Calibrating Trust in AI. A Co-Designed Checklist for Task-Specific Reliance

Shakhtar Donetsk
Politecnico di Torino
Type 3

Abstract

Situation: Workers increasingly rely on AI systems for consequential tasks, yet the conditions under which that reliance is well matched to actual system capabilities remain poorly understood.

Complication: Existing literature on trust calibration is fragmented across technical, human-factors, and regulatory literature and has not been translated into a practically usable decision tool.

Proposal: In this paper, we propose a co-designed checklist for trust calibration in AI-assisted work, derived from a prior scoping review and iteratively refined through semi-structured interviews with eight domain stakeholders (one Senior Machine Learning Engineer, one Human Resources Manager, one professor of Telecommunications Engineering, and five students from varied disciplines) and a complementary survey of 109 participants. We report the co-design process, the substantive changes between checklist versions V1 and V2, and the resulting instrument.

1 Introduction

There is a clear need for a criterion for the appropriate level of trust that employees will place in AI systems for particular tasks. On one hand, overtrust and automation bias lead users to accept outputs without independent scrutiny. On the other hand, undertrust or algorithm aversion leads users to discard useful assistance even when the system outperforms human judgment. Both failure modes entail real costs and have been documented empirically [3, 4, 5]. Despite the considerable number of studies on the factors that affect AI trust (technical reliability, explainability, regulatory environment, organisational pressure), there is a significant gap between that knowledge and the tools practitioners can use when deciding whether, and how much, to rely on an AI system, given a specific task.

This paper attempts to address that gap. Building on our prior scoping review [2], which identified five categories of trust-influencing factors and their interactions along a calibration continuum, we co-designed a general-purpose checklist that operationalizes these findings as actionable prompts.

The checklist is intended to be profession-agnostic, while profession-specific reasoning appears only in the worked use case retained from the scoping review (the university professor).

The contribution of this paper is methodological and instrumental: we describe how the checklist was built through a structured co-design process, document the changes between V1 and V2, and present the final instrument.

2 Method

2.1 Overview of the Co-design Process

The co-design process followed the iterative structure described by Madaio et al. [6]: a V1 draft is presented to stakeholders as a stimulus rather than a proposal. Session feedback is then coded and translated into traceable changes, a revised V2 is produced, and the cycle repeats until no further substantive refinements emerge.

For this study, three co-design rounds were conducted between V1 and V2, structured around three semi-structured interviews, a complementary survey, and in-class workshops running in parallel.

This multi-method design was chosen deliberately: interviews provide depth and the ability to probe context-specific reasoning, the survey provides breadth and attitudinal triangulation across a population too large to interview, and workshops allow the checklist to be tested under conditions of group deliberation that more closely resemble real deployment contexts.

No single method could have achieved all three functions, and the three streams were designed to be complementary rather than redundant.

Figure 1 summarizes the procedure:

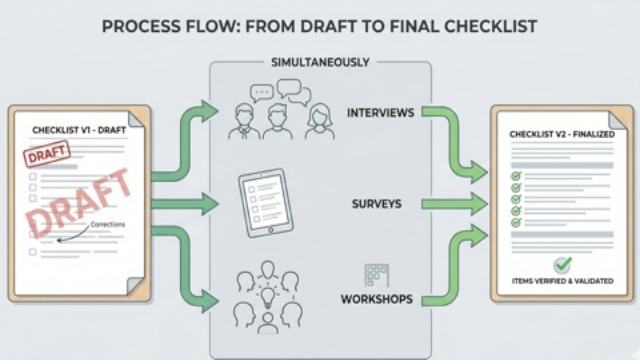

Fig. 1 Evolution from early draft (V1) to updated version (V2) (The documents in the picture have a merely illustrative purpose)

2.2 Checklist V1

V1 was constructed directly from the five-category framework developed in our Scoping Review [2]. Each framework dimension (AI Technical Structure and Quality; Human User Factors; Systemic and Regulatory Factors; Technology Provider Factors; Environment of Application) was translated into a set of concrete yes/no-style prompts and subsequently rewritten as deliberative questions, following the guidance of Madaio et al.[6] that binary items encourage compliance rather than genuine scrutiny. V1 contained 20 items organized across five sections: Task Characteristics, AI System Factors, User and Context Factors, Accountability, and Ongoing Calibration. V1 was intentionally left incomplete in a few places: a stimulus checklist should be dense enough to respond to but open enough to invite additions, so that participants engage as co-authors rather than evaluators [6].

3 Data Collection

3.1 Interview Procedure

Rationale:

Semi-structured interviews were chosen as the primary feedback mechanism because the goal was not to measure the frequency of reactions to V1 items but to understand the reasoning behind them. Trust calibration is a context-sensitive process, in which only an open conversational format allows the interviewer to follow up on why a participant would not act on a given item or what implicit assumption the item was missing. The semi-structured format provided a framework across sessions while maintaining the flexibility to pursue unanticipated responses.

Participants and Recruitment:

Eight individuals were interviewed across three sessions, recruited through direct contact and sampled within the academic environment and professional networks. The composition was designed to span three distinct stakeholder groups: domain experts responsible for consequential decisions, student users whose trust calibration is directly influenced by institutions and peers, and professional practitioners who encounter AI governance questions in operational rather than educational settings.

Session 1 involved one professor of Telecommunications Engineering, a student studying Product Design, and a student following the Management Engineering course. Session 2 involved three students from varied disciplines (Film Engineering, Management Engineering, and Biomedical Engineering). The last session (Session 3) involved two participants recruited from outside the academic environment: a Senior Machine Learning Engineer with approximately eight years of experience in credit risk and fraud detection at a fintech company, and a Human Resources Manager with approximately twelve years of experience, with a focus in the last three years towards AI-assisted applicant tracking at a manufacturing company.

Session 3 was added specifically to surface the perspective of practitioners operating under formal governance constraints and directly responsible for high-stakes decisions affecting individuals, a perspective absent from the first two sessions, which had only included educational contexts. Individual sessions were conducted separately rather than in a mixed group, following the recommendation that mixing seniority levels or roles with conflicting interests risks neglecting another group's feedback.

Protocol:

Each interview lasted approximately 20-25 minutes and followed a semi-structured guide prepared in advance. The guide was divided into three parts: (1) a warm-up in which participants described their current use of AI tools and any friction they experienced; (2) a structured walk-through of the V1 checklist, in which participants were asked for each section What is missing? What would you remove? What is hardest to act on in practice? (3) a closing asking whether, after completing the checklist, they would feel confident that their trust in the AI system was appropriately calibrated for the task at hand. One team member facilitated, while another took structured notes, and roles were rotated between sessions. Key verbatim quotes were recorded with participant consent, and all participants were informed that responses would be anonymized before use.

3.2 Survey

Rationale:

The survey served a complementary function to the interviews because the interview sample was limited to eight participants selected for stakeholder diversity rather than representativeness; it was not possible to draw population-level inferences from it alone. The survey was implemented to extend the picture and to check whether the trust-sensitivity patterns emerging from interviews were observable, typically, the relationship between perceived task stakes and willingness to rely on AI. It was not designed to test hypotheses, and its results are used illustratively rather than inferentially throughout this paper.

Participants:

The survey was administered to 109 respondents, primarily university students. The sample is not representative of the broader working population, a limitation explicitly acknowledged by the survey's contextual rather than general contribution.

Instrument:

The survey was structured into three parts: (1) demographic information; (2) general attitudes toward AI and trust, including Likert-scale items on perceived reliability, explainability, and willingness to rely on AI across task types; and (3) perceptions of AI integration in professional or educational settings. Full survey instruments and aggregated results are provided in the Appendix. In the ZIP archive mentioned below, the survey data can be found in "AI trust calibration survey.csv".

3.3 Co-design workshops

Rationale:

Workshops addressed a limitation of individual interviews, as they cannot observe how a checklist functions when used collaboratively or under social pressure. The HR Manager's feedback in Session 3 made this limitation concrete as she observed that trust in AI tools in her organization is socially contagious, shaped by what managers do and what colleagues informally state, rather than by individual assessment. Running the checklist through structured group activities allowed the team to observe whether items that seemed clear in individual sessions produced confusion or disagreement in a group context.

Procedure:

Two rounds of in-class workshops were conducted with a broader group of students. In the first round, participants evaluated V1 item by item, providing suggestions for revisions, including additions or removals, and flagging items that were theoretically important. In the second round, participants revised the intermediate version of the checklist and applied it to a profession-specific storyboard, walking through a scenario to reach a trust calibration decision. This activity enabled participants to highlight gaps that item feedback alone had not revealed. Further details on all changes attributed to the workshop activities can be found in section 4, namely Data Analysis.

3.4 Synthesis Procedure

The three data streams were integrated sequentially rather than analyzed independently.

Session 3 introduced feedback, which had been absent from the academic sessions. The ML Engineer identified gaps around model versioning, distribution shift, and verification cost, and flagged the accountability section as too thin. The HR Manager highlighted the technically distant user, specifically a non-technical practitioner with no access to the underlying system. It was noted that this group-level dynamic, already discussed in sub-section 3.3, pointed toward a structural gap in V1 that no individual-focused item currently addresses.

Survey results served as a directional check in the second stage, confirming that the two patterns are consistent with the interview findings: an increase in willingness to rely on AI and a lack of institutional guidance on AI use.

Workshop feedback was integrated last, testing whether interview-driven revisions had resolved the issues identified and surfacing any gaps invisible in individual sessions.

4 Data Analysis

4.1 Coding Procedure

After each interview session, notes were reviewed and all feedback coded into four categories: wording issues (items that were unclear or used vocabulary participants did not recognize), missing items (gaps in coverage identified by participants), structural concerns (ordering, grouping, or section logic), and feasibility concerns (items participants regarded as theoretically correct but practically unactionable in their daily work). Items flagged across multiple participants or sessions were prioritized for revision. Where a participant proposed removing an item that the team judged theoretically necessary, the item was retained and reworded to make its practical purpose explicit, with the rationale for retention noted.

The same coding framework was applied to workshop outputs, with one modification: because workshop participants worked collaboratively rather than individually, structural concerns were

weighted more heavily, as group deliberation surfaced ordering and grouping issues that individual interviews had not. Survey responses were used as a directional triangulation check rather than a primary source of revisions, confirming or qualifying patterns already identified through interviews and workshops. The workflow is summarised in **Figure 2** below.

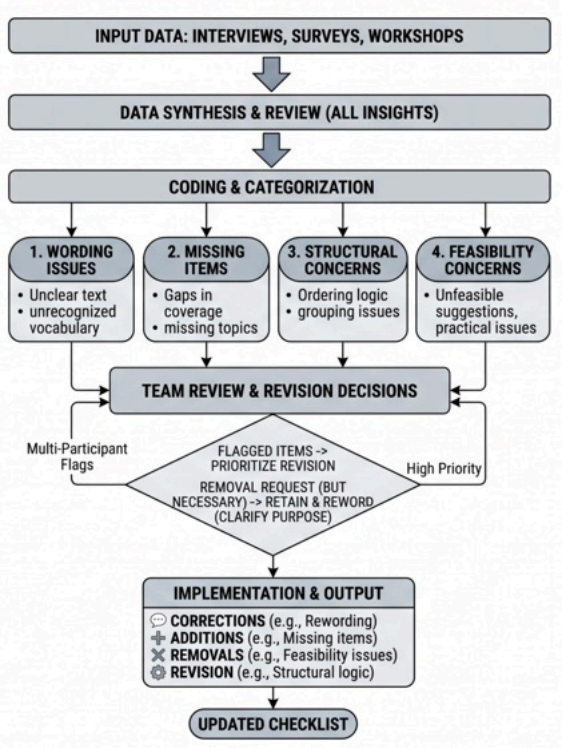

Fig. 2 Qualitative Analysis Workflow Diagram

4.2 From V1 to V2: traceable changes

The changes between V1 and V2 are distributed across all three data sources: interviews, workshops, and the survey. These span all four coding categories. The following account traces the most substantial changes to their sources, grouped by checklist section. A complete itemized log is provided in Appendix E.

Task Characteristics:

The professor in Session 1 claimed that the recoverability item treated error recovery as binary; therefore, the term “easily recoverable" was added to prompt a more calibrated response. A student in Session 1 raised the issue of confidentiality as missing, which became Task Characteristics item 7, a gap later confirmed by the HR manager in Session 3. The same student observed that V1 ignored the concepts of deadline pressure and peer behavior as drivers of AI use. This became User and Context Factors item 2, subsequently expanded when the ML Engineer in Session 3 raised verification cost as a distinct calibration factor. Both contributions were merged into a single item.

AI System Factors:

In the first session, the professor marked the subdomain relevance of the training data as missing, which later became the AI System Factors item 4. The Machine Learning Engineer in Session 3 also identified two structural gaps. The failure modes item was too generic, and as a result, V2 rewording added distribution shift and out-of-distribution behavior as the operationally relevant concern. V1 also contained no item on model versioning, and the engineer recounted a fraud detection model that silently degraded after a vendor update, continuing to output confident predictions outside its validated range. This became an AI System Factors item 3.

Jargon Removal:

Multiple participants across Sessions 2 and 3 flagged "algorithm aversion" and "active evaluator" as academic vocabulary that practitioners would not use. The HR Manager noted that the algorithm aversion self-check was unlikely to work: a reluctant user would not self-identify as irrational. V2 drops the term and reframes it as a positive, evidence-based prompt. The term “active evaluator” was retained but reframed around task-specific error recognition rather than general field familiarity. Both changes also involved structural relocation: the domain-expertise item moved from User and Context Factors to Task Characteristics (TC-6), because co-design participants consistently treated it as a property of the task–user pairing rather than a contextual factor; the algorithm-aversion self-check moved to Ongoing Calibration (OC-3), where it sits alongside other items addressing how trust should evolve over time. These section-level migrations, together with the two V1 removals documented in Appendix E, account for the reduction of User and Context Factors from five items in V1 to three in V2.

Accountability and Escalation:

The ML Engineer labeled the accountability section as too thin for its real-world weight. The existing item was reworded, and a new escalation item was added, requiring pre-committed conditions for deferring to human review. V2 renames the section accordingly.

Workshop-driven Changes:

The first workshop round led to the addition of a hybrid domain category to Task Characteristics item 3, as participants found that the technical/human-centered framing did not guide mixed tasks. Workshop Round 2’s storyboard activity revealed the absence of a prompt for practitioners whose calibration choices become norms for a wider team, prompting minor structural revisions to the Accountability section’s preamble.

Post-failure Calibration and Regulatory Items:

A student in Session 2 marked the post-failure monitoring system as too academic; consequently, V2 reframes it as a reflection prompt. The governance item drew the same concern across sessions and workshops, and was reworded to place responsibility at the individual rather than the organizational level.

4.3 Items retained against feedback

Two items were retained despite the requests for modification. The professor in Session 1 proposed moving the algorithm-aversion self-assessment to a third-party process, because a user who is being averse will not self-identify as irrational. The prompt was retained in individual form because a third-party assessment lies outside the scope of an individual decision tool. The HR Manager in Session 3 suggested that the checklist should include a separate version for technically distant users, rather than just the one for practitioners who cannot inspect failure modes or training data.

4.4 Survey as contextual check

Two survey patterns informed V2. Firstly, in this student-majority sample, willingness to rely on AI varied sharply with the perceived stakes of the task across 109 respondents, with 87% indicating that human verification should be mandatory for high-stakes output [1]. A majority of respondents reported no institutional guidance on AI use, which matches the professor's account and reinforces the idea of individual decision-making on the checklist. Full survey results are given in Appendix B.

5 Work Use Case

In this section, the aim is to demonstrate V2 by applying it to the professor of Telecommunications Engineering interviewed in Session 1, using his own account of how he uses AI tools across different tasks in his academic work. This is not an evaluation of the professor's trust calibration, but a demonstration of how the checklist surfaces the distinctions practitioners actually draw.

5.1 Scenario

The professor uses AI tools across several distinct tasks: Semantic Scholar for literature search, Claude for prose revision, and GitHub Copilot for Python scripts in his research group. He described his trust posture as deliberately differentiated, expecting retrieval tools to surface real documents and generative tools to produce plausible but potentially inaccurate text. His verification depth varies by orders of magnitude depending on the task.

5.2 How V2 applies

Task Characteristics item 1 (recoverability) immediately distinguishes his tasks. For peer review and recommendation letters, an error would be irreversible, and as he stated, using AI for such tasks "cannot be outsourced without committing a kind of fraud." For simulation scripts, errors are caught in testing with no lasting consequences. Item 5 (moral discernment) applies to peer review and recommendation letters, both of which carry moral weight he judged intrinsically non-delegable. Item 6 (domain expertise calibrated against likely error) confirms that for technical literature synthesis, he has the expertise to evaluate outputs. Still, he cannot assume his students do, which shapes the norms he sets in his research group.

AI System Factors item 2 (task-specific reliability) maps onto his distinction between retrieval and generative tools. For Semantic Scholar, reliability for surfacing real documents is established. For Claude, generating technical citations, reliability in narrow subdomains like 5G channel estimation is not established, so he verifies every factual claim. Item 3 (temporal stability) addresses his concern that vendor updates can silently alter a tool's behavior.

Accountability and Escalation item 1 (who is answerable) applies directly: for outputs leaving his research group, his name is attached, and he bears responsibility regardless of who produced the text. Item 2 (escalation conditions) is visible in his pre-commitment never to use AI for peer review or recommendation letters.

User and Context Factors item 2 (fitness-for-purpose vs pressure) is relevant because he receives no institutional AI guidance, meaning his calibration choices set norms for his entire research group. His response to fabricated references illustrates ongoing Calibration item 1 (proportional post-failure adjustment): he introduced a lab rule requiring independent verification of every AI-produced citation, a proportionate recalibration rather than wholesale rejection. The survey supports two patterns in the professor's account: 87% of 109 respondents, the majority of whom were students, indicated human verification should be mandatory for high-stakes AI tasks, and a majority reported no institutional AI guidance.

5.3 What the case reveals

Applying V2 to the professor's own tasks confirms that the items refined during co-design address the distinctions he draws in practice: task-specific reliability rather than general benchmarks, verification cost as a function of failure mode, and domain expertise as a prerequisite for active evaluation. Two limitations surfaced through this application: the absence of a prompt for users whose calibration choices set norms for others, and the insufficiency of the current domain categories for boundary tasks such as curriculum design, are discussed as open design challenges in Section 6.

6 Conclusion

As AI systems become more deeply embedded in professional workflows, the gap between theoretical trust frameworks and practical decision tools remains a real obstacle for workers. This paper addressed that gap by translating a five-category trust calibration framework into a co-designed, profession-agnostic checklist, refined through stakeholder interviews, a broad survey, and in-class workshops.

The process demonstrated that co-design does more than surface preferences: it surfaces blind spots. Practitioners contributed items that no literature review alone would have flagged, most notably silent model versioning (ASF-3), data confidentiality constraints (TC-7), and the verification cost embedded in any reliance decision (UF-2). The survey's finding, in our student-majority sample, that 87% of respondents consider human verification mandatory for high-stakes tasks [1], while a majority report receiving no institutional guidance on when to apply it, underscores that the gap V2 targets is not merely theoretical. Workers are already making these calls individually, without scaffolding.

Two structural limitations remain open and should guide the next revision cycle. First, the checklist currently addresses the individual practitioner, yet both the professor's account and the HR Manager's observations show that senior users' calibration choices propagate implicitly to entire teams; a future item should prompt users to consider whether their reliance decision will set a norm for others. Second, V2 does not yet accommodate technically distal users, those who cannot inspect failure modes, training data, or model versioning directly, and for whom several ASF items are effectively unanswerable. These are not minor omissions: they represent the two most common departures from the individual-expert model the checklist currently assumes. Addressing them would extend the instrument's reach without abandoning the profession-agnostic design that makes it broadly applicable.

References

[1] Student Survey on Trust in Artificial Intelligence. 2026. Conducted as part of this study and available in the ZIP archive mentioned below.

[2] N. Kandev, P. Nacea, N. Lodigiani, N. Moiso, R. Ferrero, T. Palena, D. Pandino, P. Montorsi, and L. Vezzù. 2026. "What factors determine whether a worker's trust in an AI system is well calibrated for a given task?" The scoping review previously conducted as part of this study is available in the ZIP archive linked below.

[3] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. Human Factors: The Journal of the Human Factors and Ergonomics Society 46, 1, 50–80. doi:10.1518/hfes.46.1.50_30392

[4] Kevin A. Hoff and Masooda N. Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society 57, 3, 407–434. doi:10.1177/0018720814547570

[5] Jason W. Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making 33, 2, 220–239. doi:10.1002/bdm.2155

[6] Michael Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20). ACM, New York, NY, USA, 1–14. doi:10.1145/3313831.3376445

A. Interview Guide

Three co-design sessions were conducted. Session 1 involved one professor of Telecommunications Engineering and two students (Product Design and Management Engineering), Session 2 involved three students from varied disciplines (Film Engineering, Management Engineering, and Biomedical Engineering), and Session 3 involved two additional participants (a Senior Machine Learning Engineer from a fintech company and a Human Resources Manager from a manufacturing company). Participants were recruited through direct contact and snowball sampling within the academic environment and professional networks. Each session lasted approximately 20-25 minutes per participant. One team member facilitated, one took structured notes, and roles were rotated between sessions. Participants were informed that their input would be anonymized before the session began.

Phase 1 - Warm-up (3 min)

- Which AI tools do you use, and for what kinds of tasks in your work or studies?
- Roughly how often do you check or verify what an AI tells you?

Phase 2 - Trust in practice (10 min)

- Can you describe a situation where you trusted an AI output and it turned out to be wrong? What happened?
- Are there tasks where you would never rely on AI, even if it performed well? What makes those different?

- Does the importance or stakes of a decision affect how carefully you check AI outputs?
- Does your institution or employer give you any guidance on when to use AI? How does that affect your choices?

Phase 3 - Checklist feedback (7 min)

- V1 of the checklist was shared with each participant on screen before this phase.
- Looking at this list, what is missing? Is there an important trust question missing here?
- Is there anything here you would remove, or that feels unrealistic in your daily work?
- Is there any item that is unclear or too abstract?
- If you had to rank the three most important items, which would they be?

Closing phase (always administered)

- Is there anything about trusting AI in your work that we have not covered?
- Participants were thanked and reminded that all responses would be anonymized before use.

B ZIP Archive

The Google Drive link provides access to a .ZIP archive containing:

- "ReadMe.txt " (Zip archive instructions)
- "AI trust calibration survey.csv" (Survey Data [1])
- "checklist_V1.pdf" (Checklist V1)
- "checklist_V2.pdf" (Checklist V2)
- "Interview_Guide.pdf" (Interview Guide)
- "Interview_report_1.pdf" (Interview Report 1)
- "Interview_report_2.pdf" (Interview Report 2)
-"Interview_report_3.pdf" (Interview Report 3)
- "shakhtar_donetsk_deliverable1". (Scoping Review [2])

This document provides transparency on how qualitative feedback from domain stakeholders and quantitative survey insights were translated from early theoretical frameworks into the final actionable checklist:

https://drive.google.com/file/d/1Z-fLayQNxgqI-IHooxvqQzMQG-mAQnaz/view?usp=sharing

C Team Members of Shakhtar Donetsk

Table 2: Team Members' names, surnames, and institutional emails

Authors		
Name	**Surname**	**Email**
Nikola	Kandev	s323849@studenti.polito.it
Petru	Nacea	s321901@studenti.polito.it
Nicolò	Lodigiani	s340758@studenti.polito.it
Nicolò	Moiso	s327721@studenti.polito.it
Riccardo	Ferrero	s336535@studenti.polito.it
Tommaso	Palena	s325536@studenti.polito.it
Davide	Pandino	s336928@studenti.polito.it
Pietro	Montorsi	s326382@studenti.polito.it
Leonardo	Vezzù	s343869@studenti.polito.it

D AI Use Disclosure

All AI tools used during this project were limited to assistance with phrasing and initial orientation during the literature search. Specifically, ChatGPT and Gemini were consulted to suggest alternative formulations for selected passages and to generate preliminary summaries of candidate sources during the early screening phase.

All sources cited in this paper were read in full by at least one team member and cross-checked by a second. No citation was included on the basis of an AI-generated summary alone. AI-suggested phrasings were reviewed and either rewritten or explicitly approved by the team members responsible for each section before inclusion. The final text of all sections reflects the team's own analysis and judgment.

E Checklist V1 to V2 Change Log

#	V1 Wording	V2 Wording	Type	Source of Change
TC-1	Have you assessed whether the AI's error is recoverable or has irreversible consequences for a stakeholder?	...easily recoverable or has irreversible consequences...	Wording	Session 1 (professor): 'recoverable' treated as binary; 'easily' added to convey degrees of recoverability and prompt proportionate scrutiny.
TC-3	Have you considered whether the task belongs to a technical domain with measurable outcomes, or to a human-centered domain?	...or a hybrid domain that combines elements of both, and establishes how to balance quantitative rigor with contextual human judgment?	Addition / Wording	Workshop Round 1: participants noted the binary framing excluded hybrid domains common in real professional work, such as curriculum design.
TC-5	Have you analyzed if the task requires moral or ethical discernment, and established protections to ensure the user does not abdicate their professional obligations to the system?	...(e.g., assessing impacts on employees, clients, or communities), ensuring you do not outsource the weight of difficult ethical trade-offs to the system?	Wording	Session 3 (HR Manager): 'stakeholder impact' language added to reflect contexts where affected parties (e.g., candidates) are third parties to the decision.
TC-6	Have you determined if the user has enough domain expertise to serve as an 'active evaluator'?	Do you have enough domain expertise to recognise a plausible-sounding but wrong output on this specific task? 'Sufficient' is calibrated against the kind of error the system is most likely to make.	Wording / Feasibility	Sessions 1–2 (multiple participants): 'active evaluator' flagged as jargon. Reanchored to a concrete, task-specific self-check.
TC-7	[Not present in V1]	Have you checked that inputs do not contain confidential, personal, or otherwise sensitive information that the tool's usage policy, contractual obligations, or applicable regulations prohibit you from sharing?	Addition	Session 1 (student): raised data confidentiality as absent. Confirmed by Session 3 (HR Manager), who noted her company policy addressed only this risk.
ASF-1	Did you check the system's failure modes, including any known biases that could affect this task?	...including whether outputs reproduce dominant training data patterns rather than reflecting a culturally and demographically diverse range of human experience?	Wording / Addition	Session 1 (Product Design student): raised that bias items should address diversity and representation in outputs, not only statistical bias. Session 3 (ML Engineer): flagged the failure mode item as too generic for operational use. Both contributions informed the rewording.
ASF-3	Have you considered whether the system provides reasonable explanations for its outputs?	[Removed; absorbed into ASF item 2 on task-specific reliability]	Removal	Session 3 (ML Engineer): 'reasonable' undefined could mean SHAP values, natural language summaries, or model cards. Item absorbed into the reliability question to avoid vagueness.
ASF-3 *(new)*	[Not present in V1]	Do you know whether the system you validated is still the same model, or have silent updates altered its behaviour since you calibrated your trust?	Addition	Session 3 (ML Engineer): recounted a fraud model that degraded silently within two weeks of deployment due to a vendor update. Model versioning absent from V1.

#	V1 Wording	V2 Wording	Type	Source of Change
ASF-4 *(new)*	[Not present in V1]	Is the model's training data genuinely relevant to your specific area of work?	Addition	Session 1 (professor): subdomain relevance of training data flagged as missing; a model trained on general text may perform poorly on narrow technical subdomains.
UF-2	[Not present in V1]	Is your decision to use AI driven by a genuine fitness-for-purpose assessment, one that accounts for the actual cost of verifying the output, or by deadline pressure, competitive expectation, or peer behaviour?	Addition	Session 1 (student): deadline pressure and peer influence were raised as unaddressed. Session 3 (ML Engineer): verification cost added as a distinct calibration factor.
ACC-1	Is it clear who holds accountability for AI-assisted decisions (the user, the system, or the deploying organisation)?	Is it explicitly clear who will be answerable if this AI-assisted decision turns out to be wrong, and would that line of accountability hold up if the output were challenged after the fact?	Wording	Session 3 (ML Engineer): accountability section flagged as too thin; item reworded to be adversarial and forward-looking rather than descriptive.
ACC-2 *(new)*	[Not present in V1]	Have you defined in advance the conditions under which you would stop using the AI and escalate to human review or a domain expert?	Addition	Session 3 (ML Engineer): explicit escalation criteria absent from V1. High-stakes decisions require pre-committed thresholds, not ad hoc judgment.
ACC-2 *(V1)*	Have human oversight mechanisms been maintained, especially for critical decisions or in high-stakes environments?	[Removed; substantive content distributed across ACC-1 (adversarial accountability framing) and ACC-2 (new escalation item)]	Removal / Restructuring	Sessions 1–3 and workshops: the original item was too broad to be actionable. Its core concern, ensuring human oversight for critical decisions, is now addressed by ACC-1's forward-looking accountability test and ACC-2's pre-committed escalation conditions, both of which operationalise oversight more concretely.
ACC-3	Does the system's use account for legislative and governance frameworks?	Are you aware of whether any professional, organisational, or regulatory standards apply to your use of this tool in this specific context, and are you operating within them?	Wording	Sessions 1–2 and workshops: original phrasing placed responsibility at the organisational level. Reworded to make it actionable for an individual practitioner.
OC-1	Have you monitored how calibration evolves after a system failure, checking if a single salient error leads to permanent algorithm aversion?	After a notable failure, have you reflected on whether your trust has adjusted proportionally, or whether you have overcorrected into permanent rejection or underreacted?	Wording / Feasibility	Session 2 (student): 'monitor calibration' framing too academic for individual practice. Rewritten as a post-failure reflection prompt; jargon removed.
OC-3	Does the user show signs of 'algorithm aversion,' tending to distrust and avoid AI output even when it demonstrably outperforms human judgment?	Are you systematically revising your level of reliance as you accumulate direct evidence of where the system succeeds and fails on tasks like this one?	Wording	Sessions 2–3 (students and HR Manager): 'algorithm aversion' flagged as academic jargon unfamiliar to practitioners. Rephrased as a positive, evidence-based self-check.
UCF-5 (V1)	Have you verified whether users understand the system's unique failure modes to ensure that a salient, non-human mistake	[Removed; substantive content redistributed across OC-1 (post-failure proportional adjustment) and	Removal / Restructuring	Sessions 2–3: the original item combined two distinct concerns: failure-mode understanding and algorithm aversion, into a single question. V2 separates these: failure-mode awareness is now implicit in ASF-1's expanded bias and failure-mode check, while the

#	V1 Wording	V2 Wording	Type	Source of Change
	does not trigger an irrational and permanent rejection of the algorithm?	OC-3 (evidence-based reliance revision)]		aversion concern is addressed by OC-1 (proportional post-failure reflection) and OC-3 (systematic evidence-based revision).

F Checklist V2

Task Characteristics (TC) :

1. Have you assessed whether the AI's error is easily recoverable or has irreversible consequences for a stakeholder?

2. Have you assessed whether the consequences of the AI error will fall on a direct stakeholder or on a third party who has not chosen to rely on the system?

3. Have you considered whether the task belongs to a technical domain with measurable outcomes, a human-centered domain, or a hybrid domain that combines elements of both, and established how to balance quantitative rigor with contextual human judgment?

4. Have you considered whether the task is well-defined or ambiguous enough to require professional interpretation that AI cannot provide?

5. Have you analyzed if the task requires moral discernment or a stakeholder impact assessment (e.g., assessing impacts on employees, clients, or communities), ensuring you do not outsource the weight of difficult ethical trade-offs to the system?

6. Do you have enough domain expertise to recognise a plausible-sounding but wrong output on this specific task, not merely general familiarity with the field? "Sufficient" is calibrated against the kind of error the system is most likely to make: if you could not spot that error, your oversight is nominal rather than real.

7. Have you checked that the inputs do not contain confidential, personal, or otherwise sensitive information that the tool's usage policy, your contractual obligations, or applicable regulations prohibit you from sharing?

AI System Factors (ASF) :

1. Have you checked the system's known failure modes and biases that could affect this specific task, including, where the task involves human contexts, whether the outputs reproduce the dominant patterns of the training data rather than reflecting a culturally and demographically diverse range of human experience?

2. Did you check if the system's reliability has been established for this specific task, rather than blindly inferring it from general performance benchmarks or different types of tasks?

3. Do you know whether the system you validated is still the same model, or have silent updates or other modifications altered its behavior since you calibrated your trust?

4. Is the model's training data genuinely relevant to your specific area of work? If this information is not obtainable, has the gap been acknowledged in your trust judgement?

5. Is the system's output format (i.e., numerical, categorical, free text) appropriate for the level of precision this task requires?

Accountability and Escalation (ACC) :

1. Is it explicitly clear who will be answerable if this AI-assisted decision turns out to be wrong (you, your team, or your organisation), and would that line of accountability hold up if the output were challenged after the fact?

2. Have you defined in advance the conditions under which you would stop using the AI and escalate to human review, peer consultation, or a domain expert, for example, when the stakes exceed a predefined threshold, the output is implausible, or the system appears to be operating outside its established competence?

3. Are you aware of whether any professional, organisational, or regulatory standards apply to your use of this tool in this specific context, and are you operating within them?

Ongoing Calibration (OC) :

1. After a notable failure, have you reflected on whether your trust has adjusted proportionally, or whether you have overcorrected into permanent rejection or underreacted and continued to trust as before?

2. Are you evaluating capability on a task-specific basis, rather than generalising reliability from one type of task to another?

3. Are you systematically revising your level of reliance as you accumulate direct evidence of where the system succeeds and fails on tasks like this one?

User and Context Factors (UF) :

1. Could long-term reliance on the system erode the judgment, reasoning, or specific skills the task is meant to exercise? In learning or formative contexts, especially, consider whether the AI is supporting your development as a professional or substituting for it.

2. Is your decision to use AI driven by a genuine fitness-for-purpose assessment, one that accounts for the actual cost of verifying the output, or by deadline pressure, competitive expectation, perceived ease, or the fact that others around you are using it?

3. Is your baseline trust in the tool driven by institutional signals, public endorsement, or popularity, rather than by a conscious evaluation of its performance on this kind of task?

G Anonymized Notes

The notes listed below were used to generate the feedback gathered from each participant during the co-design process. The participants are remaining anonymous and are denoted by the marks (P1-P8).

P1: Professor of Telecommunications Engineering - Domain expert with supervisory and evaluative responsibilities - Session 1

- ➔ TC-1: flagged that the recoverability item treated error recovery as binary and requested the addition of a qualifier to show the degrees of recovery, which led to the insertion of "easily" in V2.
- ➔ OC: proposed that the algorithm-aversion self-assessment should be moved to a third-party process; the item was retained in the individual, with a rationale noted
- ➔ Positively evaluated the checklist's differentiation between retrieval and generative tool reliability
- ➔ Confirmed the absence of institutional AI assistance

P2: Product Design student - End-user perspective - Session 1

- ➔ TC-7: raised data confidentiality as absent from V1, which led to becoming a new item in V2, requiring the users to check whether inputs contain sensitive information prohibited from sharing
- ➔ UF-2: identified deadline pressure and peer behavior as unaddressed drivers of AI reliance, which were later merged with verification cost in UF-2
- ➔ Suggested the section structure could better separate task-level and system-level concerns

P3: Management Engineering student - End-user perspective - Session 1

- ➔ TC-6: flagged "active evaluator" as jargon and suggested reanchoring the item to a concrete task-specific self-check, which was adopted in V2
- ➔ Positively evaluated the overall structure and the inclusion of ongoing calibration as a salient section

P4: Film Engineering student - End-user-perspective - Session 2

- ➔ OC-1: marked the post-failure monitoring item as too academic and requested to be reframed as a personal reflection prompt rather than a monitoring procedure, which was implemented in V2
- ➔ ACC-3: raised concerns that the governance item placed responsibility at the organizational rather than the individual level. This was reworded in V2.

P5: Management Engineering student - End-user perspective - Session 2

- ➔ Confirmed P4's concern about jargon; "algorithm aversion flagged independently is too academic.
- ➔ Noted that OC-3 should be framed as an affirmative, evidence-based prompt rather than a deficit detection question. This was adopted in V2.

P6: Biomedical Engineering student - End-user perspective - Session 2

- ➔ Positively evaluated the accountability sections' framing but noted its brevity relative to the weight of real-world accountability

P7: Senior Machine Learning Engineer - Technical practitioner under formal governance constraints - Session 3

- ➔ ASF-3: identified model versioning as entirely absent from V1; recounted a fraud detection model that silently degraded after a vendor update while continuing to output confident predictions
- ➔ ACC-1: labeled accountability section as too thin for real-world weight; item was reworded
- ➔ UF-2: added verification cost as a distinct calibration factor, merged with the Session 1 student's deadline pressure item

P8: Human Resources Manager - Professional practitioner responsible for high-stakes decisions - Session 3

- ➔ TC-5: requested that the stakeholder impact language be added to the moral discernment to reflect contexts where the affected parties (candidates) are third parties to the decision, which was adopted in V2
- ➔ TC-7: confirmed that data confidentiality gap notices in Session 1; noted her company's AI policy addressed only this risk, reinforcing its importance
- ➔ Noted that trust in AI tools within her organization is socially contagious, shaped by the manager's behavior rather than individual assessment
- ➔ Suggested the checklist to include a separate version for technically distant users who cannot inspect failure modes or training data