WHAT TO FIX

DELIVERABLE 3

A REVIEW-TO-REVISION LECTURE

Use the reviewers as a diagnostic instrument.
Turn consensus into edits.
Turn edits into evidence.

NOT A DEFENSE. A REVISION PLAN.

GRADING

HOW DELIVERABLE 3 WAS GRADED

REVIEWERS SCORE

Each review gave a numerical score.
For each team, scores were averaged across all reviews received.
Example: 10 reviews → average of 10 scores.

AVERAGE PER TEAM

The average is the team's raw peer-review signal.
One strong or weak review matters less than the consensus.

RESCALE TO 33

Highest average score = 2.2.
That maps to maximum grade = 33.
All other grades scale proportionally.

THE GRADE REWARDS CONSENSUS QUALITY, NOT INDIVIDUAL EXCUSES

DIAGNOSIS

MISTAKES MADE BY THE MAJORITY OF TEAMS

1. Too narrow stakeholder samples: same sector, same company, same course, or only students.
2. No decision rule: checklist answers do not lead to stop / revise / delegate / escalate.
3. Generality leaks: profession-specific terms appear in the “general” checklist instead of the use case.
4. Iteration is not auditable: missing V1→V2 delta tables, quotes, or feedback-to-change mapping.
5. Worked use case is narrated, not applied item-by-item with answers and final action.
6. Missing appendix evidence: full versions, guides, anonymised notes, raw feedback.
7. Checklist not operational: too long, binary, abstract, double-barrelled, or interview-like.
8. Methods under-described: recruitment, coding, conflict resolution, and field testing are vague.

FIX THE EVIDENCE CHAIN: PARTICIPANTS → FEEDBACK → CHANGE → DECISION

LECTURE RULE

HOW TO READ YOUR THREE TEAM SLIDES

CONSENSUS

These are repeated criticisms.
Treat them as the high-confidence diagnosis.

PROS / CONS

Keep what works.
But do not let strengths hide structural defects.

CHANGES

Edit the document, not just the prose.
Add artefacts, rules, examples, and tables.

YOUR TARGET: MAKE THE TOOL VERIFIABLE, GENERAL, AND ACTIONABLE

TEAM 2

10 DEGREES OF TRUST

CONSENSUS POINTS

Towards Trustworthy AI Assistants in ADAS Engineering: A Co-Designed Checklist

1. Generality violation (ADAS/ISO 26262/A-SPICE in main checklist)
2. Severe organisational/sample bias (all 5 experts from Jaguar Land Rover Italia)
3. Excessive use of binary yes/no items / item framing problems
4. Co-design procedure described but not analytically evidenced (no thematic coding shown, feedback-to-change mapping mi…
5. No decision/interpretation mechanism for the completed checklist
6. Missing appendix artefacts (interview/workshop guide, anonymised notes)

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 2

PROS AND CONS

10 DEGREES OF TRUST

KEEP

1. Genuine, well-documented iteration across four checklist versions; the V1→V4 progression shifts from pos…
2. Engagement with real industrial experts from a respected automotive firm, providing grounded and credibl…
3. The checklist conceptually covers a broad set of trust dimensions (objective metrics, subjective percept…
4. Strong scholarly transparency: AI-use disclosure table, full historical drafts in Appendix A, participan…

FIX

1. The checklist fails the generality requirement: ADAS, ISO 26262, A-SPICE, and "System Engineer" language…
2. All five domain experts come from one single company (Jaguar Land Rover Italia), creating organisational…
3. Many items are binary yes/no checkboxes or attitude statements ("I would use the system…") rather than v…
4. The co-design methodology is reported but not properly analysed: no Data Collection / Data Analysis spli…
5. No scoring rule, threshold, or interpretation mechanism: the user finishes the checklist with no idea wh…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 2

SUGGESTED CHANGES

10 DEGREES OF TRUST

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Section 3 / Checklist V4 items	Profession-specific terms (ADAS, ISO 26262, A-SPICE, "System Engineer") are embedded in the ge…	Replace them with generic wording (e.g. "applicable domain safety standards", "domain engineer in a safety-critical context") and confine specific ac…	High
Sample composition	All five industrial experts come from Jaguar Land Rover Italia, undermining external validity	Recruit at least 2–3 additional participants from different automotive or engineering firms, or from compliance/legal roles, for a future iteration	High
Checklist items (multiple)	Binary yes/no and attitude items risk compliance theatre	Convert items into observable conditions or graduated scales (e.g. importance rating 1–5, or evidence-based prompts such as "How was the ground truth…	High
Methodology section	The co-design process is narrated but not analysed	Restructure Section 2 into "Data Collection" and "Data Analysis & Synthesis"; describe thematic coding and decision rules used to accept/reject feedb…	High
Final checklist	No decision rule or scoring mechanism	Add a short interpretation section explaining what unchecked items mean (e.g. mandatory dimensions, mitigation triggers, non-adoption conditions)	High
Appendix	Required artefacts missing	Add the interview/workshop guide and the anonymised raw feedback notes, structured per session	High

PRIORITISE HIGH BEFORE POLISH

TEAM 4

CONSENSUS POINTS

TRUSTWORTHY 3

Co-Designing a Reflective Checklist for AI Task Delegation

1. Sample too small and entirely tech-focused (4 participants, software/AI backgrounds)
2. Missing legal/compliance/HR perspective in stakeholders
3. Checklist still contains IT-specific terminology (APIs, libraries, cryptography, "deployment", "debugging")
4. No decision threshold / no rule for what to do with response categories
5. Overlap and redundancy between items (verifiability vs verification cost, accountability vs escalation)
6. Lack of field testing / cross-domain validation

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 4

PROS AND CONS

TRUSTWORTHY 3

KEEP

1. Genuinely traceable iteration: each of the seven V1→V2 changes is mapped to specific participant quotes…
2. Methodologically sound stakeholder separation (technical practitioners vs business leads) avoids seniori…
3. Concrete operationalisation: items reframed from subjective self-assessment ("Am I biased?") to observab…
4. A dual worked use case (low-risk sorting vs high-risk authentication code) effectively demonstrates the…

FIX

1. Only four participants, all from software/AI startup backgrounds; no compliance, legal, HR, or non-techn…
2. Several core items still use software-specific vocabulary (APIs, libraries, knowledge bases, cryptograph…
3. The status categories (Addressed / Partially / Not Yet Addressed / Not Applicable) lead nowhere: no rule…
4. Sample asymmetry: one participant with 11 years of experience versus three with 2–3 years risks skewing…
5. Conceptual overlaps remain (Item 1.3 vs 3.1, Section 4 vs 5 organisational items).

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 4

SUGGESTED CHANGES

TRUSTWORTHY 3

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Stakeholder pool	All 4 participants are in tech; no compliance/legal/HR/non-technical voice	Recruit at least 2 additional participants from risk, compliance, legal, or a clearly non-technical profession	High
Checklist items 1.1, 1.3, 5.1	IT-specific wording (APIs, libraries, cryptography)	Replace with neutral phrasing such as "external resources or organisational knowledge bases the AI needs"; move concrete coding terms to the use case	High
Status categories	A/PA/NYA/NA labels have no decision rule	Add explicit gateway logic: e.g. "Any NYA in Sections 1–2 = do not delegate; ≥3 PA items = escalate"	High
Length / usability	20 items contradict P4's warning that long checklists are skipped	Produce a 5-item "Quick Check" for daily use; keep the full 20-item form for onboarding and high-risk cases	High
Items 1.3 and 3.1	Overlap between "verifiability" and "verification cost" caused confusion	Clearly separate them: 1.3 = is verification technically possible; 3.1 = is verification economically worthwhile	Medium
Cross-domain validation	Generality claim untested	Apply V2 to one non-tech profession (legal drafting, clinical notes, marketing copy) and report whether items still hold	Medium

PRIORITISE HIGH BEFORE POLISH

TEAM 6

CONSENSUS POINTS

JOBLESS

AI Adoption in the Workplace: A Scoping Review on Benefits, Risks, Mitigations, and Readiness Frameworks

1. Internal scoring/counting inconsistencies (51 vs 52 items; 102 vs 104 pts; T3 header says 19 items, appendix lists 18)
2. Unbalanced stakeholder representation (3 designers vs 1 engineer; missing managerial/HR/legal/compliance voices)
3. Checklist V1 not provided in full in the appendix (only thematic summary)
4. No live/field validation of the checklist
5. Excessive length (51 items) — usability concern
6. "Partial" vs "Met" scoring metric needs more rigorous definition

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 6

PROS AND CONS

JOBLESS

KEEP

1. Highly auditable iteration: Table 2 logs every V1→V2 change linked to specific participants (P1–P4) and…
2. Interdisciplinary co-design surfaced a real blind spot — domain validation (T3.5) added because two diff…
3. Robust decision logic: section minimum scores and absolute blockers prevent strong benefits from masking…
4. Nuanced 3-tier scoring (Met / Partially Met / Not Met) is a clear improvement over binary checklists.

FIX

1. Multiple numerical inconsistencies in the scoring system: 51 vs 52 items; 102 vs 104 points; T3 header r…
2. Stakeholder pool is severely unbalanced: only 1 technical engineer vs 3 designers, no managerial/HR/lega…
3. Required Checklist V1 not included in full — Appendix B is only a summary, breaking the deliverable requ…
4. Renumbering between V1 and V2 (T3.12 → T3.11) is unsignalled; cross-references in the body text become i…
5. Sessions described in Section 2.1.3 as "asynchronous structured questionnaires" — closer to surveys than…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 6

SUGGESTED CHANGES

JOBLESS

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Decision logic / Section 2.2.2 / T3 header / Appendix…	Item counts and point totals disagree (51 vs 52; 102 vs 104; T3 19 vs 18)	Recount and harmonise all four locations; align section minimums to the corrected totals	High
Appendix B	Full V1 text is missing — only summary provided	Add complete V1 (all 53 original items in full wording) as a standalone artefact	High
Stakeholder pool	Only 1 engineer; no managerial/HR/legal voice	Recruit 1–2 additional participants from technical engineering and at least one from compliance/HR/management for a future iteration	High
Item numbering	T3.12 in body becomes T3.11 in appendix with no renumbering note	Add a small V1↔V2 ID-mapping table and update every cross-reference in body and appendix	High
Section 2.1.3 / methodology	Asynchronous questionnaires mislabelled as interviews; "strongest possible grounds" overclaims	Either rename to "asynchronous structured questionnaires" and acknowledge loss of probing, or run short follow-up interviews on the most impactful it…	Medium
Worked use case (Section 3)	Verdict given without numerical evaluation	Add a section-by-section numerical breakdown (e.g. T1: 18/24, T2: 10/24) ending in the Conditional verdict	Medium

PRIORITISE HIGH BEFORE POLISH

TEAM 8

CONSENSUS POINTS

TWO GIRLS, TOO LATE

Calibrating Trust in AI for Drafting-Related Professional Tasks

1. Domain-skewed sample (legal/academic only)
2. "Quick Check" is the strongest design choice
3. No field/empirical testing in a real workflow
4. Quick Check item order is illogical (Data safe? should come earlier)
5. No clear decision rule after Quick Check
6. Examples and items still anchored in legal language

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 8

PROS AND CONS

TWO GIRLS, TOO LATE

KEEP

1. The "Quick Check" / "Full Version" two-speed architecture is genuinely innovative: it directly answers p…
2. Outstanding traceability — the Change Log (Appendix E / Table 2) plus the Per-Participant Coverage Check…
3. Move from theoretical, role/timing-coded items in V1 to plain-language, action-oriented items in V2 demo…
4. Audience Notes (replacing rigid P/M/I role codes) help students, interns, and supervisors apply the tool…

FIX

1. All five participants come from law or academia — generality claim is theoretical only.
2. Examples ("a junior lawyer", "Cassazione sentences") still anchor the Full Version in legal scenarios.
3. Quick Check ordering places "Data safe?" at position 3 even though stakeholders said data safety is foun…
4. After the Quick Check, there is no explicit rule for what to do if an answer is "No"/"Unsure": questions…
5. Coding procedure not transparent: who coded, double-coding, conflict resolution, and one raw-comment-to-…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 8

SUGGESTED CHANGES

TWO GIRLS, TOO LATE

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Sample composition	All participants from law/academia; geographic narrowness	Add at least 2 non-legal professionals (engineering, healthcare, technical writing) for a future iteration	High
Quick Check (Q1–Q5)	"Data safe?" appears as Q3 although stakeholders said it should come first	Re-order to: Allowed → Data safe → Disclose → Sources checked → Reviewed	High
Quick Check / Full Version	No decision rule after answering	Add: "If any answer is No or Unsure, stop, revise, ask a supervisor, or use a safer process"	High
Full Version examples	Examples still mostly legal-academic ("junior lawyer", "Cassazione")	Replace or supplement examples with at least one engineering, healthcare, and technical-writing case	High
Coding process	Coding procedure not transparent	Add a short paragraph: who coded, whether codes were double-checked, how disagreements were resolved, and one raw-comment → code → checklist-change e…	Medium
Audience Notes	Notes currently focus on students/interns/supervisors	Expand to cover corporate users, technical implementers, and managerial users	Medium

PRIORITISE HIGH BEFORE POLISH

TEAM 9

CONSENSUS POINTS

IN VINO VERITAI

Balancing Innovation and Identity: Operationalizing AI Ethics Through Co-Designed Checklists for SMEs

1. Checklist items are aspirational directives, not yes/no-answerable questions
2. V1→V2 iteration not traceable at item level; both versions look nearly identical in the appendices
3. Worked use case is narrated but the checklist is never operationalised on Cantina Marino Abate
4. Anonymisation requirement violated (real names of interviewees and companies)
5. No decision logic / scoring / triage mechanism on the general checklist
6. Wine-sector bias seeps into the "general" checklist; differentiation from sector-specific version is weak

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 9

IN VINO VERITAI

PROS AND CONS

KEEP

1. Exceptional empirical effort: 122-respondent survey plus four semi-structured interviews across multiple…
2. Methodologically transparent iteration narrative (V1–V5 in Figure 1), each version tied to explicit data…
3. Critical, non-techno-solutionist sensitivity: black-box risk, data drift, environmental cost, surveillan…
4. Operationalisation of ethical principles into concrete workplace practices, particularly upskilling and…

FIX

1. Final checklist items are written as imperative directives ("Maintain human oversight", "Align AI adopti…
2. Comparing Appendix B (V1) and Appendix D (final) the two are nearly identical: the co-design iteration i…
3. The worked use case for Cantina Marino Abate is descriptive only; the checklist is never actually applie…
4. Real names of interviewees and their companies appear in the appendix, breaking the anonymisation requir…
5. No scoring, no traffic light, no triage rule — every item carries equal weight; no aggregate decision is…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 9

SUGGESTED CHANGES

IN VINO VERITAI

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
General checklist (Appendix D)	All items are imperative directives, not answerable questions	Reformulate every item as an interrogative (Yes/No or short Likert), e.g. "Maintain human oversight" → "Is a designated human supervisor approving AI…	High
Appendix anonymisation	Real names of interviewees and companies are present	Replace with role-based identifiers (e.g. "Interviewee 2: Export Manager, medium-sized winery")	High
V1→V2 iteration	Item-level changes not visible; appendices nearly identical	Add a delta table per version transition: original wording, motivating quote, revised wording	High
Worked use case	Section is narrative only; checklist not actually applied	Insert a structured subsection where each item is answered Yes/No/Partial for Cantina Marino Abate, ending in a documented adoption decision	High
Decision mechanism	No scoring or triage rule	Add a traffic-light (Green/Yellow/Red) or pass-fail rubric mapping checklist outcomes to a clear recommendation	High
General vs winery checklist differentiation	Sector-specific version largely duplicates the general one	Rewrite the winery checklist around viticulture-unique items (biological timing, organic certification, drone regulations) instead of paraphrased gen…	Medium

PRIORITISE HIGH BEFORE POLISH

TEAM 10

CONSENSUS POINTS

SHAKTAR DONETSK

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

1. Small, academia-skewed sample (5 students + 1 professor)
2. Worked use case (university professor) is disproportionately long compared to method/data sections
3. Method / Data Collection sections too brief
4. Final checklist (V2) not shown in the main body; only available via external Google Drive
5. In-class workshops listed as data source but not actually analysed
6. Survey of 109 respondents underused — only briefly mentioned

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 10

PROS AND CONS

SHAKTAR DONETSK

KEEP

1. Workflow diagram in Section 3 is clear, exhaustive, and praised as a model.
2. 109-respondent survey provides a useful double-check on group hypotheses.
3. The ZIP archive with raw datasets, intermediate drafts and notes is methodologically transparent.
4. Section 4.2 cleanly maps V1→V2 changes to specific participants and reasons; Section 4.3 even reports on…

FIX

1. Sample is mostly students from the same institution; only one professor and one true "domain expert" — l…
2. Use case section (Section 5) is excessively long and reads as anecdotal background relative to the metho…
3. Method and Data Collection sections are too short and lightly described, given their importance.
4. Final V2 checklist is not displayed in the paper body — items are referenced by number only and the file…
5. In-class workshops are mentioned as a data source but no V2 change is attributed to them; their role is…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 10

SUGGESTED CHANGES

SHAKTAR DONETSK

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Main body	V2 checklist not visible in the paper	Insert a compact two-column table (V1 item	V2 item) in the bod…
Sample composition	Mostly students and only one professor	Add 2–3 interviews with non-academic workers (corporate employees, programmers, healthcare staff) for the next iteration	High
Use case (Section 5)	Disproportionately long, anecdotal	Cut by ~50% and convert into a structured walkthrough: tasks → checklist answers → decision per task	High
Method / Data Collection sections	Too brief given their methodological weight	Expand with the rationale behind each method, recruitment details, and the procedure for synthesising survey + interviews + workshops	High
Workshops	Described as data source but no change attributable to them	Either document which V2 changes came from workshops (number of participants, themes) or remove them from the data-source claim	Medium
Accountability section	Restructuring announced but no clear section in V2	Add a dedicated Accountability subsection or clarify in the text where each former item was relocated	Medium

PRIORITISE HIGH BEFORE POLISH

TEAM 11

QLEARNERSPOLITO

CONSENSUS POINTS

Trust Calibration of a Software AI Developer in Autonomous Urban Drone Navigation System

1. Generality violation: software/IT-specific language ("deployment", "debugging", "edge cases", "root cause") in the ge…
2. Sample too small / entirely tech professionals (3 software practitioners)
3. V1→V2 change log / item-level traceability missing
4. Title-use case mismatch (drone navigation in title, pedestrian detection in body)
5. Casual / unsuitable academic register in the Abstract ("when things get tricky")
6. Inconsistent response formats (Yes/No, Likert, Completed/Partially)

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 11

QLEARNERSPOLITO

PROS AND CONS

KEEP

1. Addresses a genuine stakeholder gap (developers' perspective on trust, less studied than end-users').
2. Three semi-structured interviews are audio-recorded, fully transcribed, and anonymised; verbatim quotes…
3. The connection between findings and checklist items is explicit (sections 3.1–3.6).
4. Concrete, safety-critical worked use case (pedestrian detection at Politecnico di Torino).

FIX

1. Checklist V2 uses software-engineering language ("debugging", "deployment", "edge cases", "root cause")…
2. All three participants are software professionals; no cross-sector input — the Type 3 question is about…
3. Title says "drone navigation"; body uses "pedestrian detection" — inconsistency never explained.
4. Only V1 and V2 — the minimum — and no item-level diff or change-log table shows what actually changed an…
5. Abstract uses casual language ("when things get tricky") and skips the situation/complication/proposal s…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 11

SUGGESTED CHANGES

QLEARNERSPOLITO

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Checklist V2 items	IT-specific vocabulary embedded throughout	Replace "deployment" with "implementation", "debugging" with "error analysis", "edge cases" with "atypical scenarios"; move purely technical examples…	High
Title / Use case	Title cites drone navigation; body uses pedestrian detection	Either rewrite the title to match the actual use case or restore the drone case throughout	High
Sample composition	All 3 participants are software professionals	Recruit at least 1–2 participants from a non-tech profession (healthcare, public admin, finance) and add at least one further version (V3) driven by…	High
V1→V2 change log	No item-level diff table	Add an appendix table: V1 wording	motivating particip…
Abstract	Casual register; no situation/complication/proposal structure	Rewrite in a formal scientific tone; state the V2 contribution, the target population, and the result of the iteration	High
Response formats	Yes/No, Likert, Completed/Partially mixed without explanation	Standardise to a single 1–5 Likert scale or add a clear legend explaining when each format applies	High

PRIORITISE HIGH BEFORE POLISH

TEAM 12

AI SAID SO

CONSENSUS POINTS

Co-Designing an Ethical Checklist for AI Adoption: Framework from the Workers' Perspective

1. Table 1 vs Table 2 discrepancy — items 1.5 and 1.6 logged as added but missing in the final V4
2. Workshop too small (only 3 peer participants)
3. No decision rule / scoring logic / threshold for the V4 checklist
4. Likert scale anchors poorly matched to the tool's purpose (importance vs compliance)
5. Worked use case (radiology) presented hypothetically rather than as a filled-in checklist
6. Generality only partially demonstrated (radiology bias, limited cross-domain validation)

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 12

PROS AND CONS

AI SAID SO

KEEP

1. Rigorous multi-stage lifecycle: internal prototyping (N=3) → cross-sector survey (N=27) → expert radiolo…
2. Co-design iteration is traceable via Section 6.1 and Table 1 (V1–V4 evolution with explicit stakeholder…
3. Diverse stakeholder pool (38 participants, 8 sectors, mix of students/employees/managers/experts).
4. Affirmative reflective Likert statements force deliberation and mitigate compliance-theatre.

FIX

1. Table 1 logs items 1.5 and 1.6 as added based on executive feedback, but Table 2 (final V4) only lists i…
2. Section 6.1 and Appendix Table 1 disagree on the origin/numbering of item 1.5.
3. Workshop phase had only 3 peer participants — below standard co-design size for early-stage critique.
4. Likert anchors are "1 = not at all important … 5 = essential" — measuring abstract importance rather tha…
5. No decision threshold or scoring logic: a Type-1 tool with no gateway rule.

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 12

SUGGESTED CHANGES

AI SAID SO

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Table 2 (V4)	Items 1.5 and 1.6 logged in Table 1 are missing from V4	Add the two missing items, or document and justify their removal in Section 6.1	High
Section 6.1 vs Appendix A4	Disagreement on item 1.5 origin and numbering	Reconcile the numbering and add a V1→V2→V3→V4 renumbering table	High
Likert anchors (Section 3.2)	"Not important / essential" do not match a compliance tool	Replace with an agreement scale ("Strongly disagree" → "Strongly agree") or a compliance scale ("Not met" → "Fully met")	High
Decision logic	No scoring threshold or gateway rule	Add a rule mapping aggregate responses to "Delegate / Conditional / Do not delegate" outcomes, distinguishing critical from minor items	High
Worked use case (Section 5)	Verdict given hypothetically	Show the V4 filled in item-by-item on the radiology scenario, ending in a documented verdict	High
Cross-domain demonstration	Generality claimed but only radiology is validated	Apply V4 to a profession outside radiology (legal drafting, financial analysis) and report results	High

PRIORITISE HIGH BEFORE POLISH

TEAM 13

IN CHATGPT WE TRUST

CONSENSUS POINTS

Practical Checklist for Responsible AI Integration in Workplace Tasks

1. V1 and V4 are too similar — iteration is largely cosmetic / style-level
2. V2 missing from the appendix as a standalone artefact
3. Final V4 lacks a decision/output mechanism (no scoring, no thresholds)
4. Some V4 items remain abstract / not answerable ("Contextual Explainability", "Resilience and Safeguards")
5. Generality not fully demonstrated; only one industrial design use case
6. Related Work has insufficient or absent citations

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 13

IN CHATGPT WE TRUST

PROS AND CONS

KEEP

1. Clear iterative co-design across four versions, with peer review, storyboard activity, and stakeholder i…
2. Deliberate two-cluster stakeholder design (operational workers vs strategic-managerial) surfaces genuine…
3. The thematic reorganisation from V1 (loose macro-themes) to V4 (Intended Uses, Harms, System & Data, Ove…
4. The 3D parametric modelling use case maps all five clusters to real decisions and gives the tool an hone…

FIX

1. V1 and V4 differ mostly in style; many points (economic feasibility, legality, worker skills) are essent…
2. V2 is not present as a standalone artefact in the appendix — the V1→V2→V3 chain cannot be verified.
3. The "V2 checklist" is actually a storyboard, not a checklist — terminology causes confusion.
4. V4 mixes Yes/No items with open-ended reflective prompts and gives no rule for combining them into a dec…
5. Several V4 items ("Contextual Explainability", "Resilience and Safeguards", "Workplace Well-being") are…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 13

SUGGESTED CHANGES

IN CHATGPT WE TRUST

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
V1 vs V4 comparison	The two versions look nearly identical — iteration looks cosmetic	Rewrite the most representative items so substantive content changes are visible; remove items that did not change	High
Appendix	V2 is missing as a standalone artefact	Add V2 in full, plus a short diff explaining how V1 items were merged/removed/reworded after Activity A	High
V4 / decision output	Mixed Yes/No and open-ended items, no decision rule	Add a final section mapping answers to outcomes (Full automation / Augmentation / No delegation), using a weighted or traffic-light rubric	High
Abstract reflective items	"Contextual Explainability", "Resilience and Safeguards" not directly answerable	Operationalise each with one concrete sub-question (e.g. "Has the team identified at least two failure modes and a fallback?")	High
Profession-specific items	Competitor design data and geometric hallucinations feel sector-specific	Move these into the worked use case in Section 5 and keep the general checklist neutral	High
V2 mislabelling	"V2 checklist" is a storyboard	Rename it "storyboard" and add proper caption explaining task and main takeaway	High

PRIORITISE HIGH BEFORE POLISH

TEAM 14

CONSENSUS POINTS

ME, MYSELF, AI

Co-Designing a Checklist to Assess the Risk of AI-Induced Skill Erosion in Professional Knowledge Work

1. Stakeholder diversity insufficient (4 data analysts/statisticians, recruited via the team's network)
2. Inconsistency between stated rejection of binary formats and the actual presence of binary items (2.1, 2.2, 3.2, 3.3…
3. Items 3.1 and 3.3 are essentially redundant
4. Missing artefacts in appendix (V1, interview/workshop guide, raw notes)
5. No scoring/decision mechanism / interpretation guide
6. Mandatory open notes risk survey fatigue / users skip them

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 14

PROS AND CONS

ME, MYSELF, AI

KEEP

1. Clearly documented two-phase iteration: synchronous peer review + asynchronous expert validation; V1→V2…
2. Successful operationalisation of abstract concepts (automation bias, never-skilling, tacit knowledge los…
3. Genuine generalisation by removing domain-specific anchors (SQL, R) in response to stakeholder feedback…
4. Strong systemic perspective: organisational and managerial factors framed alongside individual cognition.

FIX

1. Only 4 external validators, all data analysts/statisticians from the team's own network — convenience sa…
2. The paper rejects binary formats in theory but retains them in practice (items 2.1, 2.2, 3.2, 3.3, 3.5).
3. Items 3.1 and 3.3 ask essentially the same question (does the organisation reward accuracy over volume?)…
4. Item 1.3 mixes a Yes/No format with a 5-point scale within the same question.
5. V1 is described in the report but not provided in full alongside V2 in the appendix.

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 14

SUGGESTED CHANGES

ME, MYSELF, AI

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Stakeholder pool	All 4 validators are data analysts/statisticians; convenience-sampled	Recruit at least 2 participants from non-data professions (legal, healthcare, communications, HR) and document pilot feedback	High
Binary items (2.1, 2.2, 3.2, 3.3, 3.5)	Contradict the stated rejection of binary formats	Convert all of them into frequency/Likert scales aligned with the rest of the tool	High
Items 3.1 and 3.3	Redundant	Merge into a single item with richer response options, or differentiate (formal policy vs day-to-day managerial behaviour)	High
Item 1.3	Mixes Yes/No and 5-point scale in the same question	Drop the Yes/No portion; keep the 5-point scale	High
Appendix	Missing V1, interview/workshop guide, raw notes	Add all three as standalone artefacts so iteration can be verified	High
Decision logic	No interpretation rubric	Add a simple risk-level rubric (e.g. ≥3 "Never/Rarely" answers in Section 2 → trigger training review)	High

PRIORITISE HIGH BEFORE POLISH

TEAM 15

DOGTOOTH

CONSENSUS POINTS

Factors influencing trust in AI in the healthcare system

1. The final checklist reads as a 44-question research/interview guide, not as an operational decision tool
2. Checklist is too long (44 items)
3. Scoring rubric (Tables 4 and 5) is in the body of the paper but not embedded in the appendix checklist
4. No direct quotes from participants in the main text
5. Worked use case (dentistry) is narrated but the checklist is never actually applied to it
6. Final checklist still contains medical/profession-specific content (Q15 references)

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 15

PROS AND CONS

DOGTOOTH

KEEP

1. 11 participants — exceeds the 3–6 minimum — drawn from healthcare, biomedical engineering, business, and…
2. Clear theme-counting in Table 2 reflects participant consensus rather than individual preferences.
3. The four-dimension scoping-review structure (AI characteristics, human factors, task characteristics, or…
4. Strong conceptual mapping in Tables 4 and 5: theoretical risk conditions converted into mitigation trigg…

FIX

1. Appendix C is a 44-question research/interview guide rather than a functional operational checklist; ope…
2. No direct participant quotes in the main text — the "voice" of co-design is missing.
3. The scoring rubric is in the body (Tables 4 and 5) but completely absent from the actual checklist in th…
4. The dentistry use case (Section 6) is theoretical commentary; the checklist is never filled in for the s…
5. Final checklist still contains medical-specific language (Q15 examples; references to medical data) — br…

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 15

SUGGESTED CHANGES

DOGTOOTH

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Final checklist (Appendix C)	44-item interview-style guide rather than an operational checklist	Extract a 10–15 item operational version with closed/scalar items; move open-ended and demographic questions to a separate Interview Guide	High
Scoring rubric	Tables 4 and 5 are not embedded in the checklist	Print the scoring/decision logic directly on the checklist so a user can compute risk levels in one place	High
Worked use case (Section 6)	Theoretical commentary only — checklist never applied	Provide a filled-in mock checklist for the dentistry scenario, ending with a documented trust-calibration outcome	High
Generality	Profession-specific language (Q15 examples, medical data)	Replace medical terminology with sector-agnostic phrasing (e.g. "sensitive data", "professional standards"); move medical examples to the use case on…	High
Item Q16 / Q17 / open essay items	Open-ended items unsuitable for fast decisions	Convert into binary/scalar items whose answers can trigger high-risk flags or mitigation actions	High
Methodology	Data Collection and Data Analysis not detailed enough	Add clear "Data Collection" and "Data Analysis" subsections describing recruitment, who coded, how themes emerged, and how recurring themes were sele…	Medium

PRIORITISE HIGH BEFORE POLISH

TEAM 16

CONSENSUS POINTS

TRUSTWORTHY

Co-Designing a Responsible AI Checklist: An Operational Framework for Bankruptcy Accountants

1. Lack of empirical evidence: no direct quotes, no item-level evidence of co-design impact, no concrete feedback-to-cha…
2. No item-level V1→V4 change log (delta table missing)
3. References / Related Work section is severely lacking (only one bibliography entry)
4. Italian legal context (CCII) limits the "general" claim
5. Small expert pool (only 2 domain experts)
6. Worked use case is theoretical/narrated; not an actual application of the checklist

THE SHARED DIAGNOSIS

When multiple reviewers flag the same issue, it is no longer “opinion.”

It is the revision brief.

TEAM 16

PROS AND CONS

TRUSTWORTHY

KEEP

1. Highly structured final checklist (Table 3): each item is tied to a lifecycle phase and a responsible ro…
2. Action-verb formulation forces active engagement rather than passive ticking.
3. Strong and unique worked use case: bankruptcy accounting, grounded in real legal/procedural realities (p…
4. Four-stage iteration (V1 theoretical → V2 technical peers → V3 external stakeholders → V4 domain experts…

FIX

1. The co-design process is described but not empirically reported: no participant quotes, no item-level ev…
2. No item-level V1→V4 delta table — only narrative stage descriptions; readers cannot verify what changed…
3. References section contains effectively one entry (the team's own prior deliverable); concepts like "tru…
4. The "intentionally general" claim is undermined by deep grounding in Italian Crisis and Insolvency Code…
5. Only two domain experts in the most senior validation round — small sample for legally defensible claims.

DO NOT DELETE YOUR STRENGTHS. BUILD THE MISSING EVIDENCE AROUND THEM.

TEAM 16

SUGGESTED CHANGES

TRUSTWORTHY

LOCATION	ISSUE	SUGGESTED FIX	PRIORITY
Co-design reporting	No quotes and no item-level evidence of stakeholder impact	Add a delta table per version (V1→V2, V2→V3, V3→V4) with original wording, motivating quote/feedback, and revised wording	High
References / Related Work	Only one bibliography entry; many concepts uncited	Add a Related Work section with citations for trustworthy-AI dimensions, co-design literature, and existing checklists (IEEE EAD, NIST AI RMF, RAI Gu…	High
Generality claim	Checklist deeply anchored in Italian CCII	Either restate the tool as profession/jurisdiction-specific, or add an explicit "generalisation guide" listing items that need adaptation for other r…	High
Section 2.1 identity	Opens with "This scoping review details…" but the paper is a co-design report	Rewrite the opening to clearly state this is a co-design report built on the prior scoping review	High
Worked use case (Section 5)	Theoretical narration only	Demonstrate the checklist applied step-by-step on a concrete bankruptcy case, with explicit per-item answers and final action	High
AI Use Disclosure	Generic ("each section was manually reviewed")	Specify which sections were AI-drafted, what tools were used, and what verification was performed	High

PRIORITISE HIGH BEFORE POLISH