Co-Designing a Reflective Checklist for AI Task Delegation

Sila Gulec
s334743@studenti.polito.it
Politecnico di Torino
Italy

Ahmet Tolga Birand
s335068@studenti.polito.it
Politecnico di Torino
Italy

Giacomo Simbula
s341576@studenti.polito.it
Politecnico di Torino
Italy

Nicolò Fiori
s325013@studenti.polito.it
Politecnico di Torino
Italy

Benedetta Stellato
s341071@studenti.polito.it
Politecnico di Torino
Italy

Pietro de Francesco
s340079@studenti.polito.it
Politecnico di Torino
Italy

Behrad Galedari
s324382@studenti.polito.it
Politecnico di Torino
Italy

Kutay Gunel
s307467@studenti.polito.it
Politecnico di Torino
Italy

Jacopo Caravaggio
s339988@studenti.polito.it
Politecnico di Torino
Italy

Gabriele Gallo
s339929@studenti.polito.it
Politecnico di Torino
Italy

Abstract

As organizational tasks increasingly shift toward automated artificial intelligence systems, workers face the complex socio-technical challenge of deciding whether to fully or partially delegate their responsibilities. Our foundational scoping review synthesized this decision-making process into five core themes: Characteristics and Properties of the Task, AI Performance, Cost and Efficiency, Accountability, and Human-AI Collaboration. Abstract guidelines are notoriously difficult to operationalize, leaving workers without the necessary infrastructure to make safe, informed delegation decisions on the ground, and often leading to basic checklists that encourage passive compliance rather than critical reflection.

To bridge this gap and create a tool that effectively integrates into daily workflows, we recognized that stakeholders must be directly involved. An interactive co-design was carried out involving six participants from three stakeholder groups, including software practitioners, business leads, and legal and compliance professionals, to capture complementary operational and governance perspectives and ensure a balanced perspective. Using the developed checklist as a starting point for discussion, we were able to collect necessary feedback to identify problems in operation and improve our criteria. The resulting general decision tool moves beyond simple box-ticking, acting instead as a reflective guide designed to spark essential team discussions before an AI delegation choice is made. Refined through feedback, the resulting decision tool employs a two-tiered architecture, including a 5-item quick checklist. Finally, we demonstrate the practical value of this general checklist by applying it to a worked use case: a computer programmer using LLM-powered coding assistants like GitHub Copilot.

1 Background and Related Work

1.1 AI Task Delegation

AI task delegation consists of the partial transfer of a work activity from a human operator to an AI system. This process aims to support or automate specific tasks such as output generation, decision-making support, summarizing of information, code writing, text drafting or formulation of recommendations. This dynamic should not be considered a binary choice; rather, it develops along a spectrum of increasing autonomy. The main configuration levels include no delegation, AI assistance only, partial delegation integrated with a human review, and routine delegation associated with a lightweight verification for more standardized tasks. Indeed, choosing the correct delegation mode depends on the risk level of the operation, the characteristics of the task, the available verification protocols (to objectively check output correctness), and the management of final accountability (determining who holds liability).

1.2 Checklists for Responsible AI

Ethical principles of AI are abstract and difficult to operationalize. Checklists are created to bridge this gap, but they risk reducing complex problems to a deceptive exercise of pure compliance or box-ticking [2]. The RAI Guidelines paper further demonstrates that simpler, shorter checklist formats produce higher usability scores among practitioners [4]. In the context of AI governance, this paper focuses on developing a "Type 1 checklist", a tool designed to evaluate general task characteristics for AI delegation. A truly useful checklist must not provide pre-packaged answers, but rather stimulate discussion and critical reflection among practitioners [2]. Following this approach, the proposed tool does not act as a scoring evaluation system, but as a reflective guide to support decisions within the workflow.

1.3 Co-Designing Practical AI Governance Tools

Co-design is a collaborative approach that actively brings stakeholders into the creation of a tool. In AI governance, this method is essential for translating abstract concepts into practical resources. Without the direct involvement of practitioners, new frameworks risk being ignored or incorrectly applied [2]. As highlighted by research on AI Impact Assessment templates, collaboration with experts allows for the identification of operational obstacles and aligns tools with real workflows [3]. Engaging with stakeholders ensures that the criteria of this checklist are understandable, actionable, and effectively integrated into daily work dynamics [2]. In this study, practitioners, organizational leads, and legal and compliance professionals were interviewed separately to prevent seniority effects from influencing responses.

1.4 Starting Point: Scoping Review

The conceptual starting point of this work is a prior scoping review conducted by the authors [1]. In that document, literature was synthesized into five fundamental themes that regulate and define the boundaries of AI delegation: Characteristics and Properties of Task, Cost and Efficiency, Accountability, AI Performance and Human-AI Collaboration [1]. This thematic classification provides the structural foundation for the project: the five themes were explicitly translated into the initial set of criteria to draft Checklist V1 (Appendix A). Furthermore, the worked use case of the programmer using LLM-based code assistants derives from the same exploratory investigation [1], providing a continuous theoretical thread from the literature review to practical application.

2 Data Collection

Checklist V1 had been improved through an in-class peer session, while professional feedback was collected through six individual semi-structured interviews, each approximately 20 minutes in duration, following the co-design format of short, focused sessions across multiple stakeholder groups. The purpose of the interviews was to gather stakeholder feedback on Checklist V2, and to understand how each participant evaluated the practical risks and responsibilities involved in delegating tasks to AI. Figure 1 summarizes this process from the scoping review to Checklist V3.

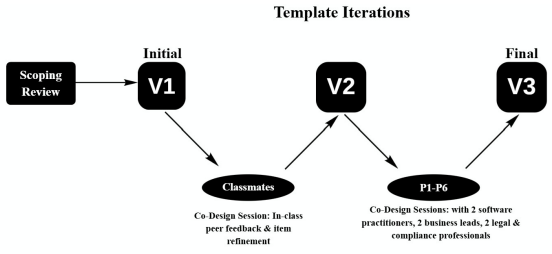

Figure 1: Co-design process from scoping review to Checklist V3.

2.1 From Scoping Review to Checklist V1

Checklist V1 was developed by translating the five themes identified in the scoping review directly into the checklist items. The five themes were Task Characteristics and Properties, AI Performance, Cost and Efficiency, Accountability, and Human-AI Collaboration. These themes were converted into a set of yes/no questions covering the key considerations within each theme. The checklist was designed as a general decision support tool and was not tailored to a specific profession or organizational context.

2.2 Checklist V1 to Checklist V2

The first revision iteration involved presenting Checklist V1 to peers in class, which served as a preparatory step shaping Checklist V2, which was later brought to professional stakeholders. During the session, the most challenging and important items were identified and discussed. This preparatory step helped clarify ambiguous wording and identify items that required restructuring before the checklist was presented to professional stakeholders. In Checklist V2, the binary format was replaced with four response options: Addressed, Partially Addressed, Not Yet Addressed, and Not Applicable. This shift was a deliberate team decision to prevent passive engagement.

2.3 Participants

Six participants were recruited from three stakeholder groups to capture complementary perspectives on AI task delegation. Group A consisted of software practitioners with direct experience using AI tools in professional development work. Group B consisted of AI business leads and consultants with organizational oversight of AI implementation. Group C consisted of legal and compliance professionals with expertise in corporate law and HR governance. The participants were kept separate during data collection to reduce the risk that differences in seniority or organizational role would suppress honest feedback (For the full participant table refer to Appendix B).

2.4 Procedure

Before the interviews, each participant received Checklist V2, the worked scenario and the structured interview questions (Appendix F) at least 24 hours in advance. Each interview began with a short icebreaker question about the participant's prior experience with AI-assisted task delegation. Participants then reviewed Checklist V2 section by section and responded to structured questions about clarity, overlap, missing items, and practical relevance. Group A and Group B were asked to apply the checklist to the two-task coding scenario: a low-risk sorting function (Task A) and a higher-risk login authentication task (Task B), to examine whether the checklist could help users determine the delegability of tasks. As the coding scenario was less directly relevant to legal and compliance work, Group C participants were asked to apply the checklist to a non-technical task, drafting a formal written warning letter to an employee based on a submitted incident report (Task C) to assess whether the checklist criteria remained meaningful outside a programming context.

2.5 Materials

Materials used during the co-design process included Checklist V2 (Appendix A), a short summary of the five themes from the scoping review, the semi-structured interview guide, the two-task

worked scenario, and the structured feedback questions. For Group C, the worked scenario was replaced by a non-technical Task C. The appendix includes the complete session guide, all checklist versions, and anonymized interview notes from all participants.

3 Data Analysis

This section documents how participant feedback was used to revise Checklist V2 to Checklist V3. After each feedback session, thematic analysis was performed on the participant feedback and common themes for revision emerged from the comments. The feedback was then directly translated into changes made to the checklist. Feedback was coded into recurring themes by grouping similar concerns raised across participants. Each theme was then traced to a specific checklist revision, ensuring that every change in V3 could be linked back to at least one participant statement. Minor language clarifications were applied directly without a separate theme. The primary issues that were revised included task context, verifiability, domain expertise, time pressure, risk classification, accountability, data protection, and escalation. These changes were made to transform Checklist V2 from a general reflective guide into a more operational decision-support tool.

3.1 Feedback and Thematic Analysis

Task readiness: context, verification, expertise, and pressure. The most common finding across all participants was that safe delegation depends on conditions surrounding the task, not only on the task's own description. P1 specified that the AI must have access to the operational context needed to complete the task correctly, including documentation, organizational conventions, approved tools, and knowledge bases, and that AI self-reports of success should not be trusted without independent testing. P2 identified verifiability as the most important checklist item, but also warned that a worker may wrongly believe that an output is verifiable simply because they can read it, while still lacking the expertise to detect any possible hidden errors. This concern applies beyond software to any task involving legal rights, privacy, HR decisions, financial implications, health, or safety. P2 also noted that overtrust cannot be self-assessed in advance, and that a checklist should instead identify observable conditions that make shallow acceptance more likely: time pressure, cognitive load, organizational pressure, lack of expertise and reliance on AI self-reports. P6 made a similar point, stating that employees should not "over-rely on AI tools without checking the output," especially when AI is used for creative work or idea generation rather than more deterministic tasks. V2 treated overtrust as a general risk, but V3 reframes it through these observable conditions. V3 also adds a time-pressure item, asking whether the worker is under conditions that may lead them to accept an okay-looking AI output, without deep review. These findings produced four changes in V3: an expanded task definition item that includes context access, a strengthened verifiability item requiring a concrete verification method, a new domain-expertise condition, and a separate time-pressure item.

Abstract phrasing, item overlap, and verification cost. Different participants found that some V1 items were too abstract or overlapping.

P2 noted that the tool-task fit lacked clear guidance, and that assessing accuracy evidence was difficult for individual workers since such data typically exists at the organizational level, instead of individual level. P2 also pointed out that verifiability and verification cost could appear similar to each other. To solve this issue, V3 separates these clearly: verifiability asks whether errors can be detected at all, while verification cost asks whether double-checking the output is actually worth the time, or manually checking is faster.

Risk classification and worst-case consequences. Participant feedback indicated that V2 lacked a clear and practical way to classify risk. P2 stated the importance of reversibility and examples to worst-case consequences, noting that an undetected error may have very different consequences, ranging from simple rework to leaked credentials, lawsuits, or production incidents. P3 recommended an explicit low, medium, and high risk classification. P4 and P5 further specified that tasks involving personal data, legal rights, employment decisions, financial implications, health, safety, or confidential business information should be treated as high risk by default. V3 therefore adds a risk classification step before delegation.

High-risk categories: confidentiality and people-related decisions. Two specific high-risk categories were absent from V2 and added in V3 based on professional stakeholder feedback. P4 argued that before using AI, the worker must check whether the task involves personal information, trade secrets, client data, or legally protected information, and that if the AI tool has not been approved for such data, delegation should not proceed. P5 extended this to the HR context, warning that tasks involving hiring, promotion, performance evaluation, salary, discipline, termination, surveillance, or employee wellbeing must be classified as high risk. In these contexts, AI may support structuring or drafting, but it should not be used to finalise decisions about people without competent human review.

Accountability and escalation. Accountability was one of the strongest themes across the feedback. P1 argued that workers still remain fully responsible for the final output even when the usage of AI is explicitly stated. P4 stated that "the AI was mistaken" is not a legally adequate explanation, and P5 made the same point in the HR context.

V2 referred broadly to developers, deployers, end users, and legal gaps, making responsibility too abstract for organizational use. In order to solve this, V3 requires a specific person or role to be identified as responsible for checking, approving, documenting, and explaining the AI-assisted output. On escalation, P4 argued that escalation paths must name exact roles such as the legal department, HR manager, data protection officer, or compliance officer, and P5 confirmed that employees must know which department to contact when AI output may be harmful or inappropriate. V3 makes escalation mandatory for high-risk or uncertain tasks and requires the escalation role to be named explicitly.

Decision rules and checklist actionability. The feedback also showed that V2's status categories did not clearly lead to a decision. Participants could record whether items were addressed or not, but the checklist did not explain what to do with those answers. This weakness was consistent with the lecture guidance that a checklist should lead to stop, revise, delegate, or escalate decisions. V3 therefore adds gateway rules: if a critical item is not yet addressed, the delegation process should stop; if multiple items are only partially

addressed, the task should be escalated; if the task is high risk and no competent reviewer is available, it should not be delegated. This makes the checklist actionable compared to the previous one, which was mainly reflective.

Checklist V2 to Checklist V3. Ten substantive changes are the direct result of participant feedback. Each change below identifies what was missing in V2, which participants raised the issue, and how it was addressed in V3

Change 1: Operational context access added (P1): V2 defined the task by inputs, outputs, and success criteria, but did not assess whether AI had the operational context necessary to successfully complete the task. V3 item 1.1 includes a context-access requirement, which asks if the AI has relevant resources. Delegation should not take place if the AI lacks the context required to do the task without guessing.

Change 2: Verifiability strengthened, domain expertise added, and verification cost separated (P1, P2): V2 treated verification as a general ability to check output with reasonable effort V3 item 1.3 requires the worker to specify a particular verification technique, item 1.4 adds a domain-expertise check, asking the worker to confirm that they have the background to recognize non-obvious mistakes. Verification cost is kept as a separate item, asking if inspection of the output still makes delegation beneficial.

Change 3: Time pressure and cognitive load made into a formal risk criterion (P2): V2 included time constraints generally but did not treat them as a risk factor that can compromise the quality of the review. V3 item 1.5 asks if the worker is under time constraints, cognitive load, or organizational pressures that may lead them to accept a plausible- looking AI output without rigorous review.

Change 4: Overtrust redefined through observable risk factors (P1, P2, P6): V2 asked workers to assess whether they might overtrust AI, which was difficult to judge in advance. P6 stated that employees over-rely on AI tools without verifying the results. V3 replaces this with observable indicators such as a lack of verification, a lack of domain expertise, time constraints, and pressure to accept faster outputs. This makes the checklist more practical because it no longer requires employees to identify their own bias explicitly.

Change 5: Risk classification added before delegation (P2, P3, P4, P5): V2 included high-risk accountability items, but it lacked a clear mechanism to classify risk initially. V3 includes a first risk triage to determine whether the task contains personal data, confidential information, or permanent harm.

Change 6: Accountability items made more concrete and role-based (P4, P5): V2 used abstract accountability language, including developers, deployers, end users, legal loopholes, and reliable human parties. V3 requires the user to specify the specific person or role in charge of the AI-assisted output. This change incorporates the feedback that the "AI made a mistake" is not an adequate accountability model.

Change 7: Escalation path made mandatory and specific (P3, P4,P5): V2 asked whether a clear legal escalation path existed, but the wording made escalation appear as an extra safety rather than a requirement for risky tasks. V3 requires escalation whenever the task is high-risk, unknown, or outside the worker's expertise. The checklist now needs the specific escalation role to be identified.

Table 1: Quick check for daily AI delegation decisions

#	Question	Status
1	Is the task clearly defined with known inputs, outputs, and success criteria?	A/PA/NYA/NA
2	Can the output be independently verified before use?	A/PA/NYA/NA
3	Does the task involve personal data, legal rights, employment decisions, or confidential information?	A/PA/NYA/NA
4	Is a specific person or role identified as responsible for the final output?	A/PA/NYA/NA
5	Is the worker free from time pressure or cognitive load that could prevent proper review?	A/PA/NYA/NA

Change 8: Confidentiality, privacy, and protected data added as explicit blockers (P4,P5): V2 did not include a section on data protection before using AI tools. V3 includes a confidentiality and data-use item, which requires the worker to determine whether the task involves protected information. If such data is involved and the AI tool has not been approved for that purpose, delegation should not proceed.

Change 9: HR and people-related decisions explicitly classified as high risk (P5): V2 did not distinguish workplace decisions about individuals as separate risk categories. P5 emphasized that such tasks carry particular risks that general high-risk items did not capture. V3 treats hiring, performance evaluation, discipline, and termination tasks as high risk.

Change 10: Redundant verification and accountability items merged or repositioned (P2, P3): V2 had overlapping items on verifiability and verification cost, and accountability. V3 separates these concepts more clearly, reduces redundancy, and gives each item a specific purpose.

Checklist V3. Checklist V3 is organized into five sections: Task Characteristics and Delegation Readiness, AI Performance and reliability, cost, Efficiency and Practical Value, Accountability, Risk, and Privacy, and Human Oversight and Escalation. Each item is answered using four response options: Addressed, Partially addressed, Not Yet Addressed, or Not Applicable. The full checklist is provided in Appendix A; this section presents the quick check, response options, decision rules, and final delegation outcomes that explain how the checklist is applied.

Quick Check: For daily use, a five-item Quick check is provided below.

(Quick Check: Use this before delegating any task to AI. If any item is marked NYA, stop and use the full checklist.)

Decision Rules: For the purposes of these rules, a task is considered high risk if it involves sensitive data, confidentiality, or other serious consequences. Critical items include all items in Section 1 and any items concerning privacy, confidentiality, accountability, or escalation:

- If any item in Section 1 is marked NYA, do not delegate yet.
- If any item on privacy, confidentiality, high-risk impact, accountability, or escalation is marked NYA, do not delegate yet.
- If three or more items are marked PA, escalate to a competent reviewer before delegation.

- If the task is high risk and no competent reviewer is available, do not delegate.
- If all critical items are A or justified NA, delegation may proceed at the level of review indicated by the checklist.

Final Delegation Outcome: After completing the checklist, select one outcome that reflects the overall picture.

Do not delegate: Use if any key items are marked Not Yet Addressed, the task is high risk in the absence of a competent reviewer, or protected data will be entered into an unapproved AI tool.

Revise before delegation: Use this if the task can be delegated but the context, verification mechanism, responsibility, or escalation path are incomplete. Before proceeding, address any gaps.

Partial delegation with review: Use this if AI can help with authoring, structuring, summarizing, or generating possibilities, but the final result must be reviewed and approved by a competent person before it is used.

Routine delegation with lightweight verification: Use this only for low-risk, reversible, clearly defined, and verifiable jobs in which the AI knows the essential context and the worker can reliably verify the outcome.

4 Worked Use Cases: Coding Assistants

We apply checklist V3 to the scenario introduced in our prior review, a programmer using an LLM-powered coding assistant for two tasks. The checklist produces different delegation outcomes for each task, showing that the same tool and worker face different delegation decisions depending on task characteristics.

4.1 Low-Risk Coding Task

Scenario: A developer uses an LLM-powered coding assistant to write a function that sorts a list of customer names alphabetically.

Checklist Assessment: This task has precisely defined inputs, outputs, and success criteria (Item 1.1) and is highly structured and deterministic (Item 1.2). The generated output can be evaluated using a comparison against expected results (Item 1.3), and the verification cost is significantly lower than completing the task manually (Item 3.1). The worker is not dealing with personal data, legal rights, or confidential information (Item 4.1), and so no expert escalation is needed (Item 5.2). Furthermore, the undetected error is minimal and easily reversible, requiring only minor revision.

Decision: Routine delegation with lightweight verification. In this case, the task can be delegated routinely as the output is easy to check, and the time saved outweighs the effort required for verification.

4.2 High-Risk Coding Task

Scenario: The same developer uses the same coding assistant to implement a login and authentication system that stores and manages user data.

Checklist Assessment: This task raises several high-risk conditions. The developer is implementing a login system that collects, stores, and manages personal data, which creates security and compliance risks (Item 4.1). As the developer has not previously worked with these requirements, the success criteria may be difficult to define in advance (Item 1.1), and the output cannot be reliably verified without relevant technical expertise (Item 1.3). The worker may also lack the domain expertise needed to identify non-obvious security errors (Item 1.4), while AI-generated code may appear plausible even when it contains serious errors (Item 2.2). A specific person or role must therefore be responsible for the final output (Item 4.4), and the escalation path should clearly name the relevant security, legal, or compliance role (Item 5.2). Before use, the output must be reviewed by a competent person with appropriate expertise (Item 5.1). Under deadline pressure, the risk of accepting an insufficiently reviewed AI output increases (Item 1.5).

Outcome: No full delegation, only limited AI assistance with mandatory expert review. AI can support drafting or suggesting options, but the final system requires human oversight and expert approval because of the operational, security, and compliance risks involved.

5 Discussion and Limitations

5.1 Practical Contribution

The main contribution of this study is Checklist V3 which is a co-designed decision framework that enables workers and organizations to determine whether a task can be delegated to an AI system and under which conditions and safeguards. Instead of a structure of full or no delegation, Checklist V3 situates the delegation process along structured categories, from no delegation through partial AI assistance to routine delegation with lightweight verification. The appropriate level of delegation is determined by a systematic assessment of task context, verifiability, domain expertise, time pressure, risk classification, accountability, data protection and escalation readiness. By maintaining this process, the checklist translates abstract responsible AI principles into specific, concrete and traceable delegation decisions.

5.2 Checklist as Reflection

V3 was explicitly designed to prevent shallow compliance. Instead of producing an automated score or a safety certificate, it requires users to justify their delegation decision in depth, asking for concrete proof, and also requires identifying missing safeguards, and escalating uncertain cases to the appropriate role or department. The checklist is also not only reflective, since feedback from various participants also demonstrated that users need clear and specific guidance on what to do when answers are incomplete. Checklist V3 transforms the input of checklist responses into one of four outcomes: do not delegate, revise before delegation, partial delegation with review, or routine delegation with lightweight verification.

5.3 Organizational Relevance

The findings also demonstrate that responsible AI delegation cannot only depend on individual judgment. The feedback consistently showed that AI use is shaped by many different factors: informal organizational norms, time pressure, and the absence of formal policy rather than deliberate governance decisions. In order to bring these into the ground, Checklist V3 functions as a diagnostic instrument for identifying where organizations need clearer definitions on their policies, and provides approved-tool guidance, data-protection protocols and escalation procedures, rather than functioning only as a task-level decision aid.

5.4 Applicability beyond software: Non-Technical Use Cases

While initially tested with coding assistants, the practical contribution of the checklist extends to non-technical disciplines, demonstrating that the criteria remain fully operational when evaluated outside a software context. For instance, to assess its wider relevance, the checklist was applied to Task C: Using an AI tool to draft a formal written warning letter to an employee based on a submitted incident report. In this scenario, while the task has clear boundaries , the checklist prompts necessary critical reflection by highlighting that the user must verify the output against current labor laws, which requires specific domain expertise. Furthermore, because the resulting document governs employee discipline, it automatically triggers a high-risk classification and necessitates careful data protection checks regarding the employee's personal data. Consequently, the tool successfully guides the user away from full automation and toward partial delegation with review: the AI can be used to generate the structural draft, but mandatory human oversight and review from the legal or compliance department is strictly required before adoption.

5.5 Limitations

This study has several limitations. The co-design process involved a small group of six participants, limiting the generalization process of the findings. The checklist was then refined through structured feedback sessions, rather than being tested in a real workplace over a long period. This means there is no guarantee on how it will hold up under heavy time pressure or daily work constraints. Finally, the checklist was only tested on software development and one basic HR task. Other fields like healthcare, finance, education and law will likely need more domain-specific safeguards.

5.6 Future Work

Future research should evaluate Checklist V3 with a larger and more diverse group of professionals under multiple sectors, including legal, compliance, data protection, healthcare, and financial services. To determine whether the checklist actually makes an impact on delegation behaviour, reduces overreliance on AI, improves escalation practices, and helps organizations identify structural gaps in their AI governance frameworks, long-term field studies will be needed.

6 Conclusion

Ensuring AI is used safely at work requires more than just high-level ethical principles. This paper proposes an operational checklist based on five theoretical dimensions of AI delegation. By co-designing this checklist with software developers, business leaders, legal and compliance professionals, we created a two-tiered decision tool, for making more deliberate decisions regarding AI delegation. Our tool has been refined through peer review and is equipped with explicit decision rules. Finally, we demonstrated its value through a real-world scenario of a programmer evaluating GitHub Copilot. While future research must test this tool across multiple diverse industries and track long-term effects, our co-designed checklist provides a strong starting point to help professionals set clearer boundaries in terms of delegating tasks to AI.

References

[1] Gulec, S., Birand, A. T., Simbula, G., Fiori, N., Stellato, B., de Francesco, P., Galedari, B., Gunel, K., Caravaggio, J., & Gallo, G. (2026). Practical criteria for AI task delegation: A scoping review of task complexity, performance reliability, and regulatory compliance. Politecnico di Torino.

[2] Madaio, M. A., Stark, L., Vaughan, J. W., & Wallach, H. (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 (CHI '20), pp. 1–14. ACM. https://doi.org/10.1145/3313831.3376445

[3] Bogucka, E., Constantinides, M., Scepanovic, S., & Quercia, D. (2025). Co-designing an AI impact assessment report template with AI practitioners and AI compliance experts. In Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society (AIES '24), pp. 168–180. ACM. https://doi.org/10.5555/3716662.3716678

[4] Constantinides, M., Bogucka, E., Quercia, D., Kallio, S., & Tahaei, M. (2024). RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. In Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '24). ACM, New York, NY, USA.

Appendix

Appendix A – Checklists

Checklist V1:

Task Characteristics and Properties: The team has described the task's inputs, expected outputs, and success criteria in simple language. Including whether the worker is already familiar with this type of task, or encountering it for the first time (unfamiliar tasks require more careful review of AI output).

[ ] Yes - [ ] No

The team has assessed whether the task is structured enough for the AI tool's capabilities.

[ ] Yes - [ ] No

The worker can verify the AI's output within a reasonable effort. For example, by inspecting the result against known criteria or running a simple test. If verification requires specialist knowledge the worker does not have, the team has identified who will provide that review.

[ ] Yes - [ ] No

If the task will be performed under time pressure, the team has considered how this effort affects the worker's ability to review AI output critically.

[ ] Yes - [ ] No

AI Performance:

The team has measured or reviewed available evidence on the AI's accuracy for this specific task type such as published benchmarks, internal test results, or documented error rates.

[ ] Yes - [ ] No

The team has checked for signs that workers may over-trust the AI, for example, by comparing a sample of AI outputs against verified correct answers and asking workers to predict the AI's error rate before revealing it.

[ ] Yes - [ ] No

The system makes errors easy to spot including confident-sounding but incorrect outputs, also called as AI hallucinations. The team has considered how the interface helps workers distinguish reliable output from plausible-looking mistakes.

[ ] Yes - [ ] No

The team does not assume that providing AI explanations alone prevents over-reliance.

[ ] Yes - [ ] No

Cost & Efficiency:

The team has estimated the time needed to verify the AI's output for this task.

[ ] Yes - [ ] No

The team acknowledges that AI output should always be reviewed before use. Delegation is pursued when the combined time of generation plus review is substantially lower than performing the task manually; but review is never skipped, even when time savings are large.

[ ] Yes - [ ] No

The organization has a plan to prevent workers from gradually losing the skills needed to do this task without AI; for example, by scheduling periodic manual practice, rotating AI-free work periods, or tracking skill assessments over time.

[ ] Yes - [ ] No

Accountability:

The team has identified who bears responsibility (developer, deployer, end-user) for errors in the AI's output for this task. The identification is done in writing and this assignment has been communicated to everyone before work begins.

[ ] Yes - [ ] No

There is no scenario where an AI-generated error falls into a gap with no accountable party.

[ ] Yes - [ ] No

The team has checked whether this task falls under a high-risk regulatory category.

[ ] Yes - [ ] No

For high-stakes tasks, documented human review occurs before AI outputs are deployed.

[ ] Yes - [ ] No

Workers have been explicitly told what they are personally responsible for when using the AI's output.

[ ] Yes - [ ] No

The team has clarified in which situations the AI system itself; rather than its developer, deployer, or user; could be considered a contributing factor to harm, and how existing liability structures address this.

[ ] Yes - [ ] No

Human-AI Collaboration:

Humans and AI work together. The AI generates suggestions, but the worker decides whether to accept, modify, or reject them. The team has defined clear boundaries; which decisions remain fully human and where AI input is advisory only.

[ ] Yes - [ ] No

The team has assessed the worker's AI literacy level, with particular attention to workers who have enough familiarity to trust the tool but not enough expertise to catch its errors. For instance, by asking workers to rate their confidence in detecting AI errors on a simple scale, or by reviewing how they handled AI output in a recent task.

[ ] Yes - [ ] No

Workers have received training not just on how to use the tool, but on how to critically evaluate its outputs for this task.

[ ] Yes - [ ] No

The workers can edit, adjust, and override the AI's output.

[ ] Yes - [ ] No

The organization allocates dedicated time for workers to review AI outputs; review is not treated as an afterthought.

[ ] Yes - [ ] No

Table 2: Checklist V2: revised checklist used during professional stakeholder feedback sessions

Section and Theme	Checklist Item	Status (A/PA/NYA/NA)
Task Characteristics & Properties		
Task Definition	Describe the task's inputs, expected outputs, and success criteria in plain language	
Task-tool fit	Determine whether the task is structured appropriately for the AI tool's capabilities.	
Verifiability	Consider whether the worker can check the AI's output with a reasonable effort.	
Time Constraint	If the task will be completed under time constraints, examine how this will affect the worker's capacity to critically evaluate AI output.	
AI Performance		
Accuracy Evidence	Examine the available evidence regarding the AI's accuracy for this particular task type.	
Error Distinguishability	Consider how the interface assists workers in distinguishing reliable output from plausible-looking errors.	
Overtrust	Determine whether employees may overtrust the AI.	
Cost & Efficiency		
Verification Cost	Estimate the time required to verify the AI's output for this task.	
Before Use Evaluation	Confirm that every AI output will be evaluated before usage.	
Skill Erosion	Ensure that the organization has a plan in place to avoid workers from gradually losing the abilities required to complete this task without AI.	
Accountability		
High-risk task assessment with regulatory bodies	Monitoring carefully if the task falls under the region's regulatory bodies as a high-risk task.	
Error responsibility delegation	Before the outputs are put into action, are the legal responsibilities communicated to developers, deployers, end users?	
Clearing the legal gaps	Is it made clear at each step, which or whom is the accountable/supportive party.	
Human double check on high-stake tasks	Are legally/ethically/technically high-stake decisions/outputs double-checked by a reliable human party?	
clear legal escalation path	In tasks where small errors are still possible and anyways delegated to A.I, is the case escalation path clear enough? Is it communicated?	
Human-AI collaboration		
Over-reliance prevention	Are the necessary guidelines provided in terms of degree of reliance on AI	
Employer/deployer AI literacy	Is it made sure that the worker/organization have the necessary knowledge on the particular collaboration that they are having with that specific AI model	
Existence of necessary commanding tools	Are there reliable control means provided for collaborating part especially in case of possible errors	
Taking in account Efficient output reviews	Is it made sure that depending on the situation, the outputs are at least periodically reviewed?	
Peer review for collaborating with AI	Are there periodic evaluation and peer reviews of colleagues proficiency and reliability on their human AI collaborations?	

(Table: Checklist version 2. The status abbreviations are respectively Addressed, Partially Addressed, Not yet Addressed, Not Applicable)

Table 3: Checklist V3: final co-designed checklist after stakeholder feedback

ID & Theme	Checklist Item	Status
Section 1: Task Characteristics and Delegation Readiness		
1.1.	Has the task been clearly defined, including the required input, the expected output, success criteria, and any organizational or external information the AI would need?	A/PA/NYA/NA
1.2.	Is the task suitable for the chosen AI tool considering whether it has clear boundaries and can be objectively checked?	A/PA/NYA/NA
1.3.	Is there a concrete way for the worker to independently verify the AI output such as trusted sources, standards, comparison, or expert review?	A/PA/NYA/NA
1.4.	Does the worker have enough domain expertise to identify non-obvious errors in the AI output?	A/PA/NYA/NA
1.5.	Is the worker operating under time pressure or cognitive load that might lead them to accept AI output without proper review?	A/PA/NYA/NA
Section 2: AI Performance and Reliability		
2.1.	Is there reliable evidence, guidance, previous testing, or accepted practice showing that this AI tool can perform the task successfully?	A/PA/NYA/NA
2.2.	Are potential AI errors visible enough for the worker to detect before the output is used?	A/PA/NYA/NA
2.3.	Will the worker avoid relying only on the AI's own claim that output is correct, tested, complete, or safe?	A/PA/NYA/NA
Section 3: Cost, Efficiency, and Practical Value		
3.1.	Does verifying the output take significantly less time and effort than completing the task manually?	A/PA/NYA/NA
3.2.	Will the AI output be reviewed before it is shared, implemented, submitted, or used?	A/PA/NYA/NA
3.3.	Is there a plan to prevent workers from losing the ability to perform or evaluate this task without AI?	A/PA/NYA/NA
Section 4: Accountability, Risk and Privacy		
4.1.	Does the task involve critical areas such as personal data, confidentiality, employment, finances, health, safety, security, or other serious consequences?	A/PA/NYA/NA
4.2.	Is there evidence that the AI tool is approved for the type of personal, client, employee, confidential, or legally protected data used in the task?	A/PA/NYA/NA
4.3.	Is the task related to recruitment, promotion, performance evaluation, salary, discipline, termination, surveillance, complaints, or employee wellbeing? If yes, it should be considered high risk.	A/PA/NYA/NA
4.4.	Has a specific person or role been identified as responsible for the final AI-assisted output or decision?	A/PA/NYA/NA
4.5.	Is it clear who reviews, approves, documents, and can explain the AI-assisted output if questioned later?	A/PA/NYA/NA
Section 5: Human Oversight and Escalation		
5.1.	For high-risk tasks, has the output been assigned to a competent reviewer with the relevant legal, technical, safety, privacy, or domain expertise?	A/PA/NYA/NA
5.2.	Is there a specific escalation path naming the exact role or department to contact if an issue occurs?	A/PA/NYA/NA
5.3.	Are there practical solutions to stop, correct, audit, reverse, or document the AI-assisted work if an error is found?	A/PA/NYA/NA
5.4.	Does the organization know how the AI-assisted output was produced, reviewed, approved, and then deployed?	A/PA/NYA/NA
5.5.	For repeated or routine AI use, is there a process for periodically reviewing outputs, worker reliance, errors, and escalation cases?	A/PA/NYA/NA

(Table: Checklist version 3. The status abbreviations are respectively Addressed, Partially Addressed, Not yet Addressed, Not Applicable)

Appendix B – Participant Table

Table 4: Participants IDs, titles, experiences, stakeholder groups, and roles in co-design

Participant ID	Job Title	Experience	Stakeholder Group	Role in Co-Design
P1	CTO of a Tech Startup	2 Years	Group A: Software practitioners	Reviewed checklist clarity and usability from a technical decision-maker perspective.
P2	Freelancer Software Engineer	3 Years	Group A: Software practitioners	Commented on practical usability and real-world applicability of checklist criteria.
P3	Consultant in Strategy & Innovation Development	11 Years	Group B: AI business leads and consultants	Commented on governance, accountability, risk criteria, and missing dimensions.
P4	Corporate Lawyer	30 Years	Group C: Legal and compliance professionals	Reviewed accountability, legal responsibility, escalation paths, and data protection items from a corporate law perspective.
P5	HR Manager (retired)	35 Years	Group C: Legal and compliance professionals	Reviewed people-related risks, fairness, confidentiality, and escalation items from an HR governance perspective.
P6	CEO/Owner of a Tech Startup	3 Years	Group B: AI business leads and consultants	Applied checklist from a product-ownership and business leadership standpoint.

(Table: Participants' IDs, titles, experiences, stakeholder groups, and roles in Co-Design; for more information refer to Appendix H)

Appendix C – Co-Design Session Guide

Individual Interview Guide

Table 5: Co-design interview method summary

Method	Six individual semi-structured interviews were conducted, approximately 20 minutes each.
Participants	6 participants from three stakeholder groups.
Stakeholder Groups	Group A (n = 2): Software practitioners who use AI tools in professional development work. Includes software engineers and technical leads with daily AI tool use. Group B (n =2): AI business leads and consultants who oversee AI delegation at an organizational level. Provides management and governance perspective. Group C (n=2): Legal and compliance professionals with expertise in corporate law and HR governance.
Why Three Groups	Group A provides task level perspective; how individual developers experience delegation decisions in practice. While Group B provides organizational perspective; how AI delegation is managed, authorized, and governed. Group C provides how AI use intersects regulatory obligations, data protection, and people related risks. Together they ensure the checklist works across the individual, organizational, and governance level.
Format Rationale	Individual interviews were preferred over group sessions because they allowed participants to react more candidly without interference from colleagues or managers.
Recording	Verbatim notes were taken during the interviews. All data stored by the project team only.
Anonymization	Participants are referred to as P1, P2, P3, P4, P5, P6 in all outputs.

Appendix D – Pre-Session Preparation:

Each participant received the following materials at least 24 hours before the interview:

Provide a brief project description: "We're creating a co-design checklist on the practical criteria for the AI task delegability. We'd appreciate your honest thoughts on an early draft."

Checklist V2: As a working draft " Please read through it once before responding. We expect to change it based on your input."

The worked scenario on the paper (For Group A and Group B: Task A, sorting customer name list; Task B, writing login/authentication code using Github Copilot; for Group C: Task C, drafting a formal written warning letter to an employee based on a submitted incident report).

The five structured questions listed in the Questions section below.

The background form (Appendix G), participants were asked to complete this form before the interview.

Appendix E – Session Structure:

Table 6: Session structure used during the co-design interviews

Step 1	Background Form	Participants answer Table A1, calibrating AI familiarity and professional context.
Step 2	Icebreaker	Participants answered one warm-up question before engaging with the checklist items: 'Describe one task you currently delegate to an AI tool and one task you would never delegate, what made you draw that line?'
Step 3	V2 Walkthrough	Participants read V2 section by section and responded to the five structured questions (see Appendix F).
Step 4	Scenario Application	Group A and Group B applied the technical scenarios (Task A and Task B), while Group C applied a non-technical scenario (Task C).
Step 5	Closing Question	Participants were asked: 'Is there anything you expected to see on this checklist that was not present?'

Appendix F – Five Structured Questions:

Table 7: Structured interview questions

Question	Text
Q1	Which items felt clear and genuinely useful to you?
Q2	Which items felt vague, impossible to answer in practice, or unnecessary; and why?
Q3	Did any two items feel like they were asking the same thing from different angles? Which ones?
Q4	Think about the tasks. Does this checklist lead you to the right decision? Is anything important missing?
Q5	Is there a criterion you would always consider before delegating a task to AI in your professional life that is not on this list?

Appendix G – Background Form:

Table 8: Participant background form template

Control Variable	Response / Option
Engineering Background / Field of Work	
Current Role (e.g., developer, manager)	
Years of Professional Experience	
AI Tools Currently Used	
Frequency of AI Tool Use for Tasks	Never / Rarely / Sometimes / Often / Daily
Familiarity with Delegating tasks to AI	None / Slight / Moderate / High / Expert
Have you ever decided whether to trust AI output at work?	Yes / No

Table 9: P1 background form

Control Variable	Response / Option
Engineering Background / Field of Work	AI Development
Current Role (e.g., developer, manager)	CTO at a startup
Years of Professional Experience	2
AI Tools Currently Used	Hermes Agent, Codex, Claude
Frequency of AI Tool Use for Tasks	Daily
Familiarity with Delegating tasks to AI	Expert
Have you ever decided whether to trust AI output at work?	Yes

Appendix H – Verbatim Feedbacks:

P1 (Age:21, Education: Bachelors Student, Sex: Male).

I think section 1 is the most important checklist section because the programmer needs to ensure that the AI has all the context it needs to finish a task. This doesn't only include task's input and output, but also context regarding documentation for libraries or APIs to use, browser capabilities if QA testing for web is required, access to bash tools if the written program is a CLI or an API that can be tested with curl. For your tasks, especially for the login system, it is important to give AI context about the libraries to be used and the documentation, because AI tends to make mistakes or hallucinate when it doesn't have context about code that has less data in its training set. Also, AI can test its code automatically and improve its mistakes if it has tool access to a staging database and its own API, it can simulate a login/register and verify it works by looking at the database. I think you should add another item for section 1 to ensure that AI has the right context or knowledge base regarding internally used coding formats, libraries and more so it can retrieve them on demand so it is aligned with the organization's current practices.

I think item 2.3 is very important as well and it is good that it is included, because sometimes people, myself included, tend to trust the abilities or the tools of AI more than it can offer. It should always be manually E2E tested or tested with a testing set properly, even if AI self reports that changes are tested successfully.

For item 5.2, I don't think this item is necessary because the code that AI generates is still code that is the same as human written code, the language is the same and the methods it uses are the same. I believe it is a level of abstraction that technology has created to enable programmers to operate on a more abstract level to write code. I think that even if AI is used, human operators/programmers should be held %100 liable no matter what happens, because at the end of the day it is their responsibility to ensure that the written code is aligned with clients' expectations, whether they use AI or not. If they choose to use AI, it is still their responsibility to QA test and read the code manually. I don't think it is required to get consent from a manager for AI generated code, because code is still code, whether it is AI generated or not. I would agree if AI generated code in a low level language like assembly, but that is almost never the case, the code generated is a high level human debuggable code.

P2 (Age: 22, Education Bachelors Student, Sex: Male).

Table 10: P2 background form

Control Variable	Response / Option
Engineering Background / Field of Work	Software Engineer
Current Role (e.g., developer, manager)	Freelancer
Years of Professional Experience	3
AI Tools Currently Used	Claude
Frequency of AI Tool Use for Tasks	Daily
Familiarity with Delegating Tasks to AI	Expert
Have you ever decided whether to trust AI output at work?	Yes

Table 11: P3 background form

Control Variable	Response / Option
Engineering Background / Field of Work	Consultant in Strategy & Innovation Development
Current Role (e.g., developer, manager)	Freelancer
Years of Professional Experience	11
AI Tools Currently Used	Claude, GPT
Frequency of AI Tool Use for Tasks	Daily
Familiarity with Delegating Tasks to AI	High
Have you ever decided whether to trust AI output at work?	Yes

1.3 verifiability, good, probably the most important item in the whole list honestly

1.4 time constraint, the point about time pressure affecting the workers capacity to evaluate ai output feels underrated, maybe add a whole separate question are you under a time constraint or cognitive load or any other pressure where you're likely to accept ok looking output without a deep review

2.2 error distinguishability relevant for task b in particular

4.2 error responsibility delegation clear and useful

1.2 task tool fit the wording felt pretty abstract, a concrete example would help

2.1 accuracy evidence hard for an individual worker to actually answer since that kind of per task data usually doesn't exist at the user level sits more naturally on the org side

2.3 overtrust tough to self assess before the fact, overtrust is the kind of bias you only notice after it has already happened

5.3 commanding tools the term wasn't immediately clear to me, concrete examples like kill switch or audit log might help

sections 4 and 5 generally some items like skill erosion ai literacy and peer review felt important but more org level than task level, useful but maybe better placed in a different section

1.3 verifiability and 3.1 verification cost one asks if you can verify and the other how long it takes

there are a couple qs that all touch on the same idea of whether a human actually checked the output

4.2 and 4.3 both seemed to ask who is accountable

maybe these could be merged or repositioned so each item has its own clear role

Task b might fail even with the questionnaire since the development might think they can verify the output because they can read the code without realising they don't know enough about cryptography to spot a subtle bug, so 1.3 could just pass through in good faith just bcs the dev might not know what he doesn't know, maybe an item about domain expertise would help close that gap, something like whether the worker has enough background to recognise non obvious errors

Reversibility, or how bad the worst case undetected error actually is, if a missed mistake just means redoing some work delegating feels low risk but if it means leaked credentials lawsuits or production incidents that's a different category entirely

P3 (Age: 37, Education: Master's Degree, Sex: Male).

What I would ADD to your checklist (likely missing)

Based on typical checklists people create, these are the elements I would almost certainly add:

Section for risk classification of questions

Low risk (general info, public content)

Medium risk (internal but non-critical info)

High risk (legal, financial, HR, safety, health)

Clear escalation path

Exactly who receives escalated cases (role, not just "human").

What I would REMOVE or REDUCE

I would likely recommend:

Removing redundant fields that ask the same info twice.

Simplifying open-ended text questions into:

short, specific questions, multiple-choice, where possible.

Eliminating vague statements like:

"ensure ethical use" replaced with specific rules (e.g., "do not generate content about protected categories...").

Direct verdict style (as if I'm your company's CEO)

If your current checklist:

clearly defines scope,

enforces risk limits and escalation,

requires specific context and sources,

is short enough to be usable,

and assigns accountability,

then I would consider it realistic and deployable.

If instead it:

is mostly conceptual or academic,

has many fields nobody will practically fill,

lacks clear "do not answer this" rules,

and has no defined escalation path or owner,

then from my perspective it is not yet realistic, and I would ask you to rework it using the structure I outlined above.

If your checklist aims to let AI answer in place of humans, it must cover at least these categories. If any of these are missing or very weak, the checklist is not "real-world ready".

P4 (Age:64, Education: Bachelors Degree, Sex: Male):

Table 12: P4 background form

Control Variable	Response / Option
Engineering Background / Field of Work	Corporate Lawyer
Current Role (e.g., developer, manager)	Corporate Lawyer at a company
Years of Professional Experience	30
AI Tools Currently Used	Chat GPT
Frequency of AI Tool Use for Tasks	Sometimes
Familiarity with Delegating Tasks to AI	Slight
Have you ever decided whether to trust AI output at work?	Yes

The checklist can be considered very helpful as it clearly understands that delegating tasks using AI is both a technical decision and a responsibility one, as well as a risk management issue. In my opinion, from the viewpoint of a corporate lawyer, the most important aspect in Checklist V2 is Section 4 "Accountability" as this is the area that causes problems for many organizations. Nevertheless, even this section seems vague enough not to be applicable in the practical setting of a company. In relation to item 4.1, "High-risk task assessment with regulatory bodies", I would not make an employee check whether the task falls within the competence of regional regulatory bodies as employees do not understand how to assess it. The item can be changed to make an employee consider whether the task involves personal data, legal rights, employment decisions, financial implications, health or safety, or any confidential business matters. Regarding the fourth item on the checklist 4.2, I agree with the statement concerning the legal responsibilities that should be explained to the developers, deployers, and end users of AI. However, I would change it a bit since in many organizations the responsible person is not always the one developing or deploying an AI solution. What I mean is that the checklist should make the user specify who is responsible for the final decision. Based on my experience, such an answer as "AI was mistaken" cannot be considered a legally adequate reason. The item 4.3 "Clearing legal gaps" is also very significant. Nonetheless, I would recommend changing it to a more specific question "Does it clarify which individual approves, reviews, documents, and takes responsibility for the output assisted by AI?" This step is needed due to the fact that legal liabilities become obscure when many parties take part in creating the output: AI vendor, company, employee, and the responsible individual. With regard to number 4.4, while I strongly concur with the need for human confirmation of high-stakes actions, the term "reliable human entity" is too general. The reviewer should be competent to assess the risk. A manager will not necessarily be the ideal reviewer in all circumstances. For instance, in case the activity involves the assessment of employees, HR professionals would come into play. Checklist item number 4.5 is one of the most crucial checklist items. However, it should be made a compulsory step rather than an optional one. If the person does not know precisely whom to call for help regarding potentially dangerous AI results, he or she should not delegate the assignment at all. In any enterprise, the absence of clarity in the process makes it irrelevant because of the time limitations of employees. The checklist should state the exact roles that have to act in such a situation: "Legal department," "HR manager," "Data protection officer," "Security team," or "Compliance officer" depending on the situation. The list should also contain the following point about confidentiality and data protection before using AI tools: Check whether your work involves personal information, trade secrets, confidential business information, data about your clients, or even some legally protected information. You should not use any data of that kind until the tool is checked by your enterprise and the terms of data protection are defined precisely because you might accidentally reveal some confidential information during work.

P5 (Age:60, Education: Masters Degree, Sex: Male).

Table 13: P5 background form

Control Variable	Response / Option
Engineering Background / Field of Work	HR Manager
Current Role (e.g., developer, manager)	Currently retired
Years of Professional Experience	35
AI Tools Currently Used	Chat GPT
Frequency of AI Tool Use for Tasks	Sometimes
Familiarity with Delegating Tasks to AI	Slight
Have you ever decided whether to trust AI output at work?	Yes

Checklist V2 is valuable, as it stresses that when delegating some task to AI, we remain responsible for it. From the standpoint of HR, it is a crucial point, as the implementation of AI systems in organizations influences the lives of employees, potential applicants, their privacy, fairness, and even trust in an organization. Nevertheless, it lacks more specific directions when making decisions regarding people. The most valuable part of Checklist V2 is related to the point of accountability. At the workplace, there must be a balance between using artificial intelligence systems to help people do their job better and not delegating one's responsibilities to AI. When AI systems assist in any aspect of employment, including hiring, promotion, performance evaluation, salary, discipline, and termination of the employment relationship, the decision of the AI system must be reviewed and approved by a competent person. Regarding item 4.1, I would make the language more clear. Most people will not be able to determine whether it is related to "regulatory bodies." The list should include a question whether the task influences recruitment process, employee data, workplace safety, salary, promotion, termination, or disciplinary actions. If yes, then the task should be classified as highly risky. Regarding items 4.2 and 4.3, I would try to change the language of these questions as well. Instead of mentioning developers, deployers, or "legal gaps," the list could simply ask about everyone's accountability: Who is checking the output, who approves, and who gives an explanation. In the HR department, the firm cannot argue that "the AI made a decision." It is still humans and the company that are responsible. Regarding 4.4, I completely agree that high-stakes tasks require human verification. But here, the reviewer needs to be competent for the task. In cases where the decision-making concerns issues relating to human resources, the reviewer must typically be the HR manager, supervisor, legal/compliance personnel, among others. It might not suffice in some cases to have the general manager review the output. The same applies to item 4.5, which talks about escalation. The employee must be able to escalate to a definite department when he or she feels that there could be something unethical or inappropriate in the output generated by the AI. As far as human resource concerns are involved, the employee should be able to escalate to human resource managers, compliance, or legal departments. Checklist V2 also fails to address these three critical HR concerns – equity, confidentiality, and transparency. AI technology might result in bias in the processes of hiring, promotions, performance appraisal, or disciplinary actions. The third concern involves AI exposing personal and confidential details of an employee if the latter enters their CV, salary, medical information, complaints, or performance into an unauthorized system. Fourthly, any decision based on AI technology must also have transparency, meaning that the process by which the decision was made can be accounted for. In general, Checklist V2 provides a solid starting point, although better protections would have to be implemented for tasks related to personnel matters. The key point here is that anything dealing with hiring, promotion, appraisal of performance, salary, termination of employment, surveillance of employees, complaints in the workplace, personal data, and welfare of employees can be considered high-risk activities. While AI can help formulate and structure the material, it cannot be used to finalize the decision-making process. Human oversight of the results is required

P6 (Age: 20, Education: High School Diploma, Sex: Male).

Table 14: P6 background form

Control Variable	Response / Option
Engineering Background / Field of Work	AI Development and Consultation
Current Role (e.g., developer, manager)	Owner/CEO of a startup
Years of Professional Experience	3
AI Tools Currently Used	Claude
Frequency of AI Tool Use for Tasks	Daily
Familiarity with Delegating Tasks to AI	Slight
Have you ever decided whether to trust AI output at work?	Yes

I think it is important for employees to not over-rely on AI tools without checking the output. I sometimes experience employees even delegating tasks to AI for creative works that I think should be done by a human. I think AI should be used for almost deterministic applications when delegating tasks. Although I know that there are many ways to do tasks in fields such as coding, with the right prompting and the delegator knowing what needs to be done and just using AI to "speed up the process". However, when it used to generate ideas either implicitly or explicitly, I think this creates quality and reliability problems. That's why I like items 2 and 5.

Appendix I – Post Session Procedure:

Debrief as a team within 24 hours of each interview, record first impressions before consulting notes in detail.

Apply thematic analysis: code all feedback into recurring themes before tracing each to a checklist revision.

For every change between V2 and V3: write one sentence tracing it to a specific participant statement (P1, P2, P3, P4, P5, P6)

For every item kept unchanged despite feedback, write one sentence explaining why it was retained.

A visibility test was conducted: a team member who had not attended the sessions was given V2 and V3 and asked to reconstruct what participants said based on the differences between the two versions alone.

Appendix J – AI Use Disclosure:

Table 15: AI use disclosure

Question	Answer
Which AI tools were used?	Claude, GPT, Notebook LLM
How were they used?	Brainstorming checklist item phrasings, drafting document templates.
What verification was applied?	All AI-generated phrasings were reviewed and revised by the team members.
Were all cited sources read by a team member?	Yes
Were participant quotes verified against session notes?	Yes. All quotes appearing in the paper and appendix were verified against session notes taken during each interview.

Appendix K – Authors

First Name: Sila
Last Name: Gulec
Email: s334743@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Ahmet Tolga
Last Name: Birand
Email: s335068@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Jacopo
Last Name: Caravaggio
E-mail: s339988@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Giacomo
Last Name: Simbula
Email: s341576@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Gabriele
Last Name: Gallo
Email: s339929@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Nicolo
Last Name: Fiori
Email: s325013@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Benedetta
Last Name: Stellato
Email: s341071@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Pietro
Last Name: de Francesco
Email: s340079@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name:Behrad
Last Name: Galedari
Email: s324382@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino

First Name: Kutay
Last Name: Gunel
Email: s307467@studenti.polito.it
Country: Italy
Affiliation: Politecnico di Torino