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

Accorroni, Casalegno, Leocata, Liao, Lieggi, Mascherin, Muratori, Mustaj, Rinaldi, Valori, Ziarati Niasar
Politecnico di Torino — Turin, Italy
Impact of AI on Occupations, A.Y. 2025/2026
Supervisor: Daniele Quercia

ABSTRACT

The growing deployment of Artificial Intelligence in professional environments raises a fundamental question: what conditions should developers consider when designing AI systems intended to perform professional tasks? Existing frameworks address AI risks and governance at an organizational level, but few provide a practical, co-designed instrument that translates workers' concerns into actionable design requirements for AI systems at the task level. To address this gap, we conducted a co-design study involving 38 participants drawn from diverse professional sectors and roles, including healthcare, finance, public administration, and technology, to iteratively develop an AI Delegability Checklist. Starting from a literature-grounded draft, we ran an in-class workshop with peer practitioners and then administered a structured survey using a five-point Likert scale, resulting in four successive versions . The final checklist is organized around six thematic dimensions: Task Moral Delegability, Accountability and Responsibility, Autonomy and Professional Identity, Transparency and Explainability, Fairness and Third-Party Harm, and Human-AI Collaboration and Governance. We demonstrate the checklist through the case of a software designer architecting an AI-assisted radiology triage system for hospital deployment. The systematic translation of checklist items into enforceable technical, organizational, and interface-level requirements illustrates how the framework shapes the design space itself, ensuring that task delegation remains ethically permissible only when all six dimensions are satisfied within a rigorously implemented shared-control model.

Keywords: AI ethics; ethical checklist; co-design; task delegation; worker autonomy; accountability; transparency; radiology

1. INTRODUCTION

The integration of Artificial Intelligence into professional environments has fundamentally reshaped how work is organized, evaluated, and performed. As AI systems become increasingly capable of executing tasks that were once the exclusive domain of human professionals, from drafting documents to supporting clinical decisions, organizations and their workers are confronted with a question that is at once practical and deeply ethical: which tasks can, and should, be delegated to an AI system? This question of AI delegability is not merely technical. It involves considerations of moral responsibility, professional identity, autonomy, fairness toward third parties, and the governance of human-machine collaboration, dimensions that existing responsible AI frameworks have only partially addressed.

Despite the proliferation of responsible AI guidelines, impact assessment tools, and ethical checklists, a significant gap remains: there is no co-designed, worker-centered instrument that helps practitioners systematically evaluate the ethical conditions under which delegating a specific task to an AI system is appropriate. Existing tools tend to be designed top-down by researchers, compliance experts, or technology companies, and primarily address the obligations of AI developers rather than the lived concerns of the workers who interact with deployed systems daily. As a result, workers often lack a structured means to articulate and communicate the conditions they consider indispensable for informing how AI systems should be designed and deployed in relation to their tasks.

To address this gap, we designed and administered a co-design study aimed at producing an AI Delegability Checklist: a structured instrument grounded in the real-world ethical expectations of workers across diverse professional sectors. Rather than deriving requirements exclusively from regulatory documents, we adopted a bottom-up co-design methodology, soliciting workers' own ethical priorities through iterative stakeholder engagement. This dual strategy, first grounding items in a literature-based framework [1], then refining them through direct participant feedback using a five-point Likert scale, reflects a commitment to value-sensitive design and to producing a tool that reflects the realities of those it is intended to serve.

Building on our prior scoping review [1], which identified six interrelated ethical criteria for AI delegation, task moral delegability (T1), accountability and responsibility (T2), autonomy and professional identity (T3), transparency and explainability (T4), fairness and third-party harm (T5), and human-AI collaboration and governance (T6), this paper presents a Type-1 co-designed checklist: given an AI-exposable task, should it be delegated to an AI system? The paper proceeds as follows: Section 2 describes the co-design process and checklist evolution; Section 3 presents the final checklist; Section 4 applies it to the medical radiology use case; Section 5 discusses findings and limitations.

2. RELATED WORK

The co-design of ethical checklists for AI systems has gained traction as a method for translating abstract principles into actionable tools. Madaio et al. [2] demonstrated through a large-scale co-design process with 48 practitioners that AI

fairness checklists are most effective when developed iteratively with the people who will use them, rather than imposed top-down. Their work established that checklist items should function as prompts for reflection rather than binary compliance boxes. Similarly, Constantinides et al. [3] developed an AI Impact Assessment Report Template through iterative co-design sessions with both AI practitioners and compliance experts, identifying four key design requirements, completeness, breadth, adaptability to uses, and adaptability to roles, that informed our own process.

The ethical conditions under which workers may legitimately delegate tasks to AI systems have been examined across multiple domains. Lubars and Tan [4] proposed a foundational delegability framework distinguishing tasks by their objective nature, reversibility of errors, and human oversight requirements. Subsequent work has identified accountability gaps arising from opaque AI systems [5], the risk of deskilling through excessive automation [6], and the importance of human-in-the-loop architectures for preserving worker agency [7]. Our checklist operationalizes these findings into a practical six-dimension tool, extending prior frameworks by grounding each dimension in direct stakeholder input collected across diverse professional sectors.

3. CO-DESIGN PROCESS

3.1 Participants and Stakeholder Groups

We involved 38 participants drawn from three distinct stakeholder groups. The first group consisted of three students from peer groups at Politecnico di Torino with backgrounds in engineering and management, recruited for the in-class co-design workshop. The second group comprised 27 external participants recruited independently by group members through personal and professional networks, spanning a wide range of sectors including healthcare, finance, public administration, information technology, commerce, energy, law, and design. Participants ranged in age from 18 to over 56, included both genders and non-binary respondents, and covered a broad spectrum of professional roles: students, employees, mid-level managers, executives, independent professionals, and directors. The third group consisted of eight professional radiologists, recruited as domain experts to validate the final checklist version within the specific high-stakes professional context identified in our scoping review. The majority of external participants reported daily or occasional use of AI tools in their current work, ensuring that responses reflected informed perspectives on AI integration in professional settings.

3.2 Procedure

The co-design process followed an iterative methodology comprising three phases, drawing on established co-design practice [2].

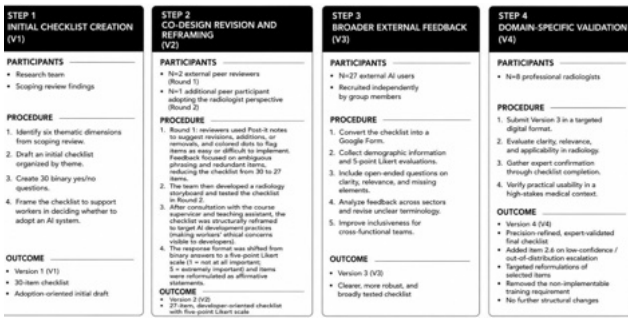

Figure 1: Four-step co-design process for the AI Delegability Checklist. Starting from a scoping review-based draft (V1), we refined the checklist through an in-class workshop with three peer participants (V2), broadened feedback via a Google Form survey with 27 external professionals (V3), and validated the final version with eight radiologists (V4).

Phase 1. Internal Draft (V1). The group developed Checklist Version 1 (V1) internally, grounded in the six thematic dimensions identified in our scoping review [1]. V1 comprised 30 binary yes/no items organized by theme. Each theme was presented on a separate A3 sheet, allowing participants to visualize the checklist structure and engage with its content in a collaborative and accessible manner. The checklist was initially framed to support workers in deciding whether to adopt an existing AI system.

Phase 2. In-Class Workshop (V1 → V2). The in-class workshop proceeded in two rounds. In Round 1, two external reviewers from a peer group collaboratively reviewed the checklist using Post-it notes to suggest revisions, additions, or removals, and colored dots to flag items perceived as easy or difficult to implement. Feedback focused primarily on ambiguous phrasing and redundant items across closely related themes, reducing the checklist from 30 to 27 items. Following Round 1, the team revised the checklist and developed a storyboard based on the radiology use case to simulate a realistic professional context for further testing. In Round 2, one additional participant from a peer group adopted the perspective of a radiologist and applied the checklist to the storyboard scenario, allowing the team to assess practical usability and contextual appropriateness of each item. Following consultations with the course supervisor and teaching assistant, the necessity for a more substantial structural reframing became evident. Drawing upon this guidance, alongside peer feedback and a closer reading of Madaio et al. [2], we concluded that the checklist should primarily target AI development practices—thereby rendering workers' ethical concerns visible to developers—rather than serving exclusively as a post-deployment adoption instrument. Accordingly, we transitioned from a binary response format to a five-point Likert scale (1 = not at all important; 5 = extremely important) and

reformulated the items into affirmative statements, resulting in Version 2 (V2).

Phase 3. Online Survey (V2 → V3). To compensate for the limited number of in-class participants and collect more heterogeneous feedback, we administered a structured online survey via Google Forms between May 3rd and May 13th, 2026. The survey comprised three sections: demographic and contextual information; five-point Likert-scale evaluation of all checklist items organized by thematic area; and open-ended qualitative questions at the end of each thematic block, inviting participants to flag unclear items or suggest missing criteria. The survey was completed by 27 external participants. Thematic analysis of qualitative responses revealed that certain technical expressions remained too ambiguous for non-engineers and cross-functional teams; language was revised accordingly, producing Version 3 (V3).

Phase 4 — Domain Expert Validation (V3 → V4). To verify the checklist's applicability within the high-stakes professional context of medical radiology, V3 was submitted to eight professional radiologists in a targeted digital format preserving the same structure as the V3 survey. The radiologists evaluated the clarity, relevance, and practical applicability of each item within their actual decision-making conditions. Their feedback identified several substantive concerns: overlap between items 1.2/1.3 and 1.5/1.6; ambiguity in item 2.5 regarding the definition and locus of performance metrics; real conceptual overlap between items 3.3 and 3.4, and between items 4.1 and 4.2; the absence in item 4.3 of an explicit requirement for out-of-distribution signalling — the highest-risk failure mode in radiology; ambiguity in the scope of T5, which conflated transparency obligations and privacy requirements across items 5.4 and 5.5; and a governance item (6.4) whose "critical training" requirement was organisational in nature and therefore not implementable at the software level. The specialists also identified a structural gap: no item addressed the system's behaviour when confidence was low or the input was anomalous. These observations prompted targeted reformulations of items 1.2, 1.4, 2.5, 3.2, 3.3, 3.4, 4.1, 4.2, 4.3, 5.4, 5.5, and 6.4, the addition of a new item 2.6 mandating automatic escalation for low-confidence or out-of-distribution cases, and the removal of the non-implementable training requirement from 6.4. No further structural changes were requested, and the revised instrument was confirmed as adequate by the radiologists. This validation phase produced Version 4 (V4): not a wholesale redesign, but a precision-refined, expert-validated final instrument.

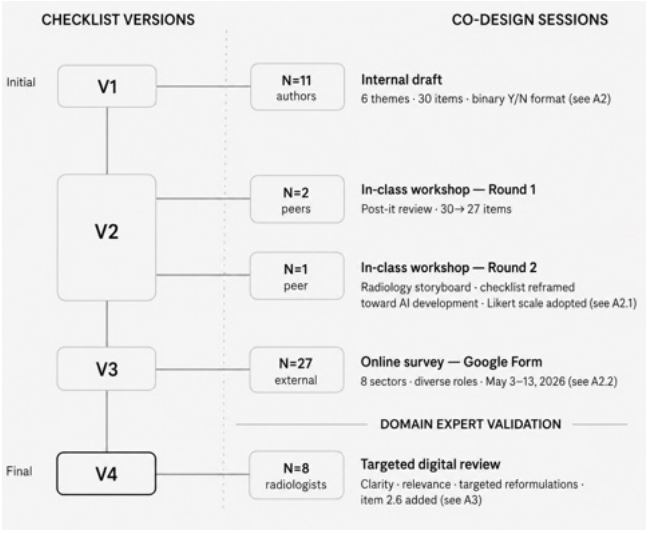

Figure 2: a four-step iterative co-design process used to refine the checklist. Starting from an internal draft (V1), it was reviewed by 11 authors and refined through two classroom workshops (2 and 1 participants), resulting in V2. A broader validation (V3) involved an online survey with 27 external participants from eight sectors (May 3–13, 2026; see Appendix A3 for more details). Finally, 8 radiologists conducted a focused digital review identifying targeted reformulations and one new item (2.6), leading to the final version (V4).

4. THE CHECKLIST

The AI Delegability Checklist is a Type-1 deliberation instrument designed to support the ethical design and development of AI systems. Unlike traditional adoption frameworks aimed at helping workers decide whether to use an existing AI system, this checklist is primarily intended for designers and developers. Its purpose is to help technical teams identify the conditions that should be considered when designing AI systems that will be integrated into professional tasks and workflows.

The checklist is organised around six thematic dimensions (T1–T6), covering task characteristics, accountability, worker autonomy, transparency, fairness, and governance. Together, these dimensions provide a set of design requirements and considerations that can guide the development of AI technologies across different sectors and use cases. Rather than functioning as a compliance or scoring mechanism, the checklist serves as a structured framework for incorporating workers' ethical concerns into system design. Although the worked use case presented in this paper focuses on software designers developing a system for medical radiology, the checklist can be applied to a wide range of AI-enabled professional contexts, maintaining applicability across a wide range of professional domains.

Lorenzo Accorroni, Luca Casalegno, Salvatore Daniel Leocata, Livio Liao, Thomas Lieggi, Jacopo Mascherin, Matteo Muratori, Redi Mustaj, Pietro Rinaldi, Marco Valori, Davide Ziarati Niasar

Checklist Items	
T1 - Task Moral Delegability - Is the task appropriate for AI delegation?	
1.1	The system is designed to handle exclusively repetitive, well-structured, and low-risk tasks.
1.2	The system automates exclusively tasks based on objective and measurable criteria, excluding activities that require subjective judgment, contextual interpretation, relational competences, or emotional involvement, without simulating or replacing human empathy.
1.3	The system architecture includes risk mitigation mechanisms such that, in the event of an algorithmic error, the consequences are contained and easily reversible.
1.4	The system is designed to incorporate human oversight in critical phases and to allow the operator to modify its operational parameters according to the context.
T2 - Accountability and Responsibility - Is it clear who is responsible?	
2.1	The software explicitly states, in the interface or terms of use, who holds legal and professional responsibility for the final result.
2.2	The system generates and displays a traceable, comprehensible audit trail explaining the AI's logical steps.
2.3	The software integrates native tools (e.g. alerts, procedural blocks) enforcing mandatory human oversight as required by current regulations (e.g. EU AI Act).
2.4	The UI guarantees the worker substantive control over the output, allowing modification, override, or rejection as easily as approval.
2.5	The tool includes an integrated feedback function to report AI errors, structurally separated from the user's performance evaluation metrics.
2.6	The system includes an automatic escalation mechanism that flags low-confidence or out-of-distribution cases and requires mandatory human review before any output is acted upon (threshold to be defined explicitly by the deployment team for each operational context, e.g. confidence score below 85%).
T3 - Autonomy, Agency and Professional Identity - Does the worker remain in control?	
3.1	The system reduces cognitive load on monotonous tasks, leaving the user creative and critical control.
3.2	The software design requires the user to engage actively with the data before the AI output is displayed, structurally preventing passive acceptance and reducing deskilling over time.
3.3	The system is designed to make the worker's contribution visible and valued, preserving their professional role and decision-making authority within the workflow.
3.4	The AI explicitly presents its output as a recommendation, structurally preventing it from being interpreted or acted upon as a final decision without human validation.
3.5	The system's functionalities were implemented following a process that actively collected and integrated workers' real operational needs.
T4 - Transparency and Explainability - Can the AI's reasoning be understood?	
4.1	The AI explicitly shows which input factors and reasoning steps led to each specific output.
4.2	The user can interrogate the system's decision criteria to verify their consistency and appropriateness for the specific case at hand.
4.3	The result contextualises statistical data by integrating qualitative and human information (e.g. clinical evidence, model limitations), including explicit signalling when the case falls outside the model's reliable operating range.
4.4	The output is translated into natural language understandable by the professional, avoiding ML jargon.
4.5	The model's behaviour is sufficiently stable and transparent to allow the expert user to predict its range of outputs.

Checklist Items	
T5 - Fairness, Bias and Third-Party Harm - Are external parties protected?	
5.1	The model has algorithmic checks actively mitigating historical biases and preventing discrimination against third parties (patients, clients, candidates).
5.2	The software respects privacy-by-design, ensuring processing without exposing or insecurely transferring sensitive third-party data.
5.3	Developers demonstrate the system was trained on broad, varied databases reflecting the diversity of the real population.
5.4	The software provides affected third parties with an accessible explanation of the output, sufficient to understand and formally challenge its validity.
5.5	The system permanently and verifiably deletes any sensitive personal data processed during the session, in compliance with applicable data protection regulations.
T6 - Human-AI Collaboration and Governance - Is the organizational structure adequate?	
6.1	The software architecture is structured for mutual collaboration, completing human work rather than replacing it.
6.2	The interface implements human-in-the-loop mechanisms (forced validation steps) and UX structured to discourage hasty approvals, avoiding deceptive patterns that bypass critical review.
6.3	The development cycle includes structured procedures to collect professionals' feedback before and during implementation.
6.4	Structured post-deployment monitoring is in place, including continuous model performance evaluation and periodic reporting to relevant stakeholders.

5. WORKED USE CASE

Profession and task. The professional role is the Software Designer. The objective is to evaluate which radiology tasks can be safely delegated to AI and to architect a system that enforces the ethical constraints identified by the checklist. The deployment context is a hospital radiology department integrated with a PACS/RIS workflow, where an AI triage assistant pre-screens incoming chest X-rays and CT scans, flagging cases by priority level before the radiologist reviews them.

AI use. A deep learning triage system embedded in the hospital's imaging pipeline, operating strictly as a decision-support tool within a shared-control paradigm. The checklist is applied by the designer as a binding specification: each dimension translates into concrete architectural and UX requirements that the system must satisfy before deployment is ethically permissible.

Framework applied.

Under T1 (Task Moral Delegability): Repetitive pre-screening is sufficiently structured to permit partial delegation. However, diagnostic interpretation requires contextual judgment that current systems cannot replicate, and errors are clinically irreversible. The designer therefore encodes a hard architectural boundary: the system may flag and prioritise, but may never produce a final diagnostic output. Full automation is ruled out at the design stage.

Under T2 (Accountability and Responsibility): The UI states at every interaction point that the radiologist retains full professional accountability. The system generates a structured, human-readable log of every inference step. An automatic escalation mechanism (item 2.6) is hardcoded: low-confidence or out-of-distribution cases are routed to mandatory human review and cannot be approved in a single click. A performance-neutral error reporting channel is built into the interface, structurally separated from any metric tied to the radiologist's output volume.

Under T3 (Autonomy, Agency and Professional Identity): To prevent passive acceptance and long-term deskilling, the designer implements a forced-interaction pattern: the AI's priority assessment is hidden until the radiologist has actively annotated key regions of the image. This preserves the clinician's diagnostic reasoning as the primary cognitive act, with the AI output functioning as a subsequent check rather than an anchor.

Under T4 (Transparency and Explainability): The interface exposes the specific input features, rendered as attention heatmaps or causal saliency maps, that produced each priority assessment. Confidence scores are translated into natural medical language, contextualised with case-specific qualitative signals. The system explicitly flags cases falling outside the model's reliable operating range, prompting heightened scrutiny rather than default trust.

Under T5 (Fairness, Bias and Third-Party Harm): Patient data is processed under strict data-masking protocols and permanently deleted post-session. Training data documentation must demonstrate demographic breadth across scanner models and patient populations, insufficient diversity is treated as a

deployment blocker. An automated bias-auditing dashboard generates periodic reports stratified by demographic subgroups, making differential error rates detectable before they produce differential clinical harm.

Under T6 (Human-AI Collaboration and Governance): The designer audits the workflow for dark UX patterns that structurally incentivise hasty approval under volume pressure. Mandatory validation steps are enforced at the workflow level. A post-deployment monitoring module tracks model performance continuously, routing anomaly reports to both technical and clinical stakeholders. Structured co-design sessions with radiologists are required prior to deployment.

Judgment. Applying the checklist as a design blueprint, the designer reaches a conditional delegation verdict. Radiology triage is not categorically undelegable, but safe delegation is only achievable within a strict shared-control paradigm that must be structurally enforced, not assumed. The six dimensions collectively constrain the design space rather than producing a binary yes/no answer: the checklist produces an architecture. Delegation is permissible only if all six sets of constraints are simultaneously satisfied. Any partial implementation, surfacing the triage output before the radiologist has engaged with the image, or omitting demographic validation of training data, renders delegation ethically impermissible regardless of clinical accuracy.

6. DISCUSSION

6.1 Co-Design Reflections

The iterative co-design process yielded several meaningful insights. The diversity of stakeholder backgrounds—spanning students, healthcare professionals, executives, and independent workers—produced complementary perspectives that enriched the checklist in ways a single-group process could not have. For instance, an energy sector executive initially prompted the inclusion of active ethical refusal mechanisms, which ultimately evolved into the strict human oversight constraints of items 1.3 and 1.4. Survey participants broadly identified the need for data volatility (item 5.5) and post-deployment monitoring (item 6.4), reflecting real professional concerns absent from the initial literature-grounded draft. Crucially, the final validation with eight expert radiologists introduced item 2.6—an automatic escalation mechanism for out-of-distribution cases—proving that domain-specific expertise is vital for uncovering structural edge cases. Another important outcome of the co-design process was the structural reframing of the checklist itself. Initially conceived as a tool to help workers evaluate whether an existing AI system should be adopted, the checklist progressively evolved into a developer-facing instrument aimed at supporting the design of AI systems. This shift allowed workers' ethical concerns to be translated into concrete design requirements that can guide technical development decisions. Consistent with this new proactive focus and the recommendations of Madaio et al. [2], shifting from a binary yes/no format to a five-point Likert scale proved essential: it transformed the checklist from a rigid compliance test into an actionable blueprint that surfaces gradations of ethical concern.

6.2 Limitations

Several limitations should be acknowledged. First, the in-class workshop involved only three participants, below the recommended minimum of six [2]; a larger or more diverse initial workshop group might have surfaced structural issues earlier in the V1 stage. Second, the checklist has not yet been tested in a live deployment context; its usability and discriminative validity in actual organizational settings remain to be empirically established. Third, the Likert-scale survey and digital validation formats cannot fully capture the relational dynamics and deliberative reasoning that emerge in live, synchronous co-design sessions, a limitation inherent to asynchronous stakeholder engagement.

6.3 Checklist as Reflection Tool, Not Compliance Instrument

Following the structural engineering checklist analogy proposed by Madaio et al. [2], our checklist is designed to prompt critical conversations rather than to certify post hoc compliance. The checklist is not intended to function as a quantitative scoring mechanism or to yield automatic recommendations. Instead, it should be understood as an indicator of whether the architectural and operational conditions for a responsible human-AI shared-control model are being addressed, without implying that ethical delegation can be guaranteed through rule satisfaction alone. The most valuable outcome of utilizing this checklist is the deliberation it generates among developers, designers, and domain professionals during the system's conceptualization, mapping out the design space rather than providing a definitive verdict. Future work should test the checklist in live organizational development cycles, involve domain-specific stakeholders from high-stakes fields like radiology from the earliest design phases, and explore whether an interactive digital version could better support the proactive, reflective function this instrument is designed to serve.

REFERENCES

[1] Accorroni, L. et al. 2026. Ethical Criteria for AI Delegation: A Scoping Review from the Workers' Perspective. Impact of AI on Occupations, Politecnico di Torino.

[2] Madaio, M.A., Stark, L., Vaughan, J.W., and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In Proceedings of CHI 2020. ACM. https://doi.org/10.1145/3313831.3376445

[3] Constantinides, M. et al. 2024. Co-Designing an AI Impact Assessment Report Template. In Proceedings of CSCW 2024. ACM. https://doi.org/10.1145/3686904

[4] Lubars, B. and Tan, C. 2019. Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability. NeurIPS 2019 Workshop on Human-Centered Machine Learning.

[5] Al Masaeid, T., Al-Adwan, A.S., and Al-Debei, M.M. 2025. AI-Augmented Decision-Making: Ethical Dimensions and Accountability Gaps. Decision Support Systems.

[6] Bankins, S. and Formosa, P. 2023. The Ethical Implications of Artificial Intelligence (AI) for Meaningful Work. Journal of Business Ethics 185, 4, 725–740. https://doi.org/10.1007/s10551-023-05339-7

[7] Lazaros, E., Tzovla, E., and Kotsiantis, S. 2026. Human-in-the-Loop Artificial Intelligence: A Systematic Review. Entropy (MDPI).

APPENDIX

A1. Interview / Workshop Guide

The following guide was used for the in-class co-design workshop (approximately 25–30 minutes).

Opening (2–3 min)

Participants were informed that the session aimed to improve a draft ethics checklist for AI task delegation. No prior knowledge of AI or ethics was required.

Checklist-specific feedback (20–25 min)

[Participants were shown the checklist]

- Are there any items you find unclear or ambiguous?
- Is there anything important that you think is missing?
- Are there any items you would remove as redundant?
- Does the Yes/No/N/A format seem adequate, or would you prefer a different response format?

Closing

Is there anything relevant you would like to add that we have not asked about?

A2. Checklist Version 1 (Internal Draft — Pre-Co-Design)

Version 1 was developed internally by the group prior to any external stakeholder involvement, grounded in the six thematic dimensions identified in the scoping review [1]. It comprised 30 binary (yes/no/N/A) items organized by theme, each presented on a separate A3 sheet allowing participants to visualize the checklist structure and engage with its content in a collaborative and accessible manner. The checklist was initially framed to support workers in deciding whether to adopt an existing AI system. It is reproduced here in its original form, including all ambiguities subsequently identified during the co-design process.

Table 1: AI Delegability Checklist (Version 1 — General)

	Question	Y / N / N/A	Notes
T1 — Task Moral Delegability			
1.1	Is the task repetitive, well-structured, and low-risk?	No	
1.2	Is the task objective and free from subjective interpretation or soft skills?		Be more specific
1.3	Is the task emotionally neutral and does it not require deep relational engagement?	Yes	
1.4	Are the consequences of an error low-stakes and more reversible?	No	
T2 — Accountability & Responsibility			
2.1	Is it clearly defined, before the AI acts, who is responsible if something goes wrong?	No	
2.2	Does the AI system produce a traceable log explaining why it generated that output?	No	
2.3	Is the delegation driven by real efficiency gains, not by a desire to avoid difficult decisions?	No	
2.4	Are there applicable regulations (e.g. EU AI Act) that require human oversight for this task?	No	Expand what you mean by regulations
2.5	Do workers have genuine control over the AI output, not just acting as formal signatories?	No	Cannot have control over output if it's about diagnosis
2.6	Can errors or anomalies in the AI's output be reported without professional repercussions?	No	
T3 — Autonomy, Agency & Professional Identity			
3.1	Does delegation reduce cognitive load and monotony, freeing time for creativity and critical thinking?	Yes	
3.2	Does delegation risk eroding my professional skills over time (deskilling)?	No	
3.3	Does delegation threaten my professional identity or the meaning of my work?	No	
3.4	Does the AI act as an augmentation tool (support) rather than a substitute for decision-making autonomy?	Yes	
3.5	Can the worker refuse the AI recommendation without fear of consequences?	No	
3.6	Was the worker involved in the decision to introduce AI, not just informed after the fact?	No	
T4 — Transparency & Explainability			
4.1	Can I understand why the AI produced a certain output?	Yes	

	Question	Y / N / N/A	Notes
4.2	Can I query the AI and verify its criteria?	No	In general yes but in this scenario no clear explanation
4.3	Is the output accompanied by human context rather than only statistical probabilities?	No	Be more specific about human context
4.4	Is the output sufficiently interpretable and understandable to the human operator?	Yes	
4.5	Is the output sufficiently predictable to allow the worker to anticipate, at least approximately, what to expect?	No	

T5 — Fairness, Bias & Third-Party Harm

5.1	Does the AI demonstrably reduce historical biases and avoid discrimination against third parties?	No	Explain more about historical biases
5.2	Can the task be delegated without transferring sensitive third-party data, or with robust privacy safeguards?	Yes	
5.3	Does the AI offer greater neutrality or procedural fairness compared to human judgement?	No	
5.4	Has the system been trained on data that is representative of the population it will be applied to?	Yes	
5.5	Can the AI decision be explained and justified in a way that allows affected third parties to understand, challenge or request review of the outcome?	No	

T6 — Human-AI Collaboration & Governance

6.1	Is the workflow structurally designed for true mutual human-AI collaboration rather than just passively replacing the human worker?	No	
6.2	Is there a formal, structural "human-in-the-loop" mechanism built directly into the workflow?	No	
6.3	Are adequate time and resources provided by the organizational governance to ensure real oversight on every important decision?	N/A	
6.4	Were the professionals actively involved in co-designing the system's integration, and did they voluntarily accept its adoption without top-down imposition?	No	

Transition V1 → V2 (In-Class Workshop)

Round 1 (2 peer reviewers): Post-it notes and coloured dots were used to suggest revisions, additions, or removals and flag items as easy or difficult to implement. Feedback focused on ambiguous phrasing and redundant items across closely related themes. The checklist was reduced from 30 to 27 items.

Round 2 (1 peer reviewer, storyboard): The revised checklist was applied to a storyboard simulating the radiology use case (Dr. Carlo Conti, 80 chest X-rays per shift, AI showing only "97% benign"). This exercise surfaced contextual gaps and confirmed the discriminative power of the checklist across T1–T6.

Structural reframing: Following supervisor and teaching assistant guidance, and informed by a closer reading of Madaio et al. [2], the checklist was reoriented from a post-deployment adoption instrument to a Type-1 deliberation tool targeting AI development practices — making workers' ethical concerns visible to developers. The binary Yes/No format was replaced by a

five-point Likert scale (1 = not at all important; 5 = extremely important) and items were reformulated as affirmative statements addressed to software designers, producing Version 2.

A2.1. Checklist Version 2 (Workshop Output — Affirmative Likert Format)

Version 2 is the direct output of the in-class workshop. Following the structural reframing described above, items were converted from binary yes/no questions into affirmative developer-facing statements evaluated on a five-point Likert scale. Three items removed during Round 1 as redundant are not reproduced here. Items are presented in their original Italian, the language used in the subsequent survey.

Table 2: AI Delegability Checklist (Version 2)

	Checklist Item (Version 2 — Likert scale 1–5)
T1 — Task Moral Delegability	
1.1	The system is designed to handle exclusively repetitive, well-structured, and low-risk tasks.
1.2	The algorithm performs tasks based on objective and measurable rules, excluding activities that require subjective interpretation, relational competences, or soft skills.
1.3	The system design automates only emotionally neutral processes, without simulating or replacing relational engagement or human empathy.
1.4	The software architecture includes risk mitigation mechanisms such that, in case of algorithmic error in this specific task, consequences are contained and easily reversible.
T2 — Accountability & Responsibility	
2.1	The software explicitly states, in the interface or terms of use, which human figure holds legal and professional responsibility for the final result.
2.2	The system can generate and display a traceable, comprehensible audit trail explaining the AI's logical steps.
2.3	The software integrates native tools (e.g., alerts, procedural blocks) that enforce or facilitate mandatory human oversight required by current regulations (e.g., EU AI Act).
2.4	The user interface (UI) guarantees the worker substantive control over the output, allowing modification, override, or rejection as easily as approval.
2.5	The tool includes an integrated feedback function (e.g., "report anomaly" button) allowing users to report AI errors without affecting their performance metrics.
T3 — Autonomy, Agency & Professional Identity	
3.1	The system is specifically optimised to reduce cognitive load on monotonous tasks, delegating routines to the machine while leaving creative and critical control to the user.
3.2	The software design encourages the user to reason about the data, avoiding excessive automation to prevent skill loss over time (deskilling).
3.3	The tool is structured to value human decisions, supporting the worker's professional identity without diminishing their role.
3.4	The AI is explicitly developed as a support tool (augmentation), offering recommendations and never definitive decisions that override human autonomy.
3.5	The system's features are implemented only following a development process that actively collects and integrates workers' feedback and real operational needs.
T4 — Transparency & Explainability	
4.1	The interface clearly shows the main causal factors (features/parameters) that led the model to generate that specific output.

	Checklist Item (Version 2 — Likert scale 1–5)
4.2	The software provides operational tools allowing the user to inspect the AI's decision criteria and verify their relevance to the specific context.
4.3	The AI output contextualises statistical data by integrating qualitative and human information (e.g., clinical evidence, model limitations in that specific case).
4.4	The output and its practical implications are translated into natural language easily understandable by the professional, avoiding machine-learning jargon.
4.5	The model's behaviour is sufficiently stable and transparent to allow an expert user to intuitively anticipate the range of actions or outputs the AI might produce.
T5 — Fairness, Bias & Third-Party Harm	
5.1	The model is equipped with algorithmic controls that actively mitigate historical biases, preventing discrimination against third parties (patients, clients, candidates).
5.2	The software architecture respects privacy-by-design, ensuring that processing occurs without exposing or unsafely transferring sensitive third-party data.
5.3	Developers demonstrate that the system was trained on large, varied databases that faithfully reflect the diversity of the real population where the software will be deployed.
5.4	The software provides simplified versions of its explanations, specifically designed to allow third parties affected by the output to understand and contest its validity.
T6 — Human-AI Collaboration & Governance	
6.1	The software architecture is structured to interface with the user in a logic of mutual collaboration, complementing human work rather than attempting to replace it entirely.
6.2	The interface operationally imposes a "human-in-the-loop" mechanism, inserting forced steps where the AI recommendation awaits a critical human validation.
6.3	The user experience (UX) is designed not to incentivise or force hasty approval (e.g., avoiding deceptive patterns), supporting accurate review by the user.
6.4	The software development cycle includes structured, continuous procedures to allow professionals to suggest changes, ensuring the tool is not experienced as an imposition.

Transition V2 → V3 (Online Survey — 27 External Participants)

Process: V2 was administered as a structured online survey via Google Forms (May 3–13, 2026) to 27 external participants spanning 14 professional sectors. Each thematic block was followed by open-ended qualitative questions inviting participants to flag unclear items or suggest missing criteria.

Language revision: Thematic analysis of qualitative responses revealed that certain technical expressions remained too ambiguous for non-engineers and cross-functional teams. All items were revised accordingly to improve accessibility and precision.

Structural changes: Two structural gaps were addressed: (a) absence of a data-deletion requirement in T5 → added item 5.5; (b) overlapping governance items in T6 (items 6.2 and 6.3 merged, new post-deployment monitoring item added as 6.4). Items 1.5 and 1.6 were also added following a survey contribution from an energy-sector executive requesting explicit ethical-refusal mechanisms.

A2.2. Checklist Version 3 (Online Survey Output)

Version 3 incorporates language simplifications and structural additions resulting from thematic analysis of 27 external respondents' qualitative feedback.

Table 3: AI Delegability Checklist (Version 3)

	Checklist Item (Version 3 — Likert scale 1–5)
T1 — Task Moral Delegability	
1.1	The system is designed to handle exclusively repetitive, well-structured, and low-risk tasks.
1.2	The algorithm performs tasks based on objective and measurable rules, excluding activities that require subjective interpretation, relational competences, or soft skills.
1.3	The system design automates only emotionally neutral processes, without simulating or replacing relational engagement or human empathy.
1.4	The software architecture includes risk mitigation mechanisms such that, in the event of an algorithmic error, the consequences are contained and easily reversible.
1.5	The system is designed to incorporate human oversight in critical phases and to allow the operator to modify its operational parameters according to the context.
1.6	The system is designed to actively reject tasks that violate predefined ethical thresholds, as the execution of such tasks would constitute an implicit precedent that legitimises their future repetition.
T2 — Accountability & Responsibility	
2.1	The software explicitly states, in the interface or terms of use, which human figure holds legal and professional responsibility for the final result.
2.2	The system can generate and display a traceable, comprehensible audit trail explaining the AI's logical steps.
2.3	The software integrates native tools (e.g., alerts, procedural blocks) that enforce or facilitate mandatory human oversight required by current regulations (e.g., EU AI Act).
2.4	The user interface guarantees the worker substantive control over the output, allowing modification, override, or rejection as easily as approval.
2.5	The tool includes an integrated feedback function (e.g., "report anomaly" button) allowing users to report AI errors without affecting their performance metrics.
T3 — Autonomy, Agency & Professional Identity	
3.1	The system is specifically optimised to reduce cognitive load on monotonous tasks, delegating routines to the machine while leaving creative and critical control to the user.
3.2	The software design encourages the user to reason about the data, avoiding excessive automation to prevent skill loss over time (deskilling).
3.3	The tool is structured to value human decisions, supporting the worker's professional identity without diminishing their role.
3.4	The AI is explicitly developed as a support tool (augmentation), offering recommendations and never definitive decisions that override human autonomy.
3.5	The system's features are implemented only following a development process that actively collects and integrates workers' feedback and real operational needs.
T4 — Transparency & Explainability	
4.1	The interface clearly shows the main causal factors (features/parameters) that led the model to generate that specific output.
4.2	The software provides operational tools allowing the user to inspect the AI's decision criteria and verify their relevance to the specific context.

	Checklist Item (Version 3 — Likert scale 1–5)
4.3	The AI output contextualises statistical data by integrating qualitative and human information (e.g., clinical evidence, model limitations in that specific case).
4.4	The output and its practical implications are translated into natural language easily understandable by the professional, avoiding machine-learning jargon.
4.5	The model's behaviour is sufficiently stable and transparent to allow an expert user to intuitively anticipate the range of actions or outputs the AI might produce.

T5 — Fairness, Bias & Third-Party Harm

5.1	The model is equipped with algorithmic controls that actively mitigate historical biases, preventing discrimination against third parties (patients, clients, candidates).
5.2	The software architecture respects privacy-by-design, ensuring that processing occurs without exposing or unsafely transferring sensitive third-party data.
5.3	Developers demonstrate that the system was trained on large, varied databases that faithfully reflect the diversity of the real population where the software will be deployed.
5.4	The software provides simplified versions of its explanations, specifically designed to allow third parties affected by the output to understand and contest its validity.
5.5	The software respects the principle of sensitive data volatility, permanently deleting from the cloud all private information temporarily saved for processing.

T6 — Human-AI Collaboration & Governance

6.1	The software architecture is structured for mutual collaboration, complementing human work rather than attempting to replace it entirely.
6.2	The interface enforces structural human-in-the-loop mechanisms (forced validation steps) and the UX is designed to discourage hasty approvals, avoiding deceptive patterns that bypass critical review by the user.
6.3	The development cycle includes structured procedures to actively collect feedback and operational needs from professionals before and during implementation, ensuring that the tool is not experienced as an imposition.
6.4	Structured post-deployment monitoring procedures are in place, including continuous evaluation of the model's behaviour in real contexts and critical training for professionals, ensuring that human-AI collaboration remains effective and aware over time.

Transition V3 → V4 (Domain Expert Validation — 8 Radiologists)

Process: V3 was submitted to eight professional radiologists in a targeted digital format preserving the same structure as the V3 survey. Radiologists evaluated clarity, relevance, and practical applicability of each item within their actual decision-making conditions.

Substantive concerns addressed: (1) Items 1.2/1.3 and Items 1.5/1.6: overlap – resolved by merging in two total new items (1.2 and 1.4). (2) Item 2.5: "without affecting performance metrics" too vague — who defines them? Reformulated to require structural separation. (3) Items 3.3/3.4: real conceptual overlap — resolved by distinct reformulations. (4) Items 4.1/4.2: significant overlap — resolved by separating reasoning-step display from criteria interrogation. (5) Item 4.3: missing out-of-distribution signalling requirement — added explicitly. (6) Items 5.4/5.5: conflated transparency and privacy — clarified into distinct obligations. (7) Item 6.4: "critical training" non-implementable at software level — removed from technical requirements.

Structural gap addressed: No item covered system behaviour for low-confidence or anomalous inputs. New item 2.6 added: automatic escalation mechanism flagging such cases for mandatory human review.

Result: V4 is not a wholesale redesign but a precision-refined, expert-validated instrument. No further structural changes were requested. Items translated into English for profession-agnostic use.

A3. Checklist Version 4 (Final — Expert-Validated)

Version 4 is the final, expert-validated instrument produced after targeted reformulations by eight professional radiologists. Relative to V3 it incorporates: reformulations of items 1.2, 1.4, 2.5, 3.2, 3.3, 3.4, 4.1, 4.2, 4.3, 5.4, 5.5, and 6.4; the addition of new item 2.6.

Table 4: AI Delegability Checklist (Version 4 — General)

	Checklist Item (Version 4 — Final Checklist)
T1 - Task Moral Delegability - Is the task appropriate for AI delegation?	
1.1	The system is designed to handle exclusively repetitive, well-structured, and low-risk tasks.
1.2	The system automates exclusively tasks based on objective and measurable criteria, excluding activities that require subjective judgment, contextual interpretation, relational competences, or emotional involvement, without simulating or replacing human empathy.
1.3	The system architecture includes risk mitigation mechanisms such that, in the event of an algorithmic error, the consequences are contained and easily reversible.
1.4	The system is designed to incorporate human oversight in critical phases and to allow the operator to modify its operational parameters according to the context.
T2 - Accountability and Responsibility - Is it clear who is responsible?	
2.1	The software explicitly states, in the interface or terms of use, who holds legal and professional responsibility for the final result.
2.2	The system generates and displays a traceable, comprehensible audit trail explaining the AI's logical steps.
2.3	The software integrates native tools (e.g. alerts, procedural blocks) enforcing mandatory human oversight as required by current regulations (e.g. EU AI Act).
2.4	The UI guarantees the worker substantive control over the output, allowing modification, override, or rejection as easily as approval.
2.5	The tool includes an integrated feedback function to report AI errors, structurally separated from the user's performance evaluation metrics.
2.6	The system includes an automatic escalation mechanism that flags low-confidence or out-of-distribution cases and requires mandatory human review before any output is acted upon (threshold to be defined explicitly by the deployment team for each operational context, e.g. confidence score below 85%).
T3 - Autonomy, Agency and Professional Identity - Does the worker remain in control?	
3.1	The system reduces cognitive load on monotonous tasks, leaving the user creative and critical control.
3.2	The software design requires the user to engage actively with the data before the AI output is displayed, structurally preventing passive acceptance and reducing deskilling over time.
3.3	The system is designed to make the worker's contribution visible and valued, preserving their professional role and decision-making authority within the workflow.
3.4	The AI explicitly presents its output as a recommendation, structurally preventing it from being interpreted or acted upon as a final decision without human validation.
3.5	The system's functionalities were implemented following a process that actively collected and integrated workers' real operational needs.
T4 - Transparency and Explainability - Can the AI's reasoning be understood?	
4.1	The AI explicitly shows which input factors and reasoning steps led to each specific output.
4.2	The user can interrogate the system's decision criteria to verify their consistency and appropriateness for the specific case at hand.
4.3	The result contextualises statistical data by integrating qualitative and human information (e.g. clinical evidence, model limitations), including explicit signalling when the case falls outside the model's reliable operating range.

Checklist Item (Version 4 — Final Checklist)	
4.4	The output is translated into natural language understandable by the professional, avoiding ML jargon.
4.5	The model's behaviour is sufficiently stable and transparent to allow the expert user to predict its range of outputs.

T5 - Fairness, Bias and Third-Party Harm - Are external parties protected?

5.1	The model has algorithmic checks actively mitigating historical biases and preventing discrimination against third parties (patients, clients, candidates).
5.2	The software respects privacy-by-design, ensuring processing without exposing or insecurely transferring sensitive third-party data.
5.3	Developers demonstrate the system was trained on broad, varied databases reflecting the diversity of the real population.
5.4	The software provides affected third parties with an accessible explanation of the output, sufficient to understand and formally challenge its validity.
5.5	The system permanently and verifiably deletes any sensitive personal data processed during the session, in compliance with applicable data protection regulations.

T6 - Human-AI Collaboration and Governance - Is the organizational structure adequate?

6.1	The software architecture is structured for mutual collaboration, completing human work rather than replacing it.
6.2	The interface implements human-in-the-loop mechanisms (forced validation steps) and UX structured to discourage hasty approvals, avoiding deceptive patterns that bypass critical review.
6.3	The development cycle includes structured procedures to collect professionals' feedback before and during implementation.
6.4	Structured post-deployment monitoring is in place, including continuous model performance evaluation and periodic reporting to relevant stakeholders.

A4. Anonymized Participant Notes

The following notes summarize the feedback collected from each participant during the co-design process. All participants are identified by anonymous codes (P1–P30). No personally identifiable information is recorded.

Workshop Participants (In-Class, May 2026)

P1 | Background: Engineering student / Technical implementer perspective | Duration: ~30 min

- Item 1.2 is considered too vague: unclear boundary between "objective" and "subjective" tasks.
- Item 2.3 needed explicit reference to specific regulations (e.g. EU AI Act) rather than generic mention.
- Yes/No/N/A format judged too rigid for governance questions; suggested a scale.
- Positively evaluated the scenario-based structure for anchoring abstract ethical questions.

P2 | Background: Management engineering student / End-user perspective | Duration: ~30 min

- Item 2.4: noted that genuine control over AI output is structurally impossible in diagnostic contexts where the AI output is opaque.
- Item 4.3: requested a more concrete definition of "human context" beyond statistical output.
- Suggested converting questions into affirmative statements to improve survey usability.
- Noted conceptual overlap between T2 and T6 items on accountability and governance.

P3 | Background: Computer engineering student / Technical implementer perspective | Duration: ~30 min

- Item 5.1: requested clearer explanation of "historical biases" with concrete examples.
- Suggested that the checklist should explicitly address the worker's right to refuse AI recommendations.
- Positively evaluated the thematic six-dimension structure as coherent and comprehensive.

Survey Participants (Google Form, May 3–13, 2026 — 27 respondents, P4–P30)

The online survey was completed by 27 external participants recruited through personal and professional networks. Respondents spanned a wide range of sectors, professional roles, and age groups. The five figures below provide a demographic overview of the survey sample, illustrating its breadth and heterogeneity.

Gender and age. The sample skewed male (70.4%) and young (48.1% aged 18–25), reflecting the recruitment channels used. A notable secondary cluster of professionals aged 46–55 (22.2%) ensured that more senior perspectives were represented alongside early-career respondents.

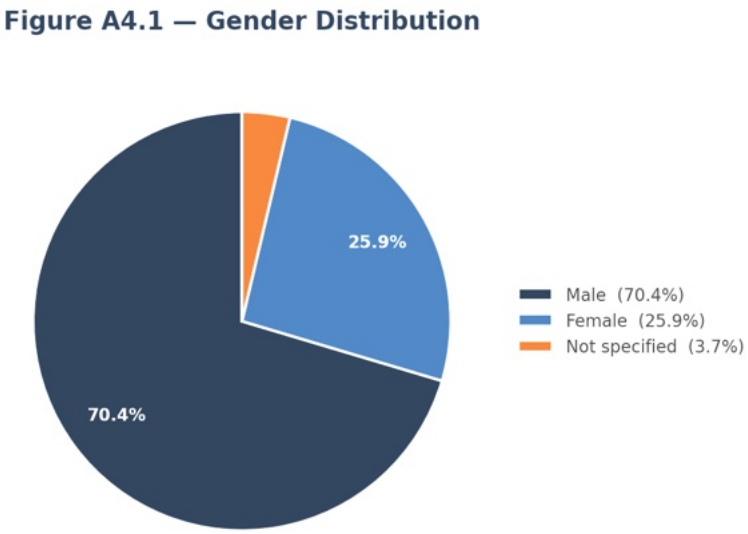

Figure A4.1 — Gender Distribution of survey respondents (n = 27).

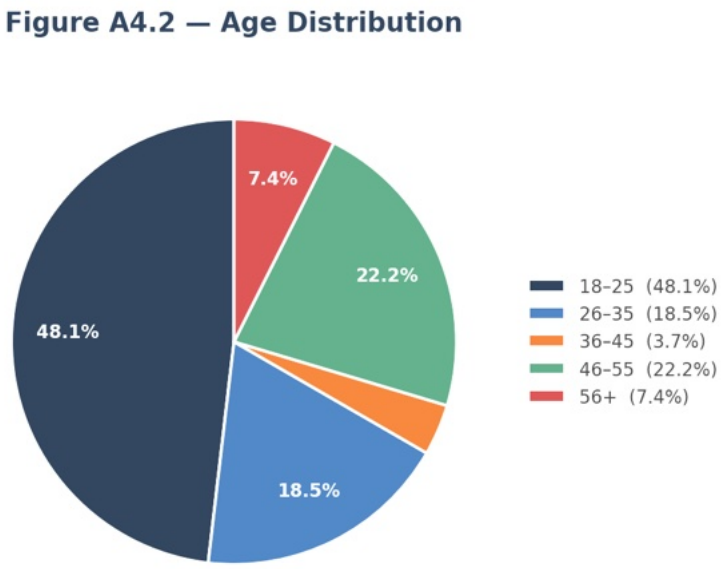

Figure A4.2 — Age distribution.

Work field. The survey drew participants from fourteen distinct professional sectors. Commerce, Healthcare, and IT & Technology were the most represented (14.8% each), followed by Industrial/Production, Finance, Education/Research, and Administration (7.4% each). The remaining sectors each contributed one respondent (3.7%), ensuring a broad and cross-sectoral perspective on the checklist items.

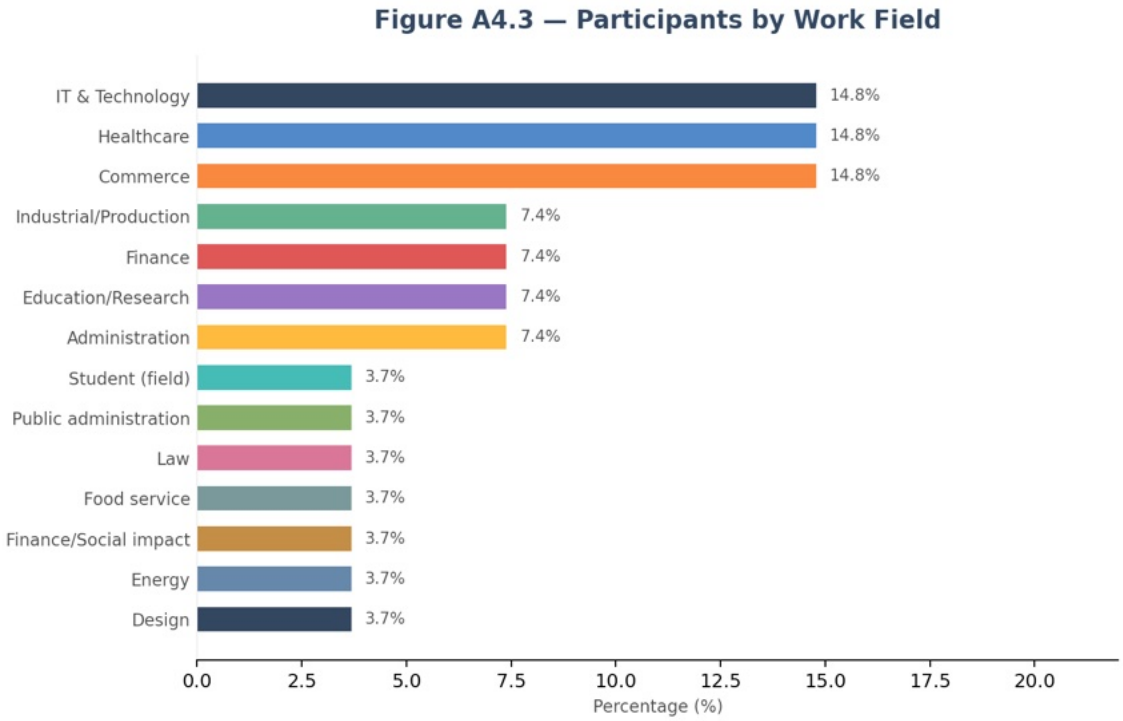

Figure A4.3 — Participants by Work Field.

Worker role. The sample covered a broad spectrum of hierarchical roles. Students constituted the largest group (37.0%), followed by employees/office workers (22.2%), executives (18.5%), freelancers (11.1%), mid-level managers (7.4%), and one president/director (3.7%). This diversity ensured that checklist items were evaluated from both operational and strategic perspectives.

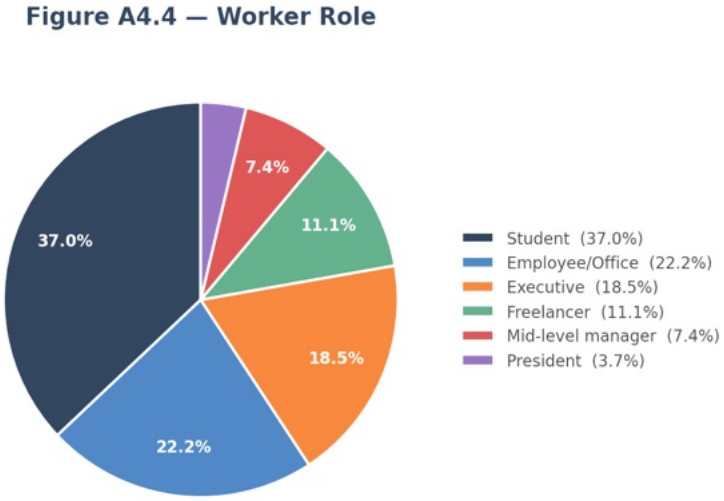

Figure A4.4 — Worker Role

AI tool usage. The majority of respondents reported active engagement with AI tools: 63.0% used them daily and 25.9% occasionally, with only 11.1% reporting no use. This high baseline familiarity with AI in professional contexts lends credibility to the Likert-scale evaluations provided, since respondents were assessing checklist items from a position of direct experience rather than hypothetical reasoning.

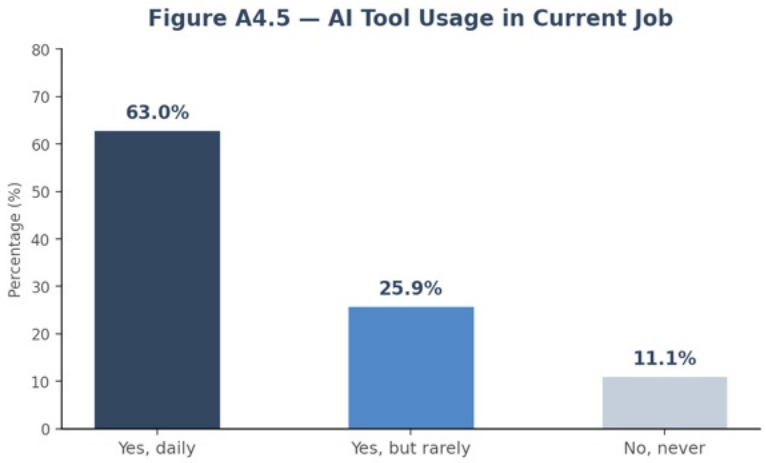

Figure A4.5 — AI Tool Usage in Current Job

A5. Checklist Evolution: V1 → V4

Table 5 summarizes the most significant revisions across the co-design process, documenting the stakeholder feedback that prompted each change. Consistent with the co-design principle that checklist evolution must be transparent and traceable [2], we record both what changed and why.

Table 5: Key checklist revisions from V1 to V4

V1 (original)	Feedback received	V4 (revised)
Binary Yes/No/N/A format across all items	Workshop participants (Round 1 + supervisor guidance): format too rigid for nuanced governance questions; Madaio et al. [2] recommend items as prompts for reflection rather than binary compliance	Five-point Likert scale on affirmative developer-facing statements (1 = not at all important; 5 = extremely important)
1.2 "Is the task objective and free from subjective interpretation or soft skills?"	Workshop Round 1: "be more specific" — boundary between objective and subjective unclear	"The system automates exclusively tasks based on objective and measurable criteria, excluding activities that require subjective judgment, contextual interpretation, relational competences, or emotional involvement"
2.3 "Are there applicable regulations requiring human oversight?"	Workshop Round 1: "expand what you mean by regulations"	Explicit reference to EU AI Act; item reframed as native enforcement tools integrated in the software
2.5 "Can errors be reported without professional repercussions?" (V1) → "feedback function not affecting performance metrics" (V3)	Radiologists (V3→V4): "without affecting performance metrics" is vague — who defines them? Not implementable concretely by a programmer	"The tool includes an integrated feedback function to report AI errors, structurally separated from the user's performance evaluation metrics" — separation is now a structural/architectural requirement
3.3 "Does delegation threaten my professional identity?" / 3.4 "Does the AI act as an augmentation tool?"	Radiologists (V3→V4): real conceptual overlap — both items addressed augmentation vs substitution without distinct focus	3.3 reformulated to focus on visibility and preservation of worker contribution; 3.4 reformulated to focus on recommendation framing and prevention of treating AI output as final decision

V1 (original)	Feedback received	V4 (revised)
4.1 "Can I understand why the AI produced a certain output?" / 4.2 "Can I query the AI and verify its criteria?"	Radiologists (V3→V4): significant overlap — both addressed explainability without distinguishing between showing reasoning and enabling verification	4.1 reformulated to specify which input factors AND reasoning steps led to the output; 4.2 reformulated to focus on user ability to interrogate and verify decision criteria for consistency and contextual appropriateness
4.3 "Is the output accompanied by human context rather than only statistical probabilities?"	Workshop Round 1: "be more specific about human context". Radiologists (V3→V4): missing explicit requirement for out-of-distribution signalling — highest-risk failure mode in radiology	"The result contextualises statistical data by integrating qualitative and human information (e.g. clinical evidence, model limitations), including explicit signalling when the case falls outside the model's reliable operating range"
5.1 "Does the AI demonstrably reduce historical biases?"	Workshop Round 1: "explain more about historical biases"	"The model has algorithmic checks actively mitigating historical biases and preventing discrimination against third parties (patients, clients, candidates)"
5.4 and 5.5 in V3 conflated transparency obligations and privacy requirements within T5	Radiologists (V3→V4): 5.4 (explanation to third parties) and 5.5 (data deletion) belong to different categories — one is transparency, one is privacy; their coexistence created confusion about what T5 was evaluating	5.4 reformulated to focus exclusively on third-party explainability and right to challenge; 5.5 reformulated to focus exclusively on permanent and verifiable deletion of session data in compliance with data protection regulations
T1 had 4 items; no item on active ethical refusal or explicit escalation for anomalous inputs	Survey (energy sector executive, V2→V3): system should actively reject ethically unacceptable tasks — items 1.5 and 1.6 added. Radiologists (V3→V4): no item addressed system behaviour when confidence was low or input anomalous — structural gap	Items 1.5 and 1.6 introduced at V3 were consolidated during expert validation into revised formulations of items 1.3 and 1.4, which now explicitly incorporate human oversight requirements. Gap on low-confidence cases addressed by new item 2.6: automatic escalation mechanism flagging low-confidence or out-of-distribution cases for mandatory human review.
T5 had 4 items; no item on data deletion	Survey (V2→V3): participants expressed concern about sensitive data retention after processing	Added item 5.5 (permanent and verifiable deletion of sensitive personal data processed during the session)
T6 had separate overlapping items on feedback collection and UX governance	Survey (V2→V3): structural overlap between governance items identified	V2 items 6.2 + 6.3 merged into single item 6.2 (human-in-the-loop mechanisms + UX anti-patterns); new item 6.4 added (structured post-deployment monitoring). Radiologists (V3→V4): "critical training" in 6.4 is organisational, not implementable at software level — removed from final formulation
V1 item 3.6 "Was the worker involved in the decision to introduce AI, not just informed after the fact?"	Round 1 workshop: item removed as part of reduction from 30 to 27 items due to overlap with governance dimension	Governance concern absorbed into V4 item 3.5: "The system's functionalities were implemented following a process that actively collected and integrated workers' real operational needs" — original yes/no framing broadened into a positive design requirement

A6. Checklist Applied to Use Case

The following table applies the V4 checklist to the medical radiology use case described in Section 5. Each item is translated from a general design requirement into a concrete technical implementation specification for the AI triage system. T1 retains four items in V4; no additional items appear in this applied version.

Table 6: V4 Checklist Applied to AI Diagnostic Triage in Medical Radiology

	Checklist Items — Applied to AI Diagnostic Triage in Medical Radiology
T1 — Task Moral Delegability — Is the task appropriate for AI delegation?	
1.1	The software pipeline must map and expose only highly structured tasks (e.g., routine pre-checking of standard chest X-rays). The system must trigger a hard-stop if a user attempts to apply it to inherently high-risk, unstructured clinical scenarios without prior security module configurations.
1.2	The algorithmic backend must rely strictly on quantitative metrics and mathematical image patterns. The designer must programmatically exclude any attempt to simulate human empathy or handle sensitive patient interactions (e.g., delivering critical diagnoses), leaving these entirely outside the software's workflow.
1.3	Because diagnostic errors are highly irreversible and high-stakes, the software must never directly write final decisions to the hospital's primary databases (RIS/PACS). The architecture must implement a temporary "staging state" where the AI output remains fully modifiable or discardable by the clinician before final commitment.
1.4	The UI must incorporate a dedicated control panel allowing the radiologist to calibrate the model's operational thresholds in real-time (e.g., adjusting the sensitivity/specificity trade-off for nodule detection algorithms depending on the clinical context).
T2 — Accountability and Responsibility — Is it clear who is responsible?	
2.1	The designer must integrate a permanent banner or a mandatory login notification within the interface explicitly stating that ultimate legal and clinical responsibility rests solely with the human professional.
2.2	The system must feature a native logging module that records every logical step taken by the AI. This audit trail must be exported in a standardized, clear format accessible to both clinicians and engineering teams for verification.
2.3	The architecture must enforce hard-coded procedural blocks that prevent the final report from being exported or transmitted unless the radiologist's digital signature is validated, ensuring native compliance with relevant regulations like the EU AI Act.
2.4	The UI layout must establish functional equivalence. Buttons to modify, override, or reject the AI's generated data must share the exact same visual prominence, click accessibility, and ease of execution as the approval button, eliminating any asymmetric friction.
2.5	A dedicated, sandboxed telemetry database must be engineered. When a clinician flags an AI misclassification, this data is routed directly to the development backlog for debugging, programmatically ensuring the report is decoupled from the worker's operational performance metrics.
2.6	The designer must build an inbound data filtering routine: if the AI's confidence score falls below a safety threshold (e.g., <80%) or if the input is flagged as anomalous or corrupted (out-of-distribution), the system must trigger a mandatory escalation loop, routing the case directly to manual human review.
T3 — Autonomy, Agency and Professional Identity — Does the worker remain in control?	
3.1	The software must automate low-cognitive-value, repetitive steps (e.g., auto-populating metadata or initial bounding box segmentation), preserving the user's interface time for high-level creative and critical diagnostic thinking.
3.2	To mitigate automation bias and passive "rubber-stamping", the designer must structure an asynchronous UI pattern: the AI's specific findings or bounding boxes must remain hidden or blurred until the radiologist interacts with the raw DICOM image first.
3.3	The final output or diagnostic report format must explicitly log and visually highlight the specific inputs or edits made by the clinician (e.g., "AI suggestion reviewed and modified by Dr. X"), ensuring their authority and contribution are preserved within the clinical record.
3.4	Every string or visual tag generated by the model must be explicitly prefixed in the UI (e.g., "AI Suggestion:", "Model Recommendation:"), programmatically preventing the system or user from treating machine data as a definitive clinical truth prior to validation.

	Checklist Items — Applied to AI Diagnostic Triage in Medical Radiology
3.5	Throughout the development sprints, the engineering team must deploy API and interface designs that directly integrate the real-world operational needs collected from clinical stakeholders, preventing top-down workflow imposition.

T4 — Transparency and Explainability — Can the AI's reasoning be understood?

4.1	The designer must embed native Explainable AI (XAI) libraries. The UI must explicitly overlay the input image with clear attention heatmaps or causal features indicating exactly which regions drove the model's classification.
4.2	The interface must offer an interactive drill-down or querying function. By clicking a specific AI finding, the clinician must be able to verify the underlying decision criteria or call up contextual benchmarks for validation.
4.3	The software must never present a naked statistical percentage (e.g., "97% benign"). The UI must contextualize this with qualitative limitations and trigger an explicit visual alert if the image parameters approach the model's reliable operating range or signal an out-of-distribution anomaly.
4.4	The backend translation layer must convert raw data outputs into standard medical and radiological terminology, eliminating machine learning jargon (e.g., avoiding terms like "loss function optimization" or "raw feature weights") from user-facing logs.
4.5	The engineering team must enforce rigorous regression testing during deployment to guarantee stable, predictable, and deterministic model behavior: visually similar inputs must produce consistent output ranges, eliminating erratic or unexpected variations for the expert user.

T5 — Fairness, Bias and Third-Party Harm — Are external parties protected?

5.1	The software backend must include automated fairness-auditing scripts that monitor and stratify error rates across different demographic indicators (e.g., age, sex, ethnicity), blocking modules that demonstrate biased performance gaps from production.
5.2	The data ingestion pipeline must include a native de-identification module. Before any imaging data leaves local hospital servers to be processed by the AI, all sensitive patient identifier metadata within the DICOM tags must be securely masked or encrypted.
5.3	The software's technical documentation dashboard (Model Cards) must display verifiable distribution data regarding the training datasets, demonstrating that the AI was built on broad, heterogeneous populations to prevent demographic blind spots.
5.4	The software must provide an accessible "Export Patient Report" feature. This generates a non-technical, plain-language document that the radiologist can hand to patients, explaining the AI's role and the criteria used to determine their triage or diagnostic path so they can formally challenge it if necessary.
5.5	The architecture must enforce a strict data-wiping routine (Garbage Collection). Upon closing a case or ending an active session, any sensitive personal data or clinical images cached in RAM or transient storage partitions must be permanently deleted in compliance with data protection laws.

T6 — Human-AI Collaboration and Governance — Is the organizational structure adequate?

6.1	The software must not be architected as a standalone, asynchronous batch processor that works "in place of" the human. The data flows must require bi-directional check-ins, positioning the system strictly as an augmentation framework that enhances the clinician's capabilities.
6.2	The interface design must completely exclude deceptive user patterns (e.g., a massive, glowing, pre-selected "Approve All" button). Critical validation steps must require conscious, non-passive UX inputs (e.g., active manual confirmations) to counteract heavy shift workloads.
6.3	The designer must build an agile, in-app feedback tool directly into the workspace. This lets radiologists instantly flag usability bugs or clinical integration friction to the engineering team during active deployment.
6.4	The system architecture must output telemetry data to a dedicated post-deployment monitoring dashboard for hospital stakeholders. This dashboard continuously tracks live model accuracy and alerts administrators to data drift or performance drops compared to laboratory baselines.

A7. Team Members

Surname	Name	Course of study	University	Email
Accorroni	Lorenzo	Aerospace Engineering	Turin Polytechnic	s341682@studenti.p…
Casalegno	Luca	Management Engineering	Turin Polytechnic	s341804@studenti.p…
Leocata	Salvatore Daniel	Computer Engineering	Turin Polytechnic	s325651@studenti.p…
Liao	Livio	Management Engineering	Turin Polytechnic	s335883@studenti.p…
Lieggi	Thomas	Energy Engineering	Turin Polytechnic	s342299@studenti.p…
Mascherin	Jacopo	Management Engineering	Turin Polytechnic	s340620@studenti.p…
Muratori	Matteo	Aerospace Engineering	Turin Polytechnic	s338365@studenti.p…
Mustaj	Redi	Management Engineering	Turin Polytechnic	s341630@studenti.p…
Rinaldi	Pietro	Automotive Engineering	Turin Polytechnic	s341323@studenti.p…
Valori	Marco	Mathematical Engineering	Turin Polytechnic	s340261@studenti.p…
Ziarati Niasar	Davide	Computer Engineering	Turin Polytechnic	s321920@studenti.p…

A8. AI Use Disclosure

In accordance with academic integrity requirements, the following table documents all uses of AI tools in the production of this co-design checklist. All AI-assisted content was reviewed, verified, and validated by the human authors of this work, who bear full responsibility for its accuracy and conclusions.

Task	**AI Tool Used**	**How AI Was Used**	**Human Verification**
Checklist items highlight	Claude, Google Gemini	Highlight the key elements for every item in order to optimize them	All items and key elements were previously created by a human member of the group and the output was then verified
Appendix drafting	Claude	Drafting of Appendix sections based on our co-design process data	All content was verified against actual group records by the human authors
Survey distribution and data aggregation	Google Forms (built-in analytics)	Automated collection of responses, basic summary statistics (means, standard deviations)	Raw response data were manually inspected by some team members; outliers and open-ended comments were thematically coded by humans
Translation	Google Translate, ChatGPT	Preliminary translation of checklist items and participant notes	All translations were reviewed and corrected by fluent English speakers in the team; technical terms were verified for consistency