computers

MDPI

*Article*

# trustSense: Measuring Human Oversight Maturity for Trustworthy AI

**Kitty Kioskli [1,*], Theofanis Fotis [1,2], Eleni Seralidou [1], Marios Passaris [1] and Nineta Polemi [1,3,*]**

[1] trustilio B.V., Vijzelstraat 68, 1017 HL Amsterdam, The Netherlands; theo.fotis@trustilio.com (T.F.); eleni.seralidou@trustilio.com (E.S.); marios.passaris@trustilio.com (M.P.)
[2] School of Education, Sport & Health Sciences, University of Brighton, Brighton BN1 9PH, UK
[3] Department of Informatics, University of Piraeus, 185 34 Piraeus, Greece
* Correspondence: kitty.kioskli@trustilio.com (K.K.); nineta.polemi@trustilio.com (N.P.)

**Abstract**

The integration of Artificial Intelligence (AI) systems into critical decision-making processes necessitates robust mechanisms to ensure trustworthiness, ethical compliance, and human oversight. This paper introduces trustSense, a novel assessment framework and tool designed to evaluate the maturity of human oversight practices in AI governance. Building upon principles from trustworthy AI, cybersecurity readiness, and privacy-by-design, trustSense employs a structured questionnaire-based approach to capture an organisation's oversight capabilities across multiple dimensions. The tool supports diverse user roles and provides tailored feedback to guide risk mitigation strategies. Its calculation module synthesises responses to generate maturity scores, enabling organisations to benchmark their practices and identify improvement pathways. The design and implementation of trustSense are grounded in user-centred methodologies, with defined personas, user flows, and a privacy-preserving architecture. Security considerations and data protection are integrated into all stages of development, ensuring compliance with relevant regulations. Validation results demonstrate the tool's effectiveness in providing actionable insights for enhancing AI oversight maturity. By combining measurement, guidance, and privacy-aware design, trustSense offers a practical solution for organisations seeking to operationalise trust in AI systems. This work contributes to the discourse on governance of trustworthy AI systems by providing a scalable, transparent, and empirically validated human maturity assessment tool.

**Keywords:** trustworthy AI; AI oversight; cybersecurity readiness; human cyber resilience; cyberpsychology; human factors

check for updates

Academic Editor: Paolo Bellavista

Received: 30 September 2025
Revised: 30 October 2025
Accepted: 1 November 2025
Published: 6 November 2025

**Citation:** Kioskli, K.; Fotis, T.; Seralidou, E.; Passaris, M.; Polemi, N. trustSense: Measuring Human Oversight Maturity for Trustworthy AI. *Computers* **2025**, *14*, 483. https://doi.org/10.3390/computers14110483

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

## 1. Introduction

Artificial Intelligence (AI) systems are increasingly integrated into decision-making processes across diverse sectors, including healthcare, transport, and finance, with the potential to significantly enhance operational efficiency and service delivery [1,2]. However, their adoption raises substantial concerns regarding trustworthiness, particularly when algorithmic decisions carry ethical, societal, and security implications [3,4]. Ensuring that AI systems operate in alignment with legal, moral, and ethical principles requires systematic evaluation frameworks that address both technical and socio-technical dimensions, while incorporating human factors into the governance of AI-driven processes [5,6].

Trustworthiness risk management in AI is closely aligned with cybersecurity risk management practices but extends beyond them by incorporating all dimensions of trustworthiness. This includes fairness, accountability, and explainability, in addition to robustness, quality, cybersecurity, and privacy-by-design principles. Previous research has demonstrated that neglecting human oversight in AI governance can exacerbate risks of bias, reduce system reliability, and undermine organisational resilience to cyber threats [7–10]. At the same time, sector-specific regulations, including the EU AI Act and the NIS2 Directive, impose stricter obligations for monitoring, auditing, and safeguarding AI systems throughout their lifecycle [11,12]. Under the EU AI Act, human oversight means that high-risk AI systems must remain under meaningful human control. This requires designing systems so humans can understand their outputs and limitations, interpret results correctly, and intervene or override decisions when needed. Oversight also involves preventing automation bias, ensuring operators are properly trained and competent, and tailoring the level of supervision to the system's risk and context. These evolving regulatory and operational requirements highlight the need for practical tools that assess and improve the maturity of human oversight in AI systems.

In this context, this paper introduces trustSense, a maturity assessment tool designed to evaluate and enhance human oversight capabilities in AI governance. Grounded in the principles of trustworthy AI, cybersecurity readiness, and privacy-by-design, trustSense provides structured questionnaires, targeted feedback, and a privacy-preserving architecture to support organisations in identifying oversight gaps and implementing corrective measures. The objective of this work is to present the conceptual design, implementation, and validation of trustSense, demonstrating its capacity to generate actionable insights that strengthen organisational readiness, promote compliance with emerging regulations, and ultimately foster greater trust in AI-enabled socio-technical systems.

## 2. Background and Related Work

The evolving landscape of AI governance and assurance is marked by the proliferation of evaluation instruments, ranging from principle-guided self-assessments to enterprise risk frameworks, each serving distinct organisational audiences and oversight objectives. The European Commission's Assessment List for Trustworthy AI (ALTAI) [13] has become a foundational benchmark for self-evaluation, offering governance-level due-diligence structured around seven requirements. Complementary frameworks such as NIST's AI Risk Management Framework (AI RMF) and its Playbook translate trustworthiness into risk-based actions, enabling organisations to Govern, Map, Measure, and Manage AI risks [7]. Similarly, the OECD's catalogue of tools and metrics aggregates a rich repository of methods and instruments intended to operationalise AI principles with appropriate fit-for-purpose selection.

- The General-Purpose AI (GPAI) published in July 2025, serves as a set of guidelines to support compliance with the AI Act until EU formal standards expected in August 2027 or later will be established [14].
- Oxford Insights (Trustworthy AI Self-Assessment) [15,16]: Oxford Insights provides a downloadable workbook-like self-assessment tool designed for public-sector policymakers to evaluate governmental AI readiness in trustworthy adoption. The format emphasises transparency, structured scoring and documentation, promoting iterative improvement within government teams. This artefact serves as an accessible, low-infrastructure governance baseline for public entities.
- Alan Turing Institute (Trustworthiness Assessment Tool) [17]: Within the Trustworthy Digital Infrastructure programme, the Alan Turing Institute's Trustworthiness Assessment Tool focuses on digital identity systems. It operationalises dimensions such as

legal compliance, security, privacy, performance, and ethics into a self-assessment with accompanying system documentation templates. While narrowly focused, it exemplifies sector-specific assurance mechanisms.

- AI4Belgium (Tool for Assessing an Organisation's AI Trustworthiness) [18]: This online instrument implements ALTAI logic in an organisational internal-assessment context, facilitating strengths and gaps identification through guided questions and alignment with EU ethical principles. It reflects community-driven adaptation of the ALTAI model for practical organisational readiness.
- University of Basel (Trust Questionnaire in the Context of AI): Researchers at the University of Basel have developed and validated psychometric questionnaires, such as variations of the Trust in Automation scale, specifically adapted for AI contexts. These instruments measure user-perceived trust and distrust dimensions, providing empirical metrics for human–AI interaction assessment.
- KPMG (AI Risk and Controls Guide-Trusted AI framework) [19,20]: KPMG's guide aligns with its Trusted AI framework, a values-driven, human-centric, and trustworthy approach, presenting an organised inventory of AI risks framed across ten ethical pillars (e.g., fairness, transparency, accountability). It supports risk identification and control design as integral components of organisational AI governance.

Beyond these principal examples, the broader ecosystem includes:

- Governmental and policy playbooks and self-assessments: Such as Canada's Algorithmic Impact Assessment and the UK's ICO AI and Data Protection Risk Toolkit [21], which operationalise regulatory compliance through structured questionnaires.
- Testing and external assurance toolkits: Such as AI Verify (testing fairness, explainability, robustness), Z-Inspection® (holistic ethical evaluation through applied-ethics processes) [22], IEEE CertifAIEd™ (certification of autonomous systems), and the Swiss Digital Trust Label for AI services.
- Technical libraries and operational dashboards: Such as IBM's AI Fairness 360 [23], Microsoft's Responsible AI Dashboard [24], and Google's Responsible AI resources, which deliver actionable model-level fairness and explainability capabilities embedded in MLOps workflows.

Recent research has also explored model-level trustworthiness interventions that address phenomena such as hallucination and reliability in multimodal systems [25], demonstrating that trustworthy AI spans both human oversight and technical mitigation dimensions. Table 1 provides a full list of relevant tools found in the literature. These instruments collectively represent three distinct yet overlapping assurance currents:

- Principle-to-practice self-assessments: Offering accessible governance baselines (e.g., ALTAI, Oxford Insights, AI4Belgium).
- Risk-and-controls frameworks: Integrating AI trustworthiness within enterprise risk governance (e.g., NIST RMF, KPMG). EC projects (FAITH [26] THEMIS [27]) have proposed AI trustworthiness risk management frameworks; e.g., the AI-TAF framework [28,29].
- Technical testing and certification ecosystems: Providing evaluative depth and external assurance (e.g., AI Verify, Z-Inspection®, Digital Trust Label).

**Table 1.** Trustworthy AI Assessment and Risk Tools.

| Tool/Framework | Provider | Type | Scope | Methodology |
| --- | --- | --- | --- | --- |
| Trustworthy AI Self-Assessment | Oxford Insights | Self-assessment | Government AI readiness and governance | Workbook with scoring |
| Trustworthiness Assessment Tool | Alan Turing Institute | Self-assessment | Digital identity systems | Structured questions + documentation templates |

**Table 1.** *Cont.*

| Tool/Framework | Provider | Type | Scope | Methodology |
|---|---|---|---|---|
| Tool for Assessing AI Trustworthiness | AI4Belgium | Self-assessment | Organisational AI governance | ALTAI-based questionnaire |
| Trust Questionnaire in AI Context | University of Basel | Psychometric survey | User trust in AI systems | Validated trust/ distrust scales |
| AI Risk and Controls Guide | KPMG (Trusted AI) | Risk and controls framework | AI risks and control design | Risk identification + control catalogue |
| Assessment List for Trustworthy AI (ALTAI) | European Commission | Self-assessment | Trustworthy AI principles compliance | 7 requirements with guidance |
| AI Risk Management Framework (AI RMF) | NIST | Risk management framework | AI risks across lifecycle | Govern/Map/ Measure/Manage |
| Catalogue of Tools and Metrics | OECD [30] | Repository | Operationalising AI principles | Curated tools and methods |
| Algorithmic Impact Assessment (AIA) | Government of Canada [31] | Impact assessment | Automated decision systems in government | Impact level + mitigation actions |
| AI Data Protection Risk Toolkit | UK ICO | Risk toolkit | GDPR compliance for AI | Step-by-step risk ID and mitigation |
| AI Verify | Singapore IMDA | Testing framework | Fairness, explainability, robustness | Software toolkit + reports |
| Z-Inspection® | Z-Inspection® Initiative | Ethics assessment process | Applied ethics for AI | Holistic evaluation process |
| IEEE CertifAIEd | IEEE | Certification programme | Ethics of autonomous systems | Assessment + certification |
| Digital Trust Label | Swiss Digital Initiative | Trust label | Digital services incl. AI | Audit against label criteria |
| AI Fairness 360 (AIF360) | IBM | Technical library | Bias detection and mitigation | Open-source metrics and algorithms |
| Responsible AI Dashboard | Microsoft | Technical dashboard | Fairness, explainability, error analysis | Integrated MLOps tools |
| Responsible AI Resources | Google | Guidance + tools | Responsible AI practices | Evaluation guidelines + tooling |
| Trustworthy AI Self-Assessment | Oxford Insights | Self-assessment | Government AI readiness and governance | Workbook with scoring |

Despite the richness of the trustworthy AI landscape, most existing tools, such as ALTAI, NIST AI RMF, and AI Verify, focus on system-level or process-level properties and external certification, rather than the maturity of organisational human-oversight capabilities. These instruments employ heterogeneous metrics, data requirements, and scoring scales that are not directly interoperable, making quantitative benchmarking infeasible within the present study. Instead, trustSense adopts a complementary and qualitative benchmarking approach that positions human oversight maturity as a measurable organisational attribute, bridging governance, risk management, and systems assurance. Through role-specific questionnaires, maturity scoring, tailored feedback, and a privacy-preserving architecture, trustSense enables organisations to assess and enhance oversight readiness. Future research will extend this work by developing harmonised indicators to enable quantitative benchmarking and cross-tool comparisons once compatible datasets become available.

## 3. The trustSense Tool

Trust in AI systems is shaped not only by the robustness of models or the reliability of datasets but also importantly by the maturity, accountability, and readiness of the human teams that design, develop, integrate, operate, and protect them.

trustSense is a powerful human assessment tool that helps organisations evaluate the maturity and preparedness of their teams working in AI and cybersecurity operations.

It allows AI technical teams to measure and strengthen their ability to manage trust-related risks in AI systems, while also assisting AI domain users in evaluating their knowledge and responsible use of these technologies. For cybersecurity defenders, trustSense serves as a resource to gauge their effectiveness in handling cyber risks and responding to incidents, and for cybersecurity investigators, it supports the examination of adversary sophistication. By delivering focused and practical insights, trustSense enables organisations to develop AI ecosystems that are more trustworthy, secure, and resilient.

The tool has been iteratively validated through pilot studies, expert input, and benchmarking against established maturity models, ensuring both contextual relevance and methodological rigour. By delivering targeted, actionable insights, trustSense enables organisations to identify strengths and address vulnerabilities, thereby developing AI ecosystems that are more trustworthy, secure, and resilient.

The tool can be accessed through the following link: https://trustsense-xu4xd.ondigitalocean.app/Index.html (accessed on 30 October 2025).

### 3.1. Scope and Roles

The scope of trustSense is defined by its ability to evaluate both the socio-technical and organisational dimensions of AI trustworthiness. Unlike approaches that concentrate exclusively on system-level features, trustSense explicitly integrates the maturity of human teams and their capacity to manage risks. This dual perspective reflects the recognition in international standards and regulatory frameworks that trust in AI depends not only on models, data and datasets but equally on the preparedness and accountability of the people and organisations that design, deploy, and safeguard them.

From a governance perspective, AI trustworthiness involves multiple roles across the AI lifecycle of design, development, deployment, and operation. These include designers (defining objectives, data requirements, and ethical constraints), developers (building and testing models), deployment participants (integrating systems into workflows and ensuring compliance), operators and monitors (supervising performance and responding to incidents), and evaluators (providing testing, assurance, and validation). These lifecycle roles provide the conceptual grounding for the trustSense framework.

To translate this model into practice, trustSense structures its assessment around four dedicated respondent categories, each supported by a tailored questionnaire:

- AI Technical Teams: responsible for system development, validation, and robustness testing, with emphasis on data quality management, adversarial resilience, and explainability.
- Domain AI Users: professionals applying AI outputs in decision-making (e.g., clinicians, financial analysts), focusing on interpretability, ethical awareness, and responsible use of AI recommendations.
- Cybersecurity Defenders: teams safeguarding AI systems against cyber threats, with emphasis on incident response readiness, resilience mechanisms, and proactive defence strategies.
- Adversary Profiling by Investigators: cybersecurity investigators (e.g., forensic analysts) complete the adversary questionnaire based on evidence from known or simulated cases. This enables the tool to profile attacker sophistication, ranging from insufficient to multi sectoral expert, by grounding the assessment in documented behaviours and capabilities rather than self-reported data.

The design of these questionnaires is grounded in prior socio-technical research by Kioskli and Polemi [25–27], whose work on incorporating psychosocial and behavioural factors into cyber risk assessment provided the foundation for trustSense's maturity dimensions.

In practice, trustSense is designed for flexible adoption across organisations of different sizes and levels of maturity. For small teams (e.g., a hospital unit of 5 clinicians using AI-based diagnostic support), contributions are usually gathered collaboratively during a meeting, with the Team Manager submitting a consensus response. This ensures that the tool captures shared practices without attributing responses to individuals. For larger teams (e.g., a financial services provider with dozens of analysts using AI risk models), trustSense is deployed through anonymous links that allow each member to respond independently. The system aggregates these inputs automatically, producing team-level scores while preserving individual anonymity (see Figure 1).

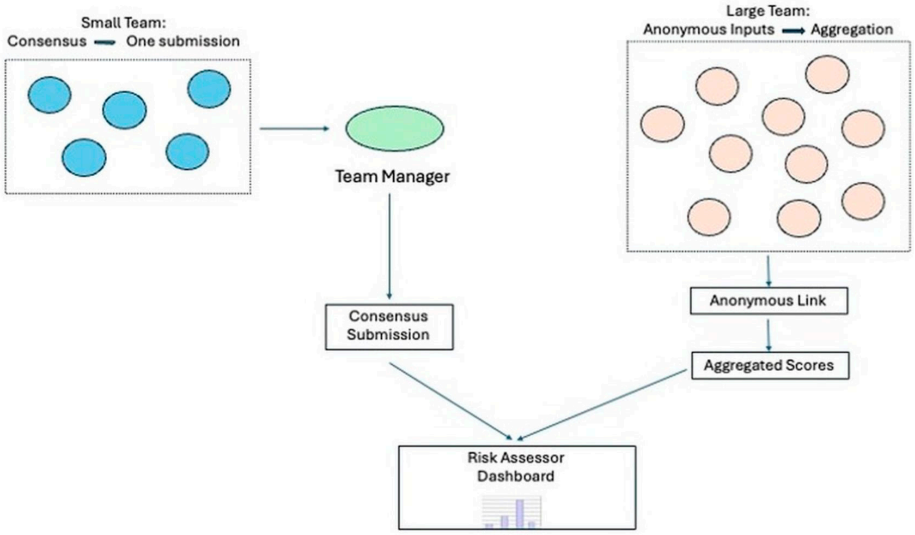


**Figure 1.** trustSense Submission Models for Small and Large Teams.

This dual approach supports inclusivity and scalability: smaller groups benefit from structured consensus-building, while larger teams gain from distributed, anonymous participation. In both cases, the Risk Assessor receives aggregated results via dashboards, which benchmark maturity and provide tailored guidance. This design ensures that trustSense can be used effectively in diverse contexts, from specialised research labs to enterprise-scale cybersecurity operations.

### *3.2. Questionnaires and Dimensions*

The trustSense tool operationalises its maturity model through four role-specific questionnaires, covering AI technical teams, domain AI users, cybersecurity defenders, and investigators responsible for adversary profiling. The four team-based questionnaires share a common foundation, diverging only in a block of technical proficiency items tailored to each role, while the adversary profiling instrument is distinct and is completed by investigators based on evidence from real or simulated incidents.

The development of these questionnaires builds directly on our previous socio-technical research [32–34], which demonstrated that incorporating psychosocial and behavioural factors into risk models produces more realistic vulnerability assessments and attacker characterisations. That earlier work established that human attitudes and behavioural traits are central to cyber risk, influencing both how organisations manage threats and how adversaries exploit vulnerabilities. Extending this line of inquiry, we subsequently reframed these constructs within a collective, team-based maturity model, thereby shifting the focus from individual-level perceptions to organisational readiness and responsibility [35].

A substantial body of empirical research further validates the inclusion of psychosocial traits in models of AI trustworthiness. Studies in psychology and human–computer interaction have shown that dispositions such as openness to experience, trust propensity, and affinity for technology shape user confidence and reliance behaviours in AI-mediated contexts [36–38]. Other work has highlighted demographic and contextual factors, including gender, digital literacy, and prior experience with automation, as additional determinants of trust [39,40]. Research in HCI further demonstrates that transparency, explainability, and adaptability are essential for calibrating trust, enabling appropriate reliance without overconfidence or scepticism [41,42]. Together, these studies reinforce the premise that organisational maturity models must account for both technical practices and human behavioural dynamics, which is precisely the orientation adopted in trustSense.

The validated questionnaire therefore measures maturity across a set of human-trait dimensions, including proactivity and threat awareness, responsibility and ethics, innovation and adaptability, resilience, collaboration and knowledge sharing, integrity, problem-solving, resource accessibility, policy adherence, motivation and commitment, privacy and compliance, and openness to interventions (Table 2). A variable technical proficiency block complements these traits: for technical teams, it captures advanced practices in data governance, robustness, and explainability; for domain users, it reflects the ability to critically interpret outputs and recognise bias; for defenders, it relates to incident detection and resilience.

**Table 2.** Human-trait maturity dimensions in the trustSense questionnaires.

| Dimension | Description |
| --- | --- |
| Proactivity and Threat Awareness | Anticipation and mitigation of technological, operational, and social AI risks. |
| Responsibility and Ethics | Collective adherence to ethical principles, standards, and accountability. |
| Innovation and Adaptability | Ability to implement new mitigation actions and revise processes after errors. |
| Resilience | Recovery capacity after incidents; maintenance of continuity and performance. |
| Collaboration and Knowledge Sharing | Structured practices for exchanging insights, training, and strengthening intelligence. |
| Integrity | Consistent adherence to legal requirements and professional codes of conduct. |
| Problem-Solving | Effective resolution of challenges through interdisciplinary collaboration. |
| Resource Accessibility | Availability of technological resources and engagement with external expertise. |
| Policy Adherence | Compliance with governance frameworks and AI-related policies. |
| Motivation and Commitment | Sustained engagement with responsible AI through training and ethical reviews. |
| Privacy and Compliance | Prioritisation of privacy protection and regulatory compliance. |
| Openness to Interventions | Willingness to accept external feedback and adapt practices (reverse-coded if resistant). |
| Technical Proficiency * | Role-specific expertise: advanced data/model practices (technical teams), critical interpretation of outputs (domain users), or cyber resilience (defenders). |

* Technical proficiency is the only dimension that varies systematically between respondent types.

The adversary profiling questionnaire complements these maturity assessments by enabling investigators to classify attacker sophistication based on evidence from threat

intelligence or red-teaming exercises. Attributes such as persistence, adaptability, technical expertise, and resource availability are scored to produce categories ranging from Insufficient to Sophisticated multi-sectoral expert [40,41]. This adversarial dimension ensures that trustworthiness assessments are contextualised not only by organisational maturity but also by the threat landscape in which AI systems operate.

### 3.3. User Interaction and Feedback

User interaction in trustSense is structured to maximise accessibility, anonymity, and practical value for organisations. Responders access the questionnaires through shared team links, which remove the need for personal accounts or identifiable information. Individual answers are processed locally in the browser and transmitted only as anonymised numeric values. These are immediately aggregated at the team level, ensuring that no individual data are stored. The designated Risk Assessor can then access the aggregated scores through dashboards, which visualise maturity levels and adversary profiles alongside targeted recommendations.

The design of the questionnaires was informed by a co-production methodology, validated through workshops and stakeholder engagement. Participants from sectors including healthcare, media, maritime and cybersecurity, reviewed questionnaire items for clarity, relevance, and completeness, leading to refinements in wording, structure, and scoring. This participatory process ensured that the tool's content was comprehensible and context-sensitive, while the final deployment of trustSense provides a stable, standardised interface rather than one that evolves in real time.

Feedback is provided in two ways. First, dashboards display aggregated team results, benchmarking scores against maturity categories and showing trends over time. These visualisations are accompanied by tailored guidance calibrated to the maturity level: teams with lower scores are directed towards training and structured workshops, while teams with higher scores receive recommendations focused on knowledge sharing, mentoring, and sustaining excellence. Second, adversary profiling, completed by investigators, contextualises organisational scores by classifying attacker sophistication (e.g., insufficient, basic, moderate, experienced, or sophisticated multi-sectoral expert).

Together, these mechanisms close the loop between assessment and organisational action. trustSense does not alter its questionnaires dynamically, but it enables teams and risk assessors to interpret results quickly, identify vulnerabilities, and plan targeted interventions.

## 4. trustSense Design and Implementation

trustSense is a lightweight, web-based assessment tool developed by trustilio to evaluate the maturity and readiness of teams engaged in AI and cybersecurity. Its design incorporates four participant profiles aligned with the Calculation Module: AI technical teams, AI domain users, cybersecurity defenders, and adversary profiles. This structure ensures that both technical conditions and human-centred factors of trustworthiness are systematically addressed. The tool applies role-specific questionnaires to assess ethical, procedural, and operational practices, generating immediate feedback through visual dashboards and tailored mitigation actions. Its implementation prioritises usability, privacy, and compliance: it operates entirely within the browser, requires no login or data storage, and adheres to GDPR principles by discarding all inputs after each session. A dedicated Section 5, outlines how trustSense outputs, such as team maturity scores and optional adversary profiles, can be integrated into AI trustworthiness risk-management processes.

### *4.1. Personas and User Flows*

Figure 2 depicts the workflow of the trustSense risk assessor: the assessor initiates the tool, chooses the relevant domain (AI or cybersecurity), and responds to structured questions regarding team maturity. trustSense subsequently generates a trustworthiness or sophistication score along with customised mitigation recommendations. The assessor reviews these outputs, applies the suggested measures where appropriate, and records the score for use in the subsequent external phase of the process.

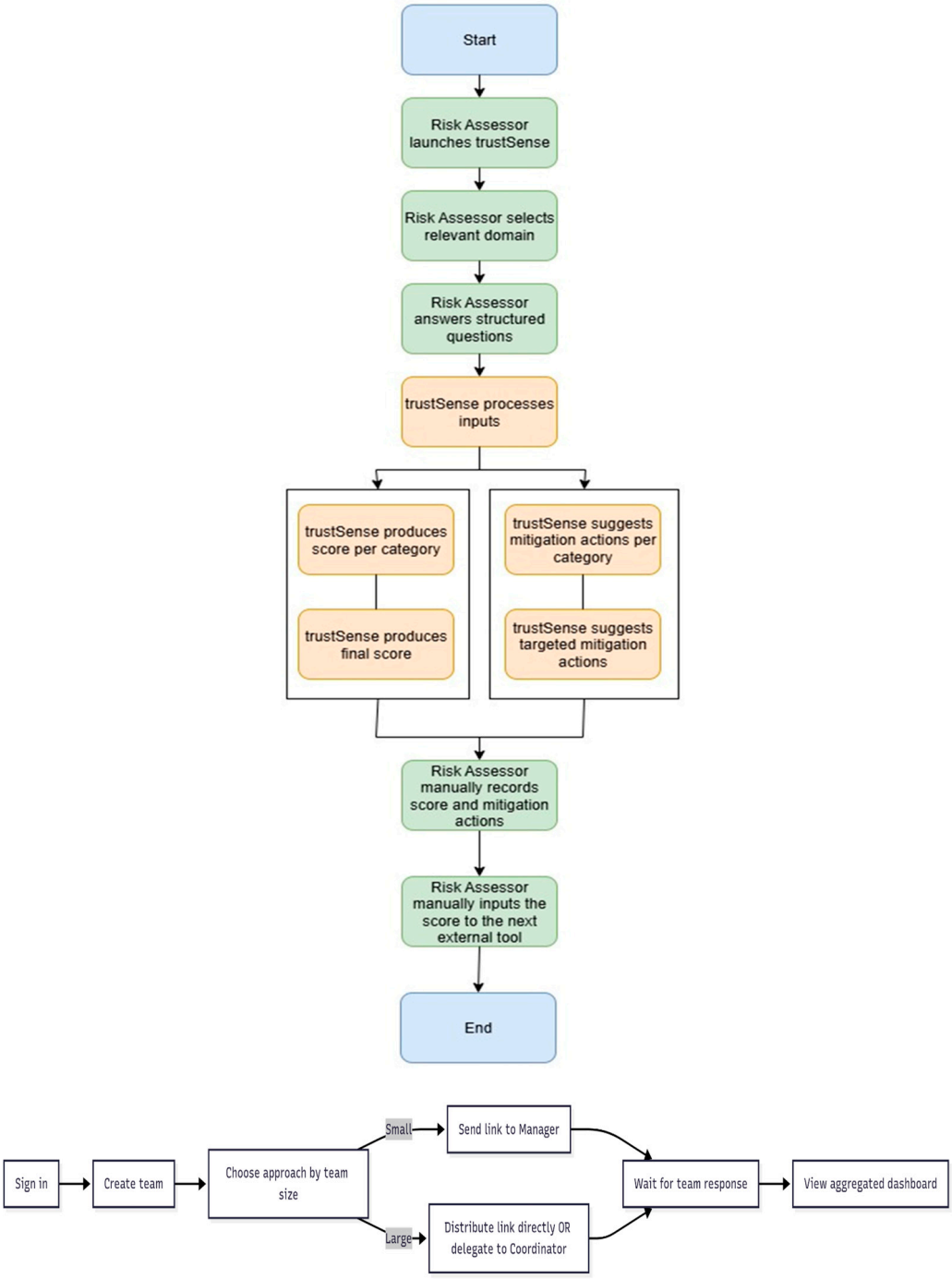


**Figure 2.** Risk assessor flow.

While Figure 2 illustrates the core pathway for the risk assessor, Figures 3–6 demonstrate how the framework embeds flexibility by adapting to different organisational structures. In smaller teams (Figure 3), decision-making remains direct and agile, whereas in

larger teams (Figures 4 and 5) the division of responsibilities between coordinators and respondents enables scalability without losing accountability. Figure 6 further emphasises the governance dimension, showing how system administrators safeguard the integrity and continuity of the process. Collectively, these flows highlight how trustSense operationalises both adaptability and oversight, ensuring that risk assessment remains robust across diverse contexts.


**Figure 3.** Team manager flow (small team).


**Figure 4.** Team coordinator flow (large team).


**Figure 5.** Team respondent flow (large team).


**Figure 6.** System administrator flow.

### *4.2. Architecture and Technology Stack*

The trustSense architecture is designed to balance scalability, data security, and anonymity, ensuring that both small and large teams can be supported without compromising integrity. This balance was achieved through a modular, cloud-native design that separates collection, aggregation, and visualisation; a two-tier operating model (offline consensus for small teams; shared anonymous links with real-time aggregation for large teams) that scales horizontally; and a managed stack (Next.js/PayloadCMS, PostgreSQL JSONB, DigitalOcean App Platform) that delivers autoscaling and resilient CI/CD. It was achieved as well by embedding privacy-by-design controls end-to-end: links carry only a teamToken; scoring occurs client-side; the API accepts only {teamToken, numeric score} and updates per-team running sums and counts while discarding individual values; no IPs/timestamps/identifiers or audit trails are retained; all transport is HTTPS; and Risk Assessor access is constrained via JWT/Firebase Auth, ensuring technical anonymity and GDPR compliance.

Figures 7–10 provide a layered view of this design, moving from a high-level system perspective to detailed data flows, decision logic, and technology choices.

As shown in Figure 7, the architecture follows a modular design that separates data collection, aggregation, and visualisation. This separation not only streamlines performance but also reinforces privacy by limiting access to raw inputs at each stage.

Figure 8 highlights how anonymity is preserved throughout the data flow: respondents interact only through a teamToken, while the API aggregates scores in real time without retaining individual-level data. This approach ensures compliance with privacy principles while still providing actionable insights at the group level.

Every survey link embeds only a teamToken, devoid of personal identifiers. The browser transmits a single numerical value, representing the respondent's score, to the server. The API then aggregates these scores by maintaining a running sum and count for each team, with raw individual scores being systematically discarded post-aggregation. Upon the Risk Assessor's access to the dashboard, the server furnishes the team's current

average score alongside relevant trend data. This information is subsequently visualised through charts and presented as targeted mitigation advice. See Figure 9.

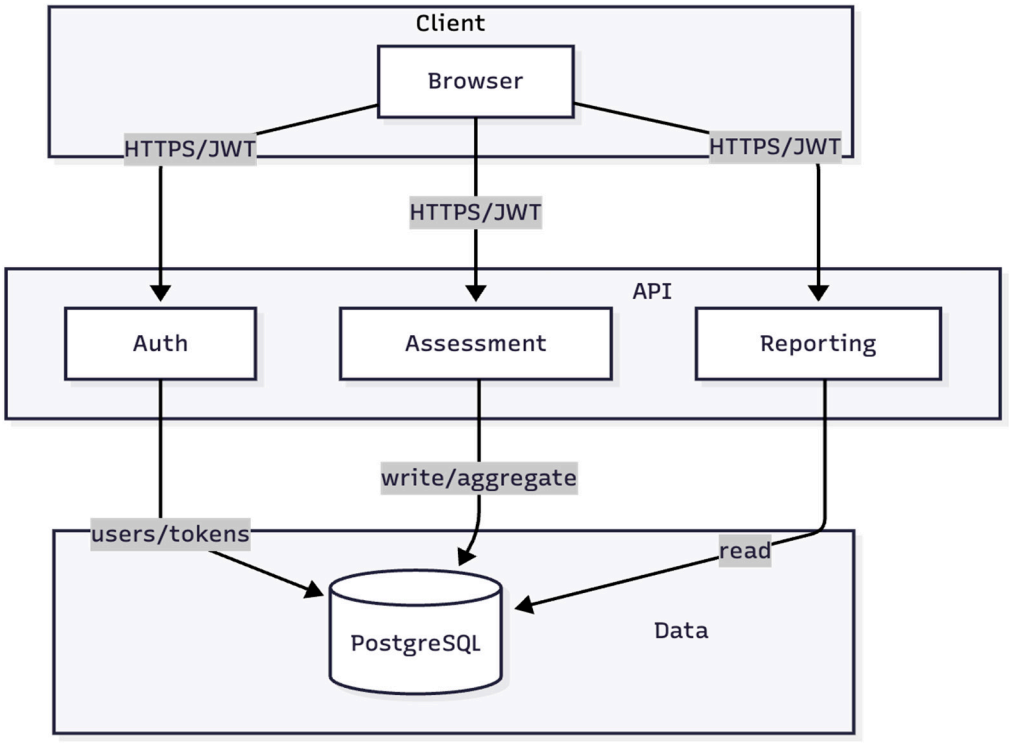


**Figure 7.** High level architecture diagram.

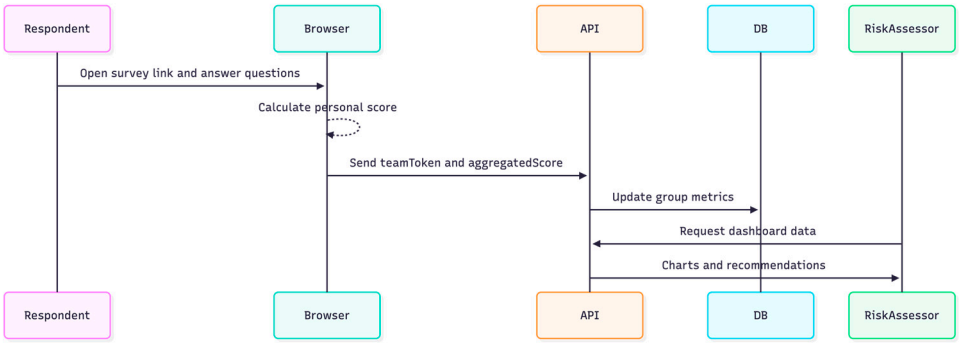


**Figure 8.** Data flow from Respondent to RiskAssessor using version 2 (Huntingdon, UK).

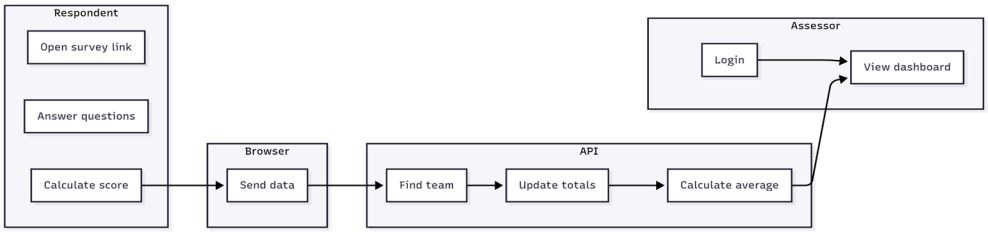


**Figure 9.** Business logic.

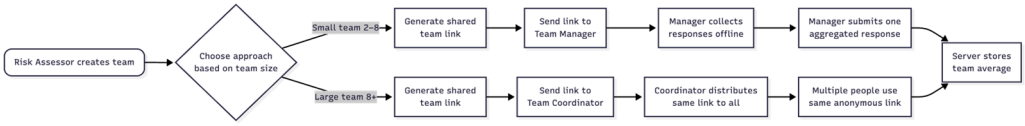


**Figure 10.** Decision path.

Figure 10 illustrates the decision-making process between a collector-based approach (suitable for small teams) and a distributed anonymous approach (ideal for large teams). Both methodologies utilise the same shared team link to guarantee complete anonymity.

The robustness of this architecture is underpinned by a carefully selected technology stack, summarised in Table 3, which ensures secure data handling, efficient performance, and maintainability across deployments.

**Table 3.** Technology Stack.

| Layer | Technology | Notes |
|---|---|---|
| Front-End | Next.js 15 (React) + TailwindCSS + shadcn/ui | Modern App Router, SSR/ISR; accessible component library |
| CMS/API | PayloadCMS v3 (Node 20) | Auto-generates REST/GraphQL and admin UI; can run inside the same Next.js custom server |
| Database | Managed PostgreSQL (DigitalOcean) | JSONB columns for flexible questionnaire schemas |
| Auth | Firebase Auth (magic-link + Google OAuth) | Off-the-shelf, battle-tested; reduces GDPR surface |
| Hosting | DigitalOcean App Platform | Simple CI/CD, autoscaling, global CDN |

### *4.3. Privacy and Security Analysis*

The proposed architecture, following comprehensive review, ensures complete anonymity and zero personal data collection. This design embeds privacy intrinsically through technical architecture, rather than relying solely on policy.

#### 4.3.1. Prohibition of Personal Data Collection

Participation in the survey process does not require the creation of user accounts; instead, respondents engage exclusively through distributed access links. The system architecture ensures that no personal identifiers are retained, with only anonymous team tokens stored in the database. Furthermore, IP addresses are deliberately excluded from both data collection and logging. Temporal information is not preserved in a form that could lead to individual identification; only aggregated team-level metrics are maintained. At no stage are names, email addresses, or other personally identifiable details collected from respondents.

#### 4.3.2. Anonymous Team Scoring Design

Shared team links are employed to ensure that all team members access the survey through identical URLs, thereby eliminating the possibility of individual tracking. Personal scores are computed locally within the browser, ensuring that raw individual values are never transmitted to the server. Only aggregated data are communicated, with individual responses discarded immediately after the aggregation process. The server retains solely the running averages, preserving the cumulative sum and count while omitting individual-level records. This design guarantees complete anonymity, such that even Risk Assessors are unable to retrieve or infer personal responses.

#### 4.3.3. Two-Tier Anonymity Approach

Small Teams (2–8 people):

In this approach, input is consolidated offline by the Team Manager, who collects contributions through meetings and discussions. A single consensus submission is then generated, ensuring that only one aggregated response is transmitted to the server. As a result, no individual traceability is possible, since the server processes solely the team-level consensus.

Large Teams (8+ people):

This method employs a shared anonymous link, ensuring that all team members access the survey through an identical URL and thereby preventing individual tracking. Individual responses are aggregated in real time, with results immediately combined into team-level averages. Due to this design, attribution of responses is technically impossible, rendering the identification of submission sources unfeasible.

#### 4.3.4. Technical Security Measures

All client–server communications are secured through enforced HTTPS encryption. Access to the Risk Assessor dashboard is protected exclusively via JWT-based authentication. The use of Firebase Authentication, a proven OAuth provider, further strengthens access control while reducing the GDPR compliance surface. Regular security assessments, including penetration testing, are conducted prior to each major release to ensure resilience against emerging threats. GDPR compliance is embedded by design, with privacy safeguarded through the technical architecture itself rather than relying solely on legal documentation.

#### 4.3.5. Data Flow Security Architecture

The data flow follows a structured path: Respondent $\rightarrow$ Browser (local scoring) $\rightarrow$ API (team token + numeric score) $\rightarrow$ Database (aggregated team data only). At no point is personal data transmitted; only team identifiers and numeric scores are exchanged. Similarly, no personal data are stored at rest, as the database retains exclusively team-level aggregates. Furthermore, no audit trails are maintained that could enable individual identification or correlation.

#### 4.3.6. Authentication Separation Model

Risk Assessors gain access to the dashboard through secure authentication mechanisms, either via Google OAuth or magic links. By contrast, respondents remain fully anonymous, as no authentication or registration is required. This strict role-based separation ensures that personal data cannot be inadvertently exposed.

#### 4.3.7. Privacy-First Design Principles

True anonymity is achieved through a system architecture in which individual identification is technically impossible rather than merely restricted by policy. Data collection is deliberately minimised, limited exclusively to essential team performance metrics. All sensitive calculations are executed client-side within the browser, ensuring they are never processed on the server. Individual responses are aggregated immediately, without any temporary storage or caching. In parallel, the iubenda platform is employed to provide comprehensive GDPR documentation, thereby embedding legal compliance within the overall design.

#### 4.3.8. Security Conclusion

The trustSense architecture implements exceptional privacy protection through its technical design. The anonymous team scoring approach ensures:

- Zero personal data collection, storage, or transmission.
- Individual responses cannot be traced to specific team members.
- System administrators cannot identify response sources.
- GDPR compliance is achieved through technical architecture, not merely legal processes.

This design surpasses typical anonymisation approaches by implementing true technical anonymity, where individual identification is architecturally impossible, thereby

establishing trustSense as one of the most privacy-preserving organisational assessment tools available.

### 4.4. Validation

The development of trustSense included a two stage validation process. In the first stage, content and face validity were established through co-production workshops, where experts and practitioners from multiple sectors reviewed the questionnaire items for clarity, completeness, and applicability. This process ensured that the maturity dimensions were contextually relevant and practically useful before technical deployment.

The second stage focused on validating the implemented platform itself. Automated testing of the trustilio trustSense tool was carried out to verify if its questionnaires on AI trustworthiness can be relied upon for accurate, consistent, and context-relevant assessments. The focus was on determining whether the tool is suitable for evaluating both individual teams and their alignment with responsible AI practices.

Instead of manual testing, the review was performed through automated scenarios. Personas representing different team types were created, and their answers were automatically submitted to the platform using Selenium-based scripts. The roles that were evaluated included the: AI Technical Team; Domain AI User Team; Cybersecurity Team and Potential Adversary (simulated attacker profiles).

To make this process systematic, a package of Python scripts was prepared for version 1. Each script corresponded to one of the roles above, and a central launcher with a graphical interface was added to make the testing process simple and repeatable.

Each role was tested with its own dedicated script:

- The AI technical team script generated different maturity profiles, submitted them to the system, and recorded the trust levels, scores, and proposed mitigation strategies.
- The Domain AI User Team script assessed how well the tool captures the maturity of non-technical AI users, extracting both the scoring and the recommendations.
- The Cybersecurity Team script provided test cases reflecting organisational readiness at different levels, collecting evaluations and improvement suggestions.
- The Potential Adversary script simulated hostile actors, gathering the attacker profiles generated by the tool.

The graphical launcher allowed evaluators to run these scenarios without technical commands, enabling batch runs and demonstrations. Each script was parameterised: the number of test cases and the output format could be defined in advance. Once the questionnaires were completed through automation, the system's responses (scores, trust levels, mitigation advice, or attacker classification) were collected and exported into Excel workbooks. This structure ensured consistency, easy comparison, and traceability.

The chosen validation approach prioritised repeatability, transparency, and non-intrusiveness, aligning with the privacy-preserving design principles of trustSense. Automated scenario-based testing was selected over traditional interview-based methods to ensure that validation could be performed without collecting any personal or contextual data from participants. This approach allowed reproducible, cross-role evaluation while maintaining full compliance with the tool's zero-data architecture. Nonetheless, we acknowledge that structured expert consensus methods, such as the Delphi technique, can provide complementary qualitative insight by refining indicator definitions, weighting logic, and interpretive validity through iterative feedback. Incorporating such a Delphi-based process represents a logical extension for future research and may be considered in the next validation cycle once cross-sector pilot data are available.

#### 4.4.1. Scoring

The testing confirmed several points:

- The scoring logic works as intended, producing results that scale naturally with input values.
- The mitigation advice adjusts appropriately: early-stage teams receive basic guidance, while advanced groups are provided with more strategic recommendations.
- Adversary profiling shows coherent mapping between input sophistication and the resulting classification.
- Outputs are suitable for supporting teams in gradual capability development.

Overall, the evaluation confirmed that the platform is well-structured, transparent in its logic, and adaptable across very different user groups.

#### 4.4.2. Insights from Team Assessments

The evaluation also looked into how the system behaves with each specific role.

##### AI Technical Teams

When technical AI teams were assessed, average scores rose consistently with experience level. Less mature groups mostly fell into the lower trust bands, while seasoned teams landed in the "high" categories; see Figure 11.

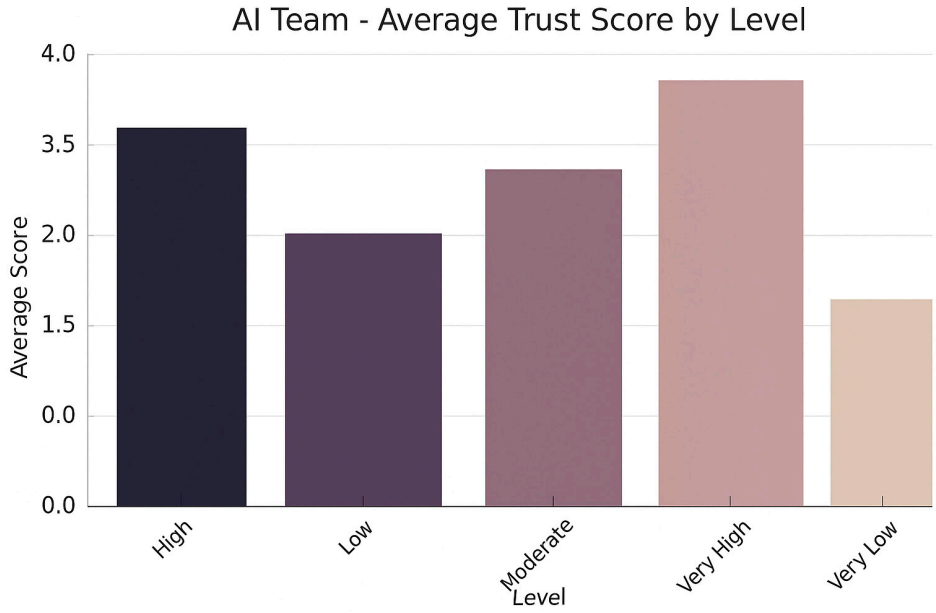


**Figure 11.** Average score by Level.

- Lower scores came with advice around ethics awareness, risk identification, and basic governance.
- Mid-range scores led to suggestions about assigning roles, building KPIs, and improving collaboration on incident response.
- High scores triggered recommendations for mentoring roles, knowledge sharing, and the use of advanced monitoring tools.

This pattern showed that the tool not only detects maturity differences but also scales its feedback from foundational to leadership-level actions.

##### Domain AI Users

Non-technical user groups displayed a similar pattern: initial levels were mostly "low," while more advanced groups clustered in "high" trust bands; see Figure 12.

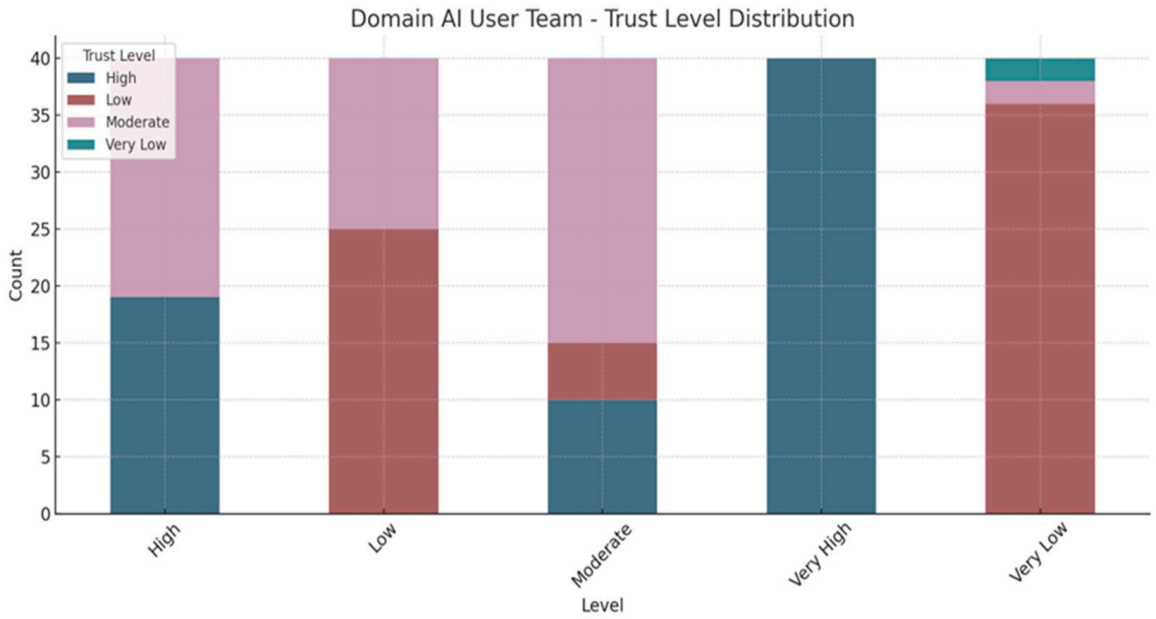


**Figure 12.** Trust Level distribution.

- At the bottom, the system emphasised building awareness of AI limitations and basic compliance practices.
- In the middle, it recommended standardised documentation and bias reviews.
- At the top, the focus shifted to scaling processes, cross-team collaboration, and governance participation.

The findings confirmed that the evaluation is sensitive to how user groups mature and provides role-appropriate guidance.

Cybersecurity Teams

In the case of cybersecurity specialists, the assessments showed that:

- Teams at the starting point need awareness programmes, introductory workshops, and simple guidelines.
- Intermediate groups benefit from structured feedback, resilience checkpoints, and open communication mechanisms.
- High-level groups were advised to expand their leadership roles, involve external peers, and sustain excellence through mentoring.

This demonstrated that the system is able to guide both novice and advanced cybersecurity teams, offering step-by-step pathways to build capacity.

Adversary Profiles

For simulated attackers, the tool did not produce trust scores or mitigation advice. Instead, it categorised adversaries according to sophistication: "insufficient," "basic," "moderate," "experienced," and "multi-sectoral expert." The distribution of results was strictly aligned with the input sophistication, confirming that the profiling logic is both consistent and predictable.

Table 4 includes an example of how the attacker personas were classified across each experience level.

**Table 4.** Attacker Personas by Experience Level.

| Input Level | Insufficient | Basic | Moderate | Experienced |
|---|---|---|---|---|
| Very Low | 9 | 1 | 0 | 0 |
| Low | 0 | 8 | 2 | 0 |
| Moderate | 0 | 0 | 10 | 0 |
| High | 0 | 0 | 0 | 10 |
| Very High | 0 | 0 | 0 | 0 |

The distribution clearly shows that trustSense consistently aligns input levels with increasingly advanced adversary profiles. For example, inputs at the very low end are mostly tagged as 'Insufficient,' whereas those at the highest level are regularly identified as 'Sophisticated' (multi-sectoral expert).

This mechanism offers security teams a structured way to think about potential threats and provides a reliable foundation for red teaming or adversary modelling exercises.

Case Study: Validation in a Healthcare Cybersecurity Context

To complement the automated scenario-based validation, a pilot deployment of trustSense was carried out within a mid-sized European healthcare provider operating an AI-assisted diagnostic tool. The objective was to assess the tool's performance in a privacy-sensitive, high-risk environment and to verify that its aggregated maturity outputs generate actionable organisational insights.

Two respondent groups participated: an AI Technical Team developing and maintaining diagnostic algorithms, and a Cybersecurity Team responsible for safeguarding patient-data systems. Each team completed its respective questionnaire through the anonymised team-link model described in Section 3.3, ensuring full compliance with the zero-data architecture. The Risk Assessor viewed only aggregated scores through the dashboard; no individual responses were stored or transmitted.

Findings: The AI Technical Team achieved a mean maturity score of 3.9 (out of 5), classified as "Intermediate–High." Strengths were observed in technical proficiency, policy adherence, and collaboration and knowledge sharing. Lower values appeared in resilience and openness to interventions, indicating limited rehearsal of contingency plans. The Cybersecurity Team averaged 3.5 ("Intermediate"), performing well on threat awareness and problem solving but showing lower motivation and commitment scores, reflecting workload fatigue. The adversary-profiling module, based on simulated incident data, identified attackers of moderate sophistication.

Interpretation: The comparative results enabled management to prioritise resilience-training workshops and cross-team mentoring initiatives. Three months after implementation, follow-up self-assessments showed improvements of 0.3–0.4 points in the previously weak dimensions, confirming that feedback from trustSense can inform targeted behavioural interventions.

Implications: This pilot confirms that trustSense can operate effectively in real-world, privacy-critical environments while maintaining anonymity and compliance. It provides initial evidence that aggregated maturity assessments translate into actionable organisational outcomes. Due to contractual and confidentiality restrictions, specific identifiers such as the organisation's name, associated tools, or partner systems cannot be disclosed. However, all data and procedures were reviewed under institutional governance to ensure legal and ethical conformity.

#### 4.4.3. Overall Takeaways

The combined validation results, including both the automated scenario testing and the applied healthcare case study presented above, provide strong evidence of the robustness and practical value of trustSense. The case study confirmed that the tool operates effectively in privacy-sensitive, high-stakes environments, generating actionable feedback that led to measurable improvements in team maturity. Together with the automated simulations, these findings demonstrate that the scoring logic, visual feedback, and mitigation guidance remain reliable across diverse roles and contexts, establishing confidence in the tool's methodological soundness and real-world applicability.

The automated review of trustSense produced several important conclusions:

1. Balanced recommendations: Feedback is calibrated to the assessed maturity level, moving from prescriptive advice for beginners to strategic actions for advanced teams.
2. Consistent results: The tool differentiates maturity stages clearly, with scores and classifications following a logical progression.
3. Practical value: It functions not only as an assessment tool but also as a roadmap for phased improvement and planning.

Taken together, these outcomes show that trustSense is a robust and flexible platform for assessing AI trustworthiness across technical, organisational, and security-focused roles.

While trustSense currently applies an equal-weight approach across its maturity dimensions to preserve transparency and avoid subjective bias, its modular structure allows future integration with traditional quantitative evaluation models that employ indicator selection and weighting, such as analytic hierarchy processes (AHP) or multi-criteria decision analysis (MCDA). Each questionnaire dimension can serve as an indicator, and domain-specific weighting schemes could be introduced to emphasise critical aspects, such as resilience or ethics, depending on organisational context. This flexibility ensures that trustSense can be combined with existing comprehensive evaluation methods while maintaining its privacy-preserving and user-centred design principles.

## 5. trustSense in Trustworthiness Assessment of AI Systems

Beyond its standalone functionality, trustSense can be embedded within risk management frameworks equally in the context of both cybersecurity and broader trustworthiness risk management, since the overall risk level (according to NIST SP 800-30 [7]) can be expressed as a function of three factors: threat, impact, and vulnerability: Risk $(\mathrm{R}) = \mathrm{T} \times \mathrm{I} \times \mathrm{V}$, where

- Threat (T): the likelihood or presence of a potential event.
- Impact (I): the magnitude or severity of consequences if the threat materialises.
- Vulnerability (V): the degree to which an asset (system, software, hardware, service, data) is susceptible to exploitation.

The risk estimation (R) provides a structured and quantifiable method for assessing risk exposure without considering the maturity of the humans in safeguarding assets and without considering the sophistication of potential attackers (e.g., knowledge gained through historical incidents or cyber threat intelligence).

In the trustworthiness risk management framework AI-TAF [28]; the trustworthiness Risk level (R) was refined by using trustSense estimates.

In particular the AI maturity level (tAIP) of the asset owner team was estimated and then the refined risk level (fR) became:

$$\mathrm{R} = \begin{cases} \mathrm{R} - 1, & \mathrm{tAIP} >= \text{ Medium } (\mathrm{M}) \\ \mathrm{R} + 1, & \mathrm{tAIP} < \text{ Medium } (\mathrm{M}) \end{cases}$$

where 1 = one level of the scale, for example, from very high to high or from low to medium.

Furthermore, the sophistication level of the potential adversary (tA) was also determined, by incorporating tA into R; the final risk level (FR) became:

$$FR = \begin{cases} R & , \quad tA = tAIP \\ R - 1, & tAIP > tA \\ R + 1, & tAIP < tA \end{cases}$$

where 1 = one level of the scale, for example, from very high to high or from low to medium.

trustSense delivers more realistic risk assessments, enabling organisations to implement tailored mitigation measures and technical controls that are targeted, affordable and aligned with their workforce's culture, skills, and behaviours. See Figure 13.

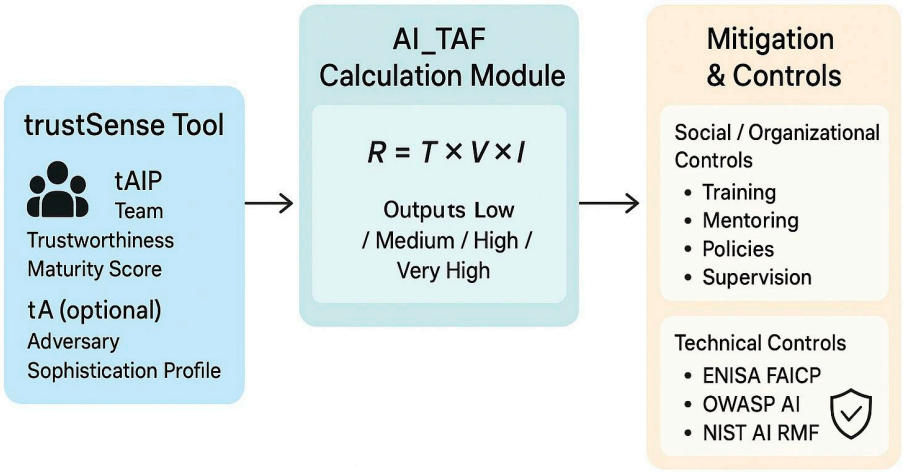


**Figure 13.** Integration of trustSense in AI_TAF Risk Governance.

## 6. Conclusions

trustSense positions human oversight maturity as a measurable organisational attribute and shows that incorporating attacker sophistication and user/team maturity into risk estimation yields more realistic, affordable controls than system-only views, especially for SMEs. In practice, the implications are human-centred: the maturity results can be used to target behaviour change (role-specific training, mentoring, and awareness), to strengthen crisis communication and psychological resilience in incident response teams, and to prioritise governance actions (e.g., staffing, escalation, and oversight) where low maturity amplifies technical risk. This aligns with prior evidence that profiling adversaries alters severity calculations and that socio-psychological factors in users materially shape risk, thereby enabling proportionate mitigation rather than blanket controls; it also complements socio-technical approaches to adversarial ML that surface human elements on both the attacker and defender sides. Organisations can embed trustSense outputs into existing risk registers and assurance routines, calibrating controls to attacker profiles and team readiness, while mapping improvements to sectoral obligations (e.g., NIS2 and health-sector conformity schemes) without compromising privacy.

Limitations include reliance on self-reported, static questionnaires (susceptible to response bias), incomplete cross-sector psychometric validation, and a short evidence window linking maturity shifts to concrete outcomes such as reduced incidents or audit findings. To address these, future work will include: (i) running multi-site, pre/post pilots (e.g., stepped-wedge) to test sensitivity to change and criterion validity; (ii) complete reliability and invariance testing across languages and roles; (iii) integrating optional, privacy-preserving telemetry (basic MLOps signals, incident logs) to triangulate scores; (iv) adding a

lightweight CTI-driven attacker-profiling plug-in that maps common threat intel to risk adjustments; (v) delivering micro-interventions grounded in behaviour models (short nudges, drills, and peer-mentoring packs) tied to specific low-scoring traits; and (vi) publishing an anonymised benchmarking corpus and control mappings (ENISA/OWASP/NIS2) such that teams can track progress and auditors can reuse evidence. These steps are feasible within routine governance cycles and will strengthen the causal link between human-factor maturity, calibrated controls, and measurable improvements in trustworthy AI operations.

**Author Contributions:** Conceptualization, K.K. and N.P.; Methodology, T.F., E.S. and N.P.; Validation, E.S.; Formal analysis, T.F. and M.P.; Resources, T.F. and E.S.; Writing—original draft, K.K.; Writing—review & editing, T.F. and E.S.; Supervision, K.K. and N.P.; Project administration, K.K. and N.P.; Funding acquisition, K.K. and N.P. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the 'Fostering Artificial Intelligence Trust for Humans towards the Optimization of Trustworthiness through Large-scale Pilots in Critical Domains' (FAITH) project, which has received funding from the European Union's Horizon Programme under grant agreement No. 101135932.

**Data Availability Statement:** The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

**Conflicts of Interest:** All authors are employed in trustilio B.V., Vijzelstraat 68, 1017 HL Amsterdam, The Netherlands. The views expressed in this paper represent only the views of the authors and not those of the European Commission or the partners in the above-mentioned project. Finally, the authors declare that there are no conflicts of interest, including any financial or personal relationships, that could be perceived as potential conflicts. trustSense® and its associated logo are registered assets of trustilio B.V. with the Benelux Office for Intellectual Property (BOIP) under application number 1528920. The use of the name and logo in this paper is solely for identification purposes and does not imply endorsement by any third party.

## References

1. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4th ed.; Pearson: London, UK, 2020.
2. European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Off. J. Eur. Union L* **2024**, *1689*, 1–139.
3. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People, an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
4. Jobin, A.; Ienca, M.; Vayena, E. The global landscape of AI ethics guidelines. *Nat. Mach. Intell.* **2019**, *1*, 389–399. [CrossRef]
5. ENISA. *Artificial Intelligence Cybersecurity Challenges*; European Union Agency for Cybersecurity: Athens, Greece, 2020. Available online: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges (accessed on 1 August 2025).
6. OECD. *OECD Framework for the Classification of AI Systems*; Organisation for Economic Co-operation and Development: Paris, France, 2022. Available online: https://oecd.ai/en/classification (accessed on 1 August 2025).
7. NIST. *AI Risk Management Framework (AI RMF 1.0) (NIST AI 100-1)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023. [CrossRef]
8. *ISO/IEC 23894:2023*; Information Technology Artificial Intelligence Guidance on Risk Management. International Organization for Standardization: Geneva, Switzerland, 2023. Available online: https://www.iso.org/standard/77304.html (accessed on 1 August 2025).
9. Stevens, T. Knowledge in the grey zone: AI and cybersecurity. *Digit. War* **2020**, *1*, 164–170. [CrossRef]
10. Kioskli, K.; Bishop, L.; Polemi, N.; Ramfos, A. Towards a human-centric AI trustworthiness risk management framework. In *Human Factors in Cybersecurity; Proceedings of the AHFE 2024 International Conference, Nice, France, 24–27 July 2024*; Moallem, A., Stephanidis, C., Eds.; AHFE Open Access: Istanbul, Turkey, 2024; Volume 127, pp. 63–73. [CrossRef]
11. European Commission. Directive (EU) 2022/2555 on Measures for a High Common Level of Cybersecurity Across the Union (NIS2 Directive). 2023. Available online: https://eur-lex.europa.eu/eli/dir/2022/2555/oj (accessed on 1 August 2025).

12. ENISA. *NIS2 Directive Overview*; European Union Agency for Cybersecurity: Athens, Greece, 2023. Available online: https://www.enisa.europa.eu/topics/awareness-and-cyber-hygiene/raising-awareness-campaigns/network-and-information-systems-directive-2-nis2 (accessed on 1 August 2025).
13. European Commission. Assessment List for Trustworthy Artificial Intelligence (ALTAI): Self-Assessment. 2020. Available online: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment (accessed on 1 August 2025).
14. The General Purpose AI Code of Practice. 2025. Available online: https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai (accessed on 1 August 2025).
15. Oxford Insights. Trustworthy AI Self-Assessment [Spreadsheet]. 2023. Available online: https://oxfordinsights.com/wp-content/uploads/2023/12/Trustworthy-AI-Self-Assessment-2023.xlsx (accessed on 1 August 2025).
16. Oxford Insights. Government AI Readiness Index 2024. Available online: https://oxfordinsights.com/wp-content/uploads/2024/12/2024-Government-AI-Readiness-Index-2.pdf (accessed on 1 August 2025).
17. Alan Turing Institute. Trustworthiness Assessment Tool (Digital Identity). 2025. Available online: https://www.turing.ac.uk/TDI/trustworthiness-assessment-tool (accessed on 1 August 2025).
18. AI4Belgium. Tool for Assessing the Trustworthiness of An Organization's AI. Available online: https://altai.ai4belgium.be/ (accessed on 15 August 2025).
19. KPMG. An Illustrative AI Risk and Controls Guide. 2025. Available online: https://kpmg.com/us/en/articles/ai-risk-and-control-guide-gated.html (accessed on 1 August 2025).
20. KPMG. AI Trust Services. 2025. Available online: https://kpmg.com/xx/en/what-we-do/services/ai/ai-trust-services.html (accessed on 1 August 2025).
21. ICO. *AI and Data Protection Risk Toolkit*; Information Commissioner's Office: Haymarket, Australia, 2022. Available online: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/ai-and-data-protection-risk-toolkit/ (accessed on 1 August 2025).
22. Z-Inspection®. Z-Inspection® Tool. Available online: https://z-inspection.org/ (accessed on 15 August 2025).
23. IBM. AI Fairness 360 (AIF360). 2019. Available online: https://research.ibm.com/blog/ai-fairness-360 (accessed on 1 August 2025).
24. Microsoft. Assess AI Systems by Using the Responsible AI Dashboard (Azure Machine Learning Documentation). 2024. Available online: https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard (accessed on 1 August 2025).
25. Li, Z.; Kong, C.; Yu, Y.; Wu, Q.; Jiang, X.; Cheung, N.-M.; Wen, B.; Kot, A.; Jiang, X. SAVER: Mitigating hallucinations in large vision-language models via style-aware visual early revision. *arXiv* **2025**, arXiv:2508.03177.
26. FAITH Project. In *Federated Artificial Intelligence Solution for Monitoring Mental Health Status After Cancer Treatment (FAITH)*; European Commission: Brussels, Belgium. Available online: https://faith-ec-project.eu/ (accessed on 1 August 2025).
27. THEMIS 5.0 Project. In *Human-Centered Trustworthiness Optimisation in Hybrid Decision Support (THEMIS 5.0)*; European Commission: Brussels, Belgium. Available online: https://www.themis-trust.eu/ (accessed on 1 August 2025).
28. Seralidou, E.; Kioskli, K.; Fotis, T.; Polemi, N. AI_TAF: A human-centric trustworthiness risk assessment framework for AI systems. *Computers* **2025**, *14*, 243. [CrossRef]
29. Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; Zhou, B. Trustworthy AI: From principles to practices. *arXiv* **2021**, arXiv:2110.01167.
30. OECD. Catalogue of Tools & Metrics for Trustworthy AI. Available online: https://oecd.ai/en/catalogue/overview (accessed on 15 August 2025).
31. Government of Canada. Algorithmic Impact Assessment (AIA). 2008, pp. 154–196. Available online: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html (accessed on 24 June 2025).
32. Kioskli, K.; Polemi, N. A socio-technical approach to cyber risk assessment. *Int. J. Electr. Comput. Eng.* **2020**, *14*, 305–309.
33. Kioskli, K.; Polemi, N. Measuring psychosocial and behavioural factors improves attack potential estimates. In Proceedings of the 15th International Conference for Internet Technology and Secured Transactions (ICITST 2020), London, UK, 8–10 December 2021; pp. 216–219. [CrossRef]
34. Kioskli, K.; Polemi, N. Estimating attackers' profiles results in more realistic vulnerability severity scores. In *Human Factors in Cybersecurity; Proceedings of the AHFE International Conference, New York, NY, USA, 24–28 July 2022*; Ahram, T., Karwowski, W., Eds.; AHFE Open Access: Istanbul, Turkey, 2022; Volume 53. [CrossRef]
35. Fotis, T.; Kioskli, K.; Seralidou, E. Charting trustworthiness: A socio-technical perspective on AI and human factors. In *Human Factors in Cybersecurity; Proceedings of the AHFE 2025 International Conference, Honolulu, HI, USA, 8–10 December 2025*; Moallem, A., Kioskli, K., Eds.; AHFE Open Access: Istanbul, Turkey, 2025; Volume 168.
36. Zhou, J.; Luo, S.; Chen, F. Effects of personality traits on user trust in human–machine collaborations. *J. Multimodal User Interfaces* **2020**, *14*, 387–400. [CrossRef]

37. Riedl, R. Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electron. Mark.* **2022**, *32*, 2021–2051. [CrossRef]
38. Kuper, A.; Krämer, N. Psychological traits and appropriate reliance: Factors shaping trust in AI. *Int. J. Hum. Comput. Interact.* **2024**, *41*, 4115–4131. [CrossRef]
39. Morana, S.; Gnewuch, U.; Jung, D.; Granig, C. The effect of anthropomorphism on investment decision-making with robo-advisor chatbots. In Proceedings of the 28th European Conference on Information Systems (ECIS), Marrakech, Morocco, 15–17 June 2020. Available online: https://aisel.aisnet.org/ecis2020_rp/63 (accessed on 1 August 2025).
40. Klumpp, M.; Zijm, H. Logistics innovation and social sustainability: How to prevent an artificial divide in human–computer interaction. *J. Bus. Logist.* **2019**, *40*, 265–278. [CrossRef]
41. Bach, T.A.; Khan, A.; Hallock, H.; Beltrão, G.; Sousa, S. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *Int. J. Hum. Comput. Interact.* **2022**, *40*, 1251–1266. [CrossRef]
42. Lee, M.; Frank, L.; Ijsselsteijn, W. Brokerbot: A cryptocurrency chatbot in the socio-technical gap of trust. *Comput. Support. Coop. Work* **2021**, *30*, 79–117. [CrossRef]


**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.