International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Trust Calibration and User Control in AI-Assisted Interfaces: Literature Review and Framework

Aryan Manoj Jadhav

Student, Department of Computer Science and Engineering, The LNM Institute of Information Technology, Jaipur, Rajasthan, India

Abstract

People should rely on artificial intelligence (AI) when it is right and on themselves when it is wrong. Interface research has pursued this calibration mainly through cues such as explanations and confidence scores, and has often measured only one of the two ways reliance fails. This integrative review synthesizes 50 sources on trust in automation, calibration measurement, interface interventions, user control, large language models (LLMs) and oversight regulation, and codes each empirical result by whether it affects over-reliance, under-reliance or both. Cues that make a specific output cheaper to check reduce over-reliance, and where under-reliance was also measured they did not raise it. Cues that add authority, such as fluent explanations, raise reliance whether or not the output is correct. User control is common but rarely tested against calibration. Control held by users consistently increased use of an AI in the studies reviewed, and it sharpened reliance when it changed what the system did. Among the studies reviewed, only control moved adaptively by the system reduced over-reliance without raising under-reliance, in a single study, and no user-held design has been tested on both error types, including with LLMs. Oversight rules such as Article 14 of the European Union's AI Act give people capacities to override and stop AI systems without requiring evidence that they use those capacities well. This review proposes a framework of seven components, three kinds of control crossed with their timing, a governance frame and seven propositions marked by the strength of their evidence, and it sets out a research agenda and guidance for designers and regulators.

Keywords: Trust Calibration, Appropriate Reliance, User Control, Human Oversight, Large Language Models, Human-AI Interaction

1. Introduction

Artificial intelligence (AI) assistance now sits inside consequential decisions, from screening and forecasting to search, writing and public administration. Its value depends less on how much people trust it than on whether that trust fits the system's actual capability. Lee and See called this fit calibration [1]. Its behavioral counterpart, appropriate reliance, means relying on the AI when it is right and on oneself when it is wrong [2]. Two errors break it: over-reliance accepts wrong outputs, and under-reliance rejects right ones.

The stakes are social as well as individual. When people over-rely, an AI's errors pass into decisions about loans, benefits, diagnoses and sentences under the appearance of human review. When they under-rely,

the benefits that justified deploying the system are lost, and the cost of checking falls on the frontline workers who must second-guess it. Regulators have answered with requirements for human oversight, but policies that assume a human will catch the machine's mistakes can legitimize flawed systems and shift blame onto operators who were never in a position to catch them [3]. Whether people can calibrate their reliance is therefore a question about accountability, labor and institutional design, as well as interface usability.

Interface research has pursued calibration mainly through cues: confidence scores, accuracy statements, explanations and onboarding [4], [5]. The results are mixed. Explanations often raise reliance whether or not the recommendation is correct [6], [7], and a meta-analysis of nine articles found no added effect of explanations on accuracy beyond the AI's advice itself [8]. Interventions that cut over-reliance frequently raise under-reliance, a trade-off that stays hidden whenever studies score only one error [9].

A second lever runs through the same literature under other names: user control. People choose whether to see advice [10], adjust an algorithm's forecasts [11], or write rules for where a model may act alone [12]. Regulation leans on control too: the European Union (EU) AI Act requires, in Article 14(4)(d) and (e), that people overseeing high-risk systems be able to override, reverse or stop them [13]. Yet control is rarely treated as a way to calibrate reliance. It is usually studied as a route to acceptance, or simply assumed to work.

1.1 The Gap

Existing reviews leave this connection open. Mehrotra et al. systematically reviewed interventions for appropriate trust, but their taxonomy has no category for user control; where control appears, it is grouped with transparency [4]. Their corpus also ends in mid-2022, before most studies of large language models (LLMs). Wischnewski et al. mapped how calibration is measured and classified interventions by timing and adaptivity, not by who holds control [5]. A recent review of reliance measures likewise finds the constructs fragmented and names users' ability to contest or change AI advice as under-explored [14].

The closest competing framework is Araujo's integrative account of trust calibration in the everyday use of conversational generative AI [15]. It links user characteristics, learned experience, system design, personalization, situational risk and institutional conditions, and it treats the object of trust as a network that includes the response, the model, the provider and the cited sources. The present review differs in three ways. It works at the level of the interface rather than everyday use. It separates three kinds of user control and when they act. And it scores every claim against both error types. The two accounts meet at governance, where Araujo's institutional conditions correspond to the governance frame proposed here.

1.2 Research Questions

Four research questions (RQ1 to RQ4) guide the review.

- RQ1. How has calibration been defined and measured, and which measures separate over-reliance from under-reliance?
- RQ2. What do interface cues do to each error type?
- RQ3. What do different kinds of user control do to each error type, and has any been tested against calibration?
- RQ4. How do these findings carry over to LLM outputs, and what do oversight frameworks assume about the people who oversee AI?

1.3 Contributions

The review makes four contributions. First, it synthesizes 50 sources by error type rather than by intervention, which exposes trade-offs that single-error studies hide. Second, it distinguishes user-held,

designer-fixed and system-moved control, crossed with whether control is set before or exercised during use. Third, it builds a framework of seven components (C1 to C7), a governance frame and seven propositions, each marked by the strength of its evidence. Fourth, it sets out specific studies for future research and concrete guidance for designers and regulators.

2. Key Concepts

2.1 Trust, Reliance and Trustworthiness

Three constructs are often run together and must be kept apart. Trust is a willingness to be vulnerable to another party's actions, based on the expectation that it will perform an action important to the trustor, irrespective of the ability to monitor or control it; that expectation rests on perceived ability, benevolence and integrity [16]. Lee and See carried this into automation as the attitude that an agent will help achieve one's goals in a situation of uncertainty and vulnerability [1]. Reliance is behavior: accepting, editing or rejecting the AI's output. Trustworthiness is a property of the system: how often, and where, its outputs are right. Trust guides reliance but does not determine it, because workload, effort and self-confidence intervene [1], [17]. Work on trust in AI specifically has been reviewed across disciplines [18]; this review concentrates on the interface and on the behavior that trust is meant to explain.

2.2 Calibration and Appropriate Reliance

Calibration is the correspondence between trust and trustworthiness [1]. Two refinements matter for interfaces. Resolution is how well trust separates situations where the AI is capable from those where it is not. Specificity is how finely trust is tied to particular functions or moments rather than to the system as a whole. Good calibration therefore needs trust that varies output by output, which a single well-chosen average does not provide.

Appropriate reliance is the behavioral counterpart, and it has two dimensions that can move independently [2].

- **Over-reliance** is following the AI when it is wrong. Automation bias is its best-known form, including omission errors (missing what the aid did not flag) and commission errors (following a wrong recommendation) [19]. Parasuraman and Riley called this failure misuse [20].
- **Under-reliance** is rejecting the AI when it is right. Algorithm aversion, the reluctance to use a better but imperfect algorithm, is its best-known form [21], [22], and it corresponds to what Parasuraman and Riley called disuse.

Schemmer et al. operationalized both on cases where a person's first answer disagrees with the AI [2]. Relative AI reliance (RAIR) is the share of cases with a wrong first answer and correct AI advice in which the person switches to the AI. Relative self-reliance (RSR) is the share of cases with a correct first answer and wrong AI advice in which the person keeps their own answer. In this review, a study scores both error types when it reports a measure of each.

2.3 Warranted Trust

Jacovi et al. add a causal test [23]. Trust is warranted only when it is caused by the system's trustworthiness with respect to an explicit or implicit contract. Trust raised by surface cues such as reputation, fluency or human-likeness is unwarranted for a contract about accuracy, even when it happens to be well placed. Warrant can be checked by manipulating trustworthiness and seeing whether trust follows.

2.4 User Control

In this review, control (component C7) is the allocation of decision authority over an AI's output between the person and the system: who decides whether the output is seen, how much it counts, and whether it

can be inspected, edited, overridden or rejected. Cues, by contrast, only inform. Parasuraman et al. treated control as allocation along a scale of automation for each stage of information processing [24]. Shneiderman separated human control from computer automation as two axes, so a highly automated system can still leave control with people [25]. The component keeps the name user control because it concerns how much authority the user holds, but that allocation can be set by different parties. Three kinds of control are distinguished by who sets it.

- **User-held:** the user chooses, for example by adjusting a forecast, requesting advice or writing delegation rules.
- **Designer-fixed:** the designer sets it in advance for everyone, for example a decide-first step on every item or a limit on how far a forecast may be adjusted.
- **System-moved:** the system shifts it adaptively, for example by deciding who answers first from a model of the user's likely correctness.

Control also differs in timing, meaning when it acts on the person. Some control is set before use, and some is exercised during use. This mirrors the distinction Wischnewski et al. draw between interventions delivered before or after use and those delivered during it [5]. The two axes are independent: a designer fixes a rule in advance, yet the rule may act once, as a bound within which the user then adjusts, or on every item during use, as a decide-first step does. Felt control must also be separated from actual control. A setting that changes how in control people feel, but not what the system does, can raise acceptance without changing outcomes [26].

Trust and control raise a conceptual puzzle. The definition of Mayer et al. places trust where monitoring and control are absent, so control might seem to replace trust rather than calibrate it. Castelfranchi and Falcone resolve the puzzle by distinguishing trust in the agent itself from the overall trust needed to rely on it [27]. Control works against the first, because relying on control signals that trust in the agent alone does not suffice, but it completes the second by covering part of what would otherwise rest on the agent. They also separate control that prevents errors from control that corrects them afterwards. Two routes follow for this review. Control can limit exposure, so that less of the decision to rely has to rest on trust in the AI. And control can add evidence, because inspecting, adjusting or delegating shows how the system behaves, as the exploratory use described by Lee and See does [1]. On this view calibration concerns the whole decision to rely, trust and control together, measured against the AI's actual capability.

3. Method

3.1 Review Type

This is an integrative review with purposive selection. Integrative reviews critique and synthesize a literature to generate new frameworks and perspectives, and they suit topics where relevant work is spread across fields that use different terms [28], [29]. That fits here: user control appears as algorithm aversion in judgment research, as cognitive forcing and delegation in human-computer interaction, and as human oversight in law and policy.

3.2 Source Selection

The corpus began from a starter list of foundational and frequently cited works on trust in automation. It was extended by backward and forward snowballing from three reviews: Mehrotra et al. [4], Wischnewski et al. [5] and Raees and Papangelis [14]. Sources were replaced as the scope sharpened, mainly in the LLM group. The corpus was frozen at 50 sources: 48 papers and two policy texts.

Sources were included if they met all of the following conditions.

- They address trust, reliance or control in human decisions supported by AI or automation.
- They contribute one of the following: a foundational construct, a measure of calibration or reliance, an interface intervention with a behavioral reliance measure, a control design, a study of reliance on LLM outputs, or an oversight framework.
- They are published in English in a peer-reviewed venue, as a technical report from a recognized research body, or as a regulation or standard.

No database search was logged, so no screening counts are reported and no PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram is given. Table 1 describes the composition of the corpus instead.

Table 1: Composition of the Corpus

Group	Sources	What the group contributes	Read in full
A. Foundations of trust in automation	9	Definitions, failure modes and the control axis	4
B. Defining and measuring calibration	7	Measures of calibration and appropriate reliance	6
C. Interface interventions	14	Cue and effort designs with behavioral reliance measures	12
D. Control and algorithm aversion	7	Control designs and design frameworks	4
E. Large language models	7	Reliance on fluent, open-ended outputs	5
F. Oversight and regulation	6	What governance assumes about overseers	3
Total	**50**		**34**

Further sources were used for positioning and for checking specific claims, and they are cited but not counted among the 50: Araujo [15], Raees and Papangelis [14], Castelfranchi and Falcone [27], McGrath et al. [30], Scharowski et al. [31] and Spain et al. [32].

3.3 Extraction and Synthesis

Each source was read into fixed fields: reference, type and basis, sample, task, AI accuracy, manipulation, trust measure, reliance measure, result, boundary conditions, mapping to the framework, and limits for AI interfaces. Policy texts were read into adapted fields: what each requires, what it assumes about people, and what the evidence says.

Three codes were applied to every empirical study. The first maps it to the seven framework components (C1 to C7). The second records the kind of any control it studies (user-held, designer-fixed or system-moved). The third records whether it reports over-reliance, under-reliance or both. Synthesis then followed a concept-centric logic [29]: findings are grouped by construct and error type, and contradictions are explained rather than listed. Appendix A summarizes every source and its codes.

3.4 Basis of Claims

Thirty-four sources were read in full. For two of them the study was read as reported in full in the first author's doctoral dissertation rather than in the published paper: Bansal et al. [33] and Okamura and Yamada [34]. For one, the full technical report underlying the published paper was read (Jian et al. [35]). The remaining 16 were read from abstracts, excerpts, report summaries or the wording of framework subcategories. Two of these, Vereschak et al. [36] and Dzindolet et al. [37], could not be obtained in full and are cited only for findings stated in their abstracts. Every proposition in Section 5.4 rests on at least one source read in full, and sources read only in part are cited for background or for claims stated in their abstracts. Appendix A marks the basis of each source. A single reviewer selected, read and coded all sources, so no inter-rater agreement is reported.

3.5 Use of a Large Language Model and Other Tools

A large language model, Claude (Anthropic), was used under the author's direction during this review for the following tasks: searching for and retrieving sources; producing summaries of sources for quicker retrieval of concepts; checking bibliographic details against publisher and DOI records; proofreading and editing the manuscript to correct any errors.

4. Findings

4.1 Measuring Calibration

Most studies measure trust rather than calibration, and the measures that can separate the two errors all need a known right answer. Both facts limit what the rest of the literature can show.

Self-report captures trust imperfectly and in isolation. The checklist of Jian et al. remains the most used scale [35], yet it was not comprehensively validated until two 2025 papers, each presenting itself as the first [30], [31]. Factor analyses repeatedly find trust and distrust to be two related factors rather than opposite poles of one [31], [32], so the scale should be scored as two subscales. Custom single items are more common still [38]. A survey of empirical methods in AI-assisted decision making found that experimental protocols rarely build in a definition of trust, so trust findings are easily overclaimed and hard to compare [36]. A validated three-item short form now allows repeated measurement during a task [30]. Kohn et al. recommend pairing self-report with behavior and creating real vulnerability, since trust with nothing at stake is not trust [38].

Attitude and behavior part ways. Across the corpus, self-reported trust moved against appropriate reliance or not at all. Higher trust went with more over-reliance [10]. Explanations changed reliance without changing trust [2]. Uncertainty expressions lowered trust intentions but not trust beliefs [39]. A tutorial that changed reliance left trust unmoved, and propensity to trust predicted trust but not reliance [40]. Trust ratings alone therefore cannot stand in for calibration.

Calibration measures compare trust with capability. Wischnewski et al. found three approaches: relative (trust compared across capability levels), correlational (trust tracking capability over time) and behavioral (reliance compared with a normative rate) [5]. With LLMs, Steyvers et al. scored people's stated confidence in each answer against its correctness, using expected calibration error and the area under the curve (AUC) for telling right answers from wrong ones [41]. The resolution and specificity described by Lee and See add that trust should vary output by output, which aggregate measures miss [1].

Separating the two errors requires an answer key. Four families of measure can tell over-reliance from under-reliance.

- Agreement or accuracy split by whether the AI was right [7], [39], [42].

- RAIR and RSR on cases where the person's first answer disagrees with the AI [2], [9], [40].
- Continuous measures for estimates: deviation from the model [11], [43] and weight of advice [44], which Bo et al. also used for estimation [9].
- Precision of the region a person delegates to the model [12].

All four need a correct answer or a computable optimum. Generative outputs, such as an essay, a plan or a block of code, usually have neither.

Most intervention studies score only one error. Among the 14 interface studies in this corpus, only four measured under-reliance, and only one reported RAIR and RSR together [40]. Trade-offs between the errors were therefore invisible in most of the evidence base. The LLM studies do better: three of seven split agreement or accuracy by whether the AI was right [7], [39], [42], one reported RAIR and RSR [9], and one scored the calibration and discrimination of people's confidence [41].

Warrant is almost never tested. Jacovi et al. define warranted trust as trust caused by trustworthiness, which requires manipulating trustworthiness and watching trust follow [23]. Jian et al. proposed the same validation for their scale: degrade a system and check that ratings and reliance fall together [35]. Few studies in the corpus do this within one design.

Control actions are rarely scored. Only two studies scored control actions against the AI's correctness, and each covered one side. Lai et al. measured the precision of the regions people delegated to a model, which captures over-reliance [12]. Dietvorst et al. measured whether people who could adjust a better model's forecasts used it more and erred less, which captures disuse [11]. No study scored control actions on both error types. This review proposes treating overrides, edits, delegation choices and setting changes as behavior in their own right, scored with the same two-error metrics, rather than as proxies for trust.

Answer to RQ1. Calibration has been defined as the fit of trust to capability, but it is mostly measured as trust alone. Only answer-key measures separate the two errors, and most interface studies have used just one side.

4.2 What Interface Cues Do to Each Error

Cues do not have a single effect. The deciding factor is whether a cue adds authority to an output, warns about it, or makes it cheaper to check. Table 2 sorts the cues in the corpus by that function.

Table 2: Cue Functions and Their Effect on Each Error Type

Cue function	Examples	Over-reliance	Under-reliance	Main evidence
Adds authority	Explanations, longer or more fluent answers, visible model coefficients	Rises	Falls	Bansal et al. 2021 [6]; Kim et al. 2025 [7]; Si et al. [42]; Steyvers et al. [41]; Poursabzi-Sangdeh et al. [43]
Enables a check	Sources, retrieved passages, salient error marks, explanations quicker to verify than the task	Falls	Falls or holds in the two studies that measured it	Vasconcelos et al. [45]; Kim et al. 2025 [7]; Si et al. [42]

Cue function	Examples	Over-reliance	Under-reliance	Main evidence
Warns	First-person hedges, a reliance disclaimer, contrastive arguments for and against	Falls	Holds (hedges, disclaimer) or rises (contrastive arguments)	Kim et al. 2024 [39]; Bo et al. [9]; Si et al. [42]
Scores each output	Confidence values	Resolution improves	Resolution improves	Zhang et al. [46]; Bansal et al. 2021 [6]
Sets expectations	Accuracy statements, onboarding, tutorials	Mixed	Mixed	Yin et al. [47]; Kocielnik et al. [26]; Cai et al. [48]; He et al. [40]

Authority cues raise reliance whether the AI is right or wrong. Explanations raised acceptance of correct and incorrect recommendations alike, and confidence alone did as well as any explanation [6]. With LLM answers, explanations raised agreement with correct answers from 67% to 78% and with incorrect ones from 78% to 83% [7]. In fact-checking, LLM explanations raised accuracy to 87% when they were right but cut it to 35% when they were wrong, below the 49% that people reached without help [42]. Longer explanations raised people's confidence without improving their ability to tell right answers from wrong ones [41]. Transparency about a model's internals can also hide its errors. In three of four experiments, people shown a regression model's coefficients were less likely than people given a black box to correct its large errors on unusual inputs; the authors attribute this to information overload or to anchoring on the model, and the effect disappeared when people were told to watch for unusual cases [43]. Across studies, a meta-analysis of nine articles found no added effect of explanations on accuracy beyond the AI's advice (standardized mean difference 0.07, not significant), although that comparison rested on only 11 studies [8].

The same explanation helps when it makes checking cheaper than solving. In a maze task, explanations cut over-reliance only when the maze was hard enough that checking the explanation was easier than solving the problem, and a salient indicator of errors brought over-reliance close to zero [45]. Feature-contribution explanations reduced under-reliance only on a task participants knew well [49], which suggests that a check works only when people have the knowledge to make it. In another study, explanations raised reliance on correct advice without lowering resistance to wrong advice [2]. The contradiction across explanation studies is therefore not noise. Explanations help when they turn into a quick check and hurt when they stand in for one.

Verification affordances help when people use them. Clickable sources raised agreement with correct LLM answers (67% to 73%) and lowered it with incorrect ones (78% to 68%) [7]. In fact-checking of adversarial claims, explanations and retrieved passages produced similar accuracy (74% and 73%, against 59% without help), but reading passages took more than twice as long, about 2.5 minutes per claim against 1 [42]. Adding passages to explanations left accuracy unchanged when the explanation was right (87% in both conditions) and did not significantly improve it when the explanation was wrong (43% against 35%). Use is the limit: 189 of 308 participants never clicked a source, although those who did were more accurate [7].

Warnings lower reliance mostly where it should fall, but arguing both sides does not. First-person hedges such as "I'm not sure, but" raised accuracy on wrong answers with little cost on right ones, even though they appeared on a random half of answers [39]. A static disclaimer was the only intervention to improve both error types on a reasoning task [9]. Both worked as general cautions rather than as information about specific outputs, and each rests on one task. Contrastive explanations, which set out the case for and against a claim, raised accuracy when the LLM's explanation was wrong (56% against 35%) but lowered it when the explanation was right (73% against 87%) [42].

Per-output confidence improves resolution but did not improve accuracy. Confidence scores led people to follow the AI more when it was confident, yet joint accuracy did not rise, because the AI's and the people's errors fell on the same cases [46]. Resolution pays off only when the person knows something the AI does not. Error boundaries that are simple and consistent help people build the mental model that makes this possible. In a game in which people chose for each item whether to accept the AI's recommendation, team performance fell as the error boundary became more complex or the task gained features, and when errors were stochastic only about half of the workers converged on the optimal choice [33].

Self-confidence decides which way reliance errs. People who overestimated their own ability under-relied on the AI (RAIR 0.27 against 0.58 for accurate self-assessors) [40]. After a tutorial that corrected self-assessment, over-estimators improved slightly on both measures (RAIR 0.24 to 0.36 and RSR 0.29 to 0.52, neither significant), while under-estimators became worse on both (RAIR 0.68 to 0.45 and RSR 0.72 to 0.30). Both changes are before-after comparisons within groups formed after the first batch of tasks, so regression to the mean may explain part of them. Lay people weighted identical advice more when it was labeled as coming from an algorithm, but national security professionals discounted all advice, relied less on the algorithm than lay people did and were less accurate as a result [44]. Higher stakes reduced over-reliance, while harder tasks increased it [45].

Answer to RQ2. Cues that add authority push the two errors in opposite directions. Cues that lower the effort of verifying a specific output reduce over-reliance, and in the two studies that also measured under-reliance they did not raise it. General warnings lower reliance mainly on wrong outputs, but the evidence rests on two studies, and contrastive arguments trade one error for the other.

4.3 What User Control Does to Each Error

Control appears in eleven designs across seven studies (Table 3). Only two of them, both system-moved, were scored on both error types. Among the studies reviewed, no user-held design was scored on both in one study, although its two halves exist in different studies.

Table 3: Control Designs Scored Against the AI's Correctness

Design	Kind	Timing	Over-reliance scored	Under-reliance scored	Result
Expectation slider (Kocielnik et al. [26])	User-held	Before use	No	No	Acceptance rose; the setting never reached the system used
Delegation rules (Lai et al. [12])	User-held	Before use	Yes (precision)	Loosely (coverage)	Beat the model alone; complementary on the model's own data only

Design	Kind	Timing	Over-reliance scored	Under-reliance scored	Result
Bounded adjustment (Dietvorst et al. 2018 [11])*	User-held	During use	No	At task level	Disuse fell; adjusters used the model more and erred less
Advice on demand (Buçinca et al. [10])	User-held	During use	Yes	No	Reported with the forcing designs, which cut over-reliance
Blind delegation (Zhang et al. [46])	User-held	During use	Partly	Partly	Delegation tracked confidence; accuracy unchanged
Per-trial automatic or manual choice (Okamura and Yamada [34])	User-held	During use	Yes (d′)	No	The measured behavior; see adaptive cues
Decide first (Buçinca et al. [10])	Designer-fixed	During use	Yes	No	Over-reliance fell; the forcing designs were trusted and liked least
Wait (Buçinca et al. [10])	Designer-fixed	During use	Yes	No	Over-reliance fell; the forcing designs did not differ from each other
Adaptive workflow (Ma et al. [50])	System-moved	During use	Yes	Yes	Over-reliance fell without under-reliance rising
Hidden advice (Ma et al. [50])	System-moved	During use	Yes	Yes	Over-reliance fell; more correct low-confidence advice rejected
Adaptive cues (Okamura and Yamada [34])	System-moved	During use	Yes (d′)	No (not tested)	Only a verbal cue raised d′ (0.92 against 0.38); accuracy not significantly higher

*The adjustment itself is user-held and exercised during use; the limit on how far it may go is designer-fixed and set before use (Table 5).

User-held control raises use, almost regardless of how much control it gives. When people could not change a statistical model's forecasts, only 32% (in a laboratory study) and 47% (online) chose to use it; when they could adjust them, 68% to 76% did [11]. Uptake barely moved between 10, 5 and 2 percentiles of leeway, and people used about half the leeway they had. Because the model was better than people, more use meant more accurate forecasts, even though the authors note that human adjustments tend to degrade algorithmic forecasts. People who had adjusted the model were more satisfied with the process and, given a later free choice, more often used the model without any adjustment (30% against 12%).

Adjustment within a bound also outperformed free use of the model: people who chose to adjust by at most 10 percentiles erred less than those who used the model freely (average error 17.90 against 20.25 percentiles). This is a remedy for disuse: seeing a model err drives people away from it, even when they have seen it outperform a human [21].

Felt control can stand in for actual control. An expectation slider in an AI scheduling assistant raised acceptance, and 87% of participants moved it [26]. Yet the version each participant used was assigned at random, so the slider never reached the system. Felt control was measured only in a study without a task, so the claim that acceptance rose through felt control is plausible rather than shown. Control that never touches the system is, at best, a lever on trust.

Control that reaches the system can calibrate. In conditional delegation, crowd workers wrote keyword rules for where a toxicity classifier could flag comments on its own [12]. Their rules raised precision above the model alone on both datasets. On the model's own data they also beat rules written without the model (average precision about 0.94 against 0.79), a complementary result. On new data, where the model was weak, they did not, and a list of the model's most influential words misled people into imprecise rules. To write a rule, participants first inspected the model's behavior on comments they searched for, so exercising control also gathered evidence.

System-moved allocation calibrated reliance. Ma et al. modeled each person's likely correctness and let the system decide who answered first [50]. That adaptive workflow cut agreement with wrong advice without lowering agreement with correct advice, and designs built on these correctness estimates outperformed both the AI alone and people alone. Hiding advice where people were likely right also cut over-reliance, but led them to reject more correct advice where the AI was unsure. In a drone inspection task in which people chose automatic or manual inspection on each trial, cues triggered by detected over-trust raised sensitivity (the signal detection index d′) only when the cue was verbal (0.92 against 0.38 without a cue), did not significantly raise accuracy, and were not tested against under-trust [34].

Designer-fixed forcing worked but was disliked. Making people decide before seeing advice, or making them wait for it, cut over-reliance, with benefits concentrated among people high in need for cognition [10]. The designs that worked were trusted and liked least. That trade-off rests on this one study: Ma et al. found no trust differences between designs [50], and Vasconcelos et al. did not analyze trust [45].

Design frameworks name control but do not test it. Guidelines for human-AI interaction recommend efficient correction, dismissal and global controls, but they were validated with design practitioners for clarity and relevance, not for effects on reliance [51]. Shneiderman separates control from automation as two axes and argues that high levels of both yield reliable, safe and trustworthy systems, without testing the claim [25]. A systematic review of algorithm aversion names decision autonomy as one of its main themes [22].

Timing may matter, though no study has varied it directly. The user-held design that cut over-reliance was set before use [12], while the one that cut disuse was exercised during use [11].

Answer to RQ3. User-held control consistently raised use and, when it changed what the system did, raised the precision of reliance. Among the studies reviewed, only system-moved control has been shown to reduce over-reliance without raising under-reliance in one study, and no user-held design has been tested on both error types.

4.4 Large Language Models

The two errors carry over to LLMs, but checking gets harder. LLMs explain by default, fluently and at length, so the cue that adds authority arrives with every answer.

Fluency drives over-reliance. In a think-aloud study with 16 participants, people agreed with 9 of 13 wrong first answers from ChatGPT, and asking follow-up questions did not raise their accuracy [7]. In the controlled follow-up, explanations raised reliance on both correct and incorrect answers. Across three models and 301 participants, people judged LLM answers more likely to be right than they were: with default explanations, the calibration error of human confidence was 0.17 to 0.29, against 0.10 to 0.15 for the models' own confidence, and people's ability to tell right answers from wrong ones was only modestly above chance (AUC about 0.59 to 0.60, against 0.75 to 0.78 for the models) [41]. Longer explanations raised confidence without improving that discrimination. In fact-checking, people over-relied on LLM explanations when they were wrong: accuracy on those claims fell to 35%, against 49% without help [42]. Cheap checks and visible flaws reduce it. Sources lowered agreement with wrong answers and raised it with right ones [7]. When an explanation happened to contradict itself, agreement with the wrong answer was lower (70% against 83%); because such inconsistencies occurred naturally in 3 of 12 wrong responses rather than being manipulated, this comparison is observational. Contrastive explanations cut over-reliance on wrong explanations but raised under-reliance on right ones, and they were no more accurate than retrieved passages; showing passages and explanations together added nothing over passages alone [42].

Most friction lowers reliance across the board. On logical-reasoning questions, a Reliance Disclaimer, Uncertainty Highlighting of tokens and an Implicit Answer all raised resistance to bad advice [9]. Only the disclaimer also kept reliance on good advice, lifting RSR from 0.44 to 0.56 and the ratio of appropriate reliance to over-reliance from 1.17 to 1.62. On estimation, the Implicit Answer raised self-reliance from 0.15 to 0.43 but cut reliance on good advice from 0.91 to 0.67. Every intervention lowered perceived helpfulness, and on the reasoning task people's confidence rose more after wrong reliance decisions than after right ones.

Uncertainty language changes confidence; whether it changes decisions depends on the task. First-person hedges raised accuracy on wrong answers at little cost on right ones, even when placed at random [39]. Steyvers et al. varied how sure each explanation sounded and found that people's confidence followed the wording closely [41]. When they simulated an assistant that matched its wording to its own confidence, by selecting the response whose explanation style matched that confidence, calibration error fell (for example from 0.26 to 0.15) and discrimination rose (AUC from about 0.59 to 0.69). People's own answers, however, were no more accurate, consistent with their lacking the knowledge to correct the model when it signaled doubt. Calibrated language helps people know when to doubt, but it does not by itself help them do better.

Two-error measurement transfers, and so does the dependence on answer keys. Agreement or accuracy split by correctness [7], [39], [42] and RAIR and RSR [9] work for LLM answers to factual, logical and estimation questions. Every LLM study in the corpus kept an answer key; writing, coding and other open-ended outputs remain unmeasured.

User control is nearly absent from this work. The only user-held acts in these studies were clicking a source and asking a follow-up question. None of the LLM studies reviewed tests a user-held control design on reliance, and a targeted search in November 2026 found no counterexample. The reviews that frame this area share the gap. Passi and Vorvoreanu define the problem as over-reliance only [52]. Liao and Wortman Vaughan set out approaches to transparency for LLMs, all of them about what to disclose rather than what users can do [53].

Answer to RQ4, first part. The findings carry over: authority cues dominate because LLMs explain by default, and cheap checks remain the lever that reduces over-reliance without raising under-reliance. Calibrated uncertainty helps people judge answers but not improve on them. Control is what does not carry over: the user-held design that sharpened reliance with classifiers depends on input regions and labels that generative outputs lack.

4.5 Oversight and Governance

Regulation and governance frameworks give overseers capacities to understand, override and stop AI systems. None checks whether those capacities produce calibrated reliance.

Article 14 of the EU AI Act is the clearest case. It requires that high-risk systems be designed so that people can effectively oversee them, and Article 14(4) lists five capacities those people must be enabled to have [13]. Table 4 sets each against the conditions for effective oversight proposed by Sterz et al. [54] and against the evidence in Sections 4.1 to 4.4.

Table 4: Article 14(4) Capacities Against the Conditions for Effective Oversight and the Evidence

Capacity	Condition it serves (Sterz et al.)	What the evidence shows	How it could be tested
(a) Understand capacities and limitations; monitor operation	Epistemic access	Global cues set expectations (Kocielnik et al. [26]), but sustained monitoring decays (Bainbridge [55])	Trust compared with capability (Section 4.1)
(b) Remain aware of automation bias	Epistemic access and self-control	Training and instructions did not reliably prevent it (Parasuraman and Manzey [19]); one disclaimer helped on one task (Bo et al. [9])	Resistance to wrong outputs (RSR)
(c) Correctly interpret output	Epistemic access	Explanations raised reliance on wrong outputs (Bansal et al. 2021 [6]; Kim et al. 2025 [7]; Si et al. [42])	RAIR and RSR with and without the interpretation tool
(d) Decide not to use, or disregard, override or reverse output	Causal power and self-control	Control raised use regardless of its amount (Dietvorst et al. 2018 [11]); overrides in practice were often harmful (Green [3])	Accuracy of overrides on both error types, not the override rate
(e) Intervene or stop the system safely	Causal power and self-control	Takeover skills erode under automation (Bainbridge [55])	Takeover performance after long automated periods

Each capacity is necessary, but none is shown to be sufficient, and Article 14 does not ask for it to be shown.

People cannot reliably do what oversight policies assume. Green coded 41 policy documents on human oversight of government algorithms [3]. They prescribe three approaches: restricting "solely" automated

decisions (20 documents), emphasizing human discretion to override (14) and requiring "meaningful" human input (7). The evidence he reviews shows that people cannot reliably perform any of these functions. Judges overrode risk-assessment recommendations toward detention and did so more harshly for Black defendants, and automation bias persisted after training. His second finding is that oversight policies legitimize flawed systems and shift blame onto frontline staff. He proposes institutional oversight instead: agencies justify in writing that an algorithm is appropriate and that any human oversight is backed by evidence, including tests of human-algorithm collaboration before deployment, and the justification is approved democratically.

Effectiveness needs more than capacity. Sterz et al. define an effective overseer as one with sufficient causal power over the system, suitable epistemic access to the situation, self-control, and intentions fitting the role [54]. They read Article 14(4)(d) and (e) as securing causal power and self-control, (a) and (c) as securing epistemic access, and (b) as touching both; Article 26(2) adds competence, training and authority. Fitting intentions have little direct equivalent in the text. They also find the list a loose collection at very different levels of abstraction, from a concrete stop button to a vague duty to understand limitations; automation bias is singled out among many relevant biases; and understanding the system's inputs is not mentioned at all. They call for empirical studies to validate oversight mechanisms and for tools to audit whether oversight is effective. The evidence in Sections 4.2 and 4.4 adds a sharper point: the capacities Article 14 secures most clearly, epistemic access through interpretation tools and causal power through overrides, are exactly those that the reviewed studies show can raise reliance on wrong outputs or raise uptake regardless of quality.

The critics move responsibility up a level. Laux proposes principles that build distrust of overseers into institutions, among them justification, periodic mandates and collective decisions [56]. Santoni de Sio and van den Hoven require that outcomes be traceable to someone in the design and operation chain who understands the system [57]. The framework of the National Institute of Standards and Technology (NIST) asks organizations, in subcategories GOVERN 3.2 and MAP 3.5, to define roles for human-AI configurations and to define, assess and document oversight processes, without naming an outcome measure [58].

Answer to RQ4, second part. Oversight frameworks assume that a person given the right capacities will catch an AI's errors. The evidence reviewed here says those capacities are necessary but not sufficient. The instrument that could check them, two-error testing before deployment, is demanded by critics but required by none of the frameworks read.

5. A Framework for Calibration with User Control

The framework joins the trust loop of Lee and See with user control and places governance around both (Figure 1) [1]. Its core is a loop of five components. The AI's actual trustworthiness (C1) reaches people through interface cues (C2). Cues shape trust (C3), and trust guides reliance (C4). Using the AI produces evidence (C5) that updates trust. Moderators (C6), such as effort, stakes and self-confidence, act on trust and on the step from trust to reliance.

User control (C7) enters the loop by two routes, following Section 2.4. It can add evidence, when exercising control shows people how the system behaves [1], and it can limit exposure, when control reduces how much of the decision to rely must rest on trust in the AI [27]. Governance sits outside the loop because it works per deployment rather than per output, and because its evidence in this corpus is

conceptual. It connects to the loop in two places: it sets the allocation of control, and it can require two-error testing before deployment.

The loop's solid links rest on experimental evidence from Sections 4.1 to 4.4. The two control routes and the governance links are hypotheses, stated as propositions in Section 5.4 with their evidence marked.

Figure 1: The Calibration Loop with User Control and the Governance Frame

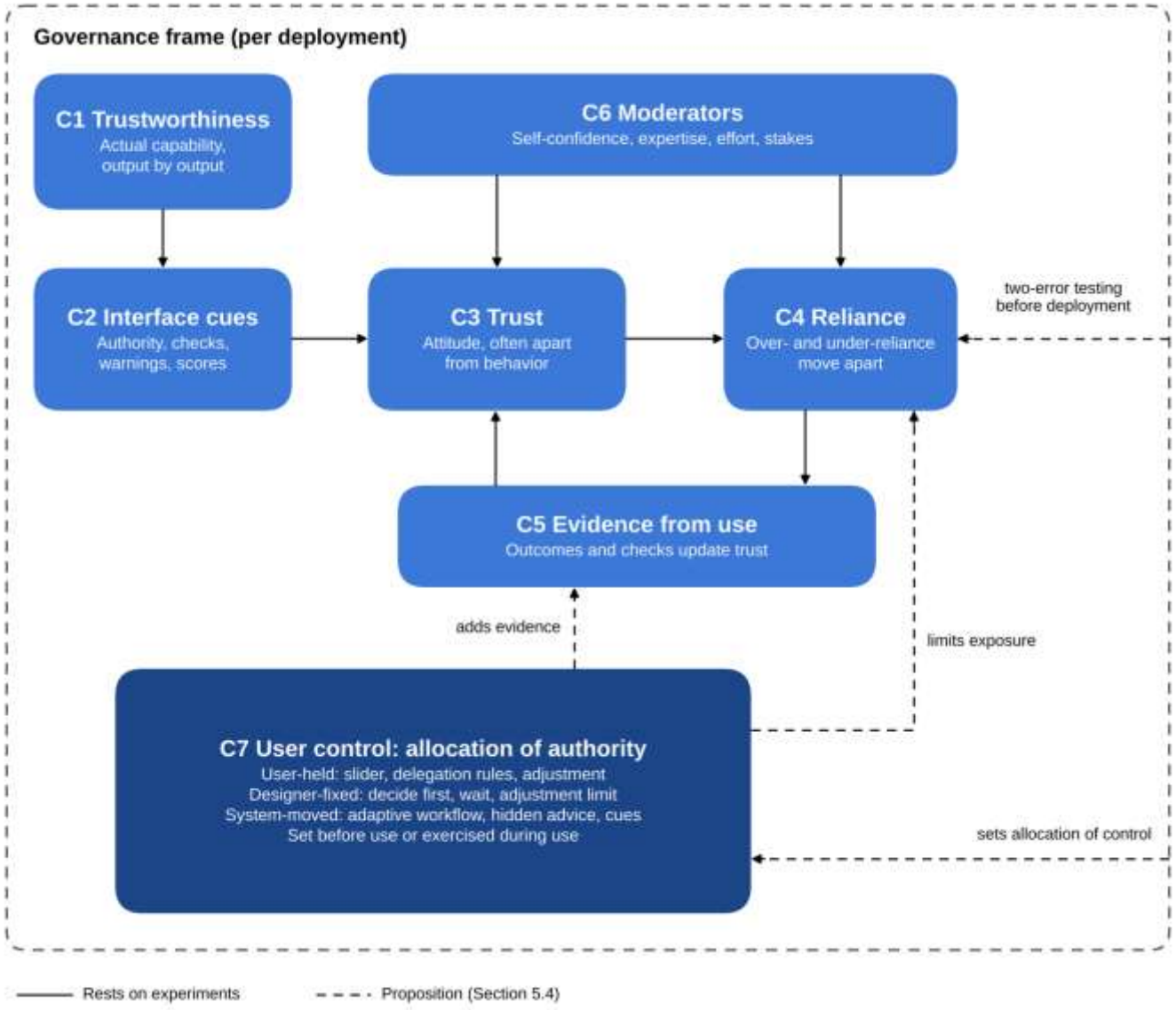


Figure 1 shows the calibration loop (C1 to C5) with its moderators (C6) and user control (C7). Solid links rest on experiments. Dashed links are propositions: control adds evidence and limits exposure, and governance acts on the loop from outside by setting the allocation of control and requiring two-error testing before deployment.

5.1 Components

- **C1 AI trustworthiness:** the system's actual capability, output by output. Error boundaries that are simple and consistent are easier to learn [33], and any cue about uncertainty is only as good as the model's own confidence [41].

- **C2 Interface cues:** what the interface shows about an output. Cues differ by function: adding authority, enabling a check, warning, scoring each output or setting expectations (Table 2).
- **C3 Trust:** the attitude, best measured as related but separate trust and distrust, and often out of step with behavior (Section 4.1).
- **C4 Reliance:** accepting, editing or rejecting an output. It fails in two directions that move independently and can be told apart only against an answer key.
- **C5 Evidence from use:** what people learn from outcomes and from checking. Seeing errors cuts trust in algorithms more than in people [21]; seeing an automated aid err lowers trust in it even when it is reliable, and explaining why errors occur restores trust whether or not that trust is warranted [37]. Lower-cost checking cuts over-reliance [7], [45].
- **C6 Moderators:** dispositional ones such as propensity to trust, and situational ones such as self-confidence, expertise, workload, effort and stakes [17], [40], [44].
- **C7 User control:** the allocation of decision authority over an output between person and system, sorted by kind and timing (Section 5.2).

5.2 Control by Kind and Timing

The three kinds of control differ in who decides, and they also differ in when control acts on the person. Table 5 places each control design from the corpus.

Table 5: Control Designs by Kind and Timing

Kind	Set before use	Exercised during use
User-held	Expectation slider (Kocielnik et al. [26]); delegation rules (Lai et al. [12])	Adjustment within a bound (Dietvorst et al. 2018 [11]); advice on demand (Buçinca et al. [10]); blind delegation (Zhang et al. [46]); per-trial choice (Okamura and Yamada [34])
Designer-fixed	Limit on how far a forecast may be adjusted (Dietvorst et al. 2018 [11])	Decide first; wait (Buçinca et al. [10])
System-moved	None found	Adaptive workflow and hidden advice (Ma et al. [50]); adaptive cues (Okamura and Yamada [34])

Designer-fixed control is always decided in advance by the designer, so the timing axis here records when the rule acts on the person. A bound on adjustments acts once, as a limit within which the user then chooses [11]. A decide-first step acts on every item during use [10]. The table also shows an empty cell: no system-moved control in the corpus is configured before use.

5.3 Governance as a Frame

Governance is drawn around the loop rather than inside it. Everything inside the loop works per user and per output and rests on experiments; governance works per deployment, and its evidence here is conceptual. Placing it inside would mix levels and imply evidence the corpus does not have.

The frame acts on the loop in two places. First, it sets the allocation of control, deciding which kinds are designer-fixed, user-held or system-moved. Article 14(3) of the EU AI Act does this when it assigns oversight measures to the provider or to the deployer [13], and NIST subcategory GOVERN 3.2 asks organizations to define roles for human-AI configurations [58]. Second, it can require two-error testing

before deployment, as Green demands [3], using the measures in Section 4.1. The institutional conditions in Araujo's account sit at the same level [15], which is where an interface-level account and an everyday-use account can be joined.

5.4 Propositions

Table 6 states seven propositions. Each is marked by the strength of its evidence, so the framework separates what the corpus shows from what it suggests.

Table 6: Propositions and Their Evidence

Proposition	Evidence status
P1. Cues that lower the cost of verifying a specific output reduce over-reliance without raising under-reliance. Cues that add authority without enabling a check raise reliance regardless of correctness. General warnings lower reliance mainly on wrong outputs.	Supported for the over-reliance effect of checks and the two-sided effect of authority cues (Vasconcelos et al. [45]; Kim et al. 2025 [7]; Bansal et al. 2021 [6]; Si et al. [42]; Steyvers et al. [41]). That checks do not raise under-reliance rests on two studies (Kim et al. 2025 [7]; Si et al. [42]). Warnings: two studies, one task each (Kim et al. 2024 [39]; Bo et al. [9]); contrastive arguments raised under-reliance (Si et al. [42]).
P2. User-held adjustment reduces disuse of a better AI regardless of how much leeway it gives; whether it also limits over-reliance is untested.	Half supported (Dietvorst et al. 2018 [11]). The other half is this review's central gap.
P3. Forcing designs cut over-reliance. Applied to every item they raise under-reliance; moved by the system to items where the user is likely right, they can avoid that cost.	Over-reliance effect supported (Buçinca et al. [10]; Bo et al. [9]; Ma et al. [50]). The under-reliance cost of uniform designs rests on one study (Bo et al. [9]), because Buçinca et al. did not measure under-reliance. The adaptive version rests on one study (Ma et al. [50]).
P4. Miscalibrated self-confidence decides which way reliance errs, so self-assessment support must be targeted: it can help over-estimators and can harm under-estimators.	Partly supported. The harm to under-estimators was significant and the help to over-estimators was not; both are before-after changes within groups selected after the first batch, so regression to the mean may explain part of them (He et al. [40]). See also Logg et al. [44] and Bo et al. [9].
P5. Control that reaches the system adds evidence from use (C7 to C5), so its calibration benefit should grow over repeated sessions.	Untested: the corpus has no repeated-session study. Lai et al. [12] show the mechanism in a single session.

Proposition	Evidence status
P6. User-held control calibrates only when it changes system behavior at the level where errors cluster, such as input regions or claims; control that does not reach the system raises acceptance only.	Partly supported with classifiers: Lai et al. [12] scored only the over-reliance side, Kocielnik et al. [26] support the second clause, and Bansal et al. 2019 [33] show that errors must cluster in learnable ways. Untested for LLMs.
P7. Override provisions without two-error testing raise uptake and shift responsibility without improving decisions.	Conceptual (Green [3]; Sterz et al. [54]), with indirect evidence for uptake (Dietvorst et al. 2018 [11]).

Each proposition names both error types or a mechanism that can be observed, so each can be tested by the studies proposed in Section 6.1.

6. Discussion

The review's central finding is an asymmetry. Cues have been tested often, mostly against one error, and the most reliable lever among them is making outputs easier to verify. Control has been tested rarely: the user-held kind raises use, and only system-moved control has been shown to reduce over-reliance without raising under-reliance in one study. This section turns that asymmetry into studies, design guidance and societal implications, and then states the review's limits.

6.1 Research Agenda

Six studies would test the propositions directly.

1. **Score a user-held design on both errors (P2, P6).** Rerun bounded adjustment on a task with item-level answer keys, or add a recall measure to conditional delegation. Report RAIR, RSR and the accuracy of each adjustment or rule, not only uptake.
2. **Cross kind with timing (Table 5).** Hold the task and the AI constant and compare user-held control set before use, user-held control exercised during use, designer-fixed forcing and system-moved allocation. Fill the empty cell with a system-moved design configured before use, for example delegation rules that the system proposes and the user accepts.
3. **Follow control across sessions (P5).** Run several sessions over weeks, vary whether control reaches the system, and track RAIR and RSR after each error the AI makes.
4. **Give LLM users control at the level of claims (P6).** Let people set rules for which kinds of claim the model must source or flag for checking, build an answer key for each claim, and score both errors claim by claim.
5. **Target self-assessment feedback (P4).** Estimate each person's self-assessment bias first, then randomize feedback separately for over- and under-estimators, include a no-feedback group for each, and measure both errors.
6. **Audit oversight in a deployment (P7).** Before and after training of the kind Article 14 envisages, compare the rate of overrides with their accuracy on both error types.

Two changes of method would help every study: log control actions together with their outcomes, and report both errors whenever an answer key exists.

6.2 Implications for Design

- **Test both errors before launch.** Build an answer-key set for the main tasks and report reliance on right and wrong outputs separately. An override rate says nothing about whether the overrides are right.
- **Put evidence within reach.** Place the source or evidence for a specific claim one click away and in view. Prefer this to longer explanations, which raise confidence without improving discrimination.
- **Warn in the first person, and calibrate where possible.** Hedges such as "I'm not sure" lowered reliance mainly on wrong answers. Tying such wording to the model's own confidence helps people judge answers, provided that confidence is accurate, but it will not help people who lack the knowledge to do better.
- **Give control that changes what the system does.** Delegation rules and bounded adjustment act on outputs; a setting that changes nothing raises acceptance without improving decisions. Bound adjustments by default and keep a separate, logged full override: bounded adjustment outperformed free use of a model in forecasting [11], but Article 14(4)(d) requires that overseers of high-risk systems be able to disregard or override output [13].
- **Apply friction where it pays.** Forcing people to decide first cuts over-reliance but is disliked and, applied to every item, can raise under-reliance. Apply it where the user is likely right or the AI is likely wrong.
- **Log control.** Record overrides, edits and delegation choices with their outcomes so that calibration can be audited later.

6.3 Societal Implications

Three implications reach beyond interface design.

The first concerns who bears the cost of checking. Over-reliance rises with task difficulty [45] and with fluent explanations [41], so when checking is expensive the burden falls on the people with the least time and expertise to carry it, often frontline workers or members of the public. Designs that make checking cheap redistribute that labor; designs that add explanation without verification shift it back onto users while appearing to inform them.

The second concerns responsibility. Override and stop provisions give people causal power, but the evidence shows that such power raises uptake regardless of quality [11] and that human overrides can add error and bias, as in risk assessments where judges' departures were harsher for Black defendants [3]. Where oversight capacities exist without evidence that they work, they risk making operators answerable for outcomes they could not have prevented. Responsibility should therefore be traced to those who allocate control and certify that oversight works [56], [57], which is the role of the governance frame.

The third concerns legitimacy. A control that makes people feel in charge without changing what the system does can raise acceptance [26], and the designs that most improve calibration, such as forcing people to decide first, are the least liked [10]. Market and political incentives may therefore favor features that build acceptance over features that improve decisions. Regulators and auditors should ask whether a control changes outcomes on both error types, rather than whether it exists or whether users like it.

6.4 Limitations

This is an integrative review with purposive selection, not a systematic review. No database search was logged, so relevant studies may be missing. The two central gap claims are therefore scoped to the studies reviewed and supported by one targeted search. A single reviewer selected, read and coded every source, so no agreement rate can be reported.

A large language model assisted with retrieval, summarization, and proofreading (Section 3.5). Sixteen of the 50 sources were read from abstracts, summaries, excerpts or subcategory wording; two of these could not be obtained in full and are cited only for findings stated in their abstracts. Two studies were read as reported in their first authors' dissertations rather than in the published papers. The positioning against Araujo rests on its abstract and a published summary [15]. Appendix A marks the basis of each source. The review covers English-language sources only.

The evidence base has its own limits. Most studies used crowd workers, low stakes and a single session, and almost all used tasks with answer keys. The LLM studies tested fixed responses from particular model versions, which may not reflect current systems. Governance evidence is conceptual and drawn from EU and United States frameworks only.

7. Conclusion

Calibrating trust in AI has mostly been pursued by telling people more. The evidence reviewed here points to a different emphasis. Cues work when they make a specific output easier to verify, and user control works when it changes what the system does, yet control has almost never been scored on both ways reliance can fail. The framework and its seven propositions turn that gap into testable studies, and the governance frame shows where regulation could require the test that oversight currently assumes.

9. References

1. J. D. Lee, K. A. See, "Trust in automation: designing for appropriate reliance", Human Factors, 2004, 46 (1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
2. M. Schemmer, N. Kühl, C. Benz, A. Bartos, G. Satzger, "Appropriate reliance on AI advice: conceptualization and the effect of explanations", Proceedings of the 28th International Conference on Intelligent User Interfaces, Association for Computing Machinery, 2023, 410–422. https://doi.org/10.1145/3581641.3584066
3. B. Green, "The flaws of policies requiring human oversight of government algorithms", Computer Law and Security Review, 2022, 45, 105681. https://doi.org/10.1016/j.clsr.2022.105681
4. S. Mehrotra, C. Degachi, O. Vereschak, C. M. Jonker, M. L. Tielman, "A systematic review on fostering appropriate trust in human-AI interaction: trends, opportunities and challenges", ACM Journal on Responsible Computing, 2024, 1 (4), Article 26. https://doi.org/10.1145/3696449
5. M. Wischnewski, N. Krämer, E. Müller, "Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions", Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2023. https://doi.org/10.1145/3544548.3581197
6. G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, "Does the whole exceed its parts? The effect of AI explanations on complementary team performance", Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2021, Article 81. https://doi.org/10.1145/3411764.3445717
7. S. S. Y. Kim, J. Wortman Vaughan, Q. V. Liao, T. Lombrozo, O. Russakovsky, "Fostering appropriate reliance on large language models: the role of explanations, sources, and inconsistencies", Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2025. https://doi.org/10.1145/3706598.3714020

8. M. Schemmer, P. Hemmer, M. Nitsche, N. Kühl, M. Vössing, "A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making", Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society, Association for Computing Machinery, 2022, 617–626. https://doi.org/10.1145/3514094.3534128
9. J. Y. Bo, S. Wan, A. Anderson, "To rely or not to rely? Evaluating interventions for appropriate reliance on large language models", Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2025, Article 905. https://doi.org/10.1145/3706598.3714097
10. Z. Buçinca, M. B. Malaya, K. Z. Gajos, "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making", Proceedings of the ACM on Human-Computer Interaction, 2021, 5 (CSCW1), Article 188. https://doi.org/10.1145/3449287
11. B. J. Dietvorst, J. P. Simmons, C. Massey, "Overcoming algorithm aversion: people will use imperfect algorithms if they can (even slightly) modify them", Management Science, 2018, 64 (3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643
12. V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, C. Tan, "Human-AI collaboration via conditional delegation: a case study of content moderation", Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2022, Article 54. https://doi.org/10.1145/3491102.3501999
13. European Union, "Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)", Official Journal of the European Union, L series, 12 July 2024. http://data.europa.eu/eli/reg/2024/1689/oj
14. M. Raees, K. Papangelis, "From trust to appropriate reliance: measurement constructs in human-AI decision-making", arXiv preprint arXiv:2604.23896, 2026. https://arxiv.org/abs/2604.23896
15. T. Araujo, "Unpacking the dynamics of generative AI use in our daily lives: towards an integrative trust calibration framework", AI and Society, 2026. https://doi.org/10.1007/s00146-026-03279-0
16. R. C. Mayer, J. H. Davis, F. D. Schoorman, "An integrative model of organizational trust", Academy of Management Review, 1995, 20 (3), 709–734. https://doi.org/10.5465/amr.1995.9508080335
17. K. A. Hoff, M. Bashir, "Trust in automation: integrating empirical evidence on factors that influence trust", Human Factors, 2015, 57 (3), 407–434. https://doi.org/10.1177/0018720814547570
18. E. Glikson, A. W. Woolley, "Human trust in artificial intelligence: review of empirical research", Academy of Management Annals, 2020, 14 (2), 627–660. https://doi.org/10.5465/annals.2018.0057
19. R. Parasuraman, D. H. Manzey, "Complacency and bias in human use of automation: an attentional integration", Human Factors, 2010, 52 (3), 381–410. https://doi.org/10.1177/0018720810376055
20. R. Parasuraman, V. Riley, "Humans and automation: use, misuse, disuse, abuse", Human Factors, 1997, 39 (2), 230–253. https://doi.org/10.1518/001872097778543886
21. B. J. Dietvorst, J. P. Simmons, C. Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err", Journal of Experimental Psychology: General, 2015, 144 (1), 114–126. https://doi.org/10.1037/xge0000033
22. J. W. Burton, M.-K. Stein, T. B. Jensen, "A systematic review of algorithm aversion in augmented decision making", Journal of Behavioral Decision Making, 2020, 33 (2), 220–239. https://doi.org/10.1002/bdm.2155

23. A. Jacovi, A. Marasović, T. Miller, Y. Goldberg, "Formalizing trust in artificial intelligence: prerequisites, causes and goals of human trust in AI", Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2021, 624–635. https://doi.org/10.1145/3442188.3445923
24. R. Parasuraman, T. B. Sheridan, C. D. Wickens, "A model for types and levels of human interaction with automation", IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans, 2000, 30 (3), 286–297. https://doi.org/10.1109/3468.844354
25. B. Shneiderman, "Human-centered artificial intelligence: reliable, safe and trustworthy", International Journal of Human-Computer Interaction, 2020, 36 (6), 495–504. https://doi.org/10.1080/10447318.2020.1741118
26. R. Kocielnik, S. Amershi, P. N. Bennett, "Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems", Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2019, 1–14. https://doi.org/10.1145/3290605.3300641
27. C. Castelfranchi, R. Falcone, "Trust and control: a dialectic link", Applied Artificial Intelligence, 2000, 14, 799–823. https://doi.org/10.1080/08839510050127560
28. R. J. Torraco, "Writing integrative literature reviews: guidelines and examples", Human Resource Development Review, 2005, 4 (3), 356–367. https://doi.org/10.1177/1534484305278283
29. J. Webster, R. T. Watson, "Analyzing the past to prepare for the future: writing a literature review", MIS Quarterly, 2002, 26 (2), xiii–xxiii.
30. M. J. McGrath, O. Lack, J. Tisch, A. Duenser, "Measuring trust in artificial intelligence: validation of an established scale and its short form", Frontiers in Artificial Intelligence, 2025, 8, 1582880. https://doi.org/10.3389/frai.2025.1582880
31. N. Scharowski, S. A. C. Perrig, N. von Felten, L. F. Aeschbach, K. Opwis, P. Wintersberger, F. Brühlmann, "To trust or distrust AI: a questionnaire validation study", Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2025, 361–374. https://doi.org/10.1145/3715275.3732025
32. R. D. Spain, E. A. Bustamante, J. P. Bliss, "Towards an empirically developed scale for system trust: take two", Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 2008, 52, 1335–1339.
33. G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, E. Horvitz, "Beyond accuracy: the role of mental models in human-AI team performance", Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 2019, 7 (1), 2–11. https://doi.org/10.1609/hcomp.v7i1.5285
34. K. Okamura, S. Yamada, "Adaptive trust calibration for human-AI collaboration", PLOS ONE, 2020, 15 (2), e0229132. https://doi.org/10.1371/journal.pone.0229132
35. J.-Y. Jian, A. M. Bisantz, C. G. Drury, "Foundations for an empirically determined scale of trust in automated systems", International Journal of Cognitive Ergonomics, 2000, 4 (1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
36. O. Vereschak, G. Bailly, B. Caramiaux, "How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies", Proceedings of the ACM on Human-Computer Interaction, 2021, 5 (CSCW2), Article 327. https://doi.org/10.1145/3476068

37. M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, "The role of trust in automation reliance", International Journal of Human-Computer Studies, 2003, 58 (6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7
38. S. C. Kohn, E. J. de Visser, E. Wiese, Y.-C. Lee, T. H. Shaw, "Measurement of trust in automation: a narrative review and reference guide", Frontiers in Psychology, 2021, 12, 604977. https://doi.org/10.3389/fpsyg.2021.604977
39. S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, J. Wortman Vaughan, "'I'm not sure, but...': examining the impact of large language models' uncertainty expression on user reliance and trust", Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2024, 822–835. https://doi.org/10.1145/3630106.3658941
40. G. He, L. Kuiper, U. Gadiraju, "Knowing about knowing: an illusion of human competence can hinder appropriate reliance on AI systems", Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2023, Article 113. https://doi.org/10.1145/3544548.3581025
41. M. Steyvers, H. Tejeda, A. Kumar, C. Belem, S. Karny, X. Hu, L. W. Mayer, P. Smyth, "What large language models know and what people think they know", Nature Machine Intelligence, 2025, 7, 221–231. https://doi.org/10.1038/s42256-024-00976-7
42. C. Si, N. Goyal, S. T. Wu, C. Zhao, S. Feng, H. Daumé III, J. Boyd-Graber, "Large language models help humans verify truthfulness, except when they are convincingly wrong", Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2024, 1459–1474. https://doi.org/10.18653/v1/2024.naacl-long.81
43. F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. Wortman Vaughan, H. Wallach, "Manipulating and measuring model interpretability", Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2021, Article 237. https://doi.org/10.1145/3411764.3445315
44. J. M. Logg, J. A. Minson, D. A. Moore, "Algorithm appreciation: people prefer algorithmic to human judgment", Organizational Behavior and Human Decision Processes, 2019, 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
45. H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Krishna, "Explanations can reduce overreliance on AI systems during decision-making", Proceedings of the ACM on Human-Computer Interaction, 2023, 7 (CSCW1), Article 129. https://doi.org/10.1145/3579605
46. Y. Zhang, Q. V. Liao, R. K. E. Bellamy, "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making", Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2020, 295–305. https://doi.org/10.1145/3351095.3372852
47. M. Yin, J. Wortman Vaughan, H. Wallach, "Understanding the effect of accuracy on trust in machine learning models", Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2019. https://doi.org/10.1145/3290605.3300509
48. C. J. Cai, S. Winter, D. Steiner, L. Wilcox, M. Terry, "'Hello AI': uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making", Proceedings of the ACM on Human-Computer Interaction, 2019, 3 (CSCW), Article 104. https://doi.org/10.1145/3359206

49. X. Wang, M. Yin, "Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making", Proceedings of the 26th International Conference on Intelligent User Interfaces, Association for Computing Machinery, 2021, 318–328. https://doi.org/10.1145/3397481.3450650
50. S. Ma, Y. Lei, X. Wang, C. Zheng, C. Shi, M. Yin, X. Ma, "Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making", Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2023, Article 759. https://doi.org/10.1145/3544548.3581058
51. S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, J. Teevan, R. Kikin-Gil, E. Horvitz, "Guidelines for human-AI interaction", Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2019, 1–13. https://doi.org/10.1145/3290605.3300233
52. S. Passi, M. Vorvoreanu, "Overreliance on AI: literature review", Technical Report MSR-TR-2022-12, Microsoft Research, 2022. https://www.microsoft.com/en-us/research/?p=854967
53. Q. V. Liao, J. Wortman Vaughan, "AI transparency in the age of LLMs: a human-centered research roadmap", Harvard Data Science Review, 2024, Special Issue 5. https://doi.org/10.1162/99608f92.8036d03b
54. S. Sterz, K. Baum, S. Biewer, H. Hermanns, A. Lauber-Rönsberg, P. Meinel, M. Langer, "On the quest for effectiveness in human oversight: interdisciplinary perspectives", Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2024, 2495–2507. https://doi.org/10.1145/3630106.3659051
55. L. Bainbridge, "Ironies of automation", Automatica, 1983, 19 (6), 775–779. https://doi.org/10.1016/0005-1098(83)90046-8
56. J. Laux, "Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act", AI and Society, 2024, 39 (6), 2853–2866. https://doi.org/10.1007/s00146-023-01777-z
57. F. Santoni de Sio, J. van den Hoven, "Meaningful human control over autonomous systems: a philosophical account", Frontiers in Robotics and AI, 2018, 5, 15. https://doi.org/10.3389/frobt.2018.00015
58. National Institute of Standards and Technology, "Artificial Intelligence Risk Management Framework (AI RMF 1.0)", NIST AI 100-1, 2023. https://doi.org/10.6028/NIST.AI.100-1