Co-Designing a Responsible AI Checklist

An Operational Framework for Bankruptcy Accountants

Trustworthy team | Deliverable 3

17/05

Politecnico di Torino
Turin, Italy

Abstract

This work explores how the growing integration of Artificial Intelligence into workplaces is reshaping professional routines, decision-making processes, and human responsibilities. Through a scoping review we identified the main factors that make AI trustworthy, organizing them into five interconnected dimensions: technical robustness, transparency, human oversight, psychological and cultural factors, and ethical governance. The research was later transformed into a practical checklist aimed at guiding AI-user interaction, promoting responsible, explainable, and human-centered use of large language models in professional contexts. Using the example of bankruptcy accounting, the study highlights the importance of maintaining human validation, explainable systems, and clear accountability while integrating AI into high-stakes decision-making environments. Overall, the project emphasizes that trustworthy AI depends not only on technological performance, but also on the quality of collaboration between humans, organizations, and intelligent systems.

Contents

1 Introduction

AI is becoming widely used: while more and more users approach to this new technology, introducing it in everyday life and workplaces, the entire world population is realising that its incredible and revolutionary computing and data analysis power is a resource we must embrace and use to make our life easier. Beyond simple automation, AI has the potential to transform the way people solve problems, make decisions, and access knowledge, allowing individuals and businesses to work more efficiently and creatively than ever before.

Lately we had the opportunity to deeply examine this new tool, trying to understand the logic behind the responses it gives us after a request, how it works, what are its limits and – most importantly – how its use can shape workplaces and working routines, impacting profoundly on how we'll have to approach the job's world in future. It became instantly clear that we were facing a radical change in working habits, leading to the elimination, replacement or generation of new jobs.

The new generations – expecially the ones that are entering in the world of work nowadays – have to get used to this technology, which is increasingly merging into workplaces. Owners, employees and clients will be more and more surrounded by these tools, and will progressively have to interact with it.

From this perspective, it becomes clear that humanity's need is no longer to ask if, when, and how this technology will become an everyday reality; rather, the new goal of society is to understand how best to collaborate to achieve the desired goals, questioning the effectiveness, efficiency, reliability, and trustworthiness of this tool.

With our professor, Mr Daniele Quercia, we had the opportunity to search for answers. We redacted a scoping review, in which we collected all the factors we identified to make AI trustworthy. We later had to deconstruct this research summarizing its contents and compose a new documents.

The new goal is to identifies the central themes around which all the identified factors revolve, finally structuring a checklist regarding AI-user interaction. Our conclusive thought was to make this list an instruction guide to AI that indicates what actions or habits are suggested or discouraged when we deal with LLMs algoritms.

To make this list as clear, understandable, impartial as possible we involved external people; we later implemented their suggestion to enrich the document, finally redacting a complete checklist on how we should interact with AI.

2 Method

2.1 Study Design: Iterative Co-Design

This scoping review details the development of a professional AI validation checklist for bankruptcy accounting. The methodology follows an iterative, four-stage co-design process intended to bridge the gap between theoretical AI ethics and the practical constraints of insolvency law.

2.2 Stage 1: Identification of Core Layers (V1)

The initial version (V1) was developed through a comprehensive literature review of AI trustworthiness. This stage established the five foundational layers of the framework: Technical Robustness, Cognitive Bridge, Control Layer, Internal Filter, and Systemic Shield.

2.3 Stage 2: Peer-Logic Validation (V2)

The checklist was refined into V2 through collaboration with technical evaluators (students in Mathematics and Mechanical Engineering). This stage focused on "peer-to-peer" testing, where individuals performing similar analytical work assessed the initial logic for technical integrity, usability, and computational consistency. The session produced two structural decisions for V2: the governance and psychological dimensions were treated as explicitly overlapping rather than parallel, and the surrounding-environment idea — that organisational improvement raises AI trust without touching the technology — was elevated from a sub-point to a design principle.

2.4 Stage 3: External Inquiry Expansion (V3)

The framework was expanded into V3 by presenting the checklist to an external group of non-experts and interdisciplinary stakeholders. This "Inquiry Expansion" phase tested the clarity of the validation questions, ensuring that the "Cognitive Bridge" (explainability) was understandable to those outside the immediate development team.

2.5 Stage 4: Sector-Specific Professionalization (V4)

The final version of both the checklist and the storyboard (V4) was refined through intensive consultation with Domain Experts, specifically court-appointed bankruptcy practitioners. This final stage calibrated the tool to the "Professional Perimeter," integrating high-stakes requirements such as *par conditio creditorum* (equal treatment of creditors), judicial source-verifiability, and strict confidentiality protocols. The expert consultation added confidentiality as a governing constraint across all five layers, formalised the four-role taxonomy, and introduced the eight lifecycle phases used to structure the final checklist.

2.6 Data Synthesis and Final Calibration

The synthesis of these four stages resulted in a tool that moves the professional from "blind trust" to "calibrated skepticism." The review concludes with a Storyboard Application, illustrating how the V4 checklist functions during a judicial liquidation, ensuring that AI-assisted decisions remain within the human-controlled "Professional Perimeter."

2.7 Summary of Methodology Evolution

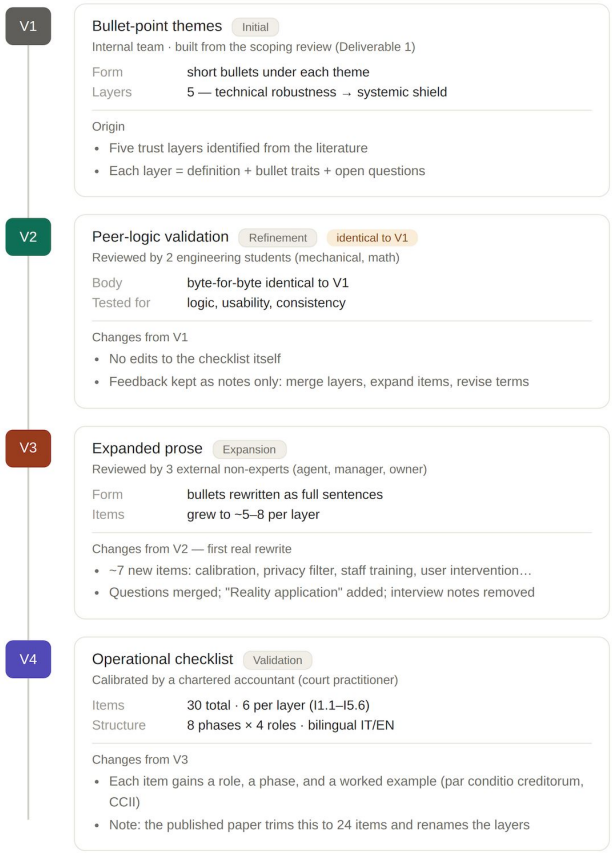

Figure 1: Visual overview of the iterative co-design process across the four checklist versions (V1–V4), showing the participant groups, form of each version, and principal changes introduced at each stage.

3 Checklist Framework

The checklist is organised around five interconnected dimensions derived from the scoping review. **Technical Robustness** addresses the accuracy, stability, and domain-appropriateness of AI outputs. **Transparency and Explainability** concerns the degree to which the system's reasoning can be inspected, traced, and communicated to third parties. **Human Oversight** covers the preservation of human agency and the insertion of validation checkpoints throughout the workflow. **Psychological and Cultural Factors** acknowledges that trust in AI is shaped not only by system performance but also by individual attitudes, workplace culture, and the perceived threat of displacement. **Ethical Governance** encompasses accountability, privacy compliance, and the organisational rules that make responsible AI delegation possible. Together, these five dimensions describe the conditions under which AI-user interaction can be considered trustworthy and human-centred. The operational checklist built on them is presented in Section 4.

Table 1: Overview of the iterative codesign process

Phase	Version	Target Audience	Focus of Refinement
Initial	V1	Theoretical Framework	Establishing the 5 Layers of Trust.
Refinement	V2	Technical Peers (Students)	Logic integrity and usability.
Expansion	V3	External Stakeholders	Clarity and universal interpretability.
Validation	V4	Sector Experts (Practitioners)	Legal calibration and professional defensibility.

Table 2: Demographics of the participants of the co-design process

Version	ID	Gender	Age	Education	Expertise
V2	P1	Male	20	Mechanical engineering	Student
V2	P2	Male	20	Mathematics engineering	Student
V3	P3	Male	52	Accounting diploma	Commercial agent
V3	P4	Female	53	Bachelor's in economics and commerce	Company manager
V3	P5	Male	54	Degree in mechanics and mechatronics	Owner, individual car service business
V4	E1	Male	54	Master's in economics and commerce	Qualified chartered accountant

4 Practical Checklist for AI-User Interaction

The checklist addresses the central question of this deliverable: *is trust in AI warranted for a given professional task, and how should that trust be calibrated?* It was produced through the four-stage co-design process described in Section 2, and it translates the five theoretical dimensions of Section 3 into concrete, role-assigned actions.

Two findings were consistent across all co-design rounds and directly shaped the checklist structure. First, governance and

psychological-cultural factors are deeply interdependent: reinforcing the surrounding environment of an AI deployment — its accountability structures, its leadership support, its error-reporting channels — improves trust in the tool without modifying the technology itself. Second, all participant groups stressed that items must function as prompts for professional reflection, not as binary compliance boxes. Every guideline therefore uses an action verb and names a responsible role, so that accountability is explicit and the item cannot be satisfied by a mechanical tick.

Each item is assigned to one of eight lifecycle phases (*Data Preparation, Configuration, Validation, Monitoring, Review, Deployment, Governance, Maintenance*) and to one of four roles: **BA** = domain expert; **IT** = technical lead; **AS** = operational staff; **MP** = strategic lead. The checklist is intentionally general: profession-specific application is reserved for the worked use case in Section 5.

Table 3: Operational checklist for responsible AI-user interaction (V4). Roles: BA = domain expert; IT = technical lead; AS = operational staff; MP = strategic lead.

ID	Guideline	Phase	Role
I1 — Technical Robustness			
I1.1	**Data quality check.** Verify that the AI processes only complete, accurate, and domain-appropriate data before starting any analysis.	Data Prep.	BA
I1.2	**Scope definition.** Configure the system to operate within a defined professional and regulatory field; exclude unrelated or generic sources.	Config.	IT
I1.3	**Stability test.** Submit equivalent inputs on separate occasions and confirm the system returns consistent outputs.	Validation	AS
I1.4	**Technical-legal validation.** Cross-check the AI's conclusions against the applicable regulatory framework before acting on them.	Review	BA
I1.5	**System integrity check.** Monitor that performance remains consistent when processing large or complex data volumes.	Monitoring	MP
I2 — Transparency and Explainability			
I2.1	**Require explanations.** Refuse to act on any output the system cannot explain step by step.	Validation	BA
I2.2	**Source inspection.** Verify independently that every source cited by the AI is real, current, and contextually relevant.	Review	AS
I2.3	**Decision justification.** Transform AI output into a traceable rationale communicable to all relevant stakeholders.	Deployment	BA
I2.4	**Document traceability.** Keep a log of all inputs and documents the AI consulted for each specific analysis.	Monitoring	IT
I2.5	**Algorithmic explainability.** Obtain from the provider documentation describing how the system's core criteria are weighted.	Config.	MP
I3 — Human Oversight			
I3.1	**Human checkpoints.** Insert a mandatory human approval stage before any AI output progresses to the next workflow step.	Review	BA
I3.2	**Skill preservation.** Perform periodic manual checks on a sample of AI outputs to prevent professional deskilling.	Monitoring	BA
I3.3	**Active intervention.** Ensure the user can correct or override any AI output at any point in the workflow.	Review	AS
I3.4	**Usability review.** Periodically evaluate whether the interface supports or hinders professional judgment.	Config.	MP
I3.5	**Stakeholder oversight.** Maintain active professional involvement throughout the process, not only at the final output stage.	Governance	MP
I4 — Psychological and Cultural Factors			
I4.1	**Trust calibration.** Use AI to support judgment, not replace it; avoid both blind reliance and unfounded rejection.	Monitoring	BA
I4.2	**Psychological safety.** Ensure the team understands that reporting AI errors is expected and never penalised.	Governance	MP
I4.3	**Expectation management.** Define explicitly what the system can and cannot do before deployment, and communicate this to all users.	Config.	MP
I4.4	**Trust feedback.** Collect periodic team assessments of the system's perceived reliability and act on the findings.	Monitoring	IT
I5 — Ethical Governance			
I5.1	**Responsibility attribution.** Formally establish that legal and professional responsibility for AI-assisted outputs rests with the human signatory, not the software.	Governance	MP
I5.2	**Ethical compliance.** Verify that the system does not produce outputs that breach professional duties or applicable regulations.	Governance	BA
I5.3	**Regulatory updates.** Implement a procedure to update AI parameters whenever the applicable legal framework changes.	Maintenance	IT
I5.4	**Reporting channel.** Create a structured procedure for logging and escalating AI errors or malfunctions to the provider or internal team.	Maintenance	AS
I5.5	**Data protection.** Confirm periodically that the system processes sensitive data in compliance with applicable privacy law.	Governance	IT

5 Worked Use Case: Bankruptcy Accounting

A bankruptcy accountant manages the insolvency procedure of a company that can no longer meet its financial obligations. The role requires reviewing accounts, reconstructing transaction histories, redistributing assets among creditors, debtors, and the state, and producing reports that are legally defensible before a judge. The stakes are high: an incorrect interpretation can harm creditors, invalidate a distribution plan, or expose the practitioner to professional liability. This context makes the trust-calibration question central — AI can accelerate analysis, but blind reliance on its outputs is not an option. Asset distribution follows the principle of *par conditio creditorum* — equal treatment of creditors according to the legal nature and priority of each claim — a non-negotiable legal constraint that the AI system must respect at every stage of the analysis.The learning process is also bidirectional: the AI learns from legal rules and professional practice, while the professional must be able to redirect the system toward more accurate legal or procedural interpretations when needed.

Applying the checklist. The following traces a single realistic sub-procedure — the detection of suspicious financial operations — through the five checklist dimensions.

Technical robustness (I1.1, I1.2, I1.4). Before starting the analysis the accountant verifies that the company's bank statements are complete and free of scanning errors (I1.1). The AI system is configured to draw exclusively from verified legal databases and official accounting manuals (I1.2, I1.4), preventing the tool from importing irrelevant precedents or generic web content into a confidential judicial file.

Transparency and explainability (I2.1, I2.2, I2.3). When the system flags a transaction as anomalous, the accountant requires a step-by-step explanation of the criteria applied before acting on the alert (I2.1). Every legal reference cited by the system is independently verified to confirm it has not been superseded by more recent case law (I2.2). The AI's reasoning is then restructured into a traceable written justification suitable for submission to the court (I2.3) — transforming a black-box flag into legally defensible evidence.

Human oversight (I3.1, I3.2). The liquidator countersigns the AI-produced draft before transmitting it to the court, maintaining a mandatory human checkpoint between algorithmic output and formal act (I3.1). A sample of flagged transactions is also manually recalculated to guard against silent systematic errors and to preserve the practitioner's own analytical competence (I3.2).

Psychological factors (I4.1, I4.3). The office explicitly scopes the AI to preliminary drafting tasks and clears final decisions — particularly those involving clawback actions — for human judgment only (I4.1). Associates are briefed at onboarding that the system cannot independently interpret ambiguous contractual clauses or ethically contested situations (I4.3), preventing the over-reliance that tends to develop after a run of correct automated outputs.

Ethical governance (I5.1, I5.2, I5.5). The engagement contract includes a clause confirming that every AI-assisted opinion is countersigned by a licensed professional before use (I5.1). Quarterly audits verify that debtor names and sensitive financial data are not being used to train external models (I5.5). When the Italian Crisis and Insolvency Code (*CCII*) is amended, the IT Manager updates the system's regulatory parameters within twenty-four hours, and the change is logged (I5.3).

Taken together, these applications show that the checklist functions not as a sequence of bureaucratic steps but as a navigational frame: it tells each role *when* to intervene, *what* to verify, and *who* bears responsibility for the outcome. The professional moves from blind trust toward what the co-design sessions consistently described as *calibrated scepticism*.

6 Discussion

Converting a scoping review into a checklist shifts the purpose of the work from knowledge collection to knowledge activation. The review established *what* conditions make AI trustworthy; the checklist operationalises *how* a practitioner can verify those conditions in daily use. This shift is not merely cosmetic: as the co-design sessions demonstrated, practitioners engage differently with a role-assigned action prompt than with an abstract principle. The presence of a responsible actor and a lifecycle phase transforms each item from a suggestion into an auditable commitment.

The co-design process itself produced findings that went beyond checklist refinement. The convergence across all participant groups on the interconnection between governance and psychological factors suggests that interventions aimed at improving organisational conditions — clearer accountability, psychological safety, accessible error-reporting — may have a larger effect on calibrated AI trust than improvements to the technology alone. This echoes a broader pattern in the human-factors literature: system reliability is necessary but not sufficient for appropriate trust; the social and organisational context in which the system is embedded matters equally.

At the same time, several limitations constrain the scope of these findings. The participant pool was small, drawn from a single national professional context, and skewed toward younger technical profiles in the earlier rounds. The worked use case addresses one specific sub-procedure within bankruptcy accounting; other tasks within the same profession, and professions in different regulatory environments, may require substantially different calibrations. The checklist should therefore be treated as a starting point rather than a definitive standard: a living document to be revised as tools evolve, regulations change, and new evidence from deployment accumulates.

7 Conclusion

Deliverable 3 shows that trustworthy AI is not only a technical issue. It depends on the interaction between system performance, explainability, human validation, psychological attitudes, and organizational responsibility. By converting the scoping review into a practical checklist, this work provides a first guide for users who want to interact with large language models in a more responsible and human-centered way.

This work set out to answer a question that the growing integration of AI into professional environments makes increasingly urgent: is trust in AI warranted for a given task, and if so, how should that trust be calibrated? The answer produced here is not a declaration but an instrument. The five-dimension framework derived from the scoping review was transformed, through four

rounds of iterative co-design, into a 30-item operational checklist that assigns concrete actions to named roles and lifecycle phases. The result moves the practitioner from passive acceptance of AI outputs toward what the co-design sessions consistently described as *calibrated scepticism*: a stance that is neither credulous nor dismissive, but structured and auditable.

Two findings from the co-design process deserve emphasis beyond the checklist itself. First, governance and psychological factors cannot be treated as separate dimensions: the organisational environment in which AI is deployed — its accountability structures, its error-reporting culture, its leadership signals — shapes individual trust as powerfully as the technology's own performance. Second, the format of a checklist item matters as much as its content. Items phrased as action prompts with a named responsible role elicited genuine engagement from all participant groups; items phrased as binary yes/no questions did not. These two findings have implications beyond bankruptcy accounting and beyond the specific tool produced here.

The interview transcripts included in the appendix reinforce both findings. Across all three interviews, participants independently converged on the same core concern: that confidentiality is not a secondary feature but a precondition for any form of professional trust. No accountability structure and no governance framework can substitute for the fundamental guarantee that sensitive documents, prompts, and conversations remain private. Equally, participants from different professional backgrounds — a commercial agent, a company manager, a car service owner, and a court-appointed insolvency practitioner — all expressed that the role of human judgment should never be fully delegated to an algorithmic system. The chartered accountant's formulation was the most precise: "I trust, but I always verify." This phrase captures the intended posture of the checklist more accurately than any theoretical definition of calibrated scepticism could.

A third implication emerges from comparing the structured interview data against the earlier peer sessions. The domain expert (E1) introduced constraints — *par conditio creditorum*, sector-specific knowledge perimeters, the eight-phase lifecycle — that no theoretical framework had anticipated. This confirms that co-design with practitioners is not merely a validation step but a discovery process: the professional context generates requirements that literature-based frameworks cannot fully predict. Future checklists in adjacent professions should therefore treat domain expert consultation not as a final calibration step but as an integral and structurally significant phase from the beginning.

The checklist should be understood as a starting point rather than a definitive standard. The participant pool was small and drawn from a single national context; the worked use case covers one sub-procedure within one profession; and the iterative versions, while effective at sharpening the tool, also expose its current limits — particularly the absence of deployment-stage testing and the reliance on a single domain expert for the final calibration round. The bankruptcy accounting use case illustrates how the framework can support professional judgment in a high-stakes context, but substantial adaptation would be required before applying it to other regulated professions.

Overall, the main contribution of this work is to provide a structured starting point for using AI more responsibly in professional decision-making. Trustworthy AI does not depend only on technical performance, but also on transparency, human oversight, ethical governance, and the professional context in which the tool is embedded. The checklist operationalises these principles into a set of concrete, role-assigned actions that encourage practitioners to move from passive reliance toward *calibrated scepticism*: combining source verification, human validation, documented reasoning, and clear responsibility allocation. Future work should extend the co-design process to larger and more diverse practitioner groups, test the checklist in live deployment settings, and investigate whether the role-and-phase structure transfers to professions operating under different regulatory frameworks. Trustworthy AI is not a property of systems alone — it is a product of the collaboration between systems, professionals, and the organisations that govern their interaction.

A Appendix

A.1 AI Tool Use Disclosure

In this project we used artificial intelligence tools collaboratively during the preparation and refinement of the deliverable. This disclosure follows the structure used in the previous scoping review and documents how AI support was combined with human verification.

Mandatory statement of verification. As required by the project guidelines, the team explicitly declares:

- All content was reviewed and approved by the authors.
- No final claim was accepted solely on the basis of an AI-generated answer.
- The checklist, use case, and conclusions remain the intellectual responsibility of the team.

Table 4: AI tool use disclosure and verification log.

Tool(s) used	Application	Verification
ChatGPT, Gemini, Claude	Brainstorming of checklist categories; wording refinement for professional users.	Team verified alignment with scoping review and use case.
ChatGPT	Reorganisation into ACM-compliant LaTeX sections; academic phrasing support.	Each section manually reviewed to preserve the authors' argument.
AI writing assistants	Formatting of tables, appendix material, disclosure text.	Final LaTeX structure checked by authors before submission.

A.2 Interview and Workshop Guide

The co-design sessions followed a structured protocol applied across all three rounds (V1→V2, V2→V3, V3→V4). Each session lasted approximately 20–45 minutes. Participants first read the current checklist version silently, then joined a facilitated discussion using the questions below.

Opening questions (all sessions):

(1) Is this instruction executable without ambiguity, or open to interpretation?
(2) Which professional role should bear responsibility for this check?
(3) What concrete example would you use to explain this item to a colleague?

Per-dimension probing questions:

Technical Robustness:

(1) Is high-quality, verified data fundamental for accurate AI responses?
(2) Does the AI operate within a defined professional perimeter?
(3) Does the platform guarantee that confidential documents and prompts remain private?
(4) Are AI outputs periodically tested for consistency and reliability?

Transparency and Explainability:

(1) Can you explain to a judge or client the logic behind the AI's results?
(2) Would you blindly accept the results provided by the AI?
(3) In the case of systematic errors, is it clear who is accountable?
(4) Does the AI show the sources used to generate its answer?
(5) Does the AI explain its reasoning without exposing confidential information?

Human Oversight:

(1) Does the human in the loop eliminate bias, or introduce a different cultural bias?
(2) Are there validation checkpoints before AI outputs enter professional documents?
(3) Can the user correct or override AI output at every stage?
(4) Does the AI provide alternative interpretations rather than autonomous decisions?

Psychological and Cultural Factors:

(1) Could increasing trust in AI lead workers in the wrong direction?
(2) Does the system encourage critical verification rather than blind reliance?
(3) Does the professional feel safe using the platform with confidential documents?
(4) Could repeated successful use increase the risk of excessive confidence?

Ethical Governance:

(1) Is it clear who is legally responsible when AI contributes to a decision?
(2) Does the organisation treat AI use as an ethical responsibility, not only efficiency?
(3) Does the AI provider formally guarantee confidentiality of projects and documents?
(4) Are there procedures for reporting errors and updating after legal changes?

A.3 Checklist Version 1 (V1)

V1 was produced directly from the scoping review. It identified five trust layers with sub-themes and exploratory questions to guide the first co-design session. No roles, phases or use-case examples were assigned at this stage.

L1 — Technical Robustness: precise and efficient tools; stability over time; quality of response; human-friendly design. *Key question:* Can quality answers be produced from poor-quality training data?

L2 — Cognitive Bridge (Transparency): XAI and interpretability; communication clarity; data provenance and rights; algorithmic accountability. *Key question:* Can you explain AI results to a judge? Would you blindly accept them?

L3 — Control Layer (Human Oversight): avoid full automation; human validation checkpoints; discourage automation bias; double-check with professional expertise. *Key question:* Does the human in the loop eliminate or reshape bias?

L4 — Internal Filter (Psychological Factors): automation bias; professional scepticism; fear and resistance; culture-shaped trust.

Key question: Could increasing trust in AI lead workers in the wrong direction?

L5 — Systemic Shield (Governance): clear accountability; legal compliance; supply-chain transparency; ethical stewardship; leadership and training; regular review. *Key question:* Is it clear who is legally responsible when AI contributes to a decision?

A.4 Checklist Version 2 (V2)

V2 expanded each layer with operational bullets and reality-application notes after two co-design sessions. The two main structural changes were: (1) governance and psychological factors treated as explicitly overlapping; (2) the surrounding-environment principle elevated to a design axiom.

L1 — Technical Robustness. Limit AI research to a defined professional and legal field; output quality depends directly on input data quality; ensure stability and consistency over time; use human-friendly interfaces; run periodic reliability checks; verify privacy settings before uploading confidential documents. *Reality:* Computational consistency lowers human error, but only with verified sources and confidentiality.

L2 — Transparency and Justification. Ensure explainability through XAI; transparency must cover reasoning, not confidential content; show documentation and case law behind every answer; verify algorithmic transparency before relying on results; inform users of data-processing agreements. *Reality:* Bankruptcy accountants must justify conclusions before judges; outputs must be traceable.

L3 — Human Oversight. Avoid full automation; AI should present options, not final decisions; human checkpoints preserve expertise; double-check outputs; train employees; allow user intervention at every stage. *Reality:* The accountant validates AI results before assuming professional responsibility.

L4 — Psychological and Cultural Factors. Over-reliance creates automation bias; calibrated scepticism ("I trust, but I always verify") is the target stance; fear lowers perceived reliability; trust is shaped by organisational culture; privacy is a key psychological precondition. *Reality:* AI performance alone does not produce well-calibrated trust.

L5 — Ethical Governance. Responsibility rests with the human signatory; comply with insolvency law, professional duties, and privacy rules; providers must formally guarantee confidentiality; maintain leadership support and error-reporting channels; conduct regular governance reviews. *Reality:* Improving surrounding governance improves AI quality as a consequence.

A.5 Anonymised Session Notes

Session 1 — V1 to V2 (Peer-Logic Validation)
Participants: P1 (Mechanical Engineering, 2nd year, regular AI user); P2 (Mathematics Engineering, 2nd year, occasional AI user).

Key feedback:

- Governance and psychological factors should be more explicitly connected; suggested treating them as overlapping dimensions.
- The surrounding-environment idea was the most original contribution.
- Several bullet points were flagged as too general; more specific language requested.
- Action: create a storyboard scenario applying all five layers to a real procedure.

Outcome: Items expanded, terminology revised, storyboard developed, true/false validation questions added.

Session 2 — V2 to V3 (External Inquiry Expansion)
Participants: P3 (Male, 52, accounting diploma, commercial agent); P4 (Female, 53, degree in economics, company manager); P5 (Male, 54, degree in mechanics, car service owner). None work in insolvency; selected to test intelligibility across backgrounds.

Key feedback:

- Items generally intelligible; "black-box" required brief explanation.
- Governance and psychological dimensions felt connected in their workplace experience.
- True/false format useful but insufficient for nuanced concerns.
- Reality-application notes requested for each layer.

Outcome: Reality-application notes added, questions refined, language simplified.

Session 3 — V3 to V4 (Domain Expert Consultation)
Participant: E1 (Male, 54, Master's in economics, qualified chartered accountant and court-appointed insolvency practitioner).

Key feedback:

- **Confidentiality** identified as the single most critical factor: no AI platform is acceptable without formal guarantees on judicial documents, prompts and conversations.
- Guiding principle: "I trust, but I always verify." Full automation never acceptable in judicial procedures.
- AI must present a range of reasoned options; the professional chooses.
- *Par conditio creditorum* flagged as a non-negotiable legal constraint.
- Roles formally assigned using four-role taxonomy; eight lifecycle phases formalised.

Outcome: V4 produced with 30 items, formal role and phase assignments, and use-case examples grounded in the Italian Crisis and Insolvency Code (CCII).

A.6 Interview Transcripts

The following transcripts record the semi-structured interviews conducted with external participants during the V3 co-design session and with the domain expert during the V4 session. All interviews were conducted in person and subsequently transcribed. Responses have been lightly edited for readability but no content has been altered.

Interview 1 — Andrea Fissore & Patrizia Sivera (Session 2, V3). Andrea Fissore, 52, sales agent (telecommunications sector); Patrizia Sivera, 53, administrative executive (private resale company). Interviewer = I; respondents referred to as A and P.

I: How often and in what way do you use AI tools?
A: Occasionally, mainly as a Google-style search engine. I try to limit my use.
P: Rarely, and very seldom in a professional context.

I: Do you think high accuracy in AI responses is linked to the quality of the data sources?
A: Absolutely. A precise answer starts with high-quality data, even if we cannot always determine whether data are truly reliable.
P: Extremely important in order to obtain reliable responses over time.

I: Does trust increase with user-friendly systems or with more technical tools?
A: I trust easier-to-approach systems more. It is natural to trust what we can understand.
P: I agree.

I: Do you blindly accept the results AI provides?
A: Usually not. I still trust the opinion of a specialist more.
P: I generally do not trust it completely; it is not always clear how it arrives at certain answers.

I: Is there transparency regarding who is responsible for an AI error?
A: No, but ultimately the responsibility lies with us for deciding whether to use the information provided.
P: It is not clear, and this raises an important point for discussion.

I: Are you aware of agreements regarding the use of personal data?
A, P: No.

I: Do you prefer full automation or human involvement in decision-making?
A: I absolutely want humans involved. We can never be certain that those programming the system are acting in good faith. I fear that the machine's "ethical code" may not be truly impartial.
P: Humans must remain in the process. Their presence reduces deskilling and prevents incorrect data from becoming the basis for further calculations. In person I can evaluate the cultural background of the decision-maker.

I: Do psychological factors influence your relationship with AI?
A: I approach it with scepticism regardless. AI has other ways—accuracy and stability—to demonstrate reliability.
P: Broader regulation and proper training could greatly influence my perception of safety and trust.

I: Do you agree that a comprehensive regulatory framework is needed to assign responsibility for AI errors?
A: Absolutely. Foundational laws assigning proper responsibility in the event of errors are extremely important.
P: I agree, though I distrust the ethics imposed on machines. Who can guarantee these laws serve ethical improvement rather than profit?

Interview 2 — Marco Vaccaneo (Session 2, V3). **Technical Robustness**
Q: Is high-quality input data fundamental? Can reliable answers be produced from poor data?
Absolutely yes. Starting from "garbage" data makes it almost impossible to obtain reliable outputs. Periodic checks on data accuracy—for instance, sampling tests using verified and certified data—are necessary.
Q: How does human-friendly design improve trust?
It reduces the cognitive barrier and puts operators at ease, increasing trust in the tool. A preliminary evaluation during the pre-purchase phase should verify that the interface is genuinely designed with human needs in mind.

Transparency and Justification
Q: Can you explain the logic behind AI results to a judge or client?
Yes, and this is crucial. It is indispensable to reconstruct the logical steps and criteria that led the AI to a specific output, especially in formal and legal contexts.
Q: Would you blindly accept AI results?
No. The real risk is cognitive laziness. Mandatory human "anti-bypass checks" must be structured to force critical review and prevent passive acceptance.
Q: In cases of systematic errors, is it clear who is accountable?
This must be managed upfront. The boundaries of liability in the event of system malfunctions must be defined and clarified contractually before the system goes live.
Q: Are you aware of the AI's agreements regarding client data?
Yes. Use of client data by AI models must be based on a clear pre-established agreement defining processing methods and limits before the system is deployed.
Q: Does the system certify that data is acquired respecting privacy and property rights?
Protecting privacy is an absolute priority. The system must offer strict guarantees against data leaks or unauthorised use of sensitive information.

Human Oversight
Q: Does the human in the loop eliminate bias, or shift it to cultural backgrounds?
Both. Externally, human presence makes AI use visible to the client, strengthening trust. Internally, it monitors employees' relationship with the tool, preventing rejection or waste of resources.

Psychological and Cultural Factors
Q: Can this theme enhance trust, or lead to deterioration?
It can increase trust, but this is not automatic. Constant monitoring is necessary to ensure users develop a healthy growth in trust without rejecting the tool.
Q: Can we assume that increasing trust always leads workers in the right direction?
No. Strict error control is indispensable. We must systematically analyse why AI fails and resolve anomalies, either manually or with support from the software manufacturer.

Ethical Governance
Q: Is it clear who is responsible when AI contributes to a decision?
This needs clarification in operational details. The first step is verifying that the system can technically analyse data specific to the relevant profession and regulatory framework.
Q: Does the organisation treat AI as an ethical responsibility?
Yes, ethical responsibility is central. The software must analyse data objectively and in the public interest, without reprocessing data to make results look artificially more correct or convenient.

Interview 3 — Pierpaolo Lollo (Session 3, V4). Qualified chartered accountant and court-appointed insolvency practitioner (E1). Interview conducted individually as part of the domain expert consultation.

Q1 (Privacy): What risks do you see in uploading corporate, accounting, or judicial documents into an AI system?
The main risks concern the handling of confidential information protected by investigative secrecy. In insolvency proceedings, documents may contain elements relevant to determining directors' liability and quantifying the "incremental financial damage." Risks include disclosure of sensitive data, indirect exposure through AI document analysis or legal opinion drafting, and the possibility that targeted searches reconstruct content that should remain confidential. In such contexts, protecting confidentiality must remain the top priority.
Q2 (Transparency): Should a professional be able to understand why AI produced a certain answer?

Understanding AI responses depends on the information available to the system. The lower the level of privacy protection, the more confidential information risks becoming public. This is particularly relevant in bankruptcy, where verification processes must remain confidential. The professional mandate includes an obligation of privacy that extends beyond what is objectively foreseeable.

Q3 (Trust): What should an AI system have to be considered trustworthy by a professional?

For use in insolvency activities, AI systems should guarantee that projects and prompt-based conversations remain private. This privacy must be officially guaranteed by the platform provider.

Q4 (Human Oversight): Are there decisions that should never be left solely to AI?

Leaving decision-making activities entirely to AI is not feasible. AI should provide a range of possible responses together with the reasoning supporting each, so the professional is informed about the basis for each determination and can make the final choice.

Q5 (Responsibility): How important is it to clarify who is responsible if AI produces an error?

Any AI error ultimately falls on the user who relied on it. Reports and opinions carry the signature of the professional, not the software. It might be appropriate to identify some form of AI responsibility for obvious errors, but this would be a difficult path to pursue.

Q6 (Bias): Is there a risk that AI produces biased evaluations influenced by training data?

Certainly, and I am fully convinced of it. AI results must come from training based on verified and demonstrably true information. Use should be sector-specific: in insolvency proceedings, the system should retrieve data only from relevant jurisprudence and from the Italian Crisis and Insolvency Code, without extending beyond those boundaries. Otherwise, responses drift into areas that make them inconsistent with the applicable legal reality.

A.7 Team Members

Team name: Trustworthy

Table 5: Team members.

Team member	Email
Saverio Medici	s322898@studenti.polito.it
Artemisia Lollo	s335817@studenti.polito.it
Andrea Antico	s326625@studenti.polito.it
Bruno Germanis	s343237@studenti.polito.it
Eugenio Costella	s342382@studenti.polito.it
Matteo Fissore	s340798@studenti.polito.it
Matteo Vaccaneo	s343865@studenti.polito.it
Alessio Nicotra	s325235@studenti.polito.it
Lorenzo Chiaramello	s338840@studenti.polito.it
Alessandro Denis Ciobanu	s311248@studenti.polito.it