The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 01, August 2024, pp. 191-202
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i01

Trust-Calibrated Human-AI Collaboration for Engineering Design Review and Program Governance: A Systematic Review

Yogesh Gawade
Industrial and Systems Engineering, Oakland University, Rochester, USA

Article Info

Article history:

Received August, 2024
Revised August, 2024
Accepted August, 2024

Keywords:

AI Risk Management;
Algorithmic Auditing;
Appropriate Reliance;
Engineering Design Review;
Explainable AI;
Human-AI Collaboration;
Program Governance;
Systematic Review;
Trust Calibration

ABSTRACT

Artificial Intelligence (AI) is now viewed as an analyst and decision support in the field of engineering design review and program governance, but the benefit of using AI as such will be contingent upon the level of human dependency to gain from system reliability. This systematic review is a synthesis of 28 sources published from 1997–2023 about human-AI interaction, trust in automation, explainable AI, AI ethics, and algorithmic auditing, and maps these sources to the stages of systematic engineering design. Each source was assigned to one of the constructs of trust, reliability outcomes, types of interventions, and governance functions. From the synthesis, reliability, transparency, and feedback are the main drivers of trust in AI, having been identified as an appropriate design target by over 18 interaction guidelines distilled from well over 150 recommendations, a review of 127 trust in automation studies, and a review of approximately 150 empirical trust in AI studies. While explanations led to higher acceptance of correct and incorrect recommendations, cognitive forcing and confidence displays resulted in better calibration at the expense of usability. Transparency was mentioned in 73 of 84 ethics guidelines and a national risk management framework was recently established, in which governance has four functions. A trust-calibrated review architecture is suggested that combines the level of automation per phase, seven formalized metrics and auditable governance gates, providing measurable criteria for implementing AI in review processes with safety and cost implications.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Name: Yogesh Gawade
Institution: Industrial and Systems Engineering, Oakland University, Rochester, USA
Email: yogeshgawade@oakland.edu

1. INTRODUCTION

Engineering design review (EDR) and program governance are decision intensive tasks that involve review of requirements, architectures, risks and schedules prior to committing resources. Increasingly, AI (artificial intelligence) tools are used for these activities to review documentation, identify nonconformities, and predict program risk, turning the reviewer from evaluator to a supervisor of algorithmic advice. More than 150 suggestions have been combined into 18 guidelines, tried by 49 design practitioners in 20 products that are used with AI, and were categorized into four temporal stages: initial use, while interacting, when the system is wrong, and over time [1].

The ability of reviewers to interrogate AI output is dependent on the notion of

explainability, namely how much the reasoning behind a models output can be made understandable to a selected target group. Explainable AI (XAI) has been categorized as models that are designed to be transparent and post-hoc methods, such as the generation of textual, visual, local, and example-based explanations, and has been defined as a key component of ethical and responsible AI, which incorporates aspects of fairness, accountability, and privacy into its deployment [2]. During a software engineering review, these categories are used to decide whether an identified flagged design defect can be assigned to a feature that is interpretable or is limited to an opaque model score.

But, the equation of explainability does not imply improved joint decisions. In the area of team performance, it examined the performance of human-AI team compared to either human agents or AI agents, finding that complementary team performance (where the human-AI team outperforms one or both agents individually) was only achieved for sentiment classification and logical reasoning tasks, and that explanations had a positive effect on the uptake of AI recommendations regardless of whether the recommendation was correct or incorrect [3]. The insight for design review boards is that a convincing explanation interface could magnify, not correct, the spread of wrong flags to approval decisions.

Deliberately delayed acceptance is a partial cure. Cognitive forcing functions, which included an on-demand piece of advice, initial independent judgment, and enforced delay, were tested with 199 participants, and found to lead to fewer reliance on erroneous AI suggestions than simple explainable designs, but the most advantageous conditions were evaluated subjectively as least advantageous [4]. In this context, the present review focuses on three questions: What are the constructs that define calibrated trust? What sorts of interventions move reliance toward appropriateness? How can these mechanisms be incorporated into design review gates and program governance structures? The contribution is a synthesis of the evidence outlined in a structured format, a mapping of the calibration mechanisms of the various phases and a series of formal metrics that will be applicable to audit.

2. LITERATURE REVIEW

2.1 Reliance Failures and Situated Trust

There are two directions of miscalibration. Across five studies, the more the decision maker observes the algorithm making errors, the more the decision maker loses faith in the algorithm, even though the algorithm is making better forecasts than the forecaster, in the case of algorithm aversion [5]. When forecasting is often inaccurate, as it is in program governance, one failed forecast could lead to the abandonment of a tool that has more success in the long run than the experts' forecast.

Trust is also influenced by the organizational context and not just properties of the models. A scenario-based study with 29 AI users and practitioners proposed social transparency as an extension of explainability, which would be characterized by the "who, what, when, and why" of earlier human interactions with an AI output, and reported to be useful for better-calibrated decisions [6]. Records that show the consensus and/or rejection of a specific AI flag by earlier engineers are therefore considered a form of transparency relevant to the governance process.Records that indicate whether a particular AI flag was accepted or rejected by previous engineers are therefore a form of transparency for the governance process.

A cautionary element is added to automation research. The automation conundrum is that operators are less likely to keep the situation awareness necessary to intervene when automation fails as automation becomes more reliable and more autonomous [7]. As shown in Figure 1, the outcome of relying can be classified into three types: appropriate reliance, overreliance, and

underreliance, as well as into two types of self-reliance: appropriate self-reliance and inappropriate reliance, the latter costs being minimized during design review.

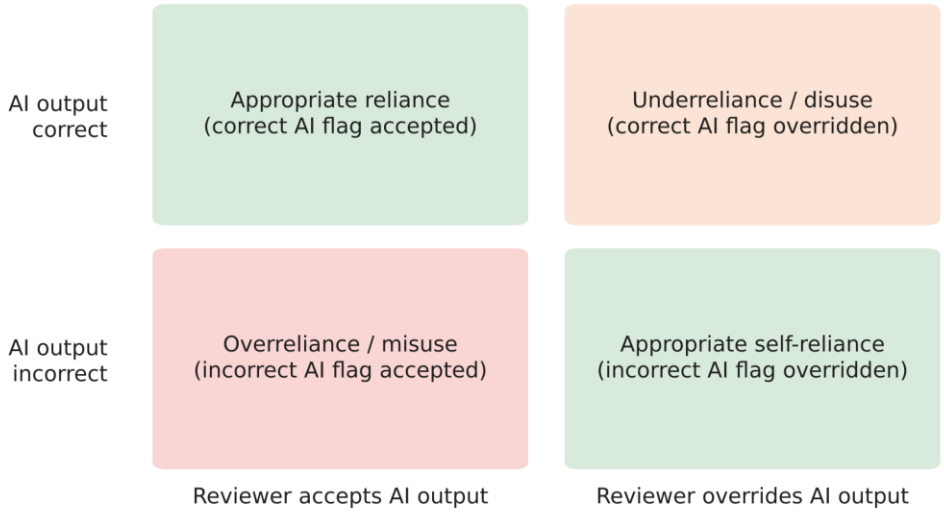


Figure 1. Reliance Outcome Matrix for AI-Assisted Design Review

Note: Green cells denote appropriate reliance; red and orange cells denote misuse and disuse, respectively. Adapted from the use, misuse, and disuse taxonomy of Parasuraman and Riley (1997).

2.2 *Ethical Principles and Determinants of Trust*

The boundaries for reliance can be set by normative frameworks. The AI4People framework succeeded in reducing the four pillars (beneficence, non-maleficence, autonomy and justice) to five principles, including explicability as an enabling principle that makes the other four assessable, and subsequently developed 20 action points for policy and practice [8]. In Figure 2 the same themes are visible in the broader landscape of guidelines, with transparency as the most often used guideline and solidarity as the least.

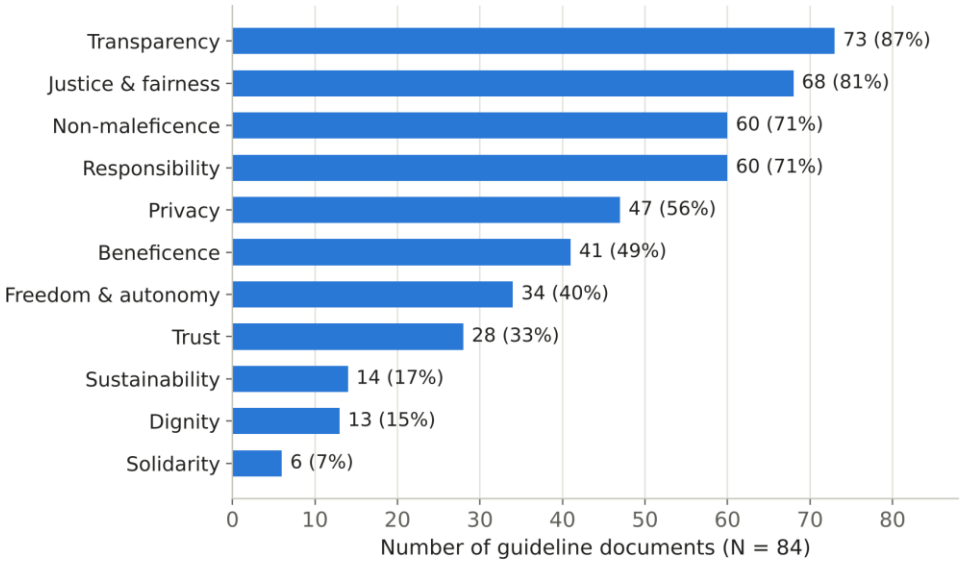


Figure 2. Frequency of Ethical Principles Across 84 AI Ethics Guidelines

Source: Adapted from Jobin, Ienca, and Vayena (2019). Percentages are relative to the 84 documents analyzed.

The previous and current literature has identified some empirical determinants: There are about 150 studies that have distinguish between cognitive trust and emotional trust with tangibility, transparency, reliability, task characteristics, and immediacy behaviors identified as the antecedents [9]. The review also segmented the types of AI representations by the type of robotics, virtual or embedded, and found that trust in embedded AI, which is the most relevant to the analytics-based review tools, often begins high and decreases as one gains experience with errors. The path highlights the importance of calibration systems that last beyond the first deployment.

2.3 Explanation Methods and Layered Trust Models

There are several approaches to the explanation problem, which can be distinguished by the problem they are trying to solve: model explanation, outcome explanation, model inspection, and transparent box design, and also differentiated by explanator type, model type and data type [10]. Outcome explanation is provided for engineering review, whereas model inspection provides program-level assurance that a classifier produces the same results for all design families, suppliers, and product generations.

Automation trust has been split into three components: dispositional trust, based on dispositional factors; situational trust, based on environmental and task factors; and learned trust, based on system experience [11]. An incremental update can be formalized as a learning process that updates the trust level towards the observed performance after each interaction, as shown in Equation (1): Tt = Tt-1 + η × (ot-1 - ot), where Tt is the trust level at the tth review cycle, ot is the observed outcome being coded as 1 for a correct output and 0 for otherwise, and η is the learning rate ranging between 0 and 1.

$$T_{t+1} = T_t + \eta(o_t - T_t) \quad (1)$$

It classifies warranted trust as trust that is derived from a real trustworthy model, and adds contractual trust, as the expectation that a model will satisfy specific stated contracts like accuracy bounds or robustness properties [12]. Table 1 brings together the various constructs presented above, including the layered and multidimensional view of the calibration, and shows that the focus of calibration is on the relation between trust and trustworthiness, rather than on the amount of trust.

Table 1. Synthesis of Trust Models and Constructs Relevant to AI-Assisted Review

Model	Core constructs	Evidence base	Implication for design review
Lee and See (2004)	Performance, process, and purpose bases; calibration, resolution, specificity	Integrative conceptual review	Match reliance to tool capability for each review task class
Hoff and Bashir (2015)	Dispositional, situational, and learned trust	127 empirical studies	Train reviewers and track learned trust over review cycles
Glikson and Woolley (2020)	Cognitive and emotional trust; tangibility, transparency, reliability, task characteristics, immediacy	Approximately 150 studies	Monitor declining trust in embedded analytics tools
Ehsan et al. (2021)	Social transparency: who, what, when, why	Scenario study, 29 participants	Expose prior accept and override histories at each gate
Jacovi et al. (2021)	Warranted versus unwarranted trust; contractual trust	Formal analysis	State accuracy and robustness contracts for gate approval

Source: Synthesized from the reviewed corpus.

3. METHODS

To emulate the scoping review of the ethics documents of AI governance, a systematic review design was adopted, in which 84 ethics guidelines were coded and 11 general principles were found: transparency in 73, justice and fairness in 68, non-maleficence and responsibility in 60 [13]. The corpus consisted of 28 journal articles, conference papers, a technical standard and a textbook which were chosen for their direct relevance to the domains of trust, reliance, explanation, ethics, auditing, and design methodology. It was necessary for the opinion pieces to be peer reviewed or institutional published to include in the sample, and for it to be in a stable digital object identifier, or for it to be an explicit treatment of human reliance on automated advice or on the governance of automated advice, opinion pieces without an analytical framework were excluded. The composition of the corpus is summarized in Table 2 and the temporal distribution of the corpus is shown in Figure 3.

Table 2. Composition of the Reviewed Corpus by Thematic Category and Source Type

Thematic category	Journal	Conference	Other	Total	Share (%)
Human-AI interaction and reliance	4	5	0	9	32.1
Trust in automation and AI	5	1	0	6	21.4
Explainability	3	1	0	4	14.3
Ethics, auditing and governance	6	1	1	8	28.6
Engineering design methodology	0	0	1	1	3.6
Total	18	8	2	28	100.0

Note: "Other" comprises one technical standard and one textbook. Publication years range from 1997 to 2023.

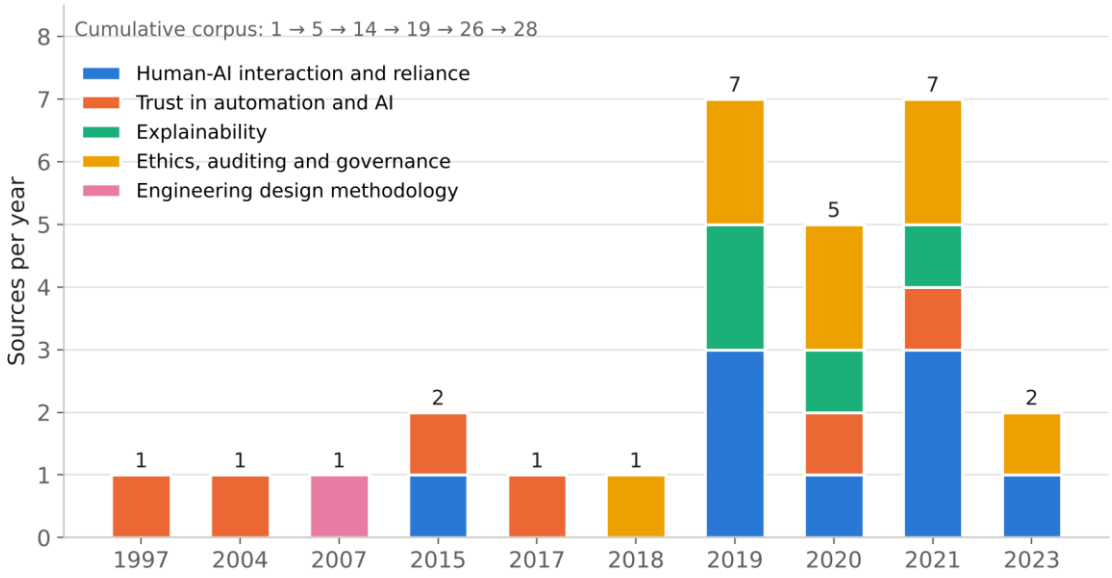


Figure 3. Temporal and Thematic Distribution of the Reviewed Corpus

Source: Compiled from the 28 reviewed sources.

Every source was coded across dimensions inspired by a design space analysis of over 100 empirical human-subject studies of AI support for decision making, which arranged the field by decision tasks, AI model and elements of support, and evaluation metrics [14]. Two additional dimensions (governance function and design review phase) were added to link behavioral observations to controls in the organization. Synthesis of evidence was done through narrative method, quantitative values were extracted from the source documents where

they were reported, and coverage scores given on a four point scale (0-3).

4. RESULTS AND DISCUSSION

4.1 Constructs and Metrics of Calibrated Trust

The basic construct is calibration, i.e. the link between trust and actual capability of automation, and resolution, i.e. the ability of trust to discriminate between levels of capability, and temporal and functional specificity [15]. There are three bases of trust: performance, process, and purpose; miscalibration is found in overtrust and/or mistrust. This relationship is shown in figure 4, with a calibration error defined by equation (2), where Ti is the normalized level of trust of the reviewers in the specific task class, and Ri is the observed reliability of the AI in that same task class.

$$\mathrm{CE} = \frac{1}{N}\sum_{i=1}^{N}|T_i - R_i| \quad (2)$$

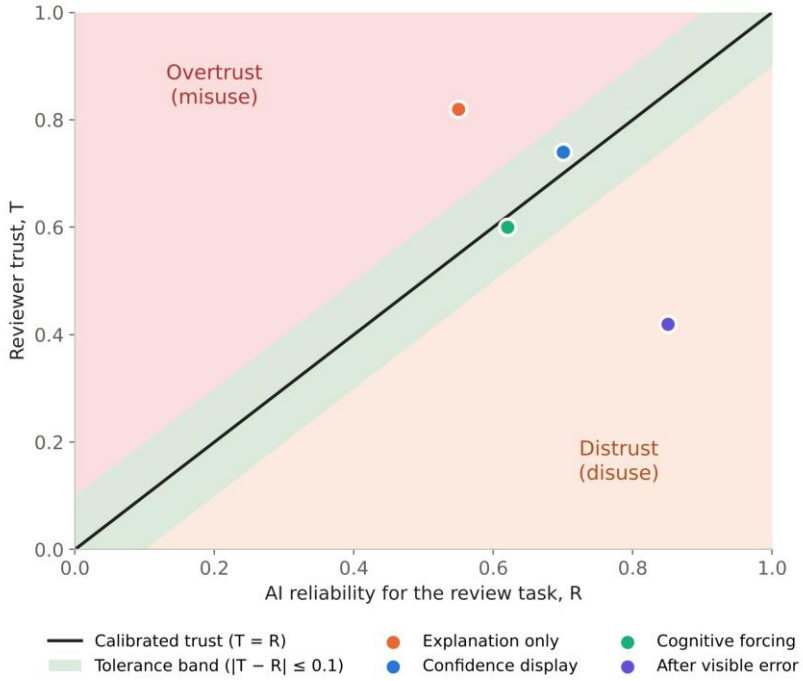


Figure 4. Trust Calibration Space for AI-Assisted Review

Note: Calibration concept after Lee and See (2004). Marker positions are illustrative of the qualitative direction of effects reported in the reviewed studies and are not measured values.

In contrast to aversion, algorithm appreciation has been observed in six experiments, in which lay subjects found algorithmic advice more convincing than human advice, while for experienced forecasters the algorithmic advice led them to be less accurate and to discount it [16]. The extent to which it is believed and relied upon is measured by the weight of advice in Equation (3), with values close to 1 signifying that the advisory value is being adopted and values close to 0 signifying that the value is not.

$$\mathrm{WOA} = \frac{J_1 - J_0}{A - J_0} \quad (3)$$

The social sciences suggest that the explanations offered by the human are contrastive, selective, and social, while mention of probabilities or statistical associations is less meaningful to the target audience than causal explanations [17]. As shown in Table 3, format for reliance interventions other than explanation formats are less studied in the reviewed reliance studies, while reliance formats that are designed around them have the least consistent calibration effects.

Table 3. Empirical Evidence on Interventions Affecting Reliance on AI Advice

Study	Task or setting	Manipulation	Principal finding	Calibration implication
Dietvorst et al. (2015)	Forecasting, five studies	Observation of algorithm errors	Algorithm abandoned after errors despite superior accuracy	Error exposure without context induces disuse
Logg et al. (2019)	Estimation and forecasting, six experiments	Source label of advice	Higher weight on algorithmic advice; experts discounted it	Expertise moderates reliance
Yin et al. (2019)	Prediction tasks, three experiments	Stated and observed accuracy	Stated accuracy effect weakened when observed accuracy was low	Disclose and monitor accuracy jointly
Zhang et al. (2020)	Income prediction	Confidence score; local explanation	Confidence improved calibration; explanation did not	Display case-level confidence
Bansal et al. (2021)	Sentiment and reasoning tasks	AI explanations	Acceptance rose regardless of correctness	Explanations alone are insufficient
Buçinca et al. (2021)	Nutrition decisions, 199 participants	Cognitive forcing functions	Overreliance reduced; lowest subjective ratings	Trade-off between calibration and usability

Source: Synthesized from the reviewed empirical studies.

4.2 Governance and Auditing Mechanisms

Ethics-based auditing at the organizational level was described as a structured process to evaluate the behavior of the entity over time and space, both in terms of consistency with the relevant ethics principles and/or norms, and considering the benefits of the procedure regularity and transparency, against conceptual, technical, economic, social, organizational and institutional constraints [18]. For program governance, auditing becomes tangible checks and balances, where reliance on AI review tools can be reviewed with traditional technical evidence like test reports and analyses.

4.3 Mapping to Engineering Design Review

The systematic engineering design process consists of four basic steps: problem or task planning and clarification, conceptual design, embodiment design and detail design, followed by evaluation and decision making steps [19]. Table 4 links these phases, along with the program gate, to appropriate AI roles, levels of automation, and calibration processes, with advisory AI with tight human control for the early phases, and highly automated rule-based conformance checking for the detail design phase. The program gate is treated separately as forecasts presented at that point have financial implications and are especially susceptible to aversion following noticeable mistakes.

Table 4. Mapping of Engineering Design Phases to Trust-Calibrated AI Support

Phase	Review question	AI role	Automation level	Calibration mechanism	Dominant failure risk
Planning and task clarification	Are requirements complete and consistent?	Requirement screening	Low (advisory)	Independent first judgment; confidence display	Overreliance on incomplete extraction

Phase	Review question	AI role	Automation level	Calibration mechanism	Dominant failure risk
Conceptual design	Which concept best meets the criteria?	Alternative ranking	Low to moderate	Contrastive explanation; social transparency	Anchoring on AI-ranked concepts
Embodiment design	Are layout, form, and materials adequate?	Simulation surrogate; risk flagging	Moderate	Accuracy tracking per task class	Out-of-distribution errors
Detail design	Do drawings and specifications conform?	Conformance checking	High with sampling audit	Sampled audit; override logging	Complacency and skill loss
Program gate	Proceed, rework, or stop?	Cost and schedule forecasting	Advisory only	Contractual accuracy bounds; documented rationale	Aversion after visible forecast error

Note: Phase structure follows Pahl et al. (2007); assignments represent the synthesis of this review.

The classic classification of human-automation interaction includes use, misuse (overreliance), disuse (neglect and/or underutilization) and abuse (automation used without considering the consequences of the operators [20]). The misuse and disuse rate can be expressed as measurable governance quantities operational indicators follow: Equations (4) and (5) define the overreliance and underreliance rates over the logged review decisions; H = 1 indicates that the AI output was accepted, and Y = 1 indicates that the AI output was correct.

$$\mathrm{OR} = P(H = 1 \mid Y = 0) \qquad (4)$$

$$\mathrm{UR} = P(H = 0 \mid Y = 1) \qquad (5)$$

An approach to the study of algorithms emerged, one that advances the idea that algorithm behaviour arises from interaction with data and users, based on the four questions of ethology, as introduced by Ingerid and psychologist [21] and known as a machine behaviour perspective. When applied to design review, this view enables on-going in-service monitoring of AI flags across projects as shown in the feedback loop of Figure 5, instead of depending on pre-deployment validation only.

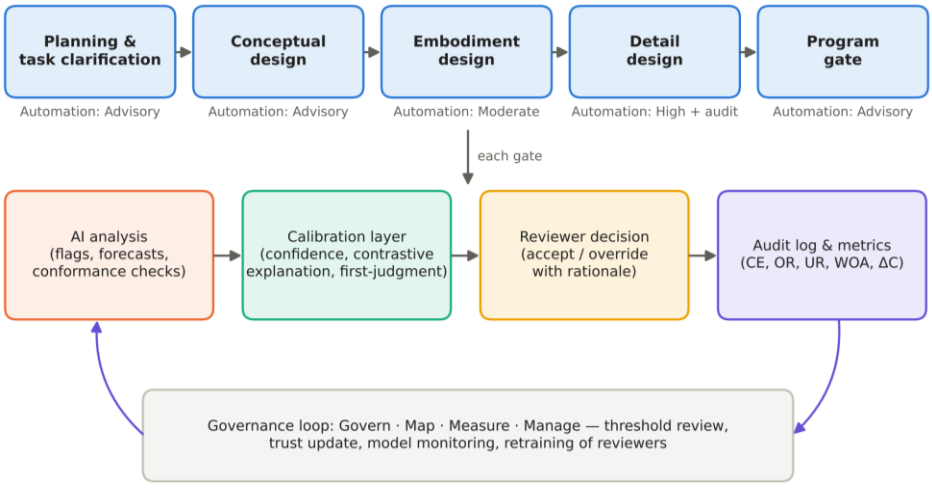


Figure 5. Trust-Calibrated Design Review and Governance Workflow

Note: Phases follow Pahl et al. (2007); governance functions follow the NIST AI RMF 1.0. CE, OR, UR, WOA, and ΔC are defined in Equations (2), (3), (4), (5), and (7).

Research on management automation suggests a paradox of automation and augmentation, where the processes of automation and augmentation are linked both across time and tasks, and where focusing solely on one results in negative consequences for the organization and society [22]. In program governance, this paradox means that successfully implementing tasks in detail design needs to be supplemented periodically with human oversight to ensure that the skills needed to audit the automation are retained and not lost. When a sample of automated decisions is periodically re-reviewed by hand, the ground truth labels for these items will also be available, which will allow for estimation of reliance metrics.

4.4 Program Governance Architecture

The internal algorithmic auditing has been designed as a sequence of four phases: Scoping, Mapping, Artifact collection, and Testing, and is accompanied by a set of documentation artifacts that are created during each of these phases, in order to close the accountability gap between development and deployment [23]. The stages are also aligned with programme gates and for each gate, the required audit artefacts are provided before an AI-assisted review result can be offered as evidence for a proceed decision.

Human centered AI redefines two-dimensional automation design space, with high human control and high computer automation mutually desired to achieve reliable, safe, and trustworthy systems [24]. This framework is compared to the interaction, ethics, auditing, and risk-management frameworks that were identified in the corpus in Table 5, and their representation across the six governance dimensions coded in this review are displayed in Figure 6. The comparison suggests that there is no single framework covering interaction design, risk measurement and auditing documentation, and this is what spurs a combination architecture.

Table 5. Comparison of Governance Frameworks Identified in the Corpus

Framework	Year	Structural elements	Primary level	Contribution to trust calibration
Guidelines for human-AI interaction	2019	18 guidelines in four stages	Interaction design	Sets expectations and supports correction
AI4People	2018	Five principles; 20 action points	Societal policy	Explicability as enabling principle
SMACTR internal audit	2020	Five audit stages	Organization (internal)	Documented artifacts at each gate
Human-centered AI	2020	Two-dimensional control-automation space	Design and organization	Preserves control at high automation
Ethics-based auditing	2021	Audit process with identified constraints	Organization and external	Verifies compliance with principles
NIST AI RMF 1.0	2023	Four functions; seven trustworthiness characteristics	Organizational lifecycle	Measures and manages trustworthiness

Source: Synthesized from the reviewed frameworks.

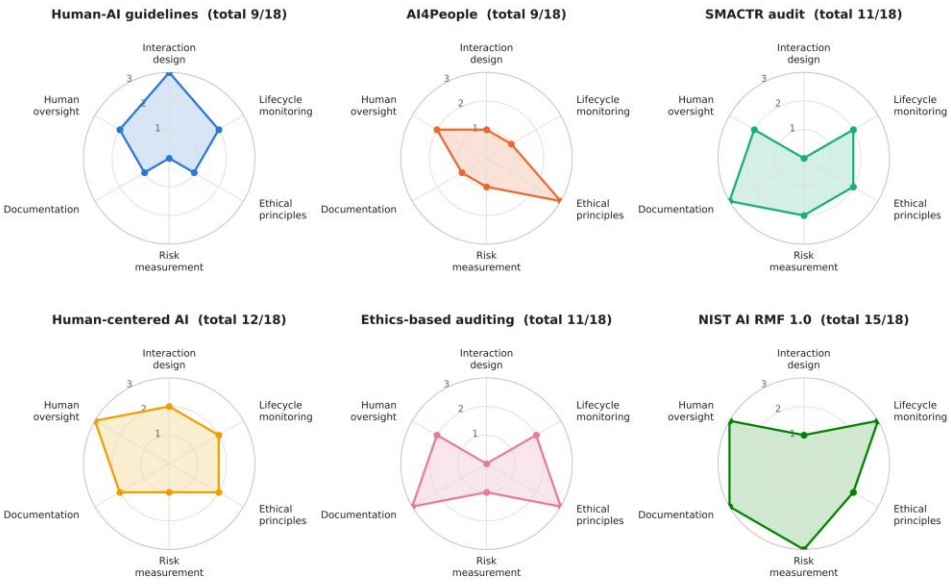


Figure 6. Coverage of Governance Dimensions by Reviewed Frameworks

Note: Scores (0 = absent, 3 = central) were assigned through qualitative coding in this review.

Risk management provides the "backbone" to the operations. NIST AI Risk Management Framework 1.0 categorizes activities across four sets of functions: Govern, Map, Measure, and Manage, and includes seven attributes of trustworthy AI: Validity and Reliability, Safety, Security and Resilience, Accountability and Transparency, Explainability and Interpretability, Privacy Enhancement, and Fairness with harmful bias managed [25]. Like the Map and Measure functions, Equation (6) sums a governance risk index across K risk items, in which the weight, likelihood, and impact are represented by wk, Lk, and Ik, respectively.

$$G = \sum_{k=1}^{K} w_k L_k I_k \,,\ \sum_{k=1}^{K} w_k = 1 \qquad (6)$$

4.5 *Measuring Calibrated Trust in Review Practice*

The remaining weak link is measurement. Reviewing 83 empirical studies on trust in AI-assisted decision making, the authors found that trust had been defined in a variety of ways across the studies, that attitudinal measures were common, and that behavioral measures were rare, and suggested that combining attitudinal and behavioral measures be done under specific conditions of vulnerability [26]. Table 6 thus defines a set of measurements including attitudinal trust and logged reliance behavior, outcome accuracy, and governance risk.

Trust is shaped by accuracy information in measurable ways: In three randomized human-subject experiments, both stated and observed model accuracy impacted trust, with the impact of stated accuracy being reduced when observed accuracy was low [27]. Deployment decisions should then be based on complementarity and not on raw accuracy, as described in Equation (7): if the value is positive, then the human-AI team performs better than the best team.

$$\Delta_C = \mathrm{Acc}_{\mathrm{HAI}} - \max(\mathrm{Acc}_{\mathrm{H}}, \mathrm{Acc}_{\mathrm{AI}}) \qquad (7)$$

Lastly, displaying case-specific confidence scores was found to improve trust calibration, whereas local feature-importance explanations did not and improved joint accuracy was not found when human and AI knowledge were not very distinct [28]. This, together with the forcing-function evidence, is used to support a layered calibration design where the confidence displays, independent first judgments, and auditable override logs are utilized simultaneously at each review gate.

Table 6. Proposed Measurement Set for Trust-Calibrated Design Review

Metric	Definition	Data source	Governance use
Learned trust (T)	Equation (1)	Repeated trust ratings per review cycle	Track trust trajectory after errors
Calibration error (CE)	Equation (2)	Trust ratings and reliability logs	Gate threshold for reviewer-tool alignment
Weight of advice (WOA)	Equation (3)	Numerical cost and schedule estimates	Detect anchoring on AI forecasts
Overreliance rate (OR)	Equation (4)	Decision logs with adjudicated outcomes	Trigger for cognitive forcing
Underreliance rate (UR)	Equation (5)	Decision logs with adjudicated outcomes	Trigger for feedback and training
Governance risk index (G)	Equation (6)	Program risk register	Prioritize Manage actions
Complementarity gain (ΔC)	Equation (7)	Accuracy of human, AI, and team	Justify AI deployment per phase

Note: Metrics combine attitudinal and behavioral indicators, consistent with recommended practice for trust evaluation.

5. CONCLUSION

This systematic review concluded that to ensure successful human-AI collaboration in engineering design review and program governance, the level of dependence of the reviewers and the level of reliability of the system must match, instead of maximal trust. Explanations are found to have a general acceptance building effect on acceptance, while confidence displays, independent first judgments and forcing functions are found to improve calibration at a moderate usability cost according to the synthesis of 28 sources. Governance frameworks come together on the concepts of transparency, accountability and lifecycle monitoring, and risk-management processes become a thematic pathway for incorporating these concepts into program gates through staged internal audit. The main contribution is a phase-specific architecture that connects levels of automation, seven quantitative calibration and risk metrics throughout the design process, and records of audit to document decisions. More effort is needed to test the proposed metrics in real design sessions, to determine threshold values of the calibration error and overreliance for each gate, and to explore the complementary performance as reviewers gain more experience using AI.

REFERENCES

[1] S. Amershi *et al.*, "Guidelines for Human-AI Interaction," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, in CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–13. doi: 10.1145/3290605.3300233.

[2] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, 2020, doi: https://doi.org/10.1016/j.inffus.2019.12.012.

[3] G. Bansal *et al.*, "Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, in CHI '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–16. doi: 10.1145/3411764.3445717.

[4] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW1, Apr. 2021, doi: 10.1145/3449287.

[5] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *J. Exp. Psychol. Gen.*, vol. 144, no. 1, pp. 114–126, 2015, doi: https://psycnet.apa.org/doi/10.1037/xge0000033.

[6] U. Ehsan, Q. V. Liao, M. Muller, M. O. Riedl, and J. D. Weisz, "Expanding Explainability: Towards Social Transparency in AI systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, in CHI '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 1–19. doi: 10.1145/3411764.3445188.

[7] M. R. Endsley, "From Here to Autonomy: Lessons Learned From Human–Automation Research," *Hum. Factors*, vol.

59, no. 1, pp. 5–27, Feb. 2017, doi: 10.1177/0018720816681350.

[8] L. Floridi *et al.*, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds Mach.*, vol. 28, no. 4, pp. 689–707, 2018, doi: 10.1007/s11023-018-9482-5.

[9] E. Glikson and A. W. Woolley, "Human trust in artificial intelligence: Review of empirical research," *Acad. Manag. Ann.*, vol. 14, no. 2, pp. 627–660, 2020, doi: https://doi.org/10.5465/annals.2018.0057.

[10] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A Survey of Methods for Explaining Black Box Models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018, doi: 10.1145/3236009.

[11] K. A. Hoff and M. Bashir, "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Hum. Factors*, vol. 57, no. 3, pp. 407–434, May 2015, doi: 10.1177/0018720814547570.

[12] A. Jacovi, A. Marasović, T. Miller, and Y. Goldberg, "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 624–635. doi: 10.1145/3442188.3445923.

[13] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019, doi: 10.1038/s42256-019-0088-2.

[14] V. Lai, C. Chen, A. Smith-Renner, Q. V. Liao, and C. Tan, "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, in FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, pp. 1369–1385. doi: 10.1145/3593013.3594087.

[15] J. D. Lee and K. A. See, "Trust in Automation: Designing for Appropriate Reliance," *Hum. Factors*, vol. 46, no. 1, pp. 50–80, Mar. 2004, doi: 10.1518/hfes.46.1.50_30392.

[16] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organ. Behav. Hum. Decis. Process.*, vol. 151, pp. 90–103, 2019, doi: https://doi.org/10.1016/j.obhdp.2018.12.005.

[17] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artif. Intell.*, vol. 267, pp. 1–38, 2019, doi: https://doi.org/10.1016/j.artint.2018.07.007.

[18] J. Mökander, J. Morley, M. Taddeo, and L. Floridi, "Ethics-Based Auditing of Automated Decision-Making Systems: Nature, Scope, and Limitations: J. Mökander et al.," *Sci. Eng. Ethics*, vol. 27, no. 4, p. 44, 2021, doi: https://doi.org/10.1007/s11948-021-00319-4.

[19] G. Pahl, W. Beitz, J. Feldhusen, and K.-H. Grote, *Engineering Design: A Systematic Approach*, 3rd ed. Springer, 2007. doi: 10.1007/978-1-84628-319-2.

[20] R. Parasuraman and V. Riley, "Humans and Automation: Use, Misuse, Disuse, Abuse," *Hum. Factors*, vol. 39, no. 2, pp. 230–253, Jun. 1997, doi: 10.1518/001872097778543886.

[21] I. Rahwan *et al.*, "Machine behaviour," *Nature*, vol. 568, no. 7753, pp. 477–486, 2019, doi: 10.1038/s41586-019-1138-y.

[22] S. Raisch and S. Krakowski, "Artificial intelligence and management: The automation–augmentation paradox," *Acad. Manag. Rev.*, vol. 46, no. 1, pp. 192–210, 2021, doi: https://doi.org/10.5465/amr.2018.0072.

[23] I. D. Raji *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44. doi: https://doi.org/10.1145/3351095.3372873.

[24] B. Shneiderman, "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy," *Int. J. Human–Computer Interact.*, vol. 36, no. 6, pp. 495–504, Apr. 2020, doi: 10.1080/10447318.2020.1741118.

[25] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, 2023. doi: https://doi.org/10.6028/NIST.AI.100-1.

[26] O. Vereschak, G. Bailly, and B. Caramiaux, "How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, no. CSCW2, pp. 1–39, 2021, doi: 10.1145/3476068.

[27] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the Effect of Accuracy on Trust in Machine Learning Models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, in CHI '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–12. doi: 10.1145/3290605.3300509.

[28] Y. Zhang, Q. V. Liao, and R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, in FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 295–305. doi: 10.1145/3351095.3372852.