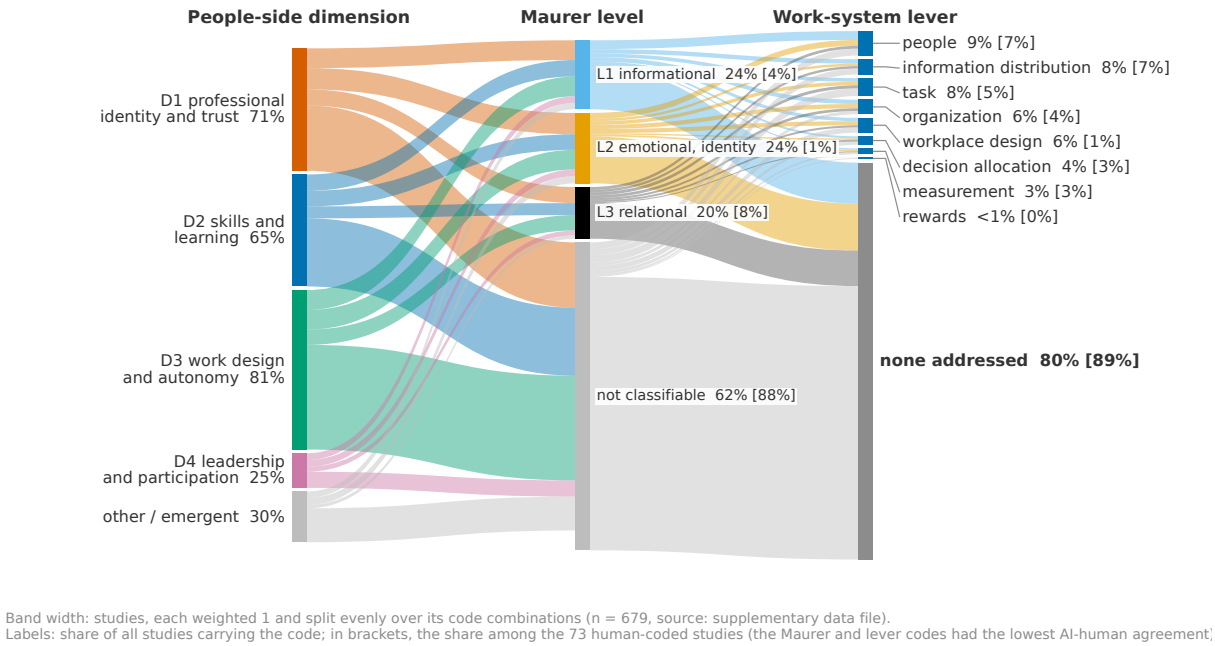


Figure 9: Most studies describe a barrier without observing an intervention. Bands run from the people-side dimension through Maurer's level of resistance to the Shea and Solomon lever changed. Node labels give the share of all 679 studies carrying the code and, in brackets, the share among the 73 human-coded studies. The sink "none addressed" includes studies that could not observe a lever change.

Table 9: Diagnosis matrix: how each level of resistance shows in each dimension, with indicators that could detect it.

	L1 "I don't get it"	L2 "I don't like it"	L3 "I don't trust you"
D1 professional identity and trust	Engineers misjudge what the tool can do; trust falls after first use. V trust-in-automation scale before and after use (Qian and Wexler, 2024); C suggestion acceptance in logs (Brown et al., 2024)	Identity-defining tasks kept from AI; AI use concealed from peers and clients. C task-level openness and use items (Choudhuri et al., 2026); E identity-threat scale (Craig et al., 2019)	Distrust of AI vendors and of management's handling of adoption. C open-ended survey codes (Määttä et al., 2026)
D2 skills and learning	Unclear use cases; users unsure how to prompt. C non-use reasons and comments on prompting (Weisz et al., 2025)	Fear of deskilling and of junior roles disappearing. C technostress creators (Tarafdar et al., 2007) used as a coding frame (Jeyam et al., 2026)	Productivity expected without time or support to learn. C interview codes (Miller et al., 2026)
D3 work design and autonomy	Verification absorbs the time saved; reasons for AI output cannot be traced. C validation-burden items (Shah et al., 2026); C interaction logs compared with self-reported verification (O'Brien, 2025)	Loss of agency under mandated tools; unwillingness to delegate enjoyed tasks. C delegation-willingness ratings (Sergeyuk et al., 2025)	Unsanctioned tools used around controls judged impractical. C interview codes (Neumann et al., 2026)
D4 leadership and participation	No stated reason or guidance; use left as a personal decision. C interview codes (Takaffoli et al., 2024)	Enforced adoption experienced as imposed. C open-ended survey codes (Määttä et al., 2026)	Sponsor judged to lack capability and credibility; management withholds decisions from the system. C readiness survey of business units (Cooper and Brem, 2024); E psychological-safety scale (Edmondson, 1999)

V validated instrument used in an included study; C indicator built within the corpus; E established instrument proposed here, used by no included study.

Table 10: Practices observed for each Shea and Solomon lever, with the strength of evidence.

Lever	Practices observed	Studies	Evidence
Organization	Internal expert group or AI unit; cross-functional AI group; external tools restricted to an approved list (Neumann et al., 2026; Peretz-Andersson et al., 2024)	9	●●
Workplace design	Internally hosted assistants and company licences; tools placed in the IDE or workspace (Wagman et al., 2025; Butler et al., 2025)	9	●●
Task	AI inserted into review, inspection or planning steps, with the human keeping the final decision (Cihan et al., 2025; Süße et al., 2023)	22	●●●
People	Training campaigns and clinics; champions and protected learning time; data scientists recruited (Miller et al., 2026; Brachman et al., 2025)	14	●●●
Rewards	None in a high- or medium-weight study	0	—
Measurement	Success metrics and performance dashboards; responsible-AI assessment before release (Peretz-Andersson et al., 2024; Parnin et al., 2025)	7	●●
Information distribution	Usage policies and guidance; prompts and outputs made visible to peers (Khojah et al., 2026; Uusitalo et al., 2024)	18	●●●
Decision allocation	Human approval over agent output; two-person review; design reviews widened to legal and security (Takerngsaksiri et al., 2025; Neumann et al., 2026; Mulgund et al., 2026)	5	●●

Studies: high- or medium-weight studies observing the change over time, in logs or experimentally. ●●● ten or more, ●● three to nine, ● one or two, — none. Ratings are the same in every dimension except People, ●● in leadership and participation (7 studies).

assistant (Butler et al., 2025) and a retrieval chatbot for commissioning technicians (Löwhagen et al., 2025).

6.4 Propositions

The ten propositions in Table 11 are hypotheses generated by the synthesis, and none has been tested in the form stated. Their verbs ("co-occurs with", "accompanies") reflect studies that describe patterns without isolating a cause. For each proposition the table counts the key studies still present in three subsets of Table 6, and P4 and P6 to P9 rest on one stratum.

P5 no longer states Maurer's upward sequence. In a laboratory study, mostly of students, repeated failures of an assistant's output escalated into frustration and distrust of the tool before participants abandoned it within one session (Tie et al., 2026). All three levels appear side by side among the developers of one company (Miller et al., 2026), and on a shop floor an initial rejection receded once workers became familiar with the system (Süße et al., 2023). For P6, randomized access raised developers' liking of AI tools (Butler et al., 2025), forced use raised later use only among sales workers (Cao et al., 2026), and six of twenty developers described the push to use AI as a subtle loss of agency (Feng et al., 2026). P10 drops Shea and Solomon's threshold of four levers, a practitioner rule of thumb (Shea and Solomon, 2020). Of the 18 studies that changed four or more levers, 13 examine systems beyond the pilot, but deployment maturity confounds the count (Section 5.5).

Table 11: Propositions, stated as hypotheses generated by the synthesis. Survey items that address them are listed in Appendix G.

	Proposition	Evidence	Key studies	Stratum	DB/ HC/ OE
P1	Seniority and expertise co-occur with distinct reliance on AI output: experienced engineers more often reject or constrain it, and novices gain more while alternating between over-reliance and avoidance	consistent; mostly software; tenure effects small	Brown et al., 2024; Qian and Wexler, 2024; Choudhuri et al., 2026; Feng et al., 2026; Löwhagen et al., 2025	SW, OE	2/2/1
P2	Trust in AI output is calibrated by verification against an independent reference and by explanation, which helped even where a reference existed; under workload, verification shrinks to proxy checks	consistent; the two routes never compared	Weisz et al., 2021; Senoner et al., 2024; Dhanorkar et al., 2026; Qian and Wexler, 2024; Shah et al., 2026	SW, OE	1/2/1
P3	Sustained use accompanies learning conditions (protected time, champions, social learning), and firm-level adoption accompanies a readiness that includes training; no study links training dose alone to sustained use	tentative; observational	Miller et al., 2026; Heimberger et al., 2026a	SW, OE	0/0/1
P4	Where work is bound to a physical process, involving the affected functions in developing the AI system and allocating tasks is associated with acceptance, at a cost in participants' effort	consistent; other engineering only; cost from one workshop	Krzywdzinski and Butollo, 2022; Amarilli et al., 2025; Yamamoto et al., 2024 for the cost	OE	2/1/3
P5	Resistance can sit at several levels at once and move between them in either direction, and an unanswered informational barrier may harden into an emotional or relational one	tentative; no longitudinal study	Tie et al., 2026; Miller et al., 2026; Süße et al., 2023; Lapointe and Rivard, 2005 outside the corpus	SW, OE	0/1/1
P6	In software teams, mandated exposure to an AI tool accompanies more favourable attitudes; the link to use comes from outside engineering, and whether a prescribed method lowers perceived agency is untested	tentative	Butler et al., 2025 on attitudes; Feng et al., 2026; Määttä et al., 2026 on pressure to use AI; Cao et al., 2026 outside engineering	SW	0/0/0
P7	Integration into the existing toolchain is associated with intention to use; that it carries individual acceptance through to organizational use is hypothesized	tentative; one survey; downstream path unmeasured	Russo, 2024	SW	1/0/0
P8	In software organizations, routing AI oversight through existing engineering reviews accompanies acceptance and continued use, though widened reviews raised coordination costs; no study compares a parallel AI governance track	consistent for acceptance; software only	Klemmer et al., 2024; Vishwarupe et al., 2026 at the adoption decision; Mulgund et al., 2026 for the cost	SW	2/0/0
P9	Controls that engineers judge impractical go together with unsanctioned use, and unclear policy with concealment of AI use	tentative; one software case for controls	Neumann et al., 2026; Khojah et al., 2026 indirect; Xia et al., 2026 knowledge workers	SW	1/0/0

	Proposition	Evidence	Key studies	Stratum	DB/ HC/ OE
P10	Aligned change across several levers co-occurs with use beyond the pilot, while training alone is hypothesized to leave use at the pilot; deployment maturity confounds the association	tentative; practitioner heuristic	Peretz-Andersson et al., 2024; Süße et al., 2023; Miller et al., 2026; Shea and Solomon, 2020	SW, OE	0/1/2

Consistent: several key studies report the pattern, none tests it; *tentative*: few studies, or partly outside engineering. Stratum: SW software, OE other domains. DB/HC/OE: key studies in the database arm, human-coded, and outside software (Table 6).

6.5 Where the evidence is missing

Table 12 pairs the gaps that block tests of the propositions with study designs. The two controlled comparisons (Section 6.3) leave P3 and P10 untested, no study crossed verification with explanation (P2), and deskilling and escalation lack longitudinal data (P1, P5; Section 5.2). Two draft propositions were withdrawn for lack of comparable studies: slower adoption at stages with deterministic verification norms, such as change management and certification sign-off, and leader credibility as a moderator of communication, never measured independently of respondents (Section 4.8). SMEs and realized organizational value are also thinly sampled (Sections 4.6 and 4.8).

7 Discussion

7.1 Theoretical implications

Adoption theory assumes a fixed artefact whose usefulness a user can judge before use, whereas AI systems learn, act with growing autonomy and are opaque to the people who check them (Berente et al., 2021). Engineers met opacity by verifying output (Section 5.1), and in one firm's agent, built around human approval of each plan and code change, use became supervision (Takerngsaksiri et al., 2025). Intention-centric models also under-explain sustained use, which the corpus seldom measures (Section 4.6). In the one survey that modelled it, perceived benefit reached intention only through workflow fit (Section 5.3). We therefore propose treating toolchain integration, verification load and controls as work-system variables, whereas TAM studies of enterprise systems treat organizational conditions such as top management support and user guidance as antecedents of intention (Hancerliogullari Koksalmis and Damar, 2022).

Most of the corpus describes resistance, and only two studies compared a lever change against a control (Sections 4.8 and 6.3). The imbalance recalls Stouten et al. (2018), who found several widely prescribed elements of practitioner change models little tested. For AI it leaves open the work-structure question that Pereira and Liker (2021) and Henderson and Salado (2024) raised for 3D scanning and MBSE.

7.2 Engineering versus the generic picture

The generic-user reviews in Table 1 predict that expertise lowers reliance (Burton et al., 2020; Mahmud et al., 2022), and the engineering evidence, mostly from software, agrees (Section 5.1), although programmers showed no default aversion to AI advice. On trust, the review of Glikson and Woolley (2020) points to transparency, whereas engineers relied on the checks they already ran on colleagues' work, and explanation improved calibration in one inspection experiment (Section 5.1). Cieslak and Valor (2024) find that the literature traces resistance mainly to job vulnerability and argue that identity, autonomy and craft matter as well. In the corpus, vulnerability appeared mostly as expectations about juniors, and identity threat as selective

delegation and concealment. Participation, a general lever for Parker and Grote (2022), has its engineering evidence in process-bound manufacturing, where it also demanded participants' effort (Section 5.4).

The software evidence may transfer poorly to physical engineering, because a released drawing or a machined tool, unlike a code change, cannot be reverted in minutes or tested automatically. The verification route of P2 is therefore likely to be slower and dearer outside software. Even among requirements engineers in regulated industries, one participant reported that verifying AI-generated design artefacts stretched a phase planned at two weeks to nearly six (Ullah Shah et al., 2025). Providers' terms leave the duty to verify with the user (Treude, 2026), and in physical engineering that duty meets certification rules and safety obligations, yet the lifecycle stages that carry them are almost empty in the corpus (Section 6.1). Development environments accept assistants as plug-ins and record every change in version control, whereas CAD, CAE and PLM environments are mostly proprietary. The integration of P7 and the review routing of P8 are therefore likely to be harder there. We read P6 to P9 as software findings and expect P1 and P2 to change form in physical work. P4 is the only proposition built on physical-process evidence.

7.3 A research agenda

Table 12 turns the gaps of Section 6.5 into study designs. Cross-sectional surveys can test associations for P1, P2, P4, P7 and P9, the last with a caveat on under-reporting, whereas P3, P5, P6, P8 and P10 need intervention, panel or organization-level designs. A replication row carries the software-based propositions into physical engineering. The last row is the empirical phase of the research programme to which this review belongs (Appendix G).

7.4 Limitations

Completeness is limited by the search and by retrieval. No information specialist reviewed the search strings with PRESS, and the people-side block lacked leadership and related terms (Section 4.8). One round of snowballing replaced the complementary databases and supplied 77 % of the corpus. With no second round and no citation-network check, the corpus inherits the neighbourhood of 157 seed studies. We expect about 135 eligible studies to be missing, 17 % of all eligible (range 121–151), most among the 236 snowball reports we could not retrieve (Table 7). The likely bias runs against leadership studies, and the thinness of leadership and participation is probably partly an artefact.

Coding reliability is the largest threat, because 606 of the 679 codings are AI drafts not individually verified. Against a blind human coder on 68 studies, their agreement ranged from $\kappa = 0.64$ for design to 0.16 for the Maurer and lever fields, and two humans also agreed poorly on constructs (0.25) and outcomes (0.22). Appraisal used a composite MMAT score and relevance-only weights for reviews (Section 3.6); Section 4.8 tests the headline results without weights.

Screening validity rests on one human reviewer and a language-model screener whose errors are systematic, so human–model agreement is not inter-rater reliability. We revised the screening criteria after inspecting where screener and reviewer disagreed during calibration, and a precision condition relaxed after scoring decided whether the final calibration round passed. That round therefore counts as development evidence (Section 3.4). For the same reason, the eligible studies expected among its unread exclusions (Table 7) probably share features; the audit estimates their number but not their kind.

Scope limits how far the propositions travel. Software engineering supplies 37 % of the studies and most high-weight ones, only English and German studies were eligible, and agentic systems, the newest, are thinly covered. A peer-review audit after inclusion replaced 20 preprints by their version of record, excluded one, and kept 14 extended abstracts and one poster, which are juried

Table 12: Research agenda: gaps, study designs to close them, and the propositions (Table 11) they would test. Withdrawn drafts: Section 6.5.

Theme	Gap type	Study design	Propositions
Lever change	Intervention	Staggered roll-out across comparable engineering teams, aligned change of several levers against training alone; logged use and output quality after the pilot, at equal deployment maturity	P3, P10
Mandate	Intervention	Randomized comparison in engineering teams of mandated exposure with mandated method; later use and perceived agency	P6
Trust calibration	Causal test	Factorial experiment crossing checkability and explanation under two workload levels; appropriate reliance as outcome	P2
Physical engineering	Replication	Field studies outside software of reliance by seniority, verification against physical tests, toolchain integration and design-review oversight	P1, P2, P6–P8
Trajectories	Temporal	Panel of novice and senior engineers with repeated unaided tests, plus process tracing of resistance episodes	P1, P5
Governance	Comparative	Matched organizations routing AI oversight through existing reviews or a parallel track; unsanctioned use measured from logs	P8, P9
Traceability stages	Context	Field studies of AI in engineering change management and certification sign-off in regulated firms	withdrawn draft (verification norms)
Leadership	Measurement	Observation or diary studies of sponsors during a roll-out, with credibility measured independently of respondents	withdrawn draft (leader credibility)
Value and SMEs	Measurement, context	Deliberate sampling of small and medium-sized firms; realized value linked to logged use	P7, P10
Construct check	Measurement	Interviews with engineers and engineering managers, and a survey of engineering organizations mapped to the propositions	associations for P1, P2, P4, P7, P9 (survey); P5 (interviews)

Table 13: Recommendations for engineering managers, the propositions (Table 11) they rest on, and the type of evidence: intervention (an engineering study with a comparison), observational (engineering studies reporting a pattern), or outside engineering.

Recommendation	Proposition	Evidence type
Read the kind of resistance before choosing an intervention (Table 9)	P5 (tentative)	Observational for the symptoms; the rule is Maurer's heuristic
Plan and staff verification; route output through existing tests and reviews, and add explanation of the output	P2	Observational, with one field experiment on explanation
Run oversight through existing engineering reviews	P8	Observational, software only
Make the approved route practical and integrate the tool into the engineering environment	P9, P7	Observational, software only
Budget process engineers' effort where work is bound to a physical process	P4	Observational, outside software only
Trial mandated exposure, with the method left to engineers, against a comparison team	P6	Mainly outside engineering
Protect time for learning from peers	P3 (tentative)	Observational, software and manufacturing
Pair novices with seniors	P1	Observational (seniority and expertise patterns); pairing untested
Track sustained use and realized value	P10 (tentative)	Observational, both strata

but lightly reviewed. The bias runs towards software practice, and the propositions probably overstate how cheaply verification-based trust carries into physical engineering (Section 7.2).

8 Implications for Engineering Managers

None of the advice below rests on a tested causal effect. Table 13 gives the proposition and the kind of evidence behind each recommendation, and advice that rests on a tentative proposition is phrased as a hypothesis.

On Maurer's account, training addresses informational resistance only. Engineers who misjudge what the tool can do lack information, but those who keep identity-defining work from it or hide their use do not, and distrust of the sponsor is a problem of leadership. The symptoms in Table 9 help tell these apart; several may coexist in one team.

In the software studies that report it, checking AI output took much of the time that generation saved (Section 5.3). A plan counting only the drafting gain overstates the benefit. Output that can be checked against tests or reviews should go through them, and explanation of the output improved trust calibration in one field experiment (Section 5.1). Checking also slips into proxies such as an agent's plan or a green test run, and in physical engineering it costs more (Section 7.2).

In software, existing code reviews and sign-offs carried AI output with the legitimacy they already had, and controls judged impractical went with unsanctioned tools (Sections 5.3 and 5.4). We would place oversight in reviews engineers already accept and make the approved route practical: tested with its users and built into the engineering environment. The evidence is from software, and Section 7.2 expects this to be harder in CAD and PLM toolchains.

Where work is bound to a physical process, involving process engineers in developing the system and allocating tasks went with acceptance but demanded their effort (Section 5.4); budget for it.

Requiring use of a tool on real work for a fixed period, while leaving the method to engineers, might raise later use (Section 5.4). In software it improved attitudes; the link to use comes from outside engineering and needs a trial against a comparison team.

Learning conditions (champions, protected time, peer learning) accompanied frequent use among developers, and no study links training dose alone to sustained use (Section 5.2). Less

experienced technicians gained most from a troubleshooting assistant in one small field experiment (Löwhagen et al., 2025), and more experienced engineers rejected AI output more often (Section 5.1). Pairing the two is plausible but unevaluated.

Sustained use and realized value are what the corpus rarely measures (Section 7.1). A roll-out can be tracked by logged use three and six months after it, by verification time and by rework.

9 Conclusion

The people-side factors identified here cannot yet be ranked by what changing them would achieve. Of the 274 field studies able to observe a lever change, 176 report none, and only two high-weight studies compared a change against a control (Sections 4.8 and 6.3). Counts of resistance by Maurer level cannot be read as prevalence (Section 4.8). For engineering managers, the most usable result is that checking AI output needs planning and staffing: in software teams it consumed much of the time that generating it saved, and evidence from other domains is too thin to say whether it holds there.

The main limitation is that most reports came from one round of snowballing from 157 database seeds, with a search that carried no leadership or culture terms, which may partly explain the thin leadership evidence. About 135 eligible studies are probably missing, and most codings are AI drafts whose agreement with human coding was lowest on the framework fields. The study most needed is a staggered roll-out across comparable teams in mechanical, electrical or systems engineering, with several work-system levers changed in some teams and training alone in others. Its outcomes would be logged use and realized value after the pilot.

Funding

This research received no specific funding.

Disclosure of interest

The authors report there are no competing interests to declare.

Data availability statement

The search strings, the deduplicated search exports, the screening decisions, the coding sheet with per-study appraisal (the data file of Supplementary Table S1), the verbatim screening and extraction prompts with their hashes, and the scripts that produce every figure and table are available at `https://osf.io/7ys25/?view_only=0e99091e533343fbbea52413a335ed60` (view-only anonymized link for review). Full texts of the included studies are not shared, because they were obtained under institutional licences.

Use of generative AI

The review used Anthropic's `claude-sonnet-5` model, through Claude Code, in two roles in the review process, each described in Sections 3.4 and 3.5 and Appendix B. First, it was a second screener at title and abstract and at full text, blind to the reviewer's decisions, with its sensitivity measured against the reviewer on blind samples and audits before its exclusions were allowed to stand. Once the audits had passed, its high-confidence exclusions stood without human reading (1,553 database and 1,924 snowball records at title and abstract, 85 snowball reports at full text); it never overruled a record the reviewer kept. Second, it drafted extraction records from the full texts, each value tied to a page and a verbatim quotation; a stratified sample was coded blind by

two human coders to measure the drafts' agreement, and the human codings replace the drafts for those studies. All prompts and model outputs are retained and deposited with the materials listed above. During manuscript preparation, the authors used Claude Opus 5.5 (Anthropic) for language editing and formatting assistance. All scientific content, including claims, analyses, and interpretations, was conceived and written by the authors, who reviewed and approved the final text. The authors reviewed and edited all generated text and take full responsibility for the content of the manuscript, including the accuracy of all references.

A Search strings and PRISMA-S report

Both databases were searched on 4 September 2026 with the strings below, reproduced as executed. Only line breaks were changed, to fit the page. The Scopus string is the planned one with the document-type, language and year limits written into the query; re-running it with the limits applied as facets instead returned the same set. For Web of Science the protocol had planned the limits as interface refinements with years 2015 to 2026; the executed query carries them inside the string and extends the upper year to 2027, so that online-first records with forward issue dates are kept. The review therefore covers 2015 to the search date.

Scopus (field TITLE-ABS-KEY)

```
TITLE-ABS-KEY (
  ( "artificial intelligence" OR "machine learning" OR "generative AI"
    OR "generative artificial intelligence" OR "large language model*"
    OR "AI agent*" OR "agentic AI" OR "AI-assisted" OR "AI-enabled"
    OR "AI-based" OR "generative design" OR copilot )
  AND
  ( engineer* OR "engineering organi*ation*" OR "engineering team*"
    OR "engineering work" OR "engineering design" OR "design engineer*"
    OR "product development" OR "product lifecycle management" OR PLM
    OR "manufacturing firm*" OR "engineering firm*" OR "R&D department*"
    OR "technical professional*" OR "systems engineering"
    OR "Industry 4.0" OR "Industry 5.0" )
  AND
  ( "technology acceptance" OR "user acceptance" OR "employee adoption"
    OR "resistance to change" OR "user resistance"
    OR "professional identity" OR "occupational identity"
    OR "role identity" OR "identity threat" OR "trust in AI"
    OR "trust in automation" OR "algorithm aversion"
    OR "algorithm appreciation" OR "appropriate reliance"
    OR "work design" OR "job design" OR "task allocation"
    OR "decision rights" OR "human-AI collaboration"
    OR "human-AI teaming" OR "change management"
    OR "organi*ational change" OR "employee participation"
    OR "employee voice" OR technostress OR "AI anxiety"
    OR "job insecurity" OR upskilling OR reskilling OR "AI literacy"
    OR deskilling OR "psychological safety" )
)
AND PUBYEAR > 2014
AND ( LIMIT-TO ( DOCTYPE , "ar" ) OR LIMIT-TO ( DOCTYPE , "cp" )
      OR LIMIT-TO ( DOCTYPE , "re" ) )
AND ( LIMIT-TO ( LANGUAGE , "English" )
      OR LIMIT-TO ( LANGUAGE , "German" ) )
```

Web of Science Core Collection (field TS)

```
TS=(
  ( "artificial intelligence" OR "machine learning" OR "generative AI"
    OR "generative artificial intelligence" OR "large language model*"
    OR "AI agent*" OR "agentic AI" OR "AI-assisted" OR "AI-enabled"
```

```
    OR "AI-based" OR "generative design" OR copilot )
  AND
  ( engineer* OR "engineering organi?ation*" OR "engineering team*"
    OR "engineering work" OR "engineering design" OR "design engineer*"
    OR "product development" OR "product lifecycle management" OR PLM
    OR "manufacturing firm*" OR "engineering firm*" OR "R&D department*"
    OR "technical professional*" OR "systems engineering"
    OR "Industry 4.0" OR "Industry 5.0" )
  AND
  ( "technology acceptance" OR "user acceptance" OR "employee adoption"
    OR "resistance to change" OR "user resistance"
    OR "professional identity" OR "occupational identity"
    OR "role identity" OR "identity threat" OR "trust in AI"
    OR "trust in automation" OR "algorithm aversion"
    OR "algorithm appreciation" OR "appropriate reliance"
    OR "work design" OR "job design" OR "task allocation"
    OR "decision rights" OR "human-AI collaboration"
    OR "human-AI teaming" OR "change management"
    OR "organi?ational change" OR "employee participation"
    OR "employee voice" OR technostress OR "AI anxiety"
    OR "job insecurity" OR upskilling OR reskilling OR "AI literacy"
    OR deskilling OR "psychological safety" )
)
AND PY=(2015-2027)
AND DT=(Article OR Review OR "Proceedings Paper")
AND LA=(English OR German)
```

Scopus does not accept `?` as a single-character wildcard inside a phrase, so *organi*ation* is truncated with an asterisk there and with a question mark in Web of Science.

PRISMA-S items

Table 14 answers the items of the PRISMA-S extension for reporting literature searches.

Table 14: PRISMA-S items for the search.

Item	Report
Databases and platforms	Scopus (scopus.com, Advanced document search, field `TITLE-ABS-KEY`); Web of Science Core Collection (webofscience.com, Advanced Search, all editions available to the subscription, field `TS`: title, abstract, author keywords, Keywords Plus). Access through Universidad Politécnica de Madrid institutional subscriptions.
Multi-database searching	Each database searched separately with its own syntax; results pooled and deduplicated.
Study registries	Not searched.
Online resources and browsing	Not searched. IEEE Xplore, ACM Digital Library, AIS eLibrary and the Design Society proceedings were named in the protocol as complementary sources but were not searched separately; citation searching was used instead.
Citation searching	One iteration of backward and forward snowballing (Wohlin, 2014), run on 25 September 2026 through the Semantic Scholar Graph API; Wohlin iterates until no new studies appear, so the single round is a departure (Section 3.8). Reference lists and citing works were harvested for every retrieved database-arm report; candidates were kept when linked to at least one of the 157 full-text includes, dated 2015 or later, and not already among the database records, giving 5,875 candidates. Missing abstracts were filled from Semantic Scholar, then OpenAlex, then Crossref.
Contacts	Authors and experts were not contacted.

Item	Report (continued)
Other methods	None. A pilot of the string on OpenAlex (title and abstract, articles from 2015) returned 369 records and was used only to size the search; its records did not enter the review.
Full search strategies	Reproduced verbatim above.
Limits and restrictions	Scopus: `PUBYEAR > 2014`; document types article, conference paper, review; English or German. Web of Science: `PY=(2015-2027)`; document types article, proceedings paper, review; English or German. All limits written into the query strings.
Search filters	No published search filter was used.
Prior work	The strings were written for this review and piloted on OpenAlex; no earlier search strategy was reused.
Updates	None; the searches were run once.
Dates of searches	Databases: 4 September 2026. Citation searching: 25 September 2026.
Peer review of the search	Not done. The strings were not reviewed by an information specialist (for example with the PRESS checklist).
Total records	Scopus 3,083 (3,089 reported by the database, 3,088 distinct identifiers exported); Web of Science 949 (950 exported); database total 4,032; citation searching 5,875; all sources 9,907.
Deduplication	Pooled export lines (4,750) deduplicated by an author-written routine in Python, in three passes: identical DOI (1,455 removed), identical normalized title with first-author surname and year (56), title similarity of at least 0.95 within the same year (3). These counts include the overlap between Scopus export windows; 796 duplicates across databases were removed, leaving 3,236 records. Snowball candidates were deduplicated against the database records before screening; seven residual duplicates found at full text were resolved to the included version.

Search execution details

Scopus reported 3,089 records after filters but refused whole-set exports and stopped paging at record 2,000. The set was exported in pages of 200 ("All documents on this page") under two opposite date sorts, newest first (ten pages) and oldest first (eight pages), giving 19 files and 3,800 lines. These contained 3,088 distinct Scopus identifiers; one record was lost at the seam between the two sort orders, where tied publication dates shift records across windows. After Scopus's internal duplicates were removed, 3,083 records remained. Web of Science Core Collection (all editions available to the subscription; the protocol had planned SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH and ESCI) returned 950 records in one full-record export, 949 after deduplication. Filters were applied inside the query strings, so both searches are reproducible from the strings alone. A pilot of the intersected string on OpenAlex (title and abstract, articles from 2015) returned 369 records, 362 of them from 2020 or later. Deduplication is reported in Table 14. 2 of the 3,236 screened records had no abstract.

B Automated screening and extraction: disclosure

Automated screening: validation in detail

Table 4 summarizes each stage; this subsection gives what the main text leaves out. The model was run between 5 and 27 September 2026 on ten records per session, whereas Khraisha et al. (2024) found accuracy fell as more records were placed in a single prompt; calibration and deployment used the same chunk size, so the measured sensitivity describes the procedure as run. In round 1 (200 records, criteria v1.0) the ungated sensitivity was 0.923 [0.797, 0.973]. In round 2 (200 fresh records) the gated rule reached 0.796 [0.664, 0.885] and the ungated rule 0.673. Between the two rounds the reviewer's keeps rose from 39 to 49, and the reviewer's use of *unsure*

from none to 11. All ten gated misses failed on the population criterion at high confidence, and several were populations the written criteria already excluded: undergraduate students, accountants in manufacturing firms, HR staff, low-literate mobile users, smart-home users and a legal analysis of copyright. The two samples were disjoint, so no record had been screened twice and intra-rater drift could be seen only in these marginals; round 3 therefore added hidden repeats. The five rulings behind criteria v2.0 admitted physical-product manufacturing, including consumer goods and food, when the organization or process is the unit of analysis; excluded non-technical functions studied in isolation; excluded student, education, general-public, domestic and convenience samples; required an explicit engineering or technical-workplace anchor for macro, legal and conceptual work; and ruled that automated scheduling is not task allocation. Round 3 screened 400 fresh records and 40 hidden repeats on a different day; the ungated rule reached 0.843 [0.750, 0.906]. The floor was lowered from 0.95 to 0.90 before scoring because a single human screener's sensitivity ranged from 0.776 to 1.000 in one published comparison (Matsui et al., 2024). The precision condition (the lower bound of the interval) was relaxed from 0.85 to 0.80 after scoring, at the reviewer's instruction; under the earlier condition the gated rule failed by 0.014.

Unread model exclusions were estimated with $K/(K + fE)$, where K is the reviewer's keeps, f the audit keep rate and E the unread exclusions. The database title-and-abstract audit gave 1.000 (lower bound 0.936). Round 3 had found 7 keeps among 249 gated exclusions, which sits poorly with an audit keep rate of zero, so a pooled estimate of 0.940 is also reported. The snowball title-and-abstract audit gave 1.000 (lower bound 0.957). The snowball keep pile of 3,851 records was read in ranked batches of 250 under a rule, fixed before any human snowball decision, to stop after two consecutive batches below 3 % keeps; the rule never fired, so the whole pile was read. At full text, the model screener was validated once, blind, against the reviewer's 404 database-arm decisions: gated 0.994 [0.965, 0.999], ungated 0.834. In the snowball arm it excluded 185 reports at high confidence; the audit of 100 gave 0.979 [0.966, 0.987], or 0.955 if all 185 are treated as unread.

Automated screening: provenance

All model runs used `claude-sonnet-5` through Claude Code, as subagents or as isolated headless sessions started outside the project repository. Title and abstract screening under criteria v2.0 used criteria hash `ce987714a14e`, system-prompt hash `77d5dd0d9580` and wrapper hash `2fab370225c2`; rounds 1 and 2 used criteria v1.0 (`06eb93bee073`) and prompt `a6e94b55f78b`. The full-text screener used rulebook `c9b3be142481` and wrapper `a8b1310010e9`, fixed in a pre-commitment (hash `bec4c06a41ac`) before any model read a full text; each report was supplied as page-marked text capped at 200,000 characters, with no access to the reviewer's decisions. Extraction drafts used rulebook `325a2be7716b` and wrapper `5305d22b9be4` for both arms. In round 1 a confidence-gated variant scored 0.974, but that threshold was chosen after the misses were seen, so the figure is in-sample and is not reported as validation. The round-3 gated rule has specificity 0.763, PPV 0.503 and Cohen's $\kappa = 0.521$ against the reviewer; κ is given for completeness, not as reliability. The snowball keep pile was ranked by model tier (include; unsure with abstract; medium- or low-confidence exclude with abstract; no abstract), then by the number of linked included studies, then by model confidence, with a seeded tie-break. Batch keep rates were 86, 90, 85, 55, 49, 55, 31, 16, 24, 15, 24, 22, 20, 28, 24 and 24 %. Abstracts were recovered for 201 of 223 title-only snowball records marked *unsure*; the reviewer re-screened them blind and excluded 135, leaving 1,449 records for retrieval. In the full-text validation, the one report the reviewer included but the gate would have dropped was a laboratory experiment on master's students, which the model coded as wrong population. Seven of the 13 snowball full-text audit misses were coded *wrong object* by the model.

Extraction agreement

The double-coded sample of 68 studies was stratified by search arm and draft appraisal weight, and the five-study calibration round that preceded it produced twelve written decision rules. Mean Cohen's κ between the two human coders was 0.31 for context, 0.44 for design and theory, 0.25 for constructs, 0.42 for the Maurer-level and lever fields, 0.22 for outcomes and 0.48 for appraisal. Between the drafts and the reviewer the same groups gave 0.48, 0.64, 0.35, 0.16, 0.42 and 0.32.

Retrieval

Open-access copies were located through OpenAlex and Semantic Scholar and downloaded by script; the remainder were sought through institutional library access. Of the 462 database reports, 58 were not retrieved. They are dominated by Taylor & Francis and Emerald journal articles and by conference papers without a DOI, while the year distributions of retrieved and unretrieved reports are close. Applying the database-arm full-text inclusion rate (157 of 404) to the unretrieved reports suggests about 23 eligible studies were lost. The 236 unretrieved snowball reports include one whose downloaded PDF was a different paper; the correct copy was held only by a publisher that blocks scripted download, and the report is counted as not retrieved.

C Included studies

The 679 included reports are listed in Supplementary Table S1, one row per report sorted by first author, with the search arm, engineering domain, design, review dimensions, Maurer levels, Shea levers, appraisal weight and whether the coding is human or an AI draft. The table runs to about twelve pages and would add several hundred studies not cited in the text to the reference list, so it is supplied as a separate supplementary file, together with the same rows as a comma-separated data file.

The corpus reflects a post-hoc audit against the peer-review criterion, run after the first round of review. Candidates were all 25 included records with a preprint DOI, a posted-content record type, or a preprint server as venue. A version of record was accepted when a Crossref or Semantic Scholar record matched the title (similarity of at least 0.90) and the first author and appeared in a journal or proceedings, or when the proceedings were checked by hand. Twenty records were replaced by that version, four already pointed to it, and one with no peer-reviewed version was excluded, leaving 679 included reports. Fourteen extended abstracts and one conference poster, juried but lightly reviewed, remain in the corpus and are flagged in Supplementary Table S1. No other criterion was revisited.

D Coding book

Each included study was coded on the fields in Table 15, one row per study. Fields marked $*$ take several values. Where a study did not report a value, it was coded *not reported*; values were never inferred. A Maurer level was assigned from evidence in the study, with the coder recording whether the assignment was explicit or inferred, and a Shea lever was coded only when the organization changed something, not when the authors recommended it.

Table 15: Extraction fields, grouped as in the coding book.

Field	Definition	Allowed values
A. Identification		
Bibliographic fields	Authors, year, title, venue, DOI, taken from the record	free text; year as issue year

Field	Definition	Allowed values
`venue_discipline`	Discipline of the publishing venue	information systems; management; human factors; engineering design; software engineering; manufacturing and operations; psychology; multidisciplinary
`open_access`	Whether the report is openly available	yes; no
`linked_reports`∗	Other publications reporting the same study	keys of those reports
B. Context		
`country`∗, `region`	Where the data were collected	ISO code, multi-country, not reported; Europe, North America, Asia, other, multi-region
`sector`∗	Industry of the studied organization	automotive; aerospace and defence; industrial machinery; electronics; software and IT; construction; energy and process; medical devices; cross-sector; other
`org_size`	Organization size	SME (under 250); large; both; not reported
`org_type`	Position of the organization in the value chain	OEM; supplier; engineering service provider; research organization; mixed
`regulated`	Safety-critical or regulated setting	yes; no; partly; not reported
`engineering_domain`∗	Engineering discipline of the work studied	mechanical; electrical; software and embedded; civil; process; systems; industrial and manufacturing; multi-domain
C. The AI in question		
`ai_type`∗	Kind of AI system studied	predictive ML; generative text or code; generative design or CAx; decision support; agentic; vision and inspection; robotics and cobots; other
`ai_named_tool`	Named product or model	free text
`ai_autonomy`	Who decides, human or AI	assistive; advisory; substitutive; agentic
`deployment_stage`	Maturity of deployment, as in the companion survey	none; pilot; first production; several in operation; broadly embedded
`lifecycle_stage`∗	Engineering lifecycle stage where the AI is used	twelve stages from requirements to planning and estimation; cross-lifecycle; not reported
D. Study design		
`design`	Research design	cross-sectional or longitudinal survey; laboratory or field experiment; single or multiple case; interview study; ethnography; mixed methods; design science; archival or log mining; simulation; conceptual; review
`n_participants`	Sample size	integer, or cases and their sizes
`informant_role`∗	Roles of the informants	individual contributor; senior or lead engineer; engineering manager; executive; AI, data or IT specialist; change or project manager; mixed
`seniority_reported`	Whether informant seniority is reported	yes; no
`time_horizon`	Temporal structure of the data	single point; repeated measures; retrospective; longitudinal (N months)
`data_source`	Origin of the data	self-report; observed or log; both
E. Theory		
`theory_used`∗	Theories the study draws on	controlled list (TAM, UTAUT, TOE, sociotechnical systems, work systems, job characteristics, identity, trust, change and resistance models, and others); none stated
`theory_role`	How the theory is used	tested; extended; sensitizing lens; proposed
`ai_specific_extension`	AI-specific constructs added to the theory	yes (which); no

Field	Definition	Allowed values
F. People-side constructs		
`dimension`∗	Review dimension of each construct	D1 professional identity and trust; D2 skills and learning; D3 work design and autonomy; D4 leadership and participation; D5 other
`constructs`∗	Constructs present in the study	controlled construct list per dimension
`measure`	How each construct was measured	named scale with source; author-developed; qualitative codes
`role_in_model`∗	Position of the construct in the model	antecedent; mediator; moderator; outcome; descriptive theme
`direction`	Direction of the reported relationship	positive; negative; curvilinear; null; mixed
`effect_detail`	The statistic, or the qualitative claim in one sentence	free text
`moderators_found`∗	Moderators reported by the study	free text
G. Analytical framework		
`maurer_level`∗	Level of resistance evidenced	L1 informational; L2 emotional and identity; L3 relational and trust; not classifiable
`maurer_explicit`, `maurer_evidence`	Whether the level is explicit or inferred, and the justifying passage	explicit; inferred by coder; quotation or paraphrase
`shea_levers`∗	Work-system levers the organization changed	organization; workplace design; task; people; rewards; measurement; information distribution; decision allocation; none
`intervention_described`, `intervention_outcome`	What the organization did, and its reported effect	free text, with evidence strength
`escalation_evidence`	Whether one resistance level produced another	yes (description); no
H. Outcomes		
`outcome_type`∗	Outcomes measured	intention; actual, sustained or intensity of use; task performance; quality; time or cost; innovation; satisfaction; well-being; identity; organizational value; survival past pilot
`adoption_operationalization`	How adoption was operationalized	intention only; self-reported use; logged use; deployment status; value realized
`value_evidence`	Evidence of realized value	none; perceived; measured
I. Appraisal and use		
`mmat_category`, `mmat_score`	MMAT design category and score (Appendix E)	qualitative; randomized; non-randomized; descriptive quantitative; mixed methods; conceptual; score 0–5
`relevance_score`	Sum of the five relevance items	0–10
`weight`	Derived weight in the synthesis	high; medium; low
`coder`, `double_coded`	Who coded the study, and whether it was double-coded	initials; yes or no
`notes`, `quotable`	Coder notes; verbatim quotations with page numbers	free text

E Appraisal instrument and per-study scores

Every included study was appraised with two instruments, and neither was used to exclude a study. In a young field with heterogeneous designs, exclusion on quality would shrink the corpus and tilt it towards one design, so appraisal was used to weight the synthesis instead.

Methodological quality

Methodological quality was appraised with the Mixed Methods Appraisal Tool, version 2018 (Hong et al., 2018). Two screening questions apply to every study: whether the research questions are

clear, and whether the collected data allow them to be addressed. Five criteria from the category matching the design (qualitative, randomized controlled, non-randomized, quantitative descriptive, or mixed methods) were then answered. MMAT has no category for archival, log-mining or simulation studies; these were appraised as quantitative descriptive, reading the sample as the data set or the simulated scenarios and missing or selectively retained data in place of nonresponse bias. An amendment of 27 September 2026, made before any weight was used and before any extraction draft was verified, fixed three derivation rules. Items answered *not applicable* leave the denominator, so the score is 5× (items met) / (applicable items), to one decimal; *cannot tell* still counts against the score, because it marks an item the report did not address. Reviews receive no MMAT score, and conceptual papers with propositions are not MMAT-scorable; both are weighted on relevance alone and capped at medium. Because the MMAT authors discourage a single summary score, the category and the per-item pattern are reported with it.

Relevance rubric

A five-item rubric, specific to this review, scores how directly each study speaks to the review question (Table 16). Each item scores 0, 1 or 2, for a total of 0 to 10.

Table 16: Relevance rubric.

Item	0	1	2
R1 Population specificity	Generic employees or a non-engineering profession	Mixed sample including engineers, not analysed separately	Engineers, designers or engineering managers, or engineers analysed as a distinct group
R2 AI in use	Hypothetical AI, a vignette, or attitudes with no deployed system	A pilot or short trial	A system deployed in real work over time
R3 Construct clarity	Factor named but not defined or measured	Defined, measured with an ad hoc instrument	Validated instrument, or rigorously coded qualitative constructs
R4 Context reporting	Sector, size and setting largely unreported	Some context reported	Sector, size, country, engineering domain and deployment stage reported
R5 Lifecycle locatability	Lifecycle position of the AI cannot be told	Inferable from the description	Explicitly stated

Derived weight

Each study's weight follows from its MMAT score and relevance score under the amended rule in Table 17, which is monotonic in both scores. Low-weight studies are still reported, but a proposition resting only on them is labelled as such and moves to the research agenda.

Table 17: Weight derivation, as amended on 27 September 2026.

Study type	Condition	Weight
Empirical (MMAT-appraised)	MMAT ≥ 4 and relevance ≥ 7	high
	MMAT ≤ 2 or relevance ≤ 4	low
	otherwise	medium
Review or conceptual	relevance ≥ 7	medium
	relevance ≤ 6	low

Per-study scores

Per-study MMAT category and score, relevance score and derived weight are given for all 679 studies in the data file that accompanies Supplementary Table S1; the weight also appears in Supplementary Table S1.

F PRISMA 2020 checklist

Table 18 gives the location of each item of the PRISMA 2020 checklist (Page et al., 2021). Where an item is not reported, the table says so.

Table 18: PRISMA 2020 checklist: where each item is reported.

Item	Topic	Where reported
Title and abstract		
1	Identified as a systematic review	Title
2	Abstract	Abstract, unstructured as the journal requires; not checked item by item against PRISMA for Abstracts
Introduction		
3	Rationale	Sections 1.1 to 1.4; prior reviews in Section 1.5 and Table 1
4	Objectives	Section 1.6 (RQ-A to RQ-E)
Methods		
5	Eligibility criteria	Section 3.2, Table 3; grouping for synthesis by the four dimensions in Section 3.7
6	Information sources	Section 3.3; databases, dates and citation searching in Table 14
7	Search strategy	Appendix A, strings as executed
8	Selection process	Section 3.4, Table 4, Fig. 3; model, prompt hashes and audits in Appendix B
9	Data collection process	Section 3.5; coding book in Appendix D; AI drafting provenance in Appendix B
10a	Outcomes sought	Appendix D, field group H; adoption measures in Section 4.6
10b	Other variables; missing information	Appendix D, field groups A–G; unreported values coded *not reported*, never inferred
11	Risk-of-bias assessment	Section 3.6 and Appendix E (MMAT 2018 and relevance rubric); the MMAT items were AI-drafted and not individually verified; agreement with human coders was measured on 68 double-coded studies ($\kappa = 0.32$ against 0.48 between humans, Section 3.5)
12	Effect measures	None pooled; direction and a one-sentence effect statement coded per construct (Appendix D); reasons for not meta-analysing in Section 3.7
13a	Eligibility for each synthesis	Section 3.7: a study enters every dimension it codes; intervention and value counts use only studies that could observe them
13b	Data preparation	No conversions; coded values used as extracted, a multi-valued field counting once per value (captions of Figs. 5 and 6)
13c	Tabulation and visual display	Section 3.7; Figs. 4 to 9; Tables 5 and 9 to 11
13d	Synthesis methods	Section 3.7: thematic and framework synthesis, evidence gap map; no meta-analysis; harvest plots and bibliometric maps planned but not produced (Section 3.8)
13e	Heterogeneity	Software engineering analysed as a separate stratum (Section 3.7)
13f	Sensitivity analyses	Section 3.7: headline results and proposition support re-tested in nine subsets, including the database arm alone
14	Reporting bias assessment	Expected missing studies estimated from unread and unretrieved records (Section 4.8); no effects were pooled, so no funnel plot or small-study test was possible

Item	Topic	Where reported (continued)
15	Certainty assessment	Appraisal weight per study (Section 3.6, Table 17) and the Evidence column of Table 11; no GRADE-type rating per outcome
Results		
16a	Study selection	Section 4.1, Fig. 2, with both arms
16b	Excluded studies	Counts by reason in Fig. 2. Reports that appeared eligible but were excluded are not listed individually
17	Study characteristics	Sections 4.2 to 4.5; per study in Supplementary Table S1 (Appendix C)
18	Risk of bias in studies	Section 4.7; results with and without weighting in Section 4.8, Table 6; per-study scores in the supplementary data file (Appendix E)
19	Results of individual studies	Reported narratively where cited in Section 5; not tabulated per study
20a	Characteristics of contributing studies	Sections 5.1 to 5.4; arm and stratum of the key studies per proposition in Table 11
20b	Results of syntheses	Section 5; Section 6
20c	Heterogeneity	Software against other engineering in Section 5.5 and Table 6 (Section 4.8)
20d	Sensitivity analyses	Section 4.8, Table 6; DB/HC/OE column of Table 11
21	Reporting biases	Expected missing studies by source in Table 7 (Section 4.8)
22	Certainty of evidence	Evidence column of Table 11
Discussion		
23a	Interpretation	Sections 7.1 and 7.2
23b	Limitations of the evidence	Section 7.4
23c	Limitations of the review process	Section 7.4; automated-screening limits in Section 3.4
23d	Implications	Practice in Section 8; research in Section 7.3, Table 12 and Section 6.5
Other information		
24a	Registration	Section 3.1: not registered before the search; the screening, extraction and synthesis protocol is deposited at `https://osf.io/7ys25/?view_only=0e99091e533343fbbea52413a335ed60` (view-only anonymized link for review)
24b	Protocol access	Section 3.1, same deposit
24c	Amendments	Section 3.8; the dated amendment log is supplementary material
25	Support	Funding statement in the declarations
26	Competing interests	Disclosure of interest in the declarations
27	Availability of data, code and materials	Data availability statement in the declarations; Supplementary Table S1 and its data file

G Propositions mapped to survey items

The propositions of Table 11 are to be examined in the empirical phase of the wider research programme, which uses a survey instrument and an interview guide. Table 19 maps each proposition to the survey items that bear on it, keeping an item only where it measures a construct named in the proposition, and names the constructs the instrument lacks. Items are identified by number and described by the construct they measure, not by their wording. Items Q6 to Q24 are five-point agreement items covering, in order: AI framed as support for expertise (Q6), sufficiency of trust in AI output for use (Q7), fit with professional role (Q8), handling of early concerns (Q9), own skills (Q10), practical relevance of training (Q11), time to learn (Q12), continuing support after roll-out (Q13), change in task allocation (Q14), appropriateness of remaining autonomy (Q15), fairness and transparency of AI-supported decisions (Q16), effect on daily work (Q17), leaders' explanation of the rationale (Q18), transformation framing (Q19), opportunity for input (Q20), active leadership support (Q21), peer exchange (Q22), learning from colleagues (Q23)

and value created (Q24). Q1 records the respondent's role, Q2 the organization's AI maturity, and Q4 the engineering process areas where AI was introduced.

Table 19: Propositions of Table 11 mapped to the survey instrument of the wider research programme. *Construct to add* names a measure the instrument does not yet contain.

	Proposition (short form)	Existing items that bear on it	Construct to add
P1	Seniority and expertise co-occur with distinct patterns of reliance on AI output	Q6, Q8 by Q1; Q1 separates engineers from managers but does not measure seniority	Years of experience in the engineering domain; checking AI output before use (reliance)
P2	Trust in AI output is calibrated by verification against an independent reference and by explanation	Q7 as the outcome only; it measures the level of trust, not its source	Perceived verifiability of the output; perceived quality of the system's explanation; perceived workload
P3	Sustained use accompanies learning conditions; no study shows training dose alone followed by sustained use	Q11, Q12, Q13, Q22, Q23 against Q24	Training hours received, in bands, so that dose is a quantity
P4	Where work is bound to a physical process, involving the affected functions in developing the system and allocating tasks is associated with acceptance, at a cost in effort	Q20 against Q24, split by Q4 into physical-process and other areas; Q14 for whether allocation changed	Time spent in participation, for the cost the proposition names
P5	Resistance can sit at several levels at once and move between them in either direction	None: movement between levels is a claim about sequence	None in the survey; the interviews reconstruct the sequence retrospectively, with recall bias stated
P6	In software teams, mandated exposure accompanies more favourable attitudes; whether a prescribed method lowers perceived agency is untested	Q15 as the agency outcome	Whether use was mandatory; whether the way of working was prescribed; frequency of personal use
P7	Toolchain integration is associated with intention to use; its link to organizational use is hypothesized	Q7, Q17 against Q24	Integration of the AI into existing tools and workflows
P8	In software organizations, routing AI oversight through existing engineering reviews accompanies acceptance and continued use	Q16 in part; fairness and transparency are not the routing of oversight	Whether AI output passed through the normal design review and sign-off
P9	Controls judged impractical go together with unsanctioned use, and unclear policy with concealment	None	Perceived practicality of the organization's AI rules; use of unsanctioned tools; disclosure of AI use, asked anonymously
P10	Aligned change across several levers co-occurs with use beyond the pilot; deployment maturity confounds the association	Lever-count index from Q11, Q14, Q16, Q18, Q20, Q22, against Q2 (beyond the pilot) and Q24	None; the index is derived from existing items

With the added constructs, a cross-sectional survey can test associations for P1, P2, P4, P7 and P9, the last with the caveat that respondents may under-report unsanctioned use and concealment even when asked anonymously. P3, P5, P6, P8 and P10 concern change over time, the effect of an intervention or an organization-level configuration, and need intervention, panel or organization-level designs; the survey items listed for them describe associations only.

References

Abbott, A. (1988). *The System of Professions: An Essay on the Division of Expert Labor*. University of Chicago Press, Chicago.

Agrawal, A., Gans, J., and Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press, Boston, MA.

Ali, W. and Khan, A. Z. (2025). Factors influencing readiness for artificial intelligence: a systematic literature review. *Data Science and Management*, 8(2):224–236.

Amarilli, F., Uboldi, S., Saraceni, F., and Tencati, L. (2025). Managing paradoxical tensions in the implementation of explainable AI for product innovation. In *2025 33rd International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 1–6, Catania, Italy. IEEE.

Amrollahi, A., Yang, J., Fazal-e Hasan, S. M., and Badreddine, B. (2026). Knowledge workers' trust and reception of generative AI's advice in complex tasks. *International Journal of Information Management*, 88:103031.

Armenakis, A. A., Harris, S. G., and Mossholder, K. W. (1993). Creating readiness for organizational change. *Human Relations*, 46(6):681–703.

Ashergawi, M., Aljuhmani, H. Y., and Alzubi, A. B. (2026). Empowering green innovation through artificial intelligence: the strategic role of organizational agility and technological readiness in the Turkish construction sector. *Engineering Management Journal*, pages 1–21.

Autor, D. H., Levy, F., and Murnane, R. J. (2003). The skill content of recent technological change: an empirical exploration. *The Quarterly Journal of Economics*, 118(4):1279–1333.

Bagočiūnaitė, R. and Endriulaitienė, A. (2026). Understanding employee attitudes toward artificial intelligence in the workplace: a systematic review of attitude definitions and measurements. *Business: Theory and Practice*, 27(1):168–178.

Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6):775–779.

Baxter, G. and Sommerville, I. (2011). Socio-technical systems: from design methods to systems engineering. *Interacting with Computers*, 23(1):4–17.

Beaudry, A. and Pinsonneault, A. (2010). The other side of acceptance: studying the direct and indirect effects of emotions on information technology use. *MIS Quarterly*, 34(4):689–710.

Benbya, H., Davenport, T. H., and Pachidi, S. (2020). Special issue editorial. artificial intelligence in organizations: current state and future opportunities. *MIS Quarterly Executive*, 19(4).

Berente, N., Gu, B., Recker, J., and Santhanam, R. (2021). Managing artificial intelligence. *MIS Quarterly*, 45(3):1433–1450.

Brachman, M., El-Ashry, A., Dugan, C., and Geyer, W. (2025). Current and future use of large language models for knowledge work. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–24.

Brown, A., D'Angelo, S., Murillo, A., Jaspan, C., and Green, C. (2024). Identifying the factors that influence trust in AI code completion. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, pages 1–9, Porto de Galinhas Brazil. ACM.

Brown, C. L. and Utley, D. R. (2019). Ambiguity aversion in engineers. *Engineering Management Journal*, 31(1):2–7.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942.

Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

Butler, J., Suh, J., Haniyur, S., and Hadley, C. (2025). Dear diary: A randomized controlled trial of generative AI coding tools in the workplace. In *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 319–329, Ottawa, ON, Canada. IEEE.

Cantamessa, M., Montagna, F., and Neirotti, P. (2012). An empirical analysis of the PLM implementation effects in the aerospace industry. *Computers in Industry*, 63(3):243–251.

Cao, X., Hu, C., Sun, J., and Zhang, D. J. (2026). How forced intervention facilitates AI adoption. *Manufacturing & Service Operations Management*, 28(4):1286–1306.

Carroll, C., Booth, A., and Cooper, K. (2011). A worked example of "best fit" framework synthesis: a systematic review of views concerning the taking of some potential chemopreventive agents. *BMC Medical Research Methodology*, 11:29.

Carter, M. and Grover, V. (2015). Me, my self, and I(T): conceptualizing information technology identity and its implications. *MIS Quarterly*, 39(4):931–957.

Challapally, A., Pease, C., Raskar, R., and Chari, P. (2025). The GenAI Divide: State of AI in Business 2025. Technical report, Massachusetts Institute of Technology.

Cherns, A. (1976). The principles of sociotechnical design. *Human Relations*, 29(8):783–792.

Choudhuri, R., Badea, C., Bird, C., Butler, J. L., DeLine, R., and Houck, B. (2026). AI where it matters: Where, why, and how developers want AI support in daily work. In *Proceedings of the IEEE/ACM 48th International Conference on Software Engineering: Software Engineering in Practice*, pages 428–439, Rio de Janeiro Brazil. ACM.

Cieslak, V. and Valor, C. (2024). Moving beyond conventional resistance and resistors: an integrative review of employee resistance to digital transformation. *Cogent Business & Management*, 12(1).

Cihan, U., Haratian, V., İçöz, A., Gül, M. K., Devran, Ö., Bayendur, E. F., Uçar, B. M., and Tüzün, E. (2025). Automated code review in practice. In *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 425–436, Ottawa, ON, Canada. IEEE.

Clegg, C. W. (2000). Sociotechnical principles for system design. *Applied Ergonomics*, 31(5):463–477.

Cooper, R. G. and Brem, A. M. (2024). The adoption of AI in new product development: Results of a multi-firm study in the US and europe. *Research-Technology Management*, 67(3):44–53.

Craig, K., Thatcher, J. B., and Grover, V. (2019). The IT identity threat: a conceptual definition and operational measure. *Journal of Management Information Systems*, 36(1):259–288.

Cui, Z., Xu, S., Gao, J., Geng, L., and Zhou, Y. (2026). Empowerment or depletion? unpacking the asymmetrical pathways from perceived autonomy to human–AI trust. *Buildings*, 16(11):2264.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340.

Day, G. S. and Shea, G. (2020). Changing the work of innovation: a systems approach. *California Management Review*, 63(1):41–60.

Dell'Acqua, F., McFowland, E., Mollick, E., Lifshitz, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2026). Navigating the jagged technological frontier: field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science*, 37(2):403–423.

Dhanorkar, S., Passi, S., and Vorvoreanu, M. (2026). Human oversight of agentic systems in practice: Examining the oversight work, challenges, and heuristics of developers using software agents. In *Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 6438–6465, Montreal QC Canada. ACM.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., and Williams, M. D. (2019). Re-examining the unified theory of acceptance and use of technology (UTAUT): towards a revised theoretical model. *Information Systems Frontiers*, 21(3):719–734.

Edgecomb, I. M., Brisco, R., Gunn, K., and Holliman, A. (2025). Artificial intelligence in engineering design: an industry perspective. *Proceedings of the Design Society*, 5:641–650.

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2):350–383.

Eigner, M. and Stelzer, R. (2009). *Product Lifecycle Management: Ein Leitfaden für Product Development und Life Cycle Management*. VDI-Buch. Springer, Berlin, Heidelberg, 2., neu bearbeitete auflage edition.

Endsley, M. R. (2023). Ironies of artificial intelligence. *Ergonomics*, 66(11):1656–1668.

Enholm, I. M., Papagiannidis, E., Mikalef, P., and Krogstie, J. (2022). Artificial intelligence and business value: a literature review. *Information Systems Frontiers*, 24(5):1709–1734.

Felder, L., Eisenreich, T., Fischer, M., Wagner, S., and Chen, C. (2026). Adoption of generative artificial intelligence in the german software engineering industry: An empirical study. In *Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering*, pages 334–344, Concordia University Montreal QC Canada. ACM.

Feng, D., Yun, B., and Yi Wang, A. (2026). From junior to senior: Allocating agency and navigating professional growth in agentic AI-mediated software engineering. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–24, Barcelona Spain. ACM.

Ford, J. D., Ford, L. W., and D'Amelio, A. (2008). Resistance to change: the rest of the story. *Academy of Management Review*, 33(2):362–377.

Garetti, M., Terzi, S., Bertacci, N., and Brianza, M. (2005). Organisational change and knowledge management in PLM implementation. *International Journal of Product Lifecycle Management*, 1(1):43–51.

Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: review of empirical research. *Academy of Management Annals*, 14(2):627–660.

Gräßler, I., Oleff, C., and Preuß, D. (2022). Proactive management of requirement changes in the development of complex technical systems. *Applied Sciences*, 12(4):1874.

Hackman, J. R. and Oldham, G. R. (1976). Motivation through the design of work: test of a theory. *Organizational Behavior and Human Performance*, 16(2):250–279.

Haefner, N., Parida, V., Gassmann, O., and Wincent, J. (2023). Implementing and scaling artificial intelligence: a review, framework, and research agenda. *Technological Forecasting and Social Change*, 197:122878.

Hancerliogullari Koksalmis, G. and Damar, S. (2022). An empirical evaluation of a modified technology acceptance model for SAP ERP system. *Engineering Management Journal*, 34(2):201–216.

Hanelt, A., Bohnsack, R., Marz, D., and Antunes Marante, C. (2021). A systematic review of the literature on digital transformation: insights and implications for strategy and organizational change. *Journal of Management Studies*, 58(5):1159–1197.

Harrir, M. M. and Triqui Sari, L. (2025). Exploring the power of artificial intelligence in supply chain management: a literature review on the artificial intelligence applications and tools used in supply chains and their distribution according to the SCOR method. *Engineering Management Journal*, 37(3):267–288.

Heimberger, H., Horvat, D., Jäger, A., and Schultmann, F. (2026a). Exploring AI adoption in manufacturing: An empirical study on effects of AI readiness. *International Journal of Production Economics*, 297:109733.

Heimberger, H., Horvat, D., and Schultmann, F. (2026b). Exploring the factors driving AI adoption in production: a systematic literature review and future research agenda. *Information Technology and Management*, 27(1):53–69.

Henderson, K. and Salado, A. (2024). The effects of organizational structure on MBSE adoption in industry: insights from practitioners. *Engineering Management Journal*, 36(1):117–143.

Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M.-P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M.-C., Vedel, I., and Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information*, 34(4):285–291.

Jallow, H., Renukappa, S., Suresh, S., and Rahimian, F. (2023). Artificial intelligence and the UK construction industry – empirical study. *Engineering Management Journal*, 35(4):420–433.

Jeyam, D., Wiedemann, A., Schwabe, G., and Güney, K. (2026). "still in the loop": Coping with technostress in DevOps teams and the impact of GenAI. In *Proceedings of the IEEE/ACM 48th International Conference on Software Engineering: Software Engineering in Practice*, pages 590–600, Rio de Janeiro Brazil. ACM.

Jussupow, E., Benbasat, I., and Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, 48(4):1575–1590.

Kellogg, K. C., Valentine, M. A., and Christin, A. (2020). Algorithms at work: the new contested terrain of control. *Academy of Management Annals*, 14(1):366–410.

Kelly, S., Kaye, S.-A., and Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, 77:101925.

Kemell, K.-K., Saarikallio, M., Nguyen-Duc, A., and Abrahamsson, P. (2025). Still just personal assistants? – a multiple case study of generative AI adoption in software organizations. *Information and Software Technology*, 186:107805.

Khojah, R., Mohamad, M., Erlenhov, L., de Oliveira Neto, F. G., and Leitner, P. (2026). Large language model company policies and policy implications in software organizations. *IEEE Software*, 43(1):64–72.

Khojah, R., Mohamad, M., Leitner, P., and de Oliveira Neto, F. G. (2024). Beyond code generation: An observational study of ChatGPT usage in software engineering practice. *Proceedings of the ACM on Software Engineering*, 1(FSE):1819–1840.

Khraisha, Q., Put, S., Kappenberg, J., Warraitch, A., and Hadfield, K. (2024). Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Research Synthesis Methods*, 15(4):616–626.

Kim, H.-W. and Kankanhalli, A. (2009). Investigating user resistance to information systems implementation: a status quo bias perspective. *MIS Quarterly*, 33(3):567–582.

King, W. R. and He, J. (2006). A meta-analysis of the technology acceptance model. *Information & Management*, 43(6):740–755.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering, version 2.3. EBSE Technical Report EBSE-2007-01, Keele University and University of Durham, Keele, UK.

Klemmer, J. H., Horstmann, S. A., Patnaik, N., Ludden, C., Burton, Jr., C., Powers, C., Massacci, F., Rahman, A., Votipka, D., Lipford, H. R., Rashid, A., Naiakshina, A., and Fahl, S. (2024). Using AI assistants in software development: A qualitative study on security practices and concerns. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 2726–2740, Salt Lake City UT USA. ACM.

Krzywdzinski, M. and Butollo, F. (2022). Combining experiential knowledge and artificial intelligence. the digital transformation of a traditional machine-building company. *management revue*, 33(2):161–184.

Lapointe, L. and Rivard, S. (2005). A multilevel model of resistance to information technology implementation. *MIS Quarterly*, 29(3):461–492.

Lee, J. D. and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80.

Li, B., Lai, E. Y., and Wang, X. S. (2026). From tools to agents: meta-analytic insights into human acceptance of AI. *Journal of Marketing*, 90(3):13–33.

Li, J., Cao, H., Lin, L., Hou, Y., Zhu, R., and El Ali, A. (2024). User experience design professionals' perceptions of generative artificial intelligence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–18, Honolulu HI USA. ACM.

Liccardo, G. and Cerchione, R. (2025). GenAI-assisted knowledge generation: A case study on human- machine collaboration through the SECI model. *European Conference on Knowledge Management*, 26(2):1217–1225.

Lines, R. (2004). Influence of participation in strategic change: resistance, organizational commitment and change goal achievement. *Journal of Change Management*, 4(3):193–215.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Long, D. and Magerko, B. (2020). What is AI literacy? competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–16, New York, NY. Association for Computing Machinery.

Luke, T., Mesa, D., Tan, L., Kuys, J., and Kocsis, A. (2026). How are professional practices adopting generative AI? the case of an engineering design and product development team. *Proceedings of the Design Society*, 6:2423–2432.

Löwhagen, N., Schwendener, P., and Netland, T. (2025). Can a troubleshooting AI assistant improve task performance in industrial contexts? *International Journal of Production Research*, pages 1–22.

Magara, T. and Phahlane, M. (2026). Unlocking IIoT potential: a systematic review of AI applications, adoption drivers, and implementation barriers. *Applied AI Letters*, 7(1):e70017.

Mahmud, H., Islam, A. K. M. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? A systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Matsui, K., Utsumi, T., Aoki, Y., Maruki, T., Takeshima, M., and Takaesu, Y. (2024). Human-comparable sensitivity of large language models in identifying eligible studies through title and abstract screening: three-layer strategy using GPT-3.5 and GPT-4 for systematic reviews. *Journal of Medical Internet Research*, 26:e52758.

Maurer, R. (1996). *Beyond the Wall of Resistance: Unconventional Strategies That Build Support for Change*. Bard Press, Austin, TX.

Maurer, R. (2010). *Beyond the Wall of Resistance: Why 70% of All Changes Still Fail — And What You Can Do About It*. Bard Press, Austin, TX, revised edition.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734.

Merhi, M. I. (2023). An evaluation of the critical success factors impacting artificial intelligence implementation. *International Journal of Information Management*, 69:102545.

Mikalef, P. and Gupta, M. (2021). Artificial intelligence capability: conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information & Management*, 58(3):103434.

Miller, C., Choudhuri, R., Ulloa, M., Haniyur, S., DeLine, R., Storey, M.-A., Murphy-Hill, E., Bird, C., and Butler, J. L. (2026). "maybe we need some more examples:" individual and team drivers of developer GenAI tool use. In *Proceedings of the 2026 IEEE/ACM 48th International Conference on Software Engineering*, pages 2582–2595, Rio de Janeiro Brazil. ACM.

Mirbabaie, M., Brünker, F., Möllmann Frick, N. R. J., and Stieglitz, S. (2022). The rise of artificial intelligence—understanding the AI identity threat at the workplace. *Electronic Markets*, 32(1):73–99.

Mulgund, P., Singh, R., and Purao, S. (2026). How gen AI is reshaping software design work: A qualitative study through an activity theory lens. In *Proceedings of the 63rd ACM Conference on Computers and People Research*, pages 42–49, Flagstaff AZ USA. ACM.

Määttä, S., Kelanti, M., and Turhan, B. (2026). Generative AI and the future of professional software development: Survey findings from finland. In *Proceedings of the 34th ACM International Conference on the Foundations of Software Engineering*, pages 1623–1631, Concordia University Montreal QC Canada. ACM.

Neumann, M., Bischof, L., Hinz, N. E., Altun, A., Stockmann, L., Schrader, D., Ahaus, A. C., Demirci, E. C., Gabel, B., Rauschenberger, M., Diebold, P., Fritzemeier, H., and Przybyłek,

A. (2026). Between policy and practice: GenAI adoption in agile software development teams. In *Lecture Notes in Business Information Processing*, pages 290–308. Springer Nature Switzerland.

Noy, S. and Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

O'Brien, G. (2025). How scientists use large language models to program. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–16, Yokohama Japan. ACM.

Oldemeyer, L., Jede, A., and Teuteberg, F. (2025). Investigation of artificial intelligence in SMEs: a systematic review of the state of the art and the main implementation challenges. *Management Review Quarterly*, 75(2):1185–1227.

Oreg, S. (2003). Resistance to change: developing an individual differences measure. *Journal of Applied Psychology*, 88(4):680–693.

Oreg, S., Vakola, M., and Armenakis, A. (2011). Change recipients' reactions to organizational change: a 60-year review of quantitative studies. *Journal of Applied Behavioral Science*, 47(4):461–524.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372:n71.

Parasuraman, R. and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3):381–410.

Parker, S. K. and Grote, G. (2022). Automation, algorithms, and beyond: why work design matters more than ever in a digital world. *Applied Psychology*, 71(4):1171–1204.

Parnin, C., Soares, G., Pandita, R., Gulwani, S., Rich, J., and Henley, A. Z. (2025). Building your own product copilot: Challenges, opportunities, and needs. In *2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 338–348, Montreal, QC, Canada. IEEE.

Pereira, R. and Liker, J. (2021). Using technology to enhance PD performance: a comparative case study 3-D scanning technology deployment. *Engineering Management Journal*, 33(2):141–154.

Pereira, V., Hadjielias, E., Christofi, M., and Vrontis, D. (2023). A systematic literature review on the impact of artificial intelligence on workplace outcomes: a multi-process perspective. *Human Resource Management Review*, 33(1):100857.

Peretz-Andersson, E., Tabares, S., Mikalef, P., and Parida, V. (2024). Artificial intelligence implementation in manufacturing SMEs: A resource orchestration approach. *International Journal of Information Management*, 77:102781.

Piderit, S. K. (2000). Rethinking resistance and recognizing ambivalence: a multidimensional view of attitudes toward an organizational change. *Academy of Management Review*, 25(4):783–794.

Qian, C. and Wexler, J. (2024). Take it, leave it, or fix it: Measuring productivity and trust in human-AI collaboration. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 370–384, Greenville SC USA. ACM.

Rafferty, A. E., Jimmieson, N. L., and Armenakis, A. A. (2013). Change readiness: a multilevel review. *Journal of Management*, 39(1):110–135.

Rahman, H., D'Cruze, R. S., Ahmed, M. U., Sohlberg, R., Sakao, T., and Funk, P. (2022). Artificial intelligence-based life cycle engineering in industrial production: a systematic literature review. *IEEE Access*, 10:133001–133015.

Regenwetter, L., Heyrani Nobari, A., and Ahmed, F. (2022). Deep generative models in engineering design: a review. *Journal of Mechanical Design*, 144(7):071704.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., and PRISMA-S Group (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, 10(1):39.

Robbes, R., Matricon, T., Degueule, T., Hora, A., and Zacchiroli, S. (2026). Agentic much? adoption of coding agents on GitHub. *ACM Transactions on Software Engineering and Methodology*, page 3822180.

Rogers, E. M. (2003). *Diffusion of Innovations*. Free Press, New York, fifth edition.

Romeo, E. and Lacko, J. (2026). Adoption and integration of AI in organizations: a systematic review of challenges and drivers towards future directions of research. *Kybernetes*, 55(3):1286–1307.

Russo, D. (2024). Navigating the complexity of generative AI adoption in software engineering. *ACM Transactions on Software Engineering and Methodology*, 33(5):1–50.

Saran, A. and White, H. (2018). Evidence and gap maps: a comparison of different approaches. *Campbell Systematic Reviews*, 14(1):1–38.

Schepers, J. and Wetzels, M. (2007). A meta-analysis of the technology acceptance model: investigating subjective norm and moderation effects. *Information & Management*, 44(1):90–103.

Selenko, E., Bankins, S., Shoss, M., Warburton, J., and Restubog, S. L. D. (2022). Artificial intelligence and the future of work: a functional-identity perspective. *Current Directions in Psychological Science*, 31(3):272–279.

Senoner, J., Schallmoser, S., Kratzwald, B., Feuerriegel, S., and Netland, T. (2024). Explainable AI improves task performance in human–AI collaboration. *Scientific Reports*, 14(1):31150.

Sergeyuk, A., Golubev, Y., Bryksin, T., and Ahmed, I. (2025). Using AI-based coding assistants in practice: State of affairs, perceptions, and ways forward. *Information and Software Technology*, 178:107610.

Shah, S. T. U., Hussein, M., Barcomb, A., Moshirpour, M., and Leung, H. (2026). Explainability and compliance in AI tools for design artifact generation: A multi-domain practitioner survey in requirements engineering. *Information and Software Technology*, 194:108066.

Shea, G. P. and Solomon, C. A. (2020). *Leading Successful Change, Revised and Updated Edition: 8 Keys to Making Change Work*. Wharton School Press, Philadelphia, revised and updated edition.

Singh, S., Misra, S. C., and Chan, F. T. S. (2020). Establishment of critical success factors for implementation of product lifecycle management systems. *International Journal of Production Research*, 58(4):997–1016.

Singla, A., Sukharevsky, A., Hall, B., Yee, L., Chui, M., and Balakrishnan, T. (2025). The state of AI in 2025: agents, innovation, and transformation. Technical report, McKinsey & Company.

Stornelli, A., Ozcan, S., and Simms, C. (2021). Advanced manufacturing technology adoption and innovation: a systematic literature review on barriers, enablers, and innovation types. *Research Policy*, 50(6):104229.

Stouten, J., Rousseau, D. M., and De Cremer, D. (2018). Successful organizational change: integrating the management practice and scholarly literatures. *Academy of Management Annals*, 12(2):752–788.

Strich, F., Mayer, A.-S., and Fiedler, M. (2021). What do I do in a world of artificial intelligence? investigating the impact of substitutive decision-making AI systems on employees' professional role identity. *Journal of the Association for Information Systems*, 22(2):304–324.

Süner-Pla-Cerdà, S., Şen, G., Kumbasar, E., Şahin, B., and Ünlü, C. E. (2026). Designer experiences and perspectives on the role of generative AI in industrial design. *AI & SOCIETY*, 41(3):2361–2384.

Süße, T., Kobert, M., and Kries, C. (2023). Human-AI interaction in remanufacturing: exploring shop floor workers' behavioural patterns within a specific human-AI system. *Labour and Industry*, 33(3):344–363.

Taherizadeh, A. and Beaudry, C. (2023). An emergent grounded theory of AI-driven digital transformation: Canadian SMEs' perspectives. *Industry and Innovation*, 30(9):1244–1273.

Takaffoli, M., Li, S., and Mäkelä, V. (2024). Generative AI in user experience design and research: How do UX practitioners, teams, and companies use GenAI in industry? In *Designing Interactive Systems Conference*, pages 1579–1593, Copenhagen Denmark. ACM.

Takerngsaksiri, W., Pasuksmit, J., Thongtanunam, P., Tantithamthavorn, C., Zhang, R., Jiang, F., Li, J., Cook, E., Chen, K., and Wu, M. (2025). Human-in-the-loop software development agents. In *2025 IEEE/ACM 47th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 342–352, Ottawa, ON, Canada. IEEE.

Tarafdar, M., Tu, Q., Ragu-Nathan, B. S., and Ragu-Nathan, T. S. (2007). The impact of technostress on role stress and productivity. *Journal of Management Information Systems*, 24(1):301–328.

Tavantzis, T., Lambiase, S., and Russo, D. (2026). Out of place in my own work: how generative AI displaces software engineers. In *Proceedings of the 20th International Symposium on Empirical Software Engineering and Measurement (ESEM 2026)*, Leibniz International Proceedings in Informatics (LIPIcs), München, Germany. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Thomas, J. and Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, 8(1):45.

Tie, J., Yao, B., Li, T., Fang, H., Ahmed, S. I., Wang, D., and Zhou, S. (2026). "should I give up now?" investigating LLM pitfalls in software engineering. *ACM Transactions on Software Engineering and Methodology*, page 3801972.

Tornatzky, L. G. and Fleischer, M. (1990). *The Processes of Technological Innovation*. Issues in Organization and Management Series. Lexington Books, Lexington, MA.

Tran, V.-T., Gartlehner, G., Yaacoub, S., Boutron, I., Schwingshackl, L., Stadelmaier, J., Sommer, I., Alebouyeh, F., Afach, S., Meerpohl, J., and Ravaud, P. (2024). Sensitivity and specificity of using GPT-3.5 Turbo models for title and abstract screening in systematic reviews and meta-analyses. *Annals of Internal Medicine*, 177(6):791–799.

Tranfield, D., Denyer, D., and Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14(3):207–222.

Treude, C. (2026). Accountable agents in software engineering: An analysis of terms of service and a research roadmap. In *Proceedings of the 3rd ACM International Conference on AI-Powered Software*, pages 31–37, Montreal QC Canada. ACM.

Trevelyan, J. (2014). *The Making of an Expert Engineer*. CRC Press, Boca Raton, FL.

Trist, E. L. and Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting. *Human Relations*, 4(1):3–38.

Ullah Shah, S. T., Hussein, M., Barcomb, A., and Moshirpour, M. (2025). Explainability as a compliance requirement: What regulated industries need from AI tools for design artifact generation. In *2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW)*, pages 13–22, Valencia, Spain. IEEE.

Uusitalo, S., Salovaara, A., Jokela, T., and Salmimaa, M. (2024). "clay to play with": Generative AI tools in UX and industrial design practice. In *Designing Interactive Systems Conference*, pages 1566–1578, Copenhagen Denmark. ACM.

Venkatesh, V. and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: four longitudinal field studies. *Management Science*, 46(2):186–204.

Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D. (2003). User acceptance of information technology: toward a unified view. *MIS Quarterly*, 27(3):425–478.

Vial, G. (2019). Understanding digital transformation: a review and a research agenda. *The Journal of Strategic Information Systems*, 28(2):118–144.

Vishwarupe, V., Flechais, I., Shadbolt, N., and Denise Jirotka, M. (2026). "to LLM, or not to LLM?": How designers and developers navigate LLMs as tools or teammates. In *Proceedings of the Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Barcelona , Spain. ACM.

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., and Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. *The International Journal of Human Resource Management*, 33(6):1237–1266.

Wagman, K. B., Dearing, M. T., and Chetty, M. (2025). Generative AI uses and risks for knowledge workers in a science organization. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, Yokohama Japan. ACM.

Webster, J. and Watson, R. T. (2002). Analyzing the past to prepare for the future: writing a literature review. *MIS Quarterly*, 26(2):xiii–xxiii.

Weisz, J. D., Kumar, S. V., Muller, M., Browne, K.-E., Goldberg, A., Heintze, K. E., and Bajpai, S. (2025). Examining the use and impact of an AI code assistant on developer productivity and experience in the enterprise. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–13, Yokohama Japan. ACM.

Weisz, J. D., Muller, M., Houde, S., Richards, J., Ross, S. I., Martinez, F., Agarwal, M., and Talamadupula, K. (2021). Perfection not required? human-AI partnerships in code translation. In *26th International Conference on Intelligent User Interfaces*, pages 402–412, College Station TX USA. ACM.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE '14)*, pages 1–10, New York, NY, USA. ACM.

Xia, Q. N., Constantinides, M., Sarkar, A., Brumby, D. P., and Cox, A. L. (2026). "if you're very clever, no one knows you've used it": The social dynamics of developing generative AI literacy in the workplace. In *Proceedings of the 5th Annual Symposium on Human-Computer Interaction for Work*, pages 1–18, Linz Austria. ACM.

Xu, Y., Wang, Y., Gao, W., Chai, Y., and Been-Lirn Duh, H. (2026). Exploring design practice with generative AI: Perspectives from AEC design professionals. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–19, Barcelona Spain. ACM.

Yamamoto, Y., Muñoz, A. A., and Sandström, K. (2024). Practical aspects of designing a human-centred AI system in manufacturing. *Procedia Computer Science*, 232:2626–2638.

Yüksel, N., Börklü, H. R., Sezer, H. K., and Canyurt, O. E. (2023). Review of artificial intelligence applications in engineering design perspective. *Engineering Applications of Artificial Intelligence*, 118:105697.

Zhan, J., Suvada, K., Xu, M., Tian, W., Cara, K. C., Wallace, T. C., and Ali, M. K. (2025). Accelerating the pace and accuracy of systematic reviews using AI: a validation study. *Systematic Reviews*, 15(1):24.